# Model-based clustering of Gaussian copulas for mixed data

Matthieu Marbac, Christophe Biernacki, Vincent Vandewalle

# Model-based clustering of Gaussian copulas for mixed data

Matthieu Marbac - Christophe Biernacki - Vincent Vandewalle

August 13, 2014

**Abstract**

A mixture model of Gaussian copulas is introduced to cluster mixed-type data (data set composed by different natures of variables). Thus, the analyze can be performed on data sets composed by any kinds of variables admitting a cumulative distribution function. Copulas are used to modelize the intra-class dependencies and to preserve any distributions for the one-dimensional margins of each component. Typically in this work, each component follows a Gaussian copula which provides one correlation coefficient per couple of variables and per class. Moreover, the one-dimensional margins of each component follow classical parametric distributions in order to facilitate the model interpretation. This model generalizes many well-known models and allows meaningful data visualization as a straightforward by-product issue. A Metropolis-within-Gibbs sampler performs the Bayesian inference by avoiding the difficulties related to the parameter estimation of the copulas with discrete margins. Experiments on simulated and real data illustrate the

model advantages: flexible parameters (one-dimensional margins and correlation matrices) associated to visualization aspects.

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

# 1 Introduction

Clustering is an efficient tool to manage large data sets since it divides the individuals into few specific classes. When it is used in the probabilistic framework, a class gathers together the individuals arisen from the same distribution. In this context, the most popular approaches modelize the data distribution with finite mixture models of parametric distributions [MP00]. The bibliography specific to homogeneous data (same nature of variables) is prolific. Among it, the Gaussian mixture models [BR93], the Poisson mixture models [KT08] and the multinomial mixture models [Goo74] are the standard models to analyze such data. Their success is due to the use of classical distributions for the mixture components. Indeed, the practitioner can easily interpret the classes. However, even if many data sets contain mixed variables (variables of different natures), there are few multivariate distributions devoted for such data. Moreover, these distributions can be scarcely any interpretable. We now present the main three models used to cluster mixed data. Note that a more detailed overview is available in [HJ11].

The *locally independent mixture model* analyzes the data by assuming that

the variables are independent given the class. It can provide meaningful results (see the applications of [Lew98, HY01]) especially when the one-dimensional margins of the components follow classical distributions. However, this model leads to severe biases when its main assumption is violated (see the application of [VHH09]). In such a case, two methods can be envisaged. The first one consists in selecting a subset of intra-class independent variables [MCMM09], but some informations contained in the data can be lost. The second method consists in using models relaxing the conditional independence assumption. We now present two of them.

The *location mixture model* [Krz93, WB99] assumes that the continuous variables follow a multivariate Gaussian distribution conditionally on both the class and the categorical variables. More precisely, its means depends on both the class and the categorical variables while its covariance matrix is only set by the class membership. This model takes into account the intra-class dependencies but it requires too many parameters. So, it was expanded by Jorgensen and Hunt [JH96, HJ99] by splitting the variables into conditionally independent blocks. Each block is composed by at most one categorical variable and follows a location model. Indeed, in a block, the categorical variable follows a full multinomial distribution and the continuous variables follow a multivariate Gaussian distribution conditionally on the categorical variable. Note that the interpretation of the classes can be complex. Indeed, for one component, the distribution of the one-dimensional margin of a continuous variable is itself a Gaussian mixture model (and not a classical distribution!). Finally, the estimation of the variable allocation into blocks is complex. The authors achieved it by an ascendant method which is sub-optimal.

The *underlying variables mixture model* [Eve88] analyzes data sets with continuous and ordinal variables. It assumes that each discrete variable arises from a latent continuous variable and that the whole continuous variables (observed and unobserved) follow a Gaussian mixture model. Thus, the distribution of the observed variables is obtained by integrating each Gaussian component on the subset of the latent variables. However, in practice, this computation is not feasible when there are more than two discrete variables. To study data sets with numerous binary variables, Morlini [Mor12] has expanded this model by estimating the scores of the latent variables from the binary variables. However, the class interpretation is done throughout the parameters related to the scores (and not related to the observed variables).

The *mixture model of Gaussian copulas* is introduced by this paper to analyze mixed data sets. Note that [SK12, MDCL13] already modelized the distribution of mixed variables by using one Gaussian copula. The proposed model expands this approach to the mixture models in order to perform the cluster analysis. Copulas [Joe97, Nel99] allow to build a multivariate model by setting, on the one hand, the distributions of the *one-dimensional margins*, and, on the other hand, the *dependency model*. Therefore, the mixture model of copulas provides *classical distributions for all the one-dimensional margins* for each components. Moreover, as each component follows a Gaussian copula [Hof07, HNW11] which modelizes its dependencies, the *intra-class dependencies* are meaningfully taken into account. Thus, a *three-level schema* 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.

This paper is organized as follows. Section 2 introduces the mixture model of Gaussian copulas and its links with well-known models. Section 3 presents the Metropolis-within-Gibbs algorithm performing the Bayesian inference. Section 4 illustrates the behavior of this algorithm and the model robustness on numerical experiments. Section 5 presents two applications of the new mixture model by clustering two real data sets. Section 6 concludes this work.

# 2 Mixture model of Gaussian copulas

## 2.1 Finite mixture model

The vector $\boldsymbol{x} = (x^1, \ldots, x^e) \in \mathbb{R}^c \times \mathcal{X}$ denotes the $e = c + d$ observed variables. Its first $c$ elements are denoted by $\boldsymbol{x}^{\mathrm{C}}$ and correspond to the subset of the continuous variables defined on the space $\mathbb{R}^c$. Its last $d$ elements are denoted by $\boldsymbol{x}^{\mathrm{D}}$ and correspond the subset of the discrete variables (integer, ordinal or binary) defined on the space $\mathcal{X}$. Note that if $x^j$ is an ordinal variable with $m_j$ modalities, then it uses a numeric coding $\{1, \ldots, m_j\}$.

Data $\boldsymbol{x}$ are assumed to arise from the mixture model of $g$ parametric distri-

butions whose the probability distribution function (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 $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ is defined on the simplex of size $g$ and groups the class proportions, where $\pi_k$ is the proportion of class $k$. The vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ groups the component parameters, where $\boldsymbol{\alpha}_k$ denotes the parameters of class $k$.

The categorical variable $z \in \{1, \ldots, g\}$ indicates the class membership of the individual but is unobserved. Moreover, it follows the multinomial distribution $\mathcal{M}_g(\pi_1, \ldots, \pi_g)$. Therefore, (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

The mixture model of Gaussian copulas assumes that each component $k$ follows a Gaussian copula. Therefore, the cumulative distribution function (cdf) of component $k$ is written as follows

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

where $\boldsymbol{\alpha}_k = (\boldsymbol{\Gamma}_k, \boldsymbol{\beta}_k)$, where $\boldsymbol{\Gamma}_k$ is a correlation matrix of size $e \times e$, where $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, \ldots, \boldsymbol{\beta}_{ke})$ and where $\boldsymbol{\beta}_{kj}$ denotes the parameters of one-dimensional margin $j$. Moreover, $u_k^j = P(x^j; \boldsymbol{\beta}_{kj})$ is the value of the cdf of one-dimensional margin $j$ for the component $k$ evaluated on $x^j$. Finally, $\Phi_e(.; \mathbf{0}, \boldsymbol{\Gamma}_k)$ is the cdf of the $e$-variate centred Gaussian distribution with correlation matrix $\boldsymbol{\Gamma}_k$ and

$\Phi_1^{-1}(.)$ is the inverse cumulative distribution function of the standard Gaussian $\mathcal{N}_1(0,1)$.

For each component, we assume also that the one-dimensional margins follow classical distributions to facilitate the interpretation. More precisely,

- If $x^j$ is *continuous*, $x^j|z=k$ follows a *Gaussian* distribution $\mathcal{N}_1(\mu_{kj},\sigma_{kj}^2)$ of mean $\mu_{kj}$ and variance $\sigma_{kj}^2$, so $\boldsymbol{\beta}_{kj}=(\mu_{kj},\sigma_{kj}^2)\in\mathbb{R}\times\mathbb{R}^{+*}$.
- If $x^j$ is *integer*, $x^j|z=k$ follows a *Poisson* distribution, so $\boldsymbol{\beta}_{kj}\in\mathbb{R}^{+*}$.
- If $x^j$ is *ordinal*, $x^j|z=k$ follows a *multinomial* distribution, so $\boldsymbol{\beta}_{kj}$ is defined on the simplex of size $m_j$.

From (2) and from the specific one-dimensional margin distributions previously explained, the pdf of component $k$ is written as follows

$$p(\boldsymbol{x};\boldsymbol{\alpha}_k) = p(\boldsymbol{x}^{\mathrm{C}};\boldsymbol{\alpha}_k)p(\boldsymbol{x}^{\mathrm{D}}|\boldsymbol{x}^{\mathrm{C}};\boldsymbol{\alpha}_k) \tag{3}$$

$$= \frac{\phi_c(\Psi(\boldsymbol{x}^{\mathrm{C}};\boldsymbol{\alpha}_k);\mathbf{0},\boldsymbol{\Gamma}_{k\mathrm{CC}})}{\prod_{j=1}^{c}\sigma_{kj}}\int_{\mathcal{S}_k(\boldsymbol{x}^{\mathrm{D}})}\phi_d(\boldsymbol{u};\boldsymbol{\mu}_k^{\mathrm{D}},\boldsymbol{\Sigma}_k^{\mathrm{D}})d\boldsymbol{u}, \tag{4}$$

where the function $\Psi(\boldsymbol{x}^{\mathrm{C}};\boldsymbol{\alpha}_k)=\left(\frac{x^j-\mu_{kj}}{\sigma_{kj}};j=1,\ldots,c\right)$ and where $\mathcal{S}_k(\boldsymbol{x}^{\mathrm{D}})=\mathcal{S}_k^{c+1}(x^{c+1})\times\ldots\times\mathcal{S}_k^{e}(x^{e})$ is the space of the antecedents of $\boldsymbol{x}^{\mathrm{D}}$ for class $k$ such as $\mathcal{S}_k^j(x^j)=]b_k^{\ominus}(x^j),b_k^{\oplus}(x^j)]$ is defined for $j=c+1,\ldots,e$ with $b_k^{\ominus}(x^j)=\Phi_1^{-1}(P(x^j-1;\boldsymbol{\beta}_{kj}))$ and $b_k^{\oplus}(x^j)=\Phi_1^{-1}(P(x^j;\boldsymbol{\beta}_{kj}))$. Moreover the correlation matrix $\boldsymbol{\Gamma}_k=\begin{bmatrix}\boldsymbol{\Gamma}_{k\mathrm{CC}} & \boldsymbol{\Gamma}_{k\mathrm{CD}}\\ \boldsymbol{\Gamma}_{k\mathrm{DC}} & \boldsymbol{\Gamma}_{k\mathrm{DD}}\end{bmatrix}$ 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. Finally, $\boldsymbol{\mu}_k^{\mathrm{D}}$ 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 $\boldsymbol{\Sigma}_k^{\mathrm{D}} = \boldsymbol{\Gamma}_{k\mathrm{DD}} - \boldsymbol{\Gamma}_{k\mathrm{DC}}\boldsymbol{\Gamma}_{k\mathrm{CC}}^{-1}\boldsymbol{\Gamma}_{k\mathrm{CD}}$.

The mixture model of Gaussian copulas involves a second latent variable $\boldsymbol{y} = (y^1, \ldots, y^e) \in \mathbb{R}^e$ such as $\boldsymbol{y}|z = k$ follows an $e$-variate centred Gaussian distribution $\mathcal{N}_e(\mathbf{0}, \boldsymbol{\Gamma}_k)$. Conditionally on $(\boldsymbol{y}, z)$, $\boldsymbol{x}$ is defined by

$$x^j = P^{-1}(\Phi_1(y^j); \boldsymbol{\beta}_{kj}),\ \forall j = 1, \ldots, e. \tag{5}$$

Thus, the generative model of the mixture model of Gaussian copulas is

- 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 (5).

For the small data sets, a better trade off between the bias and the variance of the estimate may be obtained by constraining the parameter space. 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.

The heteroscedastic (respectively homoscedastic) mixture model of Gaus-

sian copulas requires $\nu_{\text{He}}$ (respectively $\nu_{\text{Ho}}$) parameters where

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

Note finally that the mixture model of Gaussian copulas is identifiable if, at least, one variable is continuous or integer (see Appendix A).

## 2.3 Strengths of the mixture model of Gaussian copulas

**Related models** The mixture model of Gaussian copulas allows to generalize many classical mixture models, among them one can cite the following four.

- If the correlation matrices are diagonal (*i.e.* $\boldsymbol{\Gamma}_k = \boldsymbol{I}$, $\forall k = 1, \ldots, g$), then the model is equal to the locally independent mixture model.
- If all the variables are continuous (*i.e.* $c = e$ and $d = 0$), then the model is equal to a multivariate Gaussian mixture model without constraint between the parameters [BR93].
- The model 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 model is a new parametrization of the mixture model proposed by Everitt [Eve88]. Note that Everitt directly estimates the space $\mathcal{S}_k(\boldsymbol{x}^{\text{D}})$ which contains the antecedents of $\boldsymbol{x}^{\text{D}}$. Moreover, he uses a simplex algorithm to perform the

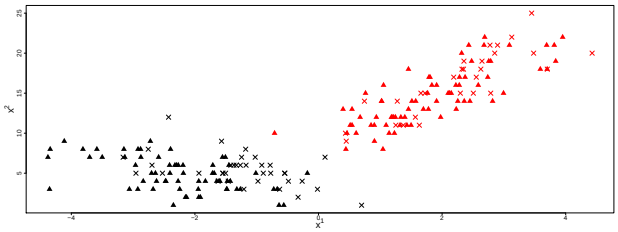
maximum likelihood inference, but this method dramatically limits the number of ordinal variables. By using the margin parameters $\boldsymbol{\beta}_{kj}$ of the components, our approach allows a Bayesian inference which avoids this drawback (see details in Section 3).

**Standardized coefficient of correlation per class** The Gaussian copula provides a friendly 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].

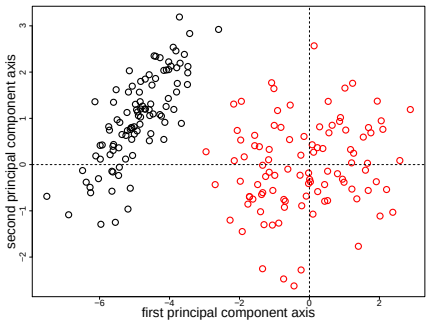
**Data visualization per class: a by-product of Gaussian copulas** By using the latent vectors of the Gaussian copulas $\boldsymbol{y}|z$, a PCA-type method allows a *visualization* of the individuals *per class* and brings out the main intra-class dependencies. Thus, the visualization of class $k$ consists in computing the coordinates $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}, z = k; \boldsymbol{\alpha}_k]$ then in projecting them on the PCA space related to the Gaussian copula of component $k$. This space is directly obtained by the spectral decomposition of $\boldsymbol{\Gamma}_k$. The individuals arisen from component $k$ follow a centred Gaussian distribution on the factorial map. Those arisen from another component have an expectation different to zero. So, if they are farther from the origin, they arise from a distribution strongly different to the distribution of class $k$. Finally, the correlation circle summarizes the intra-class correlations and avoids the direct interpretation of the correlation matrix which can be fastidious if $e$ is large. The following example illustrates

these properties.

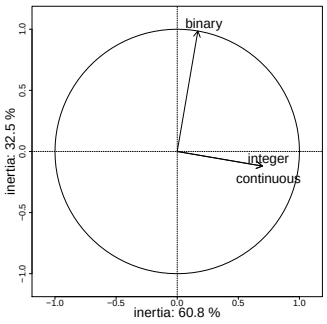
**Example.** Let one continuous, one integer and one binary arisen, in this order, from the bi-component mixture model of Gaussian copulas parametrized by $\boldsymbol{\pi} = (0.5, 0.5)$, $\boldsymbol{\beta}_{11} = (-2, 1)$, $\boldsymbol{\beta}_{12} = 5$, $\boldsymbol{\beta}_{13} = (0.5, 0.5), \boldsymbol{\beta}_{21} = (2, 1)$, $\boldsymbol{\beta}_{22} = 15$, $\boldsymbol{\beta}_{23} = (0.5, 0.5)$, $\boldsymbol{\Gamma}_1 = \begin{pmatrix} 1 & -0.4 & 0.4 \\ -0.4 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{pmatrix}$ 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 1: (a) scatterplot of the individuals described by three variables: one continuous (abscissa), one integer (ordinate) and one binary (symbol); (b) scatterplot of the individuals in the first component map of class 2; (c) variable representation in the first component map of class 2. The color indicates the class memberships.

The visualization of class 2 is presented in Figure 1. Concerning the individuals, the scatterplot shows a centred class (the red one) and a second class (the black one) located on the left side. Concerning the variables, the rep-

resentation points-out a strong intra-class correlation between the continuous variable and the integer variable.

# 3 Bayesian inference

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 arise from a mixture model of Gaussian copulas.

As pointed-out in [PCK06], the maximum likelihood inference is very difficult for a Gaussian copula with discrete margins. So, it is often replaced by the *Inference Function for Margins* method performing the inference in two steps (see Chapter 10 of [Joe97]) but which is sub-optimal. When all the variables are continuous, the fixed-point-based algorithm proposed by [SFK05] achieves the maximum likelihood estimation, but this approach is not doable for mixed data. Therefore, an EM algorithm can not be implemented because its M step would not be feasible. Moreover, if the discrete variables are numerous, its E step would be too much time consuming because it would require the difficult calculation of the integral defined in (4).

As pointed-out by [SK12], the Bayesian framework considerably simplifies the inference since it uses the latent structure of the model $(\boldsymbol{y}, z)$.

## 3.1 Prior distributions

We assume independence between the prior distributions, so

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_{k=1}^{g} \left( p(\boldsymbol{\Gamma}_k) \prod_{j=1}^{d} p(\boldsymbol{\beta}_{kj}) \right). \tag{8}$$

The classical conjugate prior distribution of the proportion vector is the Jeffreys non informative one which is the following Dirichlet distribution

$$\boldsymbol{\pi} \sim \mathcal{D}_g\left(\frac{1}{2}, \ldots, \frac{1}{2}\right). \tag{9}$$

The parameters of the one-dimensional margin of each components $\boldsymbol{\beta}_{kj}$ follows the classical conjugate prior distributions. These distributions are detailed in Appendix B. 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 $\boldsymbol{\Lambda}_k$. Indeed, $\boldsymbol{\Gamma}_k|\boldsymbol{\Lambda}_k$ is deterministic [Hof07]. So,

$$\boldsymbol{\Lambda}_k \sim \mathcal{W}^{-1}(s_0, S_0) \text{ and } \forall 1 \leq h, \ell \leq e,\ \boldsymbol{\Gamma}_k[h,\ell] = \frac{\boldsymbol{\Lambda}_k[h,\ell]}{\sqrt{\boldsymbol{\Lambda}_k[h,h]\boldsymbol{\Lambda}_k[\ell,\ell]}}, \tag{10}$$

where $(s_0, S_0)$ are two hyper-parameters. However, these parameters can not be fitted by an empirical Bayesian approach since $\boldsymbol{y}$ is not observed. To obtain uniform distribution on $]-1,1[$ for the margin distributions of each correlation coefficient, we put $s_0 = e+1$ and $S_0$ equal to the identity matrix [BMM00].

## 3.2 Gibbs and Metropolis-within-Gibbs samplers

The Bayesian inference is performed via a Gibbs sampler which is the most popular approach for the mixture models since it uses the latent structure of the data. Its stationary distribution is $p(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}|\mathbf{x})$ where $\mathbf{z} = (z_1, \ldots, z_n)$ denotes the class memberships of $\mathbf{x}$ and where $\mathbf{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ denotes the Gaussian vectors related to $\mathbf{x}$. Thus, the sequence of the generated parameters

is sampled from the marginal posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$. When a step of the Gibbs sampler is difficulty performed, it can be replaced by one iteration of a Metropolis-Hastings algorithm without changing the stationary distribution. The obtained algorithm is a Metropolis-within-Gibbs sampler whose the properties are detailed in [RC04].

**Algorithm 3.1** (The Gibbs sampler)**.** Starting from an initial value $\boldsymbol{\theta}^{(0)}$, its iteration $(r)$ performs the following four steps

$$\mathbf{z}^{(r)}, \mathbf{y}^{(r-1/2)} \sim \mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{(r-1)} \tag{11}$$

$$\boldsymbol{\beta}_{kj}^{(r)}, \boldsymbol{y}_{[rk]}^{j(r)} \sim \boldsymbol{\beta}_{kj}, \boldsymbol{y}_{[rk]}^{j}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{\jmath}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{\jmath}}^{(r)}, \boldsymbol{\Gamma}_{k}^{(r-1)} \tag{12}$$

$$\boldsymbol{\pi}^{(r)} \sim \boldsymbol{\pi}|\mathbf{z}^{(r)} \tag{13}$$

$$\boldsymbol{\Gamma}_{k}^{(r)} \sim \boldsymbol{\Gamma}_{k}|\mathbf{y}^{(r)}, \mathbf{z}^{(r)}, \tag{14}$$

where $\boldsymbol{y}_{[rk]} = \boldsymbol{y}_{\{i:z_i^{(r)}=k\}}$, $\boldsymbol{y}_i^{\bar{\jmath}(r)} = (y_i^{1(r)}, \ldots, y_i^{j-1(r)}, y_i^{j+1(r-1/2)}, \ldots, y_i^{e(r-1/2)})$ and $\boldsymbol{\beta}_{k\bar{\jmath}}^{(r)} = (\boldsymbol{\beta}_{k1}^{(r)}, \ldots, \boldsymbol{\beta}_{kj-1}^{(r)}, \boldsymbol{\beta}_{kj+1}^{(r-1)}, \ldots, \boldsymbol{\beta}_{ke}^{(r-1)})$. Note that the Gaussian variable $\mathbf{y}$ is twice sampled during one iteration of the algorithm to manage the strong dependency between $\mathbf{y}$ and $\mathbf{z}$, and between $\boldsymbol{y}_{[rk]}^{j}$ and $\boldsymbol{\beta}_{kj}$. Obviously, the stationary distribution stays unchanged. We now detail the four steps of the Gibbs sampler. Note that the difficulties of steps (11) and (12) are avoided by Metropolis-Hastings algorithms.

**Class membership and Gaussian vector sampling** The sampling from (11) is performed in two steps by using independence between the individuals

which involves that

$$p(\mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{(r-1)}) = \prod_{i=1}^{n} p(z_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)}) p(\boldsymbol{y}_i|\boldsymbol{x}_i, z_i, \boldsymbol{\theta}^{(r-1)}). \tag{15}$$

Firstly, each $z_i^{(r)}$ is independently sampled from the multinomial distribution

$$z_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)} \sim \mathcal{M}_g(t_{i1}(\boldsymbol{\theta}^{(r-1)}), \ldots, t_{ig}(\boldsymbol{\theta}^{(r-1)})), \tag{16}$$

where $t_{ik}(\boldsymbol{\theta}^{(r-1)}) = \frac{\pi_k^{(r-1)} p(\boldsymbol{x}_i; \boldsymbol{\alpha}_k^{(r-1)})}{p(\boldsymbol{x}_i; \boldsymbol{\theta}^{(r-1)})}$. Note that $t_{ik}(\boldsymbol{\theta}^{(r-1)})$ is the posterior probability that $\boldsymbol{x}_i$ has been arisen from component $k$ with the parameters $\boldsymbol{\theta}^{(r-1)}$.

Secondly, each $\boldsymbol{y}_i^{(r-1/2)}$ is independently sampled given $(\boldsymbol{x}_i, \boldsymbol{z}_i^{(r)}, \boldsymbol{\theta}^{(r-1)})$. Its first $c$ elements, denoted by $\boldsymbol{y}_i^{\mathrm{C}(r-1/2)}$, are deterministically defined by $\boldsymbol{y}_i^{\mathrm{C}(r-1/2)} = \Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_{z_i^{(r)}}^{(r-1)})$. Its last $d$ elements, denoted by $\boldsymbol{y}_i^{\mathrm{D}(r-1/2)}$, are sampled from the $d$-variate Gaussian distribution $\mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_{z_i^{(r)}}^{(r-1)})$ truncated on the space $\mathcal{S}_{z_i^{(r)}}(\boldsymbol{x}_i^{\mathrm{D}})$

$$p(\boldsymbol{y}_i^{\mathrm{D}}|\boldsymbol{x}_i, \boldsymbol{y}_i^{\mathrm{C}(r-1/2)}, z_i^{(r)}, \boldsymbol{\theta}^{(r-1)}) \propto \phi_d(\boldsymbol{y}_i^{\mathrm{D}}; \boldsymbol{\mu}_{z_i^{(r)}}^{\mathrm{D}(r-1)}, \boldsymbol{\Sigma}_{z_i^{(r)}}^{\mathrm{D}(r-1)}) \mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{D}} \in \mathcal{S}_{z_i^{(r)}}(\boldsymbol{x}_i^{\mathrm{D}})\}}, \tag{17}$$

where $\boldsymbol{\mu}_{z_i^{(r)}}^{\mathrm{D}(r-1)} = \boldsymbol{\Gamma}_{z_i^{(r)}\mathrm{DC}}^{(r-1)} \boldsymbol{\Gamma}_{z_i^{(r)}\mathrm{CC}}^{-1(r-1)} \boldsymbol{y}_i^{\mathrm{C}(r-1/2)}$.

**Remark.** The computation of $t_{ik}(\boldsymbol{\theta}^{(r-1)})$ involves to compute the integral defined in (4) which can be time consuming if $d$ is large ($d > 6$). In such a case, the sampling from (11) is replaced by one iteration of the Metropolis-Hastings algorithm detailed in Appendix C.1.

**Margin parameter and Gaussian vector sampling** The sampling from (12) is performed by using the following decomposition

$$p(\boldsymbol{\beta}_{kj}, \boldsymbol{y}^j_{[rk]}|\mathbf{x}, \boldsymbol{y}^{\bar{j}(r)}_{[rk]}, \mathbf{z}^{(r)}, \boldsymbol{\beta}^{(r)}_{k\bar{j}}, \boldsymbol{\Gamma}^{(r-1)}_k) = p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}^{\bar{j}(r)}_{[rk]}, \mathbf{z}^{(r)}, \boldsymbol{\beta}^{(r)}_{k\bar{j}}, \boldsymbol{\Gamma}^{(r-1)}_k) \times p(\boldsymbol{y}^j_{[rk]}|\mathbf{x}, \boldsymbol{y}^{\bar{j}(r)}_{[rk]}, \mathbf{z}^{(r)}, \boldsymbol{\beta}^{(r)}_{k\bar{j}}, \boldsymbol{\beta}_{kj}, \boldsymbol{\Gamma}^{(r-1)}_k). \quad (18)$$

The parameter $\boldsymbol{\beta}^{(r)}_{kj}$ is firstly sampled. The full conditional distribution of $\boldsymbol{\beta}_{kj}$ is defined up to a normalizing constant such as

$$p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}^{\bar{j}(r)}_{[rk]}, \mathbf{z}^{(r)}, \boldsymbol{\beta}^{(r)}_{k\bar{j}}, \boldsymbol{\Gamma}^{(r-1)}_k) \propto p(\boldsymbol{\beta}_{kj}) \prod_{\{i:z^{(r)}_i=k\}} p(x^j_i|\boldsymbol{y}^{\bar{j}(r)}_i, z^{(r)}_i, \boldsymbol{\Gamma}^{(r-1)}_k, \boldsymbol{\beta}_{kj}). \quad (19)$$

The distribution of $x^j_i|\boldsymbol{y}^{\bar{j}(r)}_i, z^{(r)}_i, \boldsymbol{\Gamma}^{(r-1)}_k$ with $z^{(r)}_i = k$ is defined by

$$p(x^j_i|\boldsymbol{y}^{\bar{j}(r)}_i, z^{(r)}_i, \boldsymbol{\Gamma}^{(r-1)}_k, \boldsymbol{\beta}_{kj}) = \begin{cases} \phi_1(\frac{x^j_i-\mu_{kj}}{\sigma_{kj}}; \tilde{\mu}_i, \tilde{\sigma}^2_i)/\sigma_{kj} & \text{if } 1 \leq j \leq c \\ \Phi_1(\frac{b^{\oplus}(x^j_i)-\tilde{\mu}_i}{\tilde{\sigma}_i}) - \Phi_1(\frac{b^{\ominus}(x^j_i)-\tilde{\mu}_i}{\tilde{\sigma}_i}) & \text{otherwise,} \end{cases} \quad (20)$$

where the real $\tilde{\mu}_i = \boldsymbol{\Gamma}^{(r-1)}_k[j, \bar{j}]\boldsymbol{\Gamma}^{(r-1)}_k[\bar{j}, \bar{j}]^{-1}\boldsymbol{y}^{\bar{j}(r)}_i$ is the full conditional mean of $y^j_i$, $\boldsymbol{\Gamma}_k[j, \bar{j}]$ 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 $\tilde{\sigma}^2_i$ is the full conditional variance of $y^j_i$ defined by $\tilde{\sigma}^2_i = 1 - \boldsymbol{\Gamma}^{(r-1)}_k[j, \bar{j}]\boldsymbol{\Gamma}^{(r-1)}_k[\bar{j}, \bar{j}]^{-1}\boldsymbol{\Gamma}^{(r-1)}_k[\bar{j}, j]$. As the normalizing constant of (19) is unknown, $\boldsymbol{\beta}^{(r)}_{kj}$ cannot be directly sampled. This problem is avoided by one iteration of the Metropolis-Hastings algorithm detailed in Appendix C.2.

The vector $\boldsymbol{y}^{j(r)}_{[rk]}$ is easily sampled after $\boldsymbol{\beta}^{(r)}_{kj}$. Indeed, by using independence between the individuals, the full conditional distribution of $\boldsymbol{y}^j_{[rk]}$ is explicitly

defined by

$$p(\boldsymbol{y}^j_{[rk]}|\mathbf{x}, \boldsymbol{y}^{\bar{\jmath}(r)}_{[rk]}, \mathbf{z}^{(r)}, \boldsymbol{\beta}^{(r)}_{k\bar{\jmath}}, \boldsymbol{\beta}_{kj}, \boldsymbol{\Gamma}^{(r-1)}_k) = \prod_{\{i:z_i^{(r)}=k\}} p(y^j_i|x^j_i, \boldsymbol{y}^{\bar{\jmath}(r)}_i, z^{(r)}_i, \boldsymbol{\beta}_{kj}, \boldsymbol{\Gamma}^{(r-1)}_k). \quad (21)$$

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 a Dirac distribution in $\frac{x_i^j - \mu_{kj}^{(r)}}{\sigma_{kj}^{(r)}}$. 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,

$$p(y^j_i|x^j_i, \boldsymbol{y}^{\bar{\jmath}(r)}_i, z^{(r)}_i, \boldsymbol{\beta}^{(r)}_{kj}, \boldsymbol{\Gamma}^{(r-1)}_k) = \frac{\phi_1(y^j_i; \tilde{\mu}_i, \tilde{\sigma}^2_i)}{p(x^j_i; \boldsymbol{\beta}^{(r)}_{kj})} \mathbb{1}_{\{y_i^j \in [b_k^{\ominus(r)}(x_i^j), b_k^{\oplus(r)}(x_i^j)]\}}, \quad (22)$$

where $b_k^{\ominus(r)}(x_i^j) = P(x_i^j - 1; \boldsymbol{\beta}^{(r)}_{kj})$ and $b_k^{\oplus(r)}(x_i^j) = P(x_i^j; \boldsymbol{\beta}^{(r)}_{kj})$.

So, step (12) is performed in two steps. Firstly, $\boldsymbol{\beta}^{(r)}_{kj}$ is sampled via one iteration of the Metropolis-Hastings algorithm whose the stationary distribution is $p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}^{\bar{\jmath}(r)}_{[rk]}, \mathbf{z}^{(r)}, \boldsymbol{\beta}^{(r)}_{k\bar{\jmath}}, \boldsymbol{\Gamma}_k)$. Secondly, $\boldsymbol{y}^{j(r)}_{[rk]}$ is sampled from (22).

**Vector of proportions sampling** The sampling from (13) is classical. Indeed, the conjugate Jeffreys non informative prior involves that

$$\boldsymbol{\pi}|\mathbf{z}^{(r)} \sim \mathcal{D}_g\left(n_1^{(r)} + \frac{1}{2}, \ldots, n_g^{(r)} + \frac{1}{2}\right), \quad (23)$$

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

**Correlation matrix sampling** To sample from (14), we use the approach proposed by [Hof07] in the case of semiparameteric Gaussian copula. Firstly,

a covariance matrix is generated by its explicit posterior distribution, and secondly, the correlation matrix is deduced by normalizing the covariance matrix. As $(\mathbf{y}, \mathbf{z})$ are known in this step, 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 following two steps

$$\mathbf{\Lambda}_k|\mathbf{y}^{(r)}, \mathbf{z}^{(r)} \sim \mathcal{W}^{-1}\left(s_0 + n_k^{(r-1)}, S_0 + \sum_{\{i:z_i^{(r)}=k\}} \boldsymbol{y}_i^{(r)T}\boldsymbol{y}_i^{(r)}\right), \tag{24}$$

where $\forall 1 \leq h, \ell \leq e,\ \mathbf{\Gamma}_k[h,\ell] = \frac{\mathbf{\Lambda}_k[h,\ell]}{\sqrt{\mathbf{\Lambda}_k[h,h]\mathbf{\Lambda}_k[\ell,\ell]}}$. As the homoscedastic model assumes the equality between the correlation matrices, in such a case we only sample one $\mathbf{\Lambda}$ so (24) is replaced by

$$\mathbf{\Lambda}|\mathbf{y}^{(r)}, \mathbf{z}^{(r)} \sim \mathcal{W}^{-1}\left(s_0 + n, S_0 + \sum_{i=1}^{n} \boldsymbol{y}_i^{(r)T}\boldsymbol{y}_i^{(r)}\right), \tag{25}$$

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

## 3.3 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 with probability 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 does not occurred 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

Two simulations illustrate the new model. The first simulation illustrates the relevance of the estimates by analyzing data which arise from the proposed model. The second simulation shows the robustness of the proposed model by analyzing data which arise from a mixture of Poisson distributions [KT08].

**Experiment conditions** The Gibbs sampler estimates the parameters on 100 samples for each situation. It is initialized with the maximum likelihood estimator of the locally independent model (especially relevant when the intra-class dependencies are small). Its burn-in lasts 100 iterations, then it is stopped after 1000 iterations. The estimate is computed by averaging the parameters sampled by the Gibbs algorithm. The Kullback-Leibler divergence is approximated via 10000 iterations of a Monte-Carlo method.

## 4.1 Estimation efficiency

Data are composed of one continuous variable, one integer variable and one binary variable. They are sampled from the example in Section 2.3. According to Figure 2, the estimated distribution converges to the true distribution when the sample size increases. Indeed, the Kullback-Leibler divergence of the estimated model from the true model is decreasing based on the sample size. This simulation illustrates the convergence of the estimator computed by averaging the parameters sampled by the algorithm.

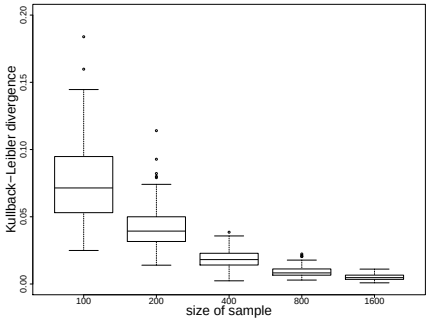



Figure 2: Decrease of the Kullback-Leibler divergence of the estimated model from the true model based on the sample size.

## 4.2 Robustness

Data are sampled from the bivariate Poisson mixture model [KT08] with $\boldsymbol{\pi} = (1/3, 2/3)$ and whose the one-dimensional margin parameters $\boldsymbol{\alpha}_k = (\lambda_{k1}, \lambda_{k2}, \lambda_{k3})$ takes the following values: $\lambda_{1h} = h$ and $\lambda_{2h} = 3 + h$, for $h = 1, 2, 3$. Its error rate is equal to 9.5%. Figure 3 shows that the mixture model of Gaussian copulas efficiently manages these data. Indeed, the Kullback-Leibler divergence almost vanishes when the size of the sample increases. Furthermore,

the error rate of the model converges to a value slightly larger than Bayes' error. We note that the parameters (one-dimensional margin parameters and correlation coefficients) are also well estimated.

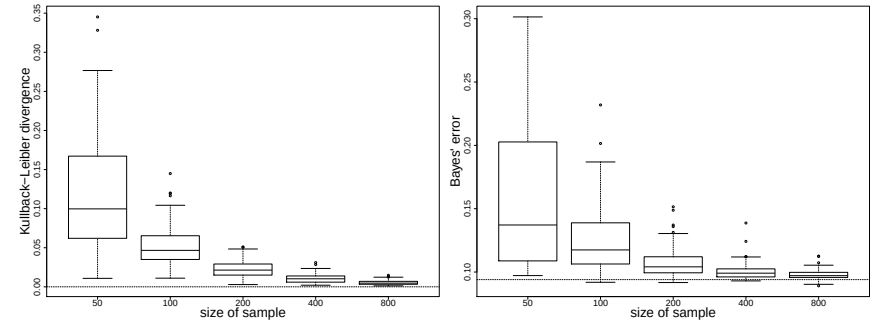



Figure 3: (a) Kullback-Leibler divergence of the estimated model from the true model; (b) Error rate of the estimated model; (c) Value of the first one-dimensional margin parameter for class 1; (d) Value of the correlation coefficient between both variables for class 1.

# 5 Applications

The analysis of two real data sets are now performed by using the mixture model of Gaussian copulas. The Gibbs sampler is used in the same conditions as in the simulations. The model selection is performed by using two infor-

mation criteria (BIC criterion [Sch78], ICL criterion [BCG00]) computed from the estimator.

## 5.1 Wine data set

**The data** Data are composed of 6497 variants of the Portuguese "Vinho Verde" wine (1599 red wines and 4898 white wines) described by eleven physicochemical continuous variables (fixed acidity, volatile acidity, citric acidity, residual sugar, chlorides, free sulphur dioxide, total density dioxide, density, pH, sulphates, alcohol) and one integer variable (note of quality) [CCA+09]. We analyze the data by concealing the color of the wines and by excluding the wine 4381 because it takes outliers.

**Model selection** Three mixture models (locally independent, heteroscedastic and homoscedastic mixture of Gaussian copulas) are fitted with different numbers of classes. Table 1 presents the values of the BIC and the ICL criteria. Both criteria distinctly select the bi-component mixture model of Gaussian copulas with free correlation matrices.

| | 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 1: Values of the BIC and ICL criteria obtained on the wine data set.

**Partition study** Table 2 presents the adjusted Rand indices and the confusion matrices which compare the estimated partitions and the color of the wines. The partition of the bi-component mixture model of Gaussian copulas with free correlation matrices is the closest to the partition of the color of the wines. These results confirm that this model best manages this data set. Moreover, this model provides well-separated classes as shown by Figure 4 which presents its PCA-type visualization per class.

| | white | red |
|---|---|---|
| c1 | 4359 | 9 |
| c2 | 538 | 1590 |

(a) Adj. Rand.: 0.68

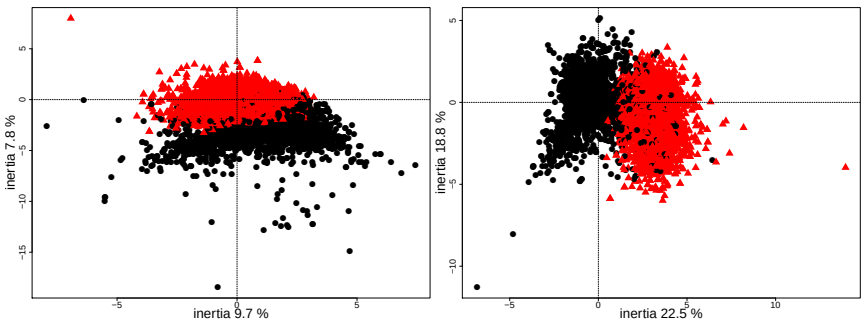
| | white | red |
|---|---|---|
| c1 | 2441 | 12 |
| c2 | 1911 | 7 |
| c3 | 545 | 1580 |

(b) Adj. Rand.: 0.30

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

(c) Adj. Rand.: 0.00

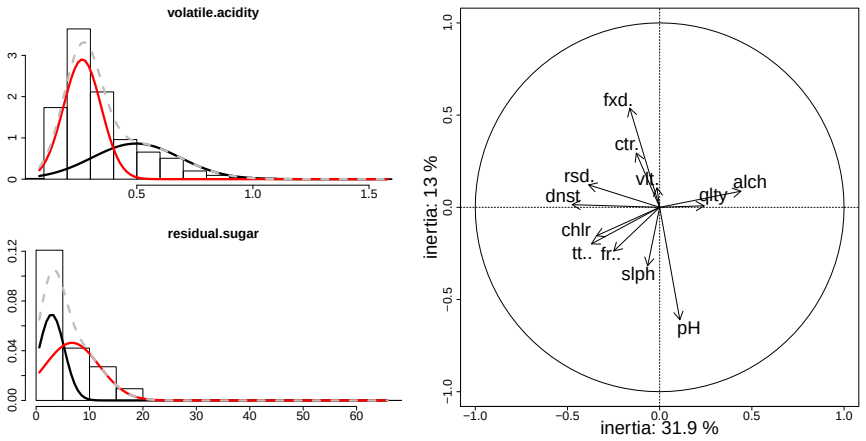
Table 2: Adjusted Rand indices and the confusion matrices between the color of the wines and the estimated partition by: (a) the bi-component heteroscedastic mixture of Gaussian copulas; (b) the tri-component homoscedastic mixture of Gaussian copulas; (c) the four-component locally independent mixture.




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

Figure 4: Visualization of the partition obtained by bi-component mixture model of Gaussian copulas with free correlation matrices for the wine data set (Class 1 is drawn by black circles and Class 2 by red triangles).

**Interpretation of the best model** A three-level interpretation (proportions, one dimensional margins and intra-class dependencies) is feasible by using the parameters summarized by Figure 5. The majority class ($\pi_1 = 0.59$), consisting primarily of white wines, is characterized by lower levels of acidity, pH, chlorides and sulphites. This class is characterized by a strong correlation between both sulphur measures opposite to a strong correlation between the density and acidity measures. The minority class ($\pi_2 = 0.41$), consisting primarily of red wines, takes larger values for both sulphur dioxide measures and the alcoholic rate. In this class, the wine quality is correlated with a large alcoholic measure and small values for the chlorides and acidity measures. Note that the wine quality of both classes is similar ($\beta_{1\text{quality}} = 5.96$ and $\beta_{2\text{quality}} = 5.58$).




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

(b) Variables in the first correlation circle given by $\boldsymbol{\Gamma}_2$

Figure 5: Summary of bi-component mixture model of Gaussian copulas with free correlation matrices. Class 1 is drawn in black and Class 2 in red.

**Conclusion** On this data, the mixture model of Gaussian copulas reduces the drawbacks of the locally independent model. By reducing the number of classes, it provides a more interpretable model which better fits the data (information criteria) and which provides a pertinent partition (adjusted Rand Index, confusion matrices, class well-separated). Finally, the estimation of the main intra-class dependencies, based on the outputs of the PCA per class, is an efficient tool to refine the interpretation.

## 5.2 Forest fire data set

**The data** Data are composed of 517 forest fires [CM07] that have occured in the northeast region of Portugal. These forest fires are described by the following 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 relating to spatial coordinates and three binary ones indicating the presence of rain, the season (summer or not summer) and the day (weekend or not weekend).

**Model selection** Table 3 presents the values of the BIC and the ICL criteria obtained by the three competing models. Both criteria distinctly select the tri-component mixture model of Gaussian copulas with equal correlation matrices.

**Partition study** Model selection is a crucial step since the three competing models lead different partitions. Indeed, these differences are highlighted in Table 4 which presents the confusing matrices.

| | 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 3: Values of the BIC and ICL criteria obtained on the forest fire data set.

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

(a)

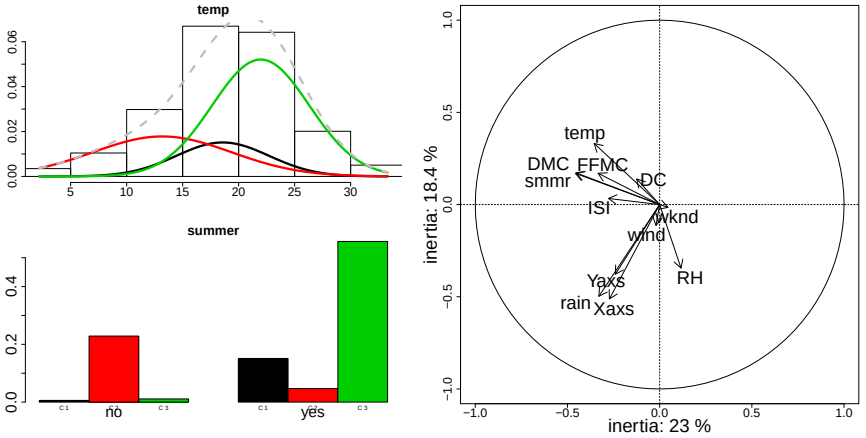
| | loc. indpt. | |
|---|---|---|
| | c1 | c2 |
| c1-homo. | 265 | 2 |
| c2-homo. | 7 | 121 |
| c3-homo. | 111 | 11 |

(b)

Table 4: 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.

**Interpretation of the best model** The three-step interpretation of the tri-component mixture model of Gaussian copulas with equal correlation matrices is presented by using the parameters summarized by Figure 6. The majority class ($\pi_1 = 0.57$) groups the fires which occured when a high temperature was coupled to a small relative humidity. Moreover, the measures of FMC, DMC and ISI are high. The second class ($\pi_2 = 0.26$) includes winter fires. These lights are developed through a strong wind and no rain. Moreover, all MFI 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 value. These fires occured when the temperature was median but when the relative humidity was high. The intra-class correlation matrix underlines the dependencies between the summer period and the high temperatures and between the 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 $\boldsymbol{\Gamma}$

Figure 6: Summary of the homoscedastic tri-component mixture of Gaussian copulas. Class 1 is displayed in green, Class 2 in red and Class 3 in black.

**Conclusion** The cluster analysis performed by the mixture of Gaussian copulas is more precise than the analysis of the locally independent model which roughly separates the summer fires from the other ones. The restrictions on the correlation matrices allows to better fit the data according to both criteria. Therefore, the homoscedastic mixture model of Gaussian copulas highlights two kinds of summer fires.

# 6 Conclusion and future extensions

The mixture model of Gaussian copulas is introduced to cluster mixed data. By using the Gaussian copulas, the one-dimensional margins of each compo-

nent follow classical distributions and the intra-class dependencies are modelized. Thus, the model can be interpreted in three steps like for the models developed for the data sets composed of one type of variable. By using the continuous latent variables of the Gaussian copulas, a PCA-type method allows a visualization of the individuals per class. Moreover, this approach provides a summary of the intra-class dependencies which can avoid the fastidious interpretation of the correlation matrices.

During the numerical experiments and during the applications, we pointed-out that this model is sufficiently robust to fit data arisen from another model. 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 of 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. However, the number of parameters required by this model can stay 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.

R-package *MixCluster* (downloadable on https://r-forge.r-project.org/R/?group_id=1939) contains code to perform the cluster analysis method described in the article. The package also contains all data sets used as examples in the article.

# 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.

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

[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.

[MCMM09] C. Maugis, G. Celeux, and M.L. Martin-Magniette. Variable selection in model-based clustering: A general variable role model-

ing. *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.

[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,*

$$\forall \boldsymbol{x} \in \mathbb{R}^c \times \mathbb{N}^d, \quad \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}; \boldsymbol{\alpha}_k) = \sum_{k=1}^{g'} \pi'_k p(\boldsymbol{x}; \boldsymbol{\alpha}'_k) \tag{26}$$

$$\Rightarrow \quad g = g',\ \boldsymbol{\pi} = \boldsymbol{\pi}',\ \boldsymbol{\alpha} = \boldsymbol{\alpha}'. \tag{27}$$

*Proof.* The identifiability of the multivariate Gaussian mixture models and of the univariate Poisson mixture model [Tei63, YS+68] involves that (26) implies

$$g = g',\ \boldsymbol{\pi} = \boldsymbol{\pi}',\ \boldsymbol{\beta}_{kj} = \boldsymbol{\beta}'_{kj} \text{ and } \boldsymbol{\Gamma}_{k\text{CC}} = \boldsymbol{\Gamma}'_{k\text{CC}}. \tag{28}$$

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 (26) 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 $(x^j, x^h) \in \gamma_t$,

$$\forall t, \quad \frac{\int_t^{a'_1} \phi(u)du}{\Phi(t)} = 0. \tag{29}$$

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 $\gamma_{(t,\xi)} = \{(x^j, x^h) \in \mathbb{N} \times \mathbb{N} : \ a_1 \in B(t, \xi)\}$. 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 $\frac{\int_{t+\xi}^{a'_1} \phi(u)du}{\Phi(t-\xi)} = 0$ and $\frac{\int_{t+\xi}^{a'_1} \phi(u)du}{\Phi(t-\xi)} > 0$. 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 (26) 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 $\boldsymbol{\Gamma}_{k\text{DD}} = \boldsymbol{\Gamma}'_{k\text{DD}}$. □

# B Prior distributions of $\boldsymbol{\beta}_k$

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

$$\sigma^2_{kj} \sim \mathcal{G}^{-1}(c_0, C_0) \text{ and } \mu_{kj}|\sigma^2_{kj} \sim \mathcal{N}_1(b_0, \sigma^2_{kj}/N_0), \tag{30}$$

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

$$\boldsymbol{\beta}_{kj} \sim \mathcal{G}(a_0, A_0). \tag{31}$$

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

$$\boldsymbol{\beta}_{kj} \sim \mathcal{D}_{m_j}\left(\frac{1}{2}, \ldots, \frac{1}{2}\right). \tag{32}$$

# C Metropolis-within-Gibbs sampler

The sampling from $\mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{(r-1)}$ and $\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}$ (defined by (11) and (19)) can be performed by one iteration of following Metropolis-Hastings algorithms. For both algorithms, the instrumental distributions assume conditional independences. So, the smaller are the intra-class dependencies of the variable $\boldsymbol{x}$, the closer of the stationary distributions are the instrumental distributions of both algorithms.

## C.1 Class membership and Gaussian vector sampling

Step (11) is performed via one iteration of the Metropolis-Hastings algorithm which independently samples each couple $(z_i, \boldsymbol{y}_i)$. Its stationary distribution is

$$p(z_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)}) \propto \pi_{z_i} p(\boldsymbol{x}_i, \boldsymbol{y}_i|z_i, \boldsymbol{\theta}^{(r-1)}). \tag{33}$$

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

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. Its first $c$ elements, denoted by $\boldsymbol{y}_i^{\star \mathrm{C}}$, are equal to $\boldsymbol{y}_i^{\star \mathrm{C}} = \Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_{z_i^\star}^{(r-1)})$. Its last $d$ elements, denoted by $\boldsymbol{y}_i^{\star \mathrm{D}}$, follows a *multivariate independent Gaussian* distribution truncated on $\mathcal{S}_{z_i^\star}(\boldsymbol{x}_i^{\mathrm{D}})$. Thus,

$$q_1(z_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)}) = \frac{1}{g} \frac{\phi_d(\boldsymbol{y}_i^{\mathrm{D}}; \mathbf{0}, \boldsymbol{I})}{\prod_{j=c+1}^{e} p(x_i^j; \boldsymbol{\beta}_{z_i j}^{(r-1)})} \mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{C}} = \Psi(\boldsymbol{x}_i^{\mathrm{C}}; \boldsymbol{\alpha}_{z_i}^{(r-1)})\}} \mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{D}} \in \mathcal{S}_{z_i}(\boldsymbol{x}_i^{\mathrm{D}})\}}. \tag{34}$$

The candidate is accepted with the probability

$$\rho_{1i}^{(r)} = \min \left\{ \frac{\pi_{z_i^\star} \phi_e(\boldsymbol{y}_i^\star; \mathbf{0}, \boldsymbol{\Gamma}_{z_i^\star}^{(r-1)})}{\pi_{z_i^{(r-1)}} \phi_e(\boldsymbol{y}_i^{(r-1)}; \mathbf{0}, \boldsymbol{\Gamma}_{z_i^{(r-1)}}^{(r-1)})} \frac{q_1(z_i^{(r-1)}, \boldsymbol{y}_i^{(r-1)}|\boldsymbol{x}_i)}{q_1(z_i^\star, \boldsymbol{y}_i^\star|\boldsymbol{x}_i)}; 1 \right\}. \tag{35}$$

Thus, at the iteration $(r)$ of the Algorithm 3.1, the sampling according to (11) is performed via one iteration of the following Metropolis-Hastings algorithm having $p(z_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)})$ as stationary distribution.

**Algorithm C.1.**

$$(z_i^\star, \boldsymbol{y}_i^\star) \sim q_1(z, \boldsymbol{y}|\boldsymbol{x}_i) \tag{36}$$

$$(z_i^{(r)}, \boldsymbol{y}_i^{(r-1/2)}) = \begin{cases} (z_i^\star, \boldsymbol{y}_i^\star) & \text{with probability } \rho_{1i}^{(r)} \\ (z_i^{(r-1)}, \boldsymbol{y}_i^{(r-1)}) & \text{with probability } 1 - \rho_{1i}^{(r)}. \end{cases} \tag{37}$$

## C.2 Margin parameter sampling

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

Thus, according to (19), the candidate $\boldsymbol{\beta}_{kj}^{\star}$ is accepted with the probability

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

Thus, at the iteration $(r)$ of the Algorithm 3.1, step (12) is performed via one iteration of the following Metropolis-Hastings algorithm whose the stationary distribution is $p(\boldsymbol{\beta}_{kj}|\boldsymbol{x}_{[rk]}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\Gamma}_k)$.

**Algorithm C.2.**

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

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