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[a,b] and Christophe Biernacki[a,b,c] and Vincent Vandewalle[a,d]
[a]Inria Lille; [b]University Lille 1; [c]CNRS; [d]EA 2694 University Lille 2

December 16, 2016

Abstract

Clustering of mixed data is important yet challenging due to a shortage of conventional distributions for such data. In this paper, we propose a mixture model of Gaussian copulas for clustering mixed data. Indeed copulas, and Gaussian copulas in particular, are powerful tools for easily modeling the distribution of multivariate variables. This model clusters data sets with continuous, integer and ordinal variables (all having a cumulative distribution function) by considering the intra-component dependencies in a similar way to the Gaussian mixture. Indeed, each component of the Gaussian copula mixture produces a correlation coefficient for each pair of variables and its univariate margins follow standard distributions (Gaussian, Poisson and ordered multinomial) depending on the nature of the variable (continuous, integer or ordinal). As an interesting by-product, this model generalizes many well-known approaches and provides tools for visualization based on its parameters. The Bayesian inference is achieved with a Metropolis-within-Gibbs sampler. The numerical experiments, on simulated and real data, illustrate the benefits of the proposed model: flexible and meaningful parametrization combined with visualization features.

keywords: Clustering; Gaussian copula; Metropolis-within-Gibbs algorithm; Mixed data; Mixture models; Visualization.

1 Introduction

In a probabilistic framework, clustering is often managed by modeling the distribution of the observed variables using finite mixture models of parametric distributions (McLachlan and Peel, 2000). A class is defined as the subset of the individuals arising from the same mixture component. The literature covering homogeneous data (composed of variables of the same type) is extensive and presents Gaussian mixture models (Banfield and Raftery, 1993), multinomial mixture models (Goodman, 1974) and Poisson mixture models (Karlis and Tsiamyrtzis, 2008) as the standard models used to cluster such data sets. The use of conventional distributions for mixture components explains the success of these models, since the components can be easily interpreted. Although many data sets contain mixed data (variables of different types), few mixture models can manage these data (Hunt and Jorgensen, 2011) due to the shortage of multivariate distributions.

The *locally independent mixture model* (Moustaki and Papageorgiou, 2005; Lewis, 1998; Hand and Yu, 2001) is a convenient approach for clustering mixed data since it assumes independence within-component between variables. Thus, each component is defined by a product of standard univariate distributions that facilitate their interpretation. However, this model can lead to severe bias when its main assumption is violated (Van Hattum and Hoijtink, 2009). Therefore, two models have been introduced to relax this assumption.

The *location mixture model* (Krzanowski, 1993; Willse and Boik, 1999) has been proposed for clustering a data set with continuous and categorical variables. It assumes that, for each component, the categorical variables follow a multinomial distribution and the continuous variables follow a multivariate Gaussian distribution conditionally on the categorical variables. Therefore, the intra-component dependencies are taken into account. However, the model requires too many parameters. Hunt and Jorgensen (1999) extended this approach by splitting the variables into within-component independent blocks. Each block contains no more than one categorical variable and follows a location model. The interpretation of this model can be complex since, for a given component, the univariate marginal of a continuous variable follows a Gaussian mixture model. Moreover, the estimation of the repartition of the variables into blocks is a difficult problem that the authors achieve with an ascending method that is sub-optimal.

The *underlying variables mixture model* (Everitt, 1988) has been proposed for clustering a data set with continuous and ordinal variables. It assumes that each ordinal variable arises from a latent continuous variable and that all continuous variables (observed and unobserved) follow a Gaussian mixture model. The distribution of observed variables is obtained by integrating each Gaussian component into the subset of latent variables. However, in practice, this computation is not feasible when there are more than two ordinal variables. In an effort to study data sets with numerous binary variables, Morlini (2012) expanded this model by estimating the scores of latent variables from those of binary variables. However, the interpretation of the mixture components can be complex since it is based on the score-related parameters (not those related to observed variables).

Previous models illustrate the difficulty for clustering mixed data with a model for which interpretation and inference are easy. Moreover, they do not take account of cases where some variables are integer. The main difficulty is due to a shortage of conventional distributions for mixed data. However, copulas are standard tools for defining multivariate distributions in a systematic way, and they therefore have good potentiality for providing a sensible answer.

Copulas (Joe, 1997; Nelsen, 1999) can be used to build a multivariate model by defining, on the one hand, the *univariate marginal distributions*, and, on the other, the *dependency model*. Recently, Smith and Khaled (2012) and Murray et al. (2013) modeled the distribution of mixed variables using one Gaussian copula. As pointed out by Pitt et al. (2006), the maximum likelihood inference is very difficult for a Gaussian copula with discrete margins. Therefore, it is often replaced by the *Inference Function for Margins* method performing the inference in two steps (Joe, 1997, 2005). When all the variables are continuous, the fixed-point-based algorithm proposed by Song et al. (2005) achieves the maximum likelihood estimation, but this approach is not doable for mixed data. Therefore, as shown by Smith and Khaled (2012), it is more convenient to work in a Bayesian framework since this simplifies the inference by using the latent structure of the model.

In this paper, the *Gaussian copula mixture model* is introduced for clustering mixed data. This new model assumes that each component follows a Gaussian copula (Hoff, 2007; Hoff et al., 2011). Thus, each component takes account of the *dependencies* in a similar way to the Gaussian mixture since it provides a correlation matrix. Moreover, the univariate margins of each component can follow *conventional distributions* to facilitate model interpretation. Hence, each component is also easily interpreted using its proportion indicating its weight, its univariate margin parameters and its correlation matrix. Finally, the continuous latent structure of the Gaussian copulas permits visualization based on a Principal Component Analysis (PCA) per component. This visualization provides a summary of the main within-component dependencies and a scatter-plot of individuals according to component parameters.

This paper is organized as follows. Section 2 introduces the Gaussian copula mixture model and explains its links to well-known models. Section 3 presents the Metropolis-within-Gibbs algorithm used to perform Bayesian inference. Section 4 illustrates the algorithm behavior and the model robustness through numerical experiments. Section 5 presents two applications of the new model by clustering real data sets. Section 6 concludes the study.

2 Mixture model of Gaussian copulas

2.1 Finite mixture model

An observation $\boldsymbol{x} = (x^1, \ldots, x^e) \in \mathbb{R}^c \times \mathcal{X}$ is composed of $e = c+d$ variables. Its first c elements, denoted by $\boldsymbol{x}^{\mathrm{c}}$, correspond to the subset of the continuous variables defined on the space $\mathbb{R}^c$. Its last d elements, denoted by $\boldsymbol{x}^{\mathrm{d}}$, 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 levels, then it uses a numeric coding $\{1, \ldots, m_j\}$. An observation is assumed to arise from the mixture model of g parametric distributions whose probability distribution function (pdf) is defined by

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

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ denotes the whole parameters. Vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ is defined on the simplex of size g and groups the component proportions, where π_k is the proportion of component k. Vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ groups the component parameters, where $\boldsymbol{\alpha}_k$ denotes the parameters of component k.

2.2 Component modeled by a Gaussian copula

The Gaussian copula mixture model considers that each component follows a Gaussian copula. Component k is also parametrized by $\boldsymbol{\alpha}_k = (\boldsymbol{\Gamma}_k, \boldsymbol{\beta}_k)$ where $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, \ldots, \boldsymbol{\beta}_{ke})$ groups the parameters of the univariate margin, $\boldsymbol{\beta}_{kj}$ being the parameters of the j-th univariate margin, and where $\boldsymbol{\Gamma}_k$ is the correlation matrix of size $e \times e$. The cumulative distribution function (cdf) of component k is written as

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

where $u_k^j = P(x^j|\boldsymbol{\beta}_{kj})$ is the value of the cdf of the univariate marginal distribution of variable j for component k evaluated at x^j, where $\Phi_e(.|\boldsymbol{0}, \boldsymbol{\Gamma}_k)$ is the cdf of the e-variate centred Gaussian distribution with correlation matrix $\boldsymbol{\Gamma}_k$ and where $\Phi_1^{-1}(.)$ is the inverse cumulative distribution function of the standard univariate Gaussian $\mathcal{N}_1(0, 1)$.

The model implies the categorical variable $z \in \{1, \ldots, g\}$ which follows the multinomial distribution $\mathcal{M}_g(\pi_1, \ldots, \pi_g)$ and which indicates the individual's component membership. In cluster analysis, the realization z is not observed while $\boldsymbol{x}$ is observed. Hence, mixture models are often interpreted as the marginal distribution of $\boldsymbol{x}$ based on the distribution of the variable pair $(\boldsymbol{x}, z)$. The Gaussian copula mixture model involves a second latent variable $\boldsymbol{y} = (y^1, \ldots, y^e) \in \mathbb{R}^e$, such that $\boldsymbol{y}|z = k$ follows an e-variate centered Gaussian distribution $\mathcal{N}_e(\boldsymbol{0}, \boldsymbol{\Gamma}_k)$. Thus, the Gaussian copula mixture can be interpreted as the marginal distribution of $\boldsymbol{x}$ based on the distribution of the variable triplet $(\boldsymbol{x}, \boldsymbol{y}, z)$. Conditionally on $(\boldsymbol{y}, z = k)$, each element of $\boldsymbol{x}$ is defined by

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

Thus, the generative model of the Gaussian copula mixture is written as

- 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).

2.3 Specific distributions for mixed-type variables

The cdf of component k defined by (2) implies the cdf of the univariate marginal distributions. Hence, it requires the definition of the distributions of the univariate margins (*i.e.* distribution of $x^j|z = k$). We use conventional parametric distributions to facilitate the component interpretation. The parameters of margin j for component k are denoted by β_{kj}. Hence,

- if x^j is continuous then $x^j|z = k$ follows a *Gaussian* distribution with mean μ_{kj} and variance σ_{kj}^2 and $\beta_{kj} = (\mu_{kj}, \sigma_{kj})$,
- if x^j is integer then $x^j|z = k$ follows a *Poisson* distribution with parameter $\beta_{kj} \in \mathbb{R}^{+*}$,
- if x^j is ordinal then $x^j|z = k$ follows an *ordered multinomial* distribution with parameter β_{kj} defined on the simplex of size m_j. Note that the order between the levels is crucial since it permits the definition of the cdf.

Given that the first c variables of $\boldsymbol{x}$ ($\boldsymbol{x}^{\mathrm{C}}$) are continuous while the last d variables ($\boldsymbol{x}^{\mathrm{D}}$) are discrete, the pdf of component k can be decomposed as

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

We use the decomposition into sub-matrices $\boldsymbol{\Gamma}_k = \begin{bmatrix} \boldsymbol{\Gamma}_{k\mathrm{CC}} & \boldsymbol{\Gamma}_{k\mathrm{CD}} \\ \boldsymbol{\Gamma}_{k\mathrm{DC}} & \boldsymbol{\Gamma}_{k\mathrm{DD}} \end{bmatrix}$, 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. Under component k, the knowledge of the continuous variable x^j implies that $y^j = \frac{x^j - \mu_{kj}}{\sigma_{kj}}$. Denoting $\boldsymbol{y}^{\mathrm{C}} = \left(\frac{x^j - \mu_{kj}}{\sigma_{kj}}; j = 1, \ldots, c\right)$, $p(\boldsymbol{x}^{\mathrm{C}}|\boldsymbol{\alpha}_k) = \frac{\phi_c(\boldsymbol{y}^{\mathrm{C}}|\mathbf{0}, \boldsymbol{\Gamma}_{k\mathrm{CC}})}{\prod_{j=1}^{c} \sigma_{kj}}$ where $\phi_c(.|\mathbf{0}, \boldsymbol{\Gamma}_{k\mathrm{CC}})$ denotes the pdf of c-variate Gaussian distribution with mean $\mathbf{0}$ correlation matrix $\boldsymbol{\Gamma}_{k\mathrm{CC}}$. If the variable j is discrete, any value y^j in the interval $\mathcal{S}_k^j(x^j) =]b_k^{\ominus}(x^j), b_k^{\oplus}(x^j)]$ produces the same observation x^j under component k, where $b_k^{\ominus}(x^j) = \Phi_1^{-1}(P(x^j - 1|\boldsymbol{\beta}_{kj}))$ and $b_k^{\oplus}(x^j) = \Phi_1^{-1}(P(x^j|\boldsymbol{\beta}_{kj}))$. Under component k, the distribution of the continuous latent variable $\boldsymbol{y}^{\mathrm{D}} = \left(y^j; j = c + 1, \ldots, e\right)$ conditionally on $\boldsymbol{y}^{\mathrm{C}}$ is a Gaussian distribution with mean $\boldsymbol{\mu}_k^{\mathrm{D}} = \boldsymbol{\Gamma}_{k\mathrm{DC}} \boldsymbol{\Gamma}_{k\mathrm{CC}}^{-1} \Psi(\boldsymbol{x}^{\mathrm{C}}; \boldsymbol{\alpha}_k)$ and covariance matrix $\boldsymbol{\Sigma}_k^{\mathrm{D}} = \boldsymbol{\Gamma}_{k\mathrm{DD}} - \boldsymbol{\Gamma}_{k\mathrm{DC}} \boldsymbol{\Gamma}_{k\mathrm{CC}}^{-1} \boldsymbol{\Gamma}_{k\mathrm{CD}}$. Thus, the pdf of component k is written as

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

where $\mathcal{S}_k(\boldsymbol{x}^{\mathrm{D}}) = \mathcal{S}_k^{c+1}(x^{c+1}) \times \ldots \times \mathcal{S}_k^{e}(x^{e})$.

Remark 2.1 (Model identifiability). The Gaussian copula mixture model is identifiable if at least one variable is continuous or integer (see Appendix A).

2.4 Strengths of the Gaussian copula mixture model

2.4.1 Related models

The Gaussian copula mixture model generalizes many conventional mixture models, including the four cases mentioned below.

- If the correlation matrices are diagonal (*i.e.* $\boldsymbol{\Gamma}_k = \boldsymbol{I}$, $\forall k = 1, \ldots, g$), then the model is equivalent to the locally independent mixture model.

- If all the variables are continuous (*i.e.* $c = e$ and $d = 0$), then the model is equivalent to the Gaussian mixture model without constraints among parameters (Banfield and Raftery, 1993). In the spirit of the homescedastic Gaussian mixture, we also propose a parsimonious version of the Gaussian copula mixture model by assuming equality between the correlation matrices over component. This model is named *homoscedastic* since the covariance matrices of the latent Gaussian variables are equal between components (*i.e.* $\boldsymbol{\Gamma}_1 = \ldots = \boldsymbol{\Gamma}_g$). The free correlation model will be now called the *heteroscedastic* model).

- The model is linked to the binned Gaussian mixture model. For example, when variables are ordinal, it is equivalent to the mixture model presented by Gouget (2006). In such cases, the model is stable through fusion of modalities.

- If the variables are both continuous and ordinal, then the model is a new parametrization of the model proposed by Everitt (1988). It should be noted that Everitt directly estimates the space $\mathcal{S}_k(\boldsymbol{x}^{\mathrm{D}})$ containing the antecedents of $\boldsymbol{x}^{\mathrm{D}}$. Moreover, he uses a simplex algorithm to perform maximum likelihood inference, but this method dramatically limits the number of ordinal variables. The new parametrization of the proposed mixture allows the univariate marginal parameters $\boldsymbol{\beta}_{kj}$ of each component to directly estimate (see details in Section 3), whereas Everitt's parametrization implies a difficult estimation of the bounds of integration. Thus, the parameter inference is easier.

2.4.2 Standardized coefficient of correlation per class

The Gaussian copula provides a user-friendly correlation coefficient for each pair of variables. Indeed, when both variables are continuous, it is equal to the upper boundary of the correlation coefficients obtained by monotonic transformation of the variables (Klaassen and Wellner, 1997). Furthermore, when both variables are discrete, it is equal to the polychoric correlation coefficient (Olsson, 1979).

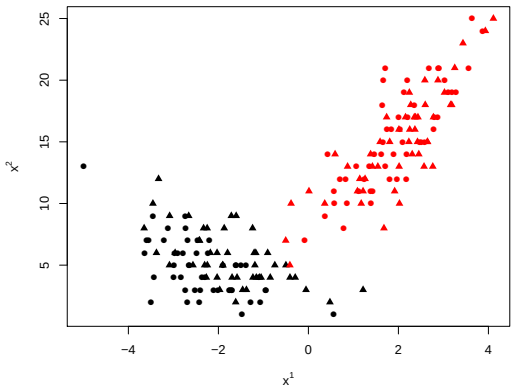
2.4.3 Data visualization per component: a by-product of Gaussian copulas

By using the latent vectors of the Gaussian copulas $\boldsymbol{y}|z$, a PCA-type method allows *visualization* of the individuals *per component* which permits the identification of main within-component dependencies. The visualization of component k is performed by computing the coordinates $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}, z = k; \boldsymbol{\alpha}_k]$ and then projecting them onto the PCA region associated with the Gaussian copula of component k. This space is obtained directly through spectral decomposition of $\boldsymbol{\Gamma}_k$. The individuals arising from component k follow a centered Gaussian distribution on this factorial map. Those arising from another component have an expectation not equal to zero. Therefore, an individual located far away from the origin arises from a distribution significantly different from the distribution of component k. Finally, the correlation circle summarizes the within-component correlations and avoids the direct interpretation of the correlation matrix $\boldsymbol{\Gamma}_k$, which can be tedious if e is large. The following example illustrates these properties.

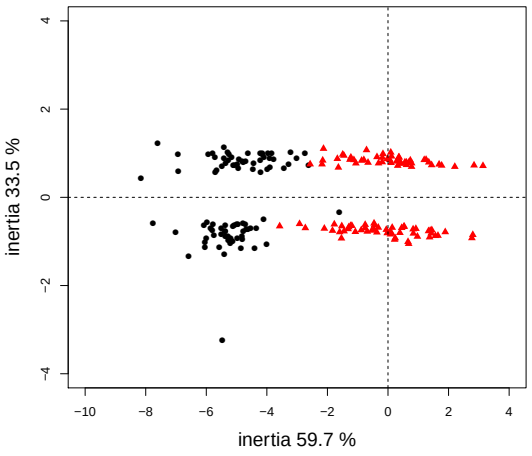
Example 2.2. Let three variables—one continuous, one integer and one binary—arise, in this order, from the bi-component Gaussian copula mixture model parametrized by

$$\boldsymbol{\pi} = (0.5, 0.5),\ \boldsymbol{\beta}_{11} = (-2, 1),\ \boldsymbol{\beta}_{12} = 5,\ \boldsymbol{\beta}_{13} = \boldsymbol{\beta}_{23} = (0.5, 0.5),\ \boldsymbol{\beta}_{21} = (2, 1),$$

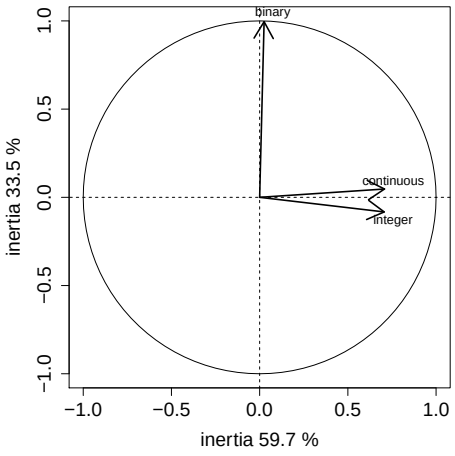
$$\boldsymbol{\beta}_{22} = 15,\ \boldsymbol{\Gamma}_1 = \begin{pmatrix} 1 & -0.4 & 0.4 \\ -0.4 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{pmatrix} \text{ and } \boldsymbol{\Gamma}_2 = \begin{pmatrix} 1 & 0.8 & 0.1 \\ 0.8 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{pmatrix}.$$



(a) Individuals described by three variables: one continuous (abscissa), one integer (ordinate) and one binary (symbol). Colors indicate the true class memberships



(b) Individuals in the first factorial map of component 2. Colors and symbols indicate the true class memberships



(c) Variables in the first factorial map of component 2

Figure 1: Example of data visualization.

Figure 1 provides an example of data visualization. Figure 1(a) shows the scatterplot of the individuals in their native space. Figure 1(b) presents the scatterplot of the individuals in the first PCA-map of the second component (red). It allows two classes to be easily distinguished: a centred one (red) and a second one (black) located on the left side. More precisely, the first axis (explained by the continuous and the integer variables) is strongly discriminative while the second axis (explained exclusively by the binary variable) is not discriminative. Figure 1(c) shows the correlation circle of the first PCA-map of the red component. It allows a strong correlation to be identified, for the red component, between the continuous and the integer variables.

3 Bayesian inference

3.1 Sampling layout on data and parameters

We observe the sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ composed of n independent realizations $\boldsymbol{x}_i \in \mathbb{R}^c \times \mathcal{X}$ assumed to arise from a Gaussian copula mixture model. As pointed out by Smith and Khaled (2012), the Bayesian framework simplifies the inference considerably since it uses the latent structure of the model $(\boldsymbol{y}, z)$. Without prior information about the data, we assume independence between the prior distributions. The proportions and the parameters of the univariate marginal distributions of each component $\boldsymbol{\beta}_{kj}$ follow the classical conjugate prior distributions (Robert, 2007). Finally, the conjugate prior of the covariance matrices is derived from an Inverse Wishart distribution as proposed by (Hoff, 2007). Details on the prior distributions are given in Appendix B.

3.2 Gibbs and Metropolis-within-Gibbs samplers

The Bayesian estimation is managed by a Gibbs sampler (described in Algorithm 3.1) which is the most popular approach for inferring 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 vector related to $\mathbf{x}$. Note that the Gaussian variable $\mathbf{y}$ is twice sampled during one iteration of the algorithm to manage the strong dependencies between $\mathbf{y}$ and $\mathbf{z}$, and between $\boldsymbol{y}^j_{[rk]} = \{y^j_i : z^{(r)}_i = k\}$ and $\boldsymbol{\beta}_{kj}$. Obviously, the stationary distribution stays unchanged. Thus, the sequence of parameters is sampled from the marginal posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$, and a consistent estimate of $\boldsymbol{\theta}$ can be obtained by taking the mean of the sampled parameters.

Algorithm 3.1 (The Gibbs sampler)**.** *Starting from an initial value* $\boldsymbol{\theta}^{(0)}$*, its iteration* (r) *consists in the following four steps* ($k \in \{1, \ldots, g\}, j \in \{1, \ldots, e\}$)

$$\mathbf{z}^{(r)}, \mathbf{y}^{(r-1/2)} \sim \mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{(r-1)} \tag{6}$$

$$\boldsymbol{\beta}^{(r)}_{kj}, \boldsymbol{y}^{j(r)}_{[rk]} \sim \boldsymbol{\beta}_{kj}, \boldsymbol{y}^j_{[rk]}|\mathbf{x}, \boldsymbol{y}^{\bar{j}(r)}_{[rk]}, \mathbf{z}^{(r)}, \boldsymbol{\beta}^{(r)}_{k\bar{j}}, \boldsymbol{\Gamma}^{(r-1)}_k \tag{7}$$

$$\boldsymbol{\pi}^{(r)} \sim \boldsymbol{\pi}|\mathbf{z}^{(r)} \tag{8}$$

$$\boldsymbol{\Gamma}^{(r)}_k \sim \boldsymbol{\Gamma}_k|\mathbf{y}^{(r)}, \mathbf{z}^{(r)} \tag{9}$$

where $\boldsymbol{y}_{[rk]} = \boldsymbol{y}_{\{i:z^{(r)}_i = k\}}$, $\boldsymbol{y}^{\bar{j}(r)}_i = (y^{1(r)}_i, \ldots, y^{j-1(r)}_i, y^{j+1(r-1/2)}_i, \ldots, y^{e(r-1/2)}_i)$ *and* $\boldsymbol{\beta}^{(r)}_{k\bar{j}} = (\boldsymbol{\beta}^{(r)}_{k1}, \ldots, \boldsymbol{\beta}^{(r)}_{kj-1}, \boldsymbol{\beta}^{(r-1)}_{kj+1}, \ldots, \boldsymbol{\beta}^{(r-1)}_{ke})$.

The samplings according to (8) and (9) are classical but the samplings from (6) and (7) are not easy. They are therefore replaced by one iteration of a Metropolis-Hastings algorithm that does not change the stationary distribution. The resulting algorithm is a Metropolis-within-Gibbs sampler (Robert and Casella, 2004) whose four steps are detailed in Appendix C.

3.3 Label switching problem

The label switching problem is generally solved by specific procedures (Stephens, 2000). However, based on the argument of Jacques and Biernacki (2014), these techniques are mainly effective when g is known.

However, when the model is used for clustering, the number of classes is unknown, and the model selection is performed using the BIC criterion (Schwarz, 1978) which simultaneously avoids the label switching phenomenon. Indeed, on the one hand, this criterion selects quite

separate classes when the sample size is small. Hence, label switching is not present (with high probability) in practice because of this class separation. On the other hand, even though it can select more classes when the sample size increases, the label switching problem does not occur since this phenomenon vanishes asymptotically.

Obviously, when the number of classes is fixed and the sample size is small, the label switching problem can occur. In such cases, we obviously advise using the procedures of Stephens (2000).

4 Simulations

In this section, two simulations are used to illustrate the new model. The first simulation shows the relevance of the estimation procedure. The second simulation illustrates the robustness of the proposed model by analyzing data sampled from a mixture of Poisson distributions.

Experiment conditions For each situation, 100 samples are generated. Parameters are estimated by taking the mean of the parameters sampled by 10^3 iterations of Algorithm 3.1 after a burn-in period of 10^2 iterations. Algorithm 3.1 is initialized with the maximum likelihood estimator of the locally independent model (particularly relevant when within-class dependencies are small). The Kullback-Leibler divergence (Kullback and Leibler, 1951) is used to compare the estimated distribution and the distribution used to sample the data. This divergence is approximated via 10^4 iterations of a Monte Carlo simulation.

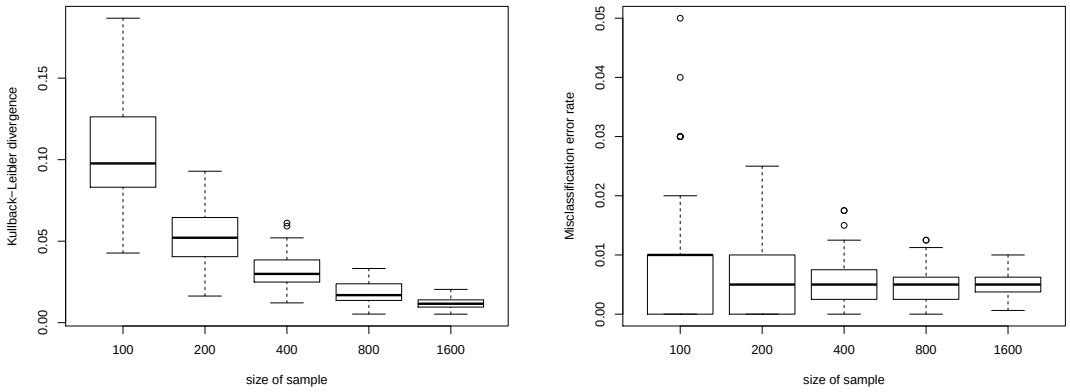
4.1 Estimation efficiency

Data sets are composed of one continuous variable, one integer variable and one binary variable. They are sampled from the distribution described in the example in Section 2.4. The results are presented in Figure 2. According to Figure 2(a), 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 decreases as a function of sample size and converges to zero. Moreover, as shown by Figure 2(b), the misclassification rate converges to the theoretical misclassification rate (equal to 0.005) when n increases. This simulation illustrates the convergence of the estimator computed by averaging the parameters sampled using the algorithm. Finally, the estimation procedure is not particularly time consuming since samples of size $n = 100$ and $n = 1600$ take 15 and 64 seconds respectively, on an Intel Core i5-3320M processor.

4.2 Robustness

Samples are generated from the bivariate Poisson mixture model (Karlis and Tsiamyrtzis, 2008) with $\boldsymbol{\pi} = (1/3, 2/3)$, whose univariate margin parameters $\boldsymbol{\alpha}_k = (\lambda_{k1}, \lambda_{k2}, \lambda_{k3})$ take on the following values: $\lambda_{1h} = h$ and $\lambda_{2h} = 3 + h$, for $h = 1, 2, 3$ (see notation detailed in Karlis and Tsiamyrtzis (2008)). Figure 3 presents the results and shows the robustness of the Gaussian copula mixture model since it efficiently manages such data sets, as detailed below.

As shown by Figure 3(a), the resulting misclassification rate converges to the theoretical misclassification rate (equal to 0.0967). Moreover, Figure 3(b) shows that the Kullback-Leibler divergence almost vanishes when the sample size increases, thus demonstrating the flexibility of the Gaussian copula mixture model. Furthermore, the resulting parameters reflect the main properties of the true distribution. Indeed, Figure 3(c) shows that the correlation coefficient between both variables, for component 1, converges to its theoretical value (equal to $\lambda_{11} + \lambda_{13} =$



(a) Kullback-Leibler divergence of the estimated model from the true model.

(b) Misclassification rate.

Figure 2: Boxplots of indicators of the good behavior of the estimation procedure for different sample sizes.

4). In the same way, Figure 3(d) shows that the univariate margin parameter of variable 1, for component 1, converges to its theoretical value (equal to $\lambda_{13}/\sqrt{(\lambda_{11}+\lambda_{13})(\lambda_{12}+\lambda_{13})} = 3/\sqrt{20} \simeq 0.67$).

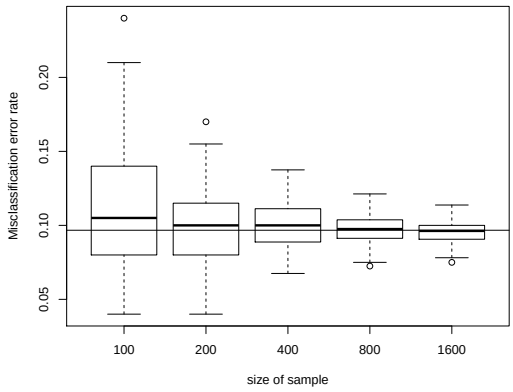
Finally, the estimation procedure is not excessively time consuming since it takes 12 and 54 seconds to analyze samples of size $n = 100$ and $n = 1600$ respectively, on an Intel Core i5-3320M processor.

5 Applications

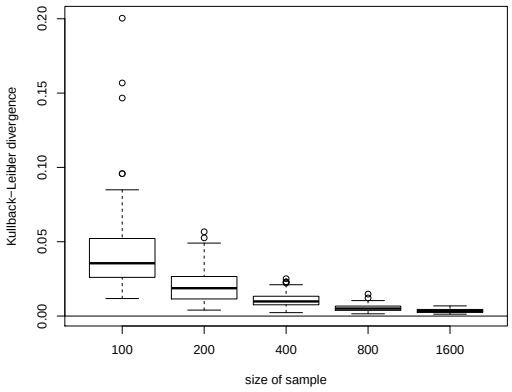
In this section, we analyze two real data sets with the proposed Gaussian copula mixture model. For each number of components, 10 runs of Gibbs sampler are performed with 1000 iterations after a burn-in period of 100 iterations.

Information criteria (BIC criterion (Schwarz, 1978), ICL criterion (Biernacki et al., 2000)...) can be used to perform the model selection. These asymptotic criteria are computed with the estimate provided by the Gibbs sampler and not with the maximum likelihood estimate. However, this procedure remains valid since the BIC criterion can be computed with the estimate of the maximum *a posteriori* (Lebarbier and Mary-Huard, 2006). Moreover, the ICL criterion can be computed by penalizing the BIC criterion with a term of entropy.

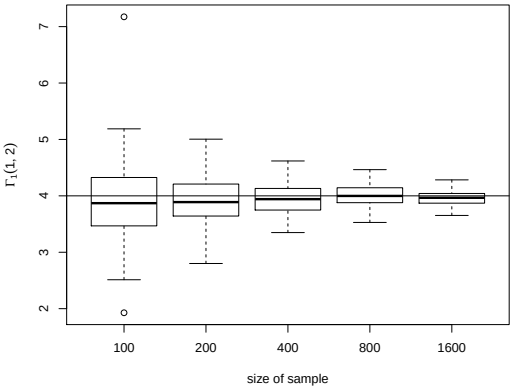
These criteria require the computation of the number of parameters. The locally independent model involves $\nu_{\text{Loc}} = (g-1) + g\sum_{j=1}^{e} \nu_j$ parameters, where ν_j is the number of parameters involved by the univariate margin distribution of one component (*i.e.* $\nu_j = 2$ if x^j is continuous, $\nu_j = 1$ if x^j is integer and $\nu_j = m_j - 1$ if x^j is ordinal with m_j modalities). The heteroscedastic Gaussian copula mixture model involves $\nu_{\text{He}} = \nu_{\text{Loc}} + g\frac{e(e-1)}{2}$ parameters and the homoscedastic Gaussian copula mixture model requires $\nu_{\text{Ho}} = \nu_{\text{Loc}} + \frac{e(e-1)}{2}$ parameters.



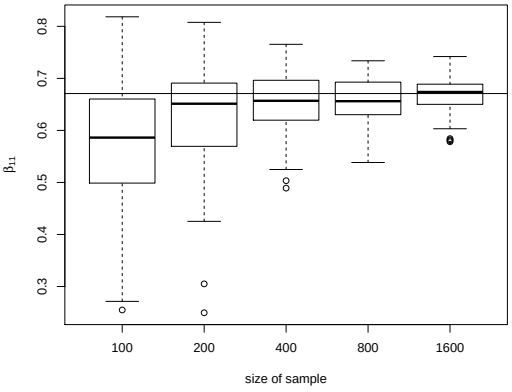
(a) Misclassification error rate



(b) Kullback-Leibler divergence from the true model



(c) Correlation coefficient between both variables for component 1



(d) univariate margin parameter of variable 1 for component 1

Figure 3: Boxplots of the indicators related to the estimated model. Values obtained with the true Poisson mixture model are indicated by the horizontal black lines.

5.1 South African Hearth data set

The data

The data are available at `http://sci2s.ugr.es/keel/dataset.php?cod=184`. This data set is a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. Many of the coronary heart disease (CHD) positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments. The class label indicates if the person has coronary heart disease (negative or positive) and is hidden for our analysis. Individuals are described by the following nine variables. The continuous variables are systolic blood pressure (sbp), cumulative tobacco (tobacco), low density lipoprotein cholesterol (ldl), adiposity, obesity and current alcohol consumption (alcohol). The integer variables are type-A behavior (typea) and age at onset (age). Finally, a binary variable indicates the presence or not of heart disease in the family history (famhist).

Model selection

Three mixture models (locally independent, heteroscedastic and homoscedastic mixture of Gaussian copulas) are fitted for various numbers of components. Table 1 presents the values of information criteria used to select the homoscedastic tri-component Gaussian copula mixture model. This model obtains the best results for fitting the data distribution (BIC) and for fitting the best partition (ICL). Moreover, this model detects less components than the locally independent model, thus its interpretation is easier. Note that the BIC criterion selects five components for the locally independent model while the ICL criterion selects four components.

	g	1	2	3	4	5
BIC	loc. indpt.	-14127.26	-13131.88	-12813.92	-12829.68	**-12738.66**
	homo.	-14724.98	-13016.09	**-12739.94**	-12774.15	-12927.45
	hetero.	-14724.98	-13076.93	**-12971.72**	-13071.92	-13253.06
ICL	loc. indpt.	-14127.26	-13144.21	-12832.12	-12887.19	**-12805.68**
	homo.	-14724.98	-13028.07	**-12762.79**	-12816.44	-12979.06
	hetero.	-14724.98	-13085.52	**-12989.06**	-13103.61	-13299.16

Table 1: Values of the BIC and ICL criteria obtained on the South African Hearth data set (best values are in bold).

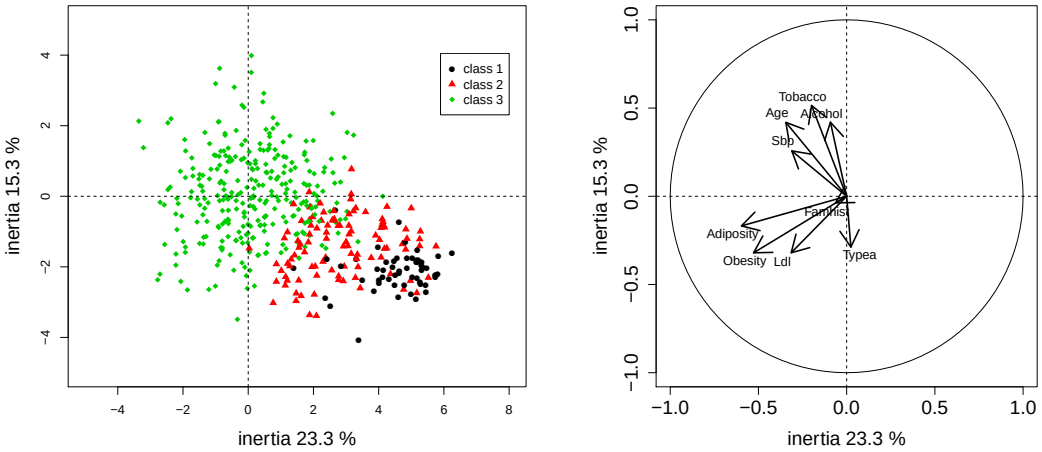
Partition study

The competing approaches overestimate the number of components since the true number of classes is two. However, the Gaussian copula mixture models (homo. and hetero.) select less classes than the locally independent mixture, which needs five components. Note that the partition provided by all of these models is strongly different to the true one since these models obtain an ARI of 0.02. We note that the equality constraint of the covariance matrix increases the value of the information criteria obtained by the homoscedastic model, but it also affects its resulting partition according to the partition resulting from the heteroscedastic model (see Table 2).

		hetero.		
		Class 1	Class 2	Class 3
homo.	Class 1	46	11	0
	Class 2	0	92	14
	Class 3	0	3	296

Table 2: Confusion matrices between the tri-component homoscedatic model (row) and the tri-component homoscedastic model (column).

Figure 4 shows the PCA visualization for component 3. The tri-component homoscedastic Gaussian copula mixture model provides three well-separated classes as shown by the factorial representation of Figure 4(a). We can see that class 2 (red triangles) is an "intermediate" class, while class 1 (red dots) is strongly different to class 3 (black dots). We now detail the model interpretation.



(a) Scatterplot of the individuals

(b) Correlation circle

Figure 4: Visualization using parameters of component 3 for the tri-component homoscedastic Gaussian copula mixture model.

Interpretation of best-fit model

The PCA visualization (Figure 4) shows that the individuals of class 3 have riskier behaviour than the others (high tobacco and alcohol consumption, older population, high level of obesity).

A three-level interpretation (proportions, univariate marginal distributions and within-class dependencies) is done using the model parameters. The main characteristics of the variables are summarized in Figure 5:

- Class 1 (*weak-risk behaviours*): this is the smallest class ($\pi_1 = 0.07$). It contains the young individuals with low alcohol and tobacco consumption. This class groups 57 individuals where only one has a coronary heart disease.
- Class 2 (*moderate-risk behaviours*): this class is moderated in size ($\pi_2 = 0.24$). This class is composed of individuals which high alcohol consumption. The other variables take intermediate values. Among the 106 individuals belonging to this class, 20 have coronary heart disease.
- Class 3 (*high-risk behaviours*): this is the biggest class ($\pi_3 = 0.69$). It contains the individuals with the highest risk behaviour. Of the 299 individuals in this class, 139 have coronary heart disease.

Finally, for all the components, age and the consumption of alcohol and tobacco are strongly linked (see Figure 4(b)). Moreover, adiposity and obesity are also strongly linked.

Conclusion

For such data, the Gaussian copula mixture model reduces the drawbacks of the locally independent model. By decreasing the number of components, it yields a more interpretable model that better fits the data (BIC criterion) and provides a pertinent partition (ICL criterion). Finally, the estimation of main within-class dependencies, based on PCA outputs per component, is an efficient tool for refining the interpretation.

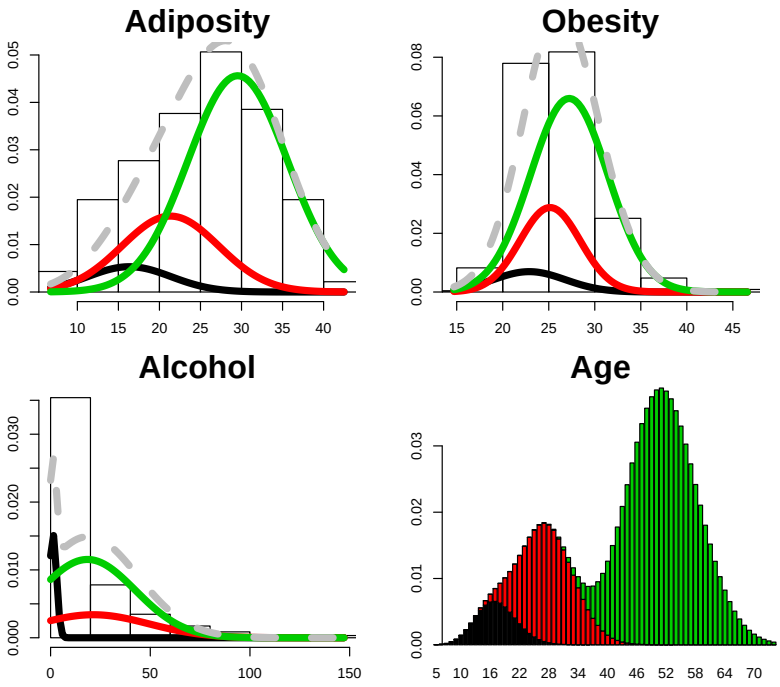


Figure 5: One dimensional marginal distributions for the South African Hearth data: whole distribution (gray), component 1 (black), component 2 (red) and component 3 (green) of the tri-component homoscedastic Gaussian copula mixture model.

5.2 Forest fire data set

The data

The data are composed of 517 forest fires that have occurred in the northeast region of Portugal (Cortez and Morais, 2007). These forest fires are described by the following meteorological variables: seven continuous variables (four fire weather index (FWI) system variables, *i.e.* fine fuel moisture code (FFMC), duff moisture code (DMC), drought code (DC), initial season index (ISI), and three meteorological variables, *i.e.* temperature (Temp), relative humidity (RH) and wind) and three binary variables indicating the presence of rain, the season (summer or other) and the day of the week (weekend or other).

Model selection

Table 3 presents the values of information criteria used to distinctly select the heteroscedastic tri-component Gaussian copula mixture model. Note that the locally independent model degenerates with five components.

The heteroscedastic model with three components obtains better values for the information criteria than the locally independent model since it models the within-component dependencies. Moreover, as shown by Table 4, these within-component dependencies influence the resulting partition since four individuals are affiliated in different classes by both models.

Interpretation of best-fit model

The three-step interpretation of the heteroscedastic tri-component Gaussian copula mixture model is presented using the parameters summarized in Figure 6:

	g	1	2	3	4	5
BIC	loc. indpt.	-15152.95	-14164.51	**-13990.27**	-14068.92	NA
	homo.	-14401.80	**-13751.82**	-13927.05	-13986.90	-13755.69
	hetero	-14401.80	-13781.86	**-13680.67**	-13846.63	-13745.84
ICL	loc. indpt.	-15152.95	-14170.97	**-14022.49**	-14131.22	NA
	homo.	-14401.80	**-13756.76**	-13956.68	-14070.49	-13774.41
	hetero	-14401.80	-13785.33	**-13682.68**	-13885.11	-13776.81

Table 3: Values of the BIC and ICL criteria obtained on the forest fire data set (best values are in bold).

		hetero.		
		Class 1	Class 2	Class 3
loc. indpt.	Class 1	33	1	0
	Class 2	1	402	0
	Class 3	2	0	78

Table 4: Confusion matrix for the partitions provided by the locally independent model with three components (rows) and the heteroscedastic Gaussian copula mixture model with three components (columns).

- Class 1 (*unpredictable fires*): this is the smallest class ($\pi_1 = 0.09$). It contains fires that are difficult to predict since they occur with small values of the four FWI system variables (especially ISI). These fires occur during all the year but only when the weather is dry.
- Class 2 (*predictable summer fires*): this is the biggest class ($\pi_2 = 0.78$). This class is composed of fires occurring mainly in summer. They appear with high values of the four FWI system variables and a high temperature.
- Class 3 (*winter fires*): this is a class of moderate size ($\pi_3 = 0.13$). It contains fires occurring in winter, so with low temperature. They occur in dry weather and very small values of the four FWI systems (except ISI).

The correlation matrices highlight dependencies between the summer period and high temperatures, and between the FFMC and DMC values (see Figure 7(b) for component 3). Finally, it should be noted that the variable indicating the day of the week roughly follows the same distribution for all three classes.

The results of the PCA done according to component 3 is shown in Figure 7. Obviously, Figure 7(a) shows that the individuals belonging to component 3 are strongly different to the other ones. Indeed, few individuals belonging to component 2 are visible in this map. The other individuals are too far away from the origin. Thus, the distribution of component 3 is strongly different from the other distributions.

Conclusion

The cluster analysis obtained with the Gaussian copula mixture model is more accurate than the one obtained with the locally independent model. Indeed, it provides a meaningful model which needs less components. This model sheds light on three kinds of fires: fires predictable with the FWI system (class 2), fires occurring in winter (class 3) and unpredictable fires (class 1).

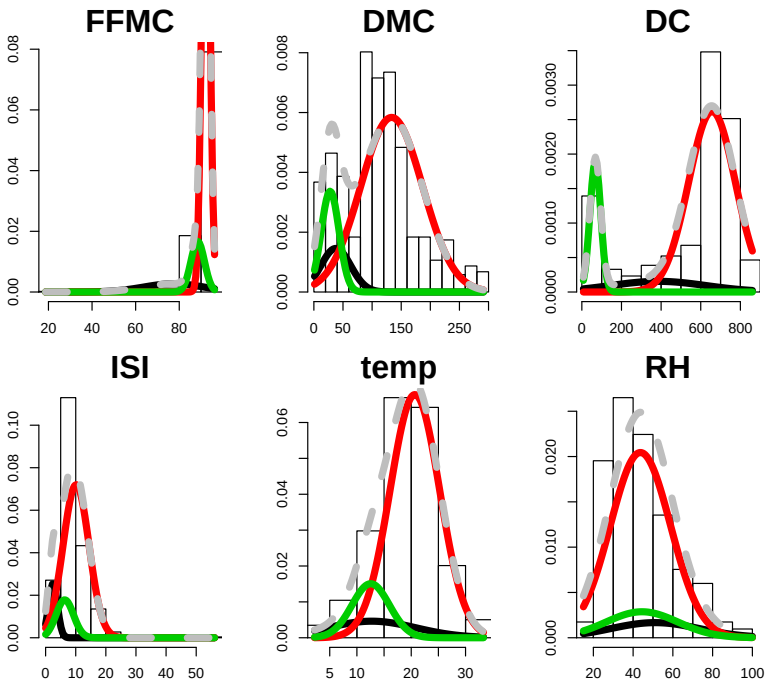
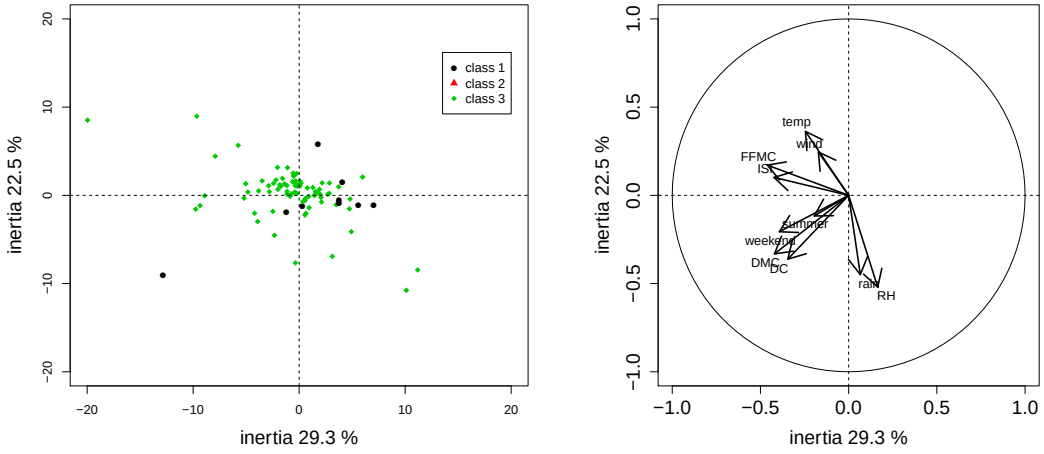


Figure 6: One dimensional marginal distributions for the forest fire data: whole distribution (gray), component 1 (black), component 2 (red) and component 3 (green) of the tri-component homoscedastic Gaussian copula mixture model.



(a) Scatterplot of the individuals

(b) Correlation circle

Figure 7: Visualization based on the parameters of component 3 for the tri-component heteroscedastic Gaussian copula mixture model.

6 Conclusion and future extensions

A Gaussian copula mixture model has been introduced and used to cluster mixed data. Using Gaussian copulas, the univariate marginal distributions of each component follow conventional distributions, and within-class dependencies are effectively modelled. Thus, the model results can be easily interpreted. Using the continuous latent variables of Gaussian copulas, a PCA-type method allows for component-based visualization of individuals. Moreover, this approach

provides a summary of within-component dependencies, which avoids tedious interpretation of correlation matrices.

In the description of numerical experiments and applications, we pointed out that this model is sufficiently robust to fit data obtained from another model. Furthermore, it can reduce the bias produced by the locally independent model (*e.g.* reduction of the number of components).

The number of parameters increases with the number of components and number of variables, particularly due to the correlation matrices of the Gaussian copulas. In order to overcome this drawback, we have proposed a homoscedastic version of the model which assumes equality between correlation matrices. However, the number of parameters of this model is still a quadratic function of the number of variables. Therefore, more parsimonious correlation matrices could be proposed in future studies for clustering high-dimensional mixed data.

Since the distribution of all the variables is modeled, this model could be used to manage data sets with missing values. By assuming that values are missing at random, the Gibbs sampler could also be adapted, but the underlying principle would remain roughly the same.

Finally, the proposed model cannot cluster non-ordinal categorical variables having more than two modalities. In such cases, the cumulative distribution function is not defined. An artificial order between modalities could be added to define a cumulative distribution function, but this method presents three potential difficulties that require attention: 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 verified.

MixCluster (`https://r-forge.r-project.org/R/?group_id=1939`) is an R package which performs the cluster analysis method described in the article. It also contains the data sets used in this paper.

References

Banfield, J. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pages 803–821.

Barnard, J., McCulloch, R., and Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1312.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Cortez, P. and Morais, A. (2007). A data mining approach to predict forest fires using meteorological data. *Associação Portuguesa para a Inteligência Artificial (APPIA)*.

Everitt, B. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, 6(5):305–309.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer.

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

Gouget, C. (2006). *Utilisation des modèles de mélange pour la classification automatique de données ordinales*. PhD thesis, Université de Technologie de Compiègne.

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

Hoff, P. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pages 265–283.

Hoff, P., Niu, X., and Wellner, J. (2011). Information bounds for Gaussian copulas. *arXiv preprint arXiv:1110.3572*.

Hunt, L. and Jorgensen, M. (1999). Theory & Methods: Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, 41(2):154–171.

Hunt, L. and Jorgensen, M. (2011). Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):352–361.

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

Joe, H. (1997). *Multivariate models and dependence concepts*, volume 73. CRC Press.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419.

Karlis, D. and Tsiamyrtzis, P. (2008). Exact Bayesian modeling for bivariate Poisson data and extensions. *Statistics and Computing*, 18(1):27–40.

Klaassen, C. and Wellner, J. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77.

Krzanowski, W. (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification*, 10(1):25–49.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86.

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

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

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

Morlini, I. (2012). 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.

Moustaki, I. and Papageorgiou, I. (2005). Latent class models for mixed variables with applications in archaeometry. *Computational Statistics & Data Analysis*, 48(3):659–675.

Murray, J., Dunson, D., Carin, L., and Lucas, J. (2013). Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665.

Nelsen, R. (1999). *An introduction to copulas*. Springer.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460.

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

Raftery, A. (1996). Hypothesis testing and model selection. In *Markov chain Monte Carlo in practice*, pages 163–187. Springer.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer.

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.

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

Smith, M. and Khaled, M. (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association*, 107(497):290–303.

Song, P. X.-K., Fan, Y., and Kalbfleisch, J. D. (2005). Maximization by parts in likelihood inference. *Journal of the American Statistical Association*, 100(472):1145–1158.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Teicher, H. (1963). Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, pages 1265–1269.

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

Willse, A. and Boik, R. (1999). Identifiable finite mixtures of location models for clustering mixed-mode data. *Statistics and Computing*, 9(2):111–121.

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

A Proof of model identifiability

Model identifiability is proved by two propositions. The first proposition proves 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 Gaussian copula mixture model is weakly identifiable (Teicher, 1963) if the variables are continuous and integer ones (*i.e. *the univariate marginal distributions of the components are Gaussian or Poisson distributions). Thus,*

$$\forall \boldsymbol{x} \in \mathbb{R}^c \times \mathbb{N}^d, \quad \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}|\boldsymbol{\alpha}_k) = \sum_{k=1}^{g'} \pi'_k p(\boldsymbol{x}|\boldsymbol{\alpha}'_k) \tag{10}$$

$$\Rightarrow \quad g = g',\ \boldsymbol{\pi} = \boldsymbol{\pi}',\ \boldsymbol{\alpha} = \boldsymbol{\alpha}'. \tag{11}$$

Proof. The identifiability of the multivariate Gaussian mixture models and of the univariate Poisson mixture model (Teicher, 1963; Yakowitz et al., 1968) means that (10) implies

$$g = g',\ \boldsymbol{\pi} = \boldsymbol{\pi}',\ \boldsymbol{\beta}_{kj} = \boldsymbol{\beta}'_{kj} \text{ and } \boldsymbol{\Gamma}_{k\text{CC}} = \boldsymbol{\Gamma}'_{k\text{CC}}. \tag{12}$$

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 $a'_k = \frac{b_k^{\oplus}(x^j) - \rho'_k v_k}{\sqrt{1-\rho_k'^2}}$. Without loss of generality, we order the components such that $\sigma_{kj} > \sigma_{k+1j}$ and if $\sigma_{kj} = \sigma_{k+1j}$ then $\mu_{kj} > \mu_{k+1j}$, then (10) 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 $x^h \to \infty$ such that $(x^j, x^h) \in \gamma_t$,

$$\forall t, \quad \frac{\int_t^{a'_1} \phi(u)du}{\Phi(t)} = 0. \tag{13}$$

Thus $a'_1 = a_1$, so $\rho'_1 = \rho_1$. Repeating this argument for $k = 2, \ldots, g$ and for all the couples (j,h), we conclude that $\boldsymbol{\Gamma}_{k\text{CD}} = \boldsymbol{\Gamma}'_{k\text{CD}}$.

When both variables are integer, we use the same argument with $\gamma_{(t,\xi)} = \{(x^j, x^h) \in \mathbb{N} \times \mathbb{N} : a_1 \in B(t,\xi)\}$. Note that if $\rho_1 \neq \rho'_1$ then $\exists n_0$ such that $\forall x^j > n_0\ a'_1 > t+\xi$. Letting $x^h \to \infty$ such that $(x^j, x^h) \in \gamma_{(t,\xi)}$, 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 $\boldsymbol{\Gamma}_{k\text{DD}} = \boldsymbol{\Gamma}'_{k\text{DD}}$. □

Proposition A.2 (Identifiability of the Gaussian copula mixture model)**.** *The Gaussian copula mixture model is weakly identifiable (Teicher, 1963) 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 extended to the other cases. We now show that $\boldsymbol{\Gamma}_{k\text{CD}} = \boldsymbol{\Gamma}'_{k\text{CD}}$ and $\boldsymbol{\Gamma}_{k\text{DD}} = \boldsymbol{\Gamma}'_{k\text{DD}}$.

Let $j = 1$ and let $h \in \{2, 3\}$. We note $\rho_k = \boldsymbol{\Gamma}_k(j,h)$, $\rho'_k = \boldsymbol{\Gamma}'_k(j,h)$, $v_k = \Phi_1^{-1}(P(x^j; \boldsymbol{\beta}_{kj}))$, $\varepsilon_k(x^j) = \pi_k \frac{\phi(v_k;0,1)}{\sigma_{kj}}$, $a_k = \frac{b_k^{\oplus}(x^j) - \rho_k v_k}{\sqrt{1-\rho_k^2}}$ and $a'_k = \frac{b_k'^{\oplus}(x^j) - \rho'_k v_k}{\sqrt{1-\rho_k'^2}}$. Without loss of generality, we order the components such that $\sigma_{kj} > \sigma_{[k+1]j}$ and if $\sigma_{kj} = \sigma_{[k+1]j}$ then $\mu_{kj} > \mu_{[k+1]j}$. Note that (10) 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 obtained by tending x^1 to $-\infty$ if $\rho_k < 0$. 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

We assume independence between the parameters as follows

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}) \prod_{k=1}^{g} \left(p(\boldsymbol{\Gamma}_k) \prod_{j=1}^{e} p(\boldsymbol{\beta}_{kj}) \right). \tag{14}$$

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

$$\boldsymbol{\pi} \sim \mathcal{D}_g \left(\frac{1}{2}, \ldots, \frac{1}{2} \right). \tag{15}$$

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

$$\sigma^2_{kj} \sim \mathcal{G}^{-1}(c_0, C_0) \text{ and } \mu_{kj}|\sigma^2_{kj} \sim \mathcal{N}_1(b_0, \sigma^2_{kj}/N_0), \tag{16}$$

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 Raftery (1996), 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

$$\boldsymbol{\beta}_{kj} \sim \mathcal{G}(a_0, A_0). \tag{17}$$

According to Frühwirth-Schnatter (2006), 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 means that

$$\boldsymbol{\beta}_{kj} \sim \mathcal{D}_{m_j} \left(\frac{1}{2}, \ldots, \frac{1}{2} \right). \tag{18}$$

The conjugate prior of a covariance matrix is the Inverse Wishart distribution denoted by $\mathcal{W}^{-1}(.,.)$. Therefore, 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 (Hoff, 2007). So,

$$\boldsymbol{\Lambda}_k \sim \mathcal{W}^{-1}(s_0, S_0) \text{ and } \forall 1 \leq h, \ell \leq e,\ \boldsymbol{\Gamma}_k[h,\ell] = \frac{\boldsymbol{\Lambda}_k[h,\ell]}{\sqrt{\boldsymbol{\Lambda}_k[h,h]\boldsymbol{\Lambda}_k[\ell,\ell]}}, \tag{19}$$

where (s_0, S_0) are two hyper-parameters. However, an empirical Bayesian approach cannot be fitted to these parameters since $\boldsymbol{y}$ is not observed. Uniform distribution on $]-1,1[$ is also obtained for the margin distributions of each correlation coefficient by setting $s_0 = e + 1$ and S_0 equal to the identity matrix (Barnard et al., 2000).

C Metropolis-within-Gibbs sampler

This section explains the four samplings used in Algorithm 3.1. The first two samplings are difficult to perform directly, so they are done by one iteration of two Metropolis-Hastings algorithms. For both Metropolis-Hastings algorithms, the instrumental distributions assume conditional independence between parameters. So the smaller the within-class dependencies are, the closer the instrumental distributions of both algorithms are to the stationary distributions. Finally, the last two samplings used in Algorithm 3.1 are classical.

C.1 Class membership and Gaussian vector sampling

The aim is to sample from (6) but the sampling from $\mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{(r-1)}$ cannot be performed directly. So it is achieved by one iteration of the following Metropolis-Hastings algorithms. The sampling from (6) is performed in two steps using independence between the individuals, which implies that

$$p(\mathbf{z}, \mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{(r-1)}) = \prod_{i=1}^{n} p(z_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)}) p(\boldsymbol{y}_i|\boldsymbol{x}_i, z_i, \boldsymbol{\theta}^{(r-1)}). \tag{20}$$

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

$$z_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)} \sim \mathcal{M}_g(t_{i1}(\boldsymbol{\theta}^{(r-1)}), \ldots, t_{ig}(\boldsymbol{\theta}^{(r-1)})), \tag{21}$$

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 $\boldsymbol{x}_i$ arises 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 $\mathcal{S}_{z_i^{(r)}}(\boldsymbol{x}_i^{\mathrm{D}})$

$$p(\boldsymbol{y}_i^{\mathrm{D}}|\boldsymbol{x}_i, \boldsymbol{y}_i^{\mathrm{C}(r-1/2)}, z_i^{(r)}, \boldsymbol{\theta}^{(r-1)}) \propto \phi_d(\boldsymbol{y}_i^{\mathrm{D}}; \boldsymbol{\mu}_{z_i^{(r)}}^{\mathrm{D}(r-1)}, \boldsymbol{\Sigma}_{z_i^{(r)}}^{\mathrm{D}(r-1)}) \mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{D}} \in \mathcal{S}_{z_i^{(r)}}(\boldsymbol{x}_i^{\mathrm{D}})\}}, \tag{22}$$

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

The computation of $t_{ik}(\boldsymbol{\theta}^{(r-1)})$ involves the calculation of the integral defined in (5) which can be time consuming if d is large ($d > 6$). In such cases, the sampling from (6) is replaced by one iteration of the Metropolis-Hastings algorithm, which independently samples each couple $(z_i, \boldsymbol{y}_i)$. Its stationary distribution is

$$p(z_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)}) \propto \pi_{z_i} p(\boldsymbol{x}_i, \boldsymbol{y}_i|z_i, \boldsymbol{\theta}^{(r-1)}). \tag{23}$$

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

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 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 $\boldsymbol{y}_i^{\star\mathrm{D}}$, follow a *multivariate independent Gaussian* distribution truncated on $\mathcal{S}_{z_i^\star}(\boldsymbol{x}_i^{\mathrm{D}})$. Thus,

$$q_1(z_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(r-1)}) = \frac{1}{g} \frac{\phi_d(\boldsymbol{y}_i^{\mathrm{D}}|\mathbf{0}, \boldsymbol{I}) \mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{D}} \in \mathcal{S}_{z_i}(\boldsymbol{x}_i^{\mathrm{D}})\}}}{\prod_{j=c+1}^{e} p(x_i^j; \boldsymbol{\beta}_{z_i j}^{(r-1)})} \mathbb{1}_{\{\boldsymbol{y}_i^{\mathrm{C}} = \Psi(\boldsymbol{x}_i^{\mathrm{C}}|\boldsymbol{\alpha}_{z_i}^{(r-1)})\}}. \tag{24}$$

The candidate is accepted with the probability

$$\rho_{1i}^{(r)} = \min \left\{ \frac{q_1(z_i^{(r-1)}, \boldsymbol{y}_i^{(r-1)}|\boldsymbol{x}_i)}{q_1(z_i^\star, \boldsymbol{y}_i^\star|\boldsymbol{x}_i)} \frac{\pi_{z_i^\star} \phi_e(\boldsymbol{y}_i^\star; \mathbf{0}, \boldsymbol{\Gamma}_{z_i^\star}^{(r-1)})}{\pi_{z_i^{(r-1)}} \phi_e(\boldsymbol{y}_i^{(r-1)}; \mathbf{0}, \boldsymbol{\Gamma}_{z_i^{(r-1)}}^{(r-1)})}; 1 \right\}. \tag{25}$$

Thus, at iteration (r) of Algorithm 3.1, the sampling according to (6) 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 its stationary distribution.

Algorithm C.1.

$$(z_i^\star, \boldsymbol{y}_i^\star) \sim q_1(z, \boldsymbol{y}|\boldsymbol{x}_i) \tag{26}$$

$$(z_i^{(r)}, \boldsymbol{y}_i^{(r-1/2)}) = \begin{cases} (z_i^\star, \boldsymbol{y}_i^\star) & \textit{with probability } \rho_{1i}^{(r)} \\ (z_i^{(r-1)}, \boldsymbol{y}_i^{(r-1)}) & \textit{with probability } 1-\rho_{1i}^{(r)}. \end{cases} \tag{27}$$

C.2 Margin parameter and Gaussian vector sampling

The aim is to sample from (7) but the sampling from $\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\Gamma}_k$ cannot be performed directly. So, it is done by means by one iteration of the following Metropolis-Hastings algorithms. The sampling from (7) is performed using the following decomposition

$$\begin{aligned} p(\boldsymbol{\beta}_{kj}, \boldsymbol{y}_{[rk]}^j|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}) &= p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}) \\ &\quad \times p(\boldsymbol{y}_{[rk]}^j|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\beta}_{kj}, \boldsymbol{\Gamma}_k^{(r-1)}). \end{aligned} \tag{28}$$

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

$$p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}) \propto p(\boldsymbol{\beta}_{kj}) \prod_{\{i:z_i^{(r)}=k\}} p(x_i^j|\boldsymbol{y}_i^{\bar{j}(r)}, z_i^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}, \boldsymbol{\beta}_{kj}). \tag{29}$$

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

$$p(x_i^j|\boldsymbol{y}_i^{\bar{j}(r)}, z_i^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}, \boldsymbol{\beta}_{kj}) = \begin{cases} \phi_1(\frac{x_i^j-\mu_{kj}}{\sigma_{kj}}; \tilde{\mu}_i, \tilde{\sigma}_i^2)/\sigma_{kj} & \text{if } 1 \leq j \leq c \\ \Phi_1(\frac{b^\oplus(x_i^j)-\tilde{\mu}_i}{\tilde{\sigma}_i}) - \Phi_1(\frac{b^\ominus(x_i^j)-\tilde{\mu}_i}{\tilde{\sigma}_i}) & \text{otherwise,} \end{cases} \tag{30}$$

where the real $\tilde{\mu}_i = \boldsymbol{\Gamma}_k^{(r-1)}[j,\bar{j}]\boldsymbol{\Gamma}_k^{(r-1)}[\bar{j},\bar{j}]^{-1}\boldsymbol{y}_i^{\bar{j}(r)}$ 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 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 $\tilde{\sigma}_i^2 = 1 - \boldsymbol{\Gamma}_k^{(r-1)}[j,\bar{j}]\boldsymbol{\Gamma}_k^{(r-1)}[\bar{j},\bar{j}]^{-1}\boldsymbol{\Gamma}_k^{(r-1)}[\bar{j},j]$. As the normalizing constant of (29) is unknown, $\boldsymbol{\beta}_{kj}^{(r)}$ cannot be directly sampled. This problem is avoided by one iteration of the Metropolis-Hastings algorithm. The instrumental distribution of this 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 a 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 (29), the candidate $\boldsymbol{\beta}_{kj}^\star$ is accepted with the probability

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

Thus, at iteration (r) of Algorithm 3.1, step (7) is performed via one iteration of the following Metropolis-Hastings algorithm whose 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.

$$\boldsymbol{\beta}_{kj}^\star \sim q_2(\boldsymbol{\beta}_{kj}|\mathbf{x},\mathbf{z}) \tag{31}$$

$$\boldsymbol{\beta}_{kj}^{(r)} = \begin{cases} \boldsymbol{\beta}_{kj}^\star & \textit{with probability } \rho_2^{(r)} \\ \boldsymbol{\beta}_{kj}^{(r-1)} & \textit{with probability } 1-\rho_2^{(r)}. \end{cases} \tag{32}$$

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

$$p(\boldsymbol{y}_{[rk]}^{j}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\beta}_{kj}, \boldsymbol{\Gamma}_k^{(r-1)}) = \prod_{\{i:z_i^{(r)}=k\}} p(y_i^j|x_i^j, \boldsymbol{y}_i^{\bar{j}(r)}, z_i^{(r)}, \boldsymbol{\beta}_{kj}, \boldsymbol{\Gamma}_k^{(r-1)}). \tag{33}$$

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 at $\frac{x_i^j - \mu_{kj}^{(r)}}{\sigma_{kj}^{(r)}}$. 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 such as,

$$p(y_i^j|x_i^j, \boldsymbol{y}_i^{\bar{j}(r)}, z_i^{(r)}, \boldsymbol{\beta}_{kj}^{(r)}, \boldsymbol{\Gamma}_k^{(r-1)}) = \frac{\phi_1(y_i^j; \tilde{\mu}_i, \tilde{\sigma}_i^2)}{p(x_i^j; \boldsymbol{\beta}_{kj}^{(r)})} \mathbb{1}_{\{y_i^j \in [b_k^{\ominus(r)}(x_i^j), b_k^{\oplus(r)}(x_i^j)]\}}, \tag{34}$$

So, step (7) is performed in two stages. First, $\boldsymbol{\beta}_{kj}^{(r)}$ is sampled via one iteration of the Metropolis-Hastings algorithm of which the stationary distribution is $p(\boldsymbol{\beta}_{kj}|\mathbf{x}, \boldsymbol{y}_{[rk]}^{\bar{j}(r)}, \mathbf{z}^{(r)}, \boldsymbol{\beta}_{k\bar{j}}^{(r)}, \boldsymbol{\Gamma}_k)$. Second, $\boldsymbol{y}_{[rk]}^{j(r)}$ is sampled from (34).

C.3 Proportion vector sampling

The aim is to sample from (8). The sampling from (8) is classical. Indeed, the conjugate Jeffreys non informative prior means that

$$\boldsymbol{\pi}|\mathbf{z}^{(r)} \sim \mathcal{D}_g\left(n_1^{(r)} + \frac{1}{2}, \ldots, n_g^{(r)} + \frac{1}{2}\right), \tag{35}$$

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

C.4 Correlation matrix sampling

The aim is to sample from (9). To sample from (9), we use the approach proposed by Hoff (2007) in the case of semiparameteric Gaussian copulas. First, a covariance matrix is generated by its explicit posterior distribution, and second, 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 $\boldsymbol{\Gamma}_k|\mathbf{y}^{(r)}, \mathbf{z}^{(r)}$ is performed by the following two steps

$$\boldsymbol{\Lambda}_k|\mathbf{y}^{(r)}, \mathbf{z}^{(r)} \sim \mathcal{W}^{-1}\left(s_0 + n_k^{(r-1)}, S_0 + \sum_{\{i:z_i^{(r)}=k\}} \boldsymbol{y}_i^{(r)T} \boldsymbol{y}_i^{(r)}\right), \tag{36}$$

where $\forall 1 \leq h, \ell \leq e,\ \boldsymbol{\Gamma}_k[h,\ell] = \frac{\boldsymbol{\Lambda}_k[h,\ell]}{\sqrt{\boldsymbol{\Lambda}_k[h,h]\boldsymbol{\Lambda}_k[\ell,\ell]}}$. As the homoscedastic model assumes equality between the correlation matrices, in this case we only sample one $\boldsymbol{\Lambda}$ so (36) is replaced by

$$\boldsymbol{\Lambda}|\mathbf{y}^{(r)}, \mathbf{z}^{(r)} \sim \mathcal{W}^{-1}\left(s_0 + n, S_0 + \sum_{i=1}^n \boldsymbol{y}_i^{(r)T} \boldsymbol{y}_i^{(r)}\right), \tag{37}$$

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