A semiparametric and location-shift copula-based mixture model

Gildas Mazo

A semiparametric and location-shift copula-based mixture model

Gildas Mazo

Université Catholique de Louvain
Belgium

Abstract

Modeling of distributions mixtures has rested on Gaussian distributions and/or a conditional independence hypothesis for a long time. Only recently have researchers begun to construct and study broader generic models without appealing to such hypotheses. Some of these extensions use copulas as a tool to build flexible models, as they permit to model the dependence and the marginal distributions separately. But this approach also has drawbacks. First, the practitioner has to make more arbitrary choices, and second, marginal misspecification may loom on the horizon. This paper aims at overcoming these limitations by presenting a copula-based mixture model which is semiparametric. Thanks to a location-shift hypothesis, semiparametric estimation, also, is feasible, allowing for data adaptation without any modeling effort.

Keywords: location; shift; copula; mixture; clustering; semiparametric; nonparametric.

1 Introduction

The modeling of a mixture of distributions has long rested upon Gaussian distributions [23] and it is only recently that researchers have started to construct and study broader generic models [7, 18, 21, 17, 20, 29]. Among these extensions, models featuring copulas are still rare, but certainly promising [20, 17, 29]. Indeed, copulas allow for building very flexible models, as they permit to handle the marginal distributions and the dependence separately.

Let h be a mixture model density. It is of the form

$$h(x_1, \ldots, x_d) = \sum_{z=1}^{K} \pi_z h_z(x_1, \ldots, x_d), \tag{1}$$

where K is the number of groups, and for $z = 1, \ldots, K$, h_z and π_z are the conditional density and the weight of the z-th group respectively. The π_z satisfy $\pi_z \geq 0$ and $\sum_z \pi_z = 1$. A copula-based mixture model is simply a standard mixture model in which the conditional density h_z has been decomposed into the copula and the marginals, that is,

$$h_z(x_1, \ldots, x_d) = c_z \left\{H_{1z}(x_1), \ldots, H_{dz}(x_d)\right\} \prod_{j=1}^{d} h_{jz}(x_j), \tag{2}$$

where c_z is the copula in the z-th group, and H_{jz} and h_{jz} denote the distribution function and the density of the j-th variable of interest in the z-th group respectively. Such a decomposition is always possible as long as the marginals are continuous. It is sometimes called the copula decomposition or Sklar's decomposition, in view of Sklar's theorem [28]. For more details about copulas in general, see e.g. [24, 15, 9] or the Appendix.

Thus, pluging (2) into (1), we get

$$h(x_1, \ldots, x_d) = \sum_{z=1}^{K} \pi_z c_z \{H_{1z}(x_1), \ldots, H_{dz}(x_d)\} \prod_{j=1}^{d} h_{jz}(x_j). \tag{3}$$

As a matter of fact, as long as the marginals are continuous, any standard mixture model (1) can be re-written as in (3). Nevertheless, it is wise to reserve the term copula-based mixture model only to those models which make explicit use of formula (3).

In order to build a parametric copula-based mixture model, $d \times K + d$ parametric families have to be chosen. In practice, this is quite a large number of choices and therefore one often assumes that all the marginals come from the same parametric family. But then this restriction can be too strong for applications. This issue, in particular, was pointed out in [29].

In this paper, we aim at overcoming these limitations by presenting a new copula-based model where there is no need to parametrize the marginal distributions. It is a semiparametric model with parametric copulas but nonparametric marginals. In this respect it echos the common semiparametric copula models of the "nonmixture" literature, see e.g. [10]. In each dimension, we assume the existence of a symmetric distribution whose location shifts according to group assignment. As a result, this symmetric distribution can be estimated nonparametrically and therefore can adapt to many types of distributions with no modeling effort.

This paper is organized as follows. Section 2 presents the new location-shift semiparametric copula-based mixture model. Section 3 deals with estimation. Section 4 illustrates the model's features. Section 5 discusses the possibility of relaxing the location-shift assumption. A general Discussion closes the paper.

2 The model

Let $(X_1, \ldots, X_d)$ be the vector of interest and let Z be the group (or cluster, or class) assignment. For instance $Z = 1$ means that $(X_1, \ldots, X_d)$ belongs to the first group. The number of clusters is denoted by K, so that $Z \in \{1, \ldots, K\}$. Let H_{jz} and h_{jz} denote respectively the distribution function and the density of X_j given $Z = z$. Let h denote the density of $(X_1, \ldots, X_d)$ and define $\pi_z = P(Z = z)$. $\mathcal{R}$ stands for the real line.

For all $j = 1, \ldots, d$, we assume

$$H_{jz}(x_j) = G_j\left(x_j - \mu_{jz}\right), \qquad x_j, \mu_{jz} \in \mathcal{R}, \tag{4}$$

so that

$$h(x_1, \ldots, x_d) = \sum_{z=1}^{K} \pi_z c_z \left\{G_1\left(x_1 - \mu_{1z}\right), \ldots, G_d\left(x_d - \mu_{dz}\right)\right\} \prod_{j=1}^{d} g_j\left(x_j - \mu_{jz}\right) \tag{5}$$

where G_j and g_j are respectively the distribution function and the density of a symmetric distribution, that is, $G_j(x_j) = 1 - G_j(-x_j)$ for all continuity points x_j. This location-shift hypothesis (4) implies that X_j, $j = 1, \ldots, d$, is assumed to have support $(-\infty, +\infty)$. If it is not, then the data have to be distorted to achieve unboundedness. Note that the support of X_j given $Z = z$ does not depend on z and is equal to $(-\infty, +\infty)$. Hypothesis (4) means that the marginal distributions in a cluster z and a cluster z' differ only by a shift of location. Put differently, we assume, given $Z = z$, that $X_j = Y_j + \mu_{jz}$ where $Y_j \sim G_j$ and Y_j is independent of Z. Note, however, that $(Y_1, \ldots, Y_d)$ is *not* independent of Z (since its copula may depend on Z).

Several arguments can be made in favor of Hypothesis (4). First, it is intuitively clear, and therefore the user is expected to know whether if he/she is ready to accept it or not. This is certainly a plus. Second, conventional mixture models satisfy a location-shift hypothesis, as a structured version of the standard Gaussian mixture model, in which (4) holds with

$$g_j(x_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{1}{2}\left(\frac{x_j}{\sigma_j}\right)^2\right], \qquad x_j \in \mathcal{R},\ \sigma_j > 0$$

To play the role of g_j, other symmetric distributions can be called for, as for instance, the student distribution

$$g_j(x_j) = \frac{\Gamma((\nu_j + 1)/2)}{\sqrt{\nu_j \pi}\Gamma(\nu_j/2)} \left(1 + \frac{x_j^2}{\nu_j}\right)^{-(\nu_j+1)/2}, \qquad \nu_j > 2,\ x_j \in \mathcal{R}, \tag{6}$$

or the Laplace distribution

$$g_j(x_j) = \frac{1}{2b_j} \exp\left(-\frac{|x_j|}{b_j}\right), \qquad b_j > 0,\ x_j \in \mathcal{R}. \tag{7}$$

Last but not least, hypothesis (4) allows for nonparametric estimation, as explained in Section 3.

3 Estimation

This section presents an EM-like algorithm to estimate the unknown parameter $\boldsymbol{\phi} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{G})$ with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\boldsymbol{\mu} = (\mu_{11}, \ldots, \mu_{d1}, \ldots, \mu_{1K}, \ldots, \mu_{dK})$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, and $\mathbf{G} = (G_1, \ldots, G_d)$ of the model (5), in which we assumed that each copula c_z depend on a (possibly multivariate) parameter $\boldsymbol{\theta}_z$. The parameters G_j, $j = 1, \ldots, d$, are infinite-dimensional and therefore standard EM algorithms [5, 23, 25] are unfeasible. There are however refinements on which we can rest in order to perform estimation. One such refinement is given in [4], where a stochastic and semiparametric EM-like algorithm is applied to a univariate semiparametric location-shift mixture model. Capitalizing on their results, we also present in Section 3.2 a stochastic and semiparametric EM-like algorithm allowing for semiparametric estimation in our model (5). But first, the ideas underlying the standard EM algorithm are recalled in Section 3.1.

3.1 A brief recap of the EM algorithm

Generally, consider a model of the form

$$h(x_1, \ldots, x_d; \boldsymbol{\phi}) = \sum_{z=1}^{K} \pi_z h_z(x_1, \ldots, x_d; \boldsymbol{\chi}_z),$$

where $\boldsymbol{\phi} = (\boldsymbol{\pi}, \boldsymbol{\chi})$ with $\boldsymbol{\chi} = (\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_K)$ is the parameter to be estimated. Let $\mathbf{X}^{(i)}$, $i = 1, \ldots, n$, with $\mathbf{X}^{(i)} = (X_1^{(i)}, \ldots, X_d^{(i)})$, be a random sample (independent and identically distributed variables) and $\mathbf{x}^{(i)}$, $i = 1, \ldots, n$, with $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})$, be a realization of it. Likewise, let $Z^{(i)}$, $i = 1, \ldots, n$, be the associated unobserved sample of the group assignment variables. If the complete data $(\mathbf{x}^{(i)}, z^{(i)})$, $i = 1, \ldots, n$, were observed, we could maximize the log-likelihood

$$L(\mathbf{X}, \mathbf{Z}; \boldsymbol{\phi}) = \sum_{i=1}^{n} \log h(\mathbf{X}^{(i)}, Z^{(i)}; \boldsymbol{\phi}),$$

where $\mathbf{X}$ stands for $(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)})$, $\mathbf{Z}$ stands for $(Z^{(1)}, \ldots, Z^{(n)})$, and $h(\mathbf{X}^{(i)}, Z^{(i)}; \boldsymbol{\phi})$ denotes the density of $(\mathbf{X}^{(1)}, Z^{(1)})$ at $(\mathbf{X}^{(i)}, Z^{(i)})$ assuming the true parameter is $\boldsymbol{\phi}$. Since we do not have the data $\mathbf{Z}$, the idea is to replace the log-likelihood $L(\mathbf{X}, \mathbf{Z}; \boldsymbol{\phi})$ by its expectation given the data, $E[L(\mathbf{x}, \mathbf{Z}; \boldsymbol{\phi})|\mathbf{X} = \mathbf{x}]$, as the objective function to maximize over $\boldsymbol{\phi}$. For two points $\boldsymbol{\phi}$ and $\boldsymbol{\phi}'$ in the parameter set, define $Q(\boldsymbol{\phi}|\boldsymbol{\phi}') = E_{\boldsymbol{\phi}'}[L(\mathbf{x}, \mathbf{Z}; \boldsymbol{\phi})|\mathbf{X} = \mathbf{x}]$ where $E_{\boldsymbol{\phi}'}$ denotes the expectation under the parameter $\boldsymbol{\phi}'$. The EM algorithm is an iterative two-step procedure. Given an estimate $\boldsymbol{\phi}^t$ of $\boldsymbol{\phi}$ at the t-th step of the algorithm, one follows the two steps given below.

1. *E step.* Compute $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^t) \equiv E_{\boldsymbol{\phi}^t}[L(\mathbf{x}, \mathbf{Z}; \boldsymbol{\phi})|\mathbf{X} = \mathbf{x}]$.

2. *M step.* Set $\boldsymbol{\phi}^{t+1} = \arg\max_{\boldsymbol{\phi}} Q(\boldsymbol{\phi}|\boldsymbol{\phi}^t)$.

These two steps are repeated until the obtained estimates become stable.

Remarkably, for convergence properties of the EM algorithm to hold, finding the exact maximizer is not needed. It is required only to find a parameter $\boldsymbol{\phi}^*$ such that

$$Q(\boldsymbol{\phi}^*|\boldsymbol{\phi}^t) \geq Q(\boldsymbol{\phi}^t|\boldsymbol{\phi}^t).$$

For more details, see for instance [25, 23, 5, 22].

3.2 The estimation procedure

Let $\boldsymbol{\phi}^t = (\boldsymbol{\pi}^t, \boldsymbol{\mu}^t, \boldsymbol{\theta}^t, \mathbf{G}^t)$ be the list of parameters of interest at the t-th step, where $\boldsymbol{\mu}^t = (\mu_{11}^t, \ldots, \mu_{d1}^t, \ldots, \mu_{1K}^t, \ldots, \mu_{dK}^t)$, $\boldsymbol{\pi}^t = (\pi_1^t, \ldots, \pi_K^t)$, $\boldsymbol{\theta}^t = (\boldsymbol{\theta}_1^t, \ldots, \boldsymbol{\theta}_K^t)$ and $\mathbf{G}^t = (G_1^t, \ldots, G_d^t)$. The density of $Z^{(1)}$ given $\mathbf{X}^{(1)} = \mathbf{x}^{(i)}$ at z under the parameter $\boldsymbol{\phi}^t$ is written as

(8)
$$h(z|\mathbf{x}^{(i)}; \boldsymbol{\phi}^t) = \frac{\pi_z^t c_z \left(G_1^t\left(x_1^{(i)} - \mu_{1z}^t\right), \ldots, G_d^t\left(x_d^{(i)} - \mu_{dz}^t\right); \boldsymbol{\theta}_z^t\right) \prod_{j=1}^d g_j^t\left(x_j^{(i)} - \mu_{jz}^t\right)}{h(x_1^{(i)}, \ldots, x_d^{(i)}; \boldsymbol{\phi}^t)},$$

where the function h in the denominator was given in (5). According to the formal EM algorithm given in Section 3.1, we should maximize over $\boldsymbol{\phi}$ the objective function

$$\begin{aligned}(9) \qquad Q(\boldsymbol{\phi}|\boldsymbol{\phi}^t) =& E\left[\sum_{i=1}^n \left\{\log h(\mathbf{x}^{(i)}|Z^{(i)}; \boldsymbol{\phi}) + \log P(Z^{(i)} = z)\right\} |\mathbf{X} = \mathbf{x}\right] \\ =& \sum_{z,i} \Big[\log c_z \left\{G_1(x_1^{(i)} - \mu_{1z}), \ldots, G_d(x_d^{(i)} - \mu_{dz}); \boldsymbol{\theta}_z\right\} \\ & + \sum_{j=1}^d \log g_j(x_j^{(i)} - \mu_{jz}) + \log \pi_z\Big] h(z|\mathbf{x}^{(i)}; \boldsymbol{\phi}^t),\end{aligned}$$

which involves an infinite dimensional parameter. A solution of this problem being unknown, we instead borrow ideas from [4] and propose a semiparametric and stochastic EM-like algorithm, in which passing from the t-th state $\boldsymbol{\phi}^t$ to the $(t+1)$-th state $\boldsymbol{\phi}^{t+1}$ involves the following steps.

1. *E step.* Compute $h(z|\mathbf{x}^{(i)}; \boldsymbol{\phi}^t)$ for $i = 1, \ldots, n$ and $z = 1, \ldots, K$ by using (8).

2. *S step.*

   (a) Define $\tilde{Z}^{t+1}(\mathbf{u})$, $\mathbf{u} \in \mathcal{R}^d$ to follow a multinomial distribution with probabilities $P(Z^{(1)} = z|\mathbf{X}^{(1)} = \mathbf{u}; \boldsymbol{\phi}^t)$, $z = 1, \ldots, K$.

   (b) Given the data $\mathbf{x}^{(i)}$, generate a sample $\tilde{Z}^{t+1}(\mathbf{x}^{(i)})$, $i = 1, \ldots, n$.

   (c) Put $\tilde{x}_j^{(i),t+1} = x_j^{(i)} - \mu_{j,\tilde{Z}^{t+1}(\mathbf{x}^{(i)})}^t$ for $i = 1, \ldots, n$, $j = 1, \ldots, d$.

(d) Update the symmetric distributions by computing kernel estimates

$$\hat{g}_j(u) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{u - \tilde{x}_j^{(i),t+1}}{h_n}\right), \qquad \hat{G}_j(u) = \int_{-\infty}^{u} \hat{g}_j(s)\, ds$$

where K is some kernel density and h_n is some bandwidth.

(e) Symmetrize $g_j^{t+1}(u) \equiv \{\hat{g}_j(u) + \hat{g}_j(-u)\}/2$

3. *M step.*

(a) Update the cluster weights

$$\pi_z^{t+1} = \frac{1}{n} \sum_{i=1}^{n} h(z|\mathbf{x}^{(i)}; \boldsymbol{\phi}^t)$$

(b) Update the location parameters

$$\mu_{jz}^{t+1} = \frac{\sum_{i=1}^{n} x_j^{(i)} h(z|\mathbf{x}^{(i)}; \boldsymbol{\phi}^t)}{\sum_{i=1}^{n} h(z|\mathbf{x}^{(i)}; \boldsymbol{\phi}^t)}$$

(c) Update the copula parameters; for $z = 1, \ldots, K$,

$$\boldsymbol{\theta}_z^{t+1} = \arg\max_{\boldsymbol{\theta}_z} \sum_i \log c_z \left\{ G_1^{t+1}(x_1^{(i)} - \mu_{1z}^{t+1}), \ldots, G_d^{t+1}(x_d^{(i)} - \mu_{dz}^{t+1}); \boldsymbol{\theta}_z \right\}$$

In S step (c), a sample $\tilde{x}_j^{(i),t+1}$, $i = 1, \ldots, n$ is constructed. Note that if $\mathbf{X}^{(i)}$ is distributed according to h with parameter $\boldsymbol{\phi}^t$, then $\tilde{x}_j^{(i),t+1}$, $i = 1, \ldots, n$, constitutes a sample of G_j^t by Lemma 1 in [4]. In S step (d), the bandwidth h_n for the kernel density estimates can be chosen following one of the numerous methods available in the literature. See e.g. [16] for a review of bandwidth selection methods and [27] for a book on nonparametric density estimation. Note that the statistical software `R` (`https://www.r-project.org/`) uses plug-in methods [30] by default.

We now explain the heuristic underlying the derivation of the proposed algorithm. First, note that, because the term $\log \pi_z$ is isolated in the expression of Q in (9), the formula in M step (a) for the update of the cluster weights π_z is the same as in the common EM algorithm. To proceed, we use a heuristic which mimics the strategy employed in many copula (non-mixture) models: we first estimate the marginals and then plug their estimates into the part of the likelihood which involves the copula term. See for instance [10] and [14] Chapter 10.

The objective function corresponding to the j-th marginal is given by

$$Q_j(\boldsymbol{\phi}_j|\boldsymbol{\phi}^t) = \sum_{i,z} [\log g_j(x_j^{(i)} - \mu_{jz}) + \log \pi_z] h(z|\mathbf{x}^{(i)}; \boldsymbol{\phi}^t),$$

where $\boldsymbol{\phi}_j = (\{\mu_{jz}\}_{z=1}^K, g_j)$; note that we have, in the marginal parameter vector $\boldsymbol{\phi}_j$, deliberately forgotten the mixing proportions π_z, since, in view of the above remark, we know how to update them. The above objective function corresponds exactly to the univariate model of [4] except that the original weights

$h_j(z|x_j^{(i)};\boldsymbol{\phi}_j)$ have been replaced by the actual weights $h(z|\mathbf{x}^{(i)};\boldsymbol{\phi}^t)$ of the multivariate model. Therefore, following the method of [4] but with the approriate weights yields the updates for $\{\mu_{jz}\}_{z=1}^K$ in M step (b) and g_j in the S step. Finally, the updated estimates for μ_{jz} and g_j are pluged in into the arguments of the copulas c_z in (9), and the maximization over $\boldsymbol{\theta}_z$ can be performed as in M step (c).

Initialization In order to obtain a first estimate $\boldsymbol{\phi}^0$, we did the following. We obtained a first clustering of the observations by applying a nonparametric clustering procedure, as for instance the k-means algorithm [12]. The k-means algorithm provides with the centers of groups, which we use for the μ_{jz}^0. Then a sample of G_j^0 can be constructed as

$$\cup_z\{x_j^{(i)} - \mu_{jz}^0 : \text{the } i\text{-th observation was classified in the } z\text{-th cluster}\}.$$

Finally, $\boldsymbol{\theta}_z^0$ is computed using data that were classified within the z-th group and using the semiparametric procedure of [10].

The above estimation procedure assumed that the copula families and the number of mixture components were known. To relax these assumptions, we may adapt the methods of the standard literature on parametric mixture models, see e.g. [23] Chapter 6. In the literature, a common method is to compare the Akaike Information Criterion [26, 1, 2], also known as AIC, across all models and for a different number of groups. Recall that the AIC is equal (strictly speaking, proportional) to the maximum log-likelihood minus the number of free parameters. In our case, however, it is not so clear what the number of free parameters is because our model involves an infinite dimensional parameter set. Nevertheless, if we wish to select a model among all semiparametric models of the form (5), then the only choice to make is about the copula families. Thus, among several semiparametric models with different copula families and a different number of clusters, we propose to select the model which maximizes the pseudo-AIC

(10) (maximum log-likelihood)-(number of free copula parameters).

It should be noted that this strategy is, at the moment, a heuristic, because the theoretical properties of the EM-like algorithm have not been established yet and therefore the assumptions on which are based the AIC of Akaike are not met. That being said, we found through a simulation study that the pseudo-AIC criterion (10) performs well, see Section 4.2.

4 Illustrations

This section provides with three illustrations of the location-shift semiparametric copula-based mixture model (5) at work. The first illustration compares this model to the standard Gaussian mixture model. It shows that we can fit non-Gaussian marginals and that marginal estimation can be robust with respect to copula misspecification. The second illustration investigates the effect of noise and the convergence speeds of the standard EM algorithm and its semiparametric and stochastic version. Finally the third illustration assesses the usefulness of the proposed pseudo-AIC criterion (10) for model selection when the copula families or the number of clusters is unknown.

4.1 A first illustration

We simulated $n = 300$ observations of dimension $d = 2$ under the model (5) with $K = 3$ groups and the following parameters: $(\mu_{11}, \mu_{21}) = (0, 3)$, $(\mu_{12}, \mu_{22}) = (3, 0)$, $(\mu_{13}, \mu_{23}) = (-3, 0)$ and $\pi_1 = \pi_2 = \pi_3 = 1/3$. The symmetric distributions G_1 and G_2 were respectively chosen to be a Student distribution with 4 degrees of freedom (6) and a Laplace distribution (7) such that its variance equals 1/2. Finally, the copulas were Gaussian copulas, that is,

$$c_z(u_1, u_2; \theta_z) = \frac{1}{\sqrt{1-\theta_z^2}} \exp\left\{ -\frac{1}{2(1-\theta_z^2)} \left[(z_1^2 + z_2^2)\theta_z^2 - 2z_1 z_2 \theta_z \right] \right\},$$

where $u_1, u_2 \in [0, 1]$, z_j is the quantile of order u_j of the standard normal distribution. The parameters were $\theta_1 = 1/2$, $\theta_2 = 0$, $\theta_3 = -1/2$.

Based on these simulated data, inference was performed with the semiparametric and stochastic EM-like algorithm of Section 3.2 under three different models: the standard Gaussian mixture model (hereafter denoted by GMM for Gaussian Mixture Model), our model with Gaussian copulas (SPG for Semi-Parametric Gauss) and our model with Frank copulas (SPF SemiParametric Frank), the formula for Frank copulas being

$$c_z(u_1, u_2; \theta_z) = \frac{\partial^2 C_z(u_1, u_2; \theta_z)}{\partial u_1 \partial u_2}, \qquad \text{where}$$
$$C_z(u_1, u_2; \theta_z) = -\frac{1}{\theta_z} \log\left(1 + \frac{(e^{-\theta_z u_0} - 1)(e^{-\theta_z u_j} - 1)}{e^{-\theta_z} - 1} \right), \qquad u_1, u_2 \in [0, 1],$$

where $\theta_z \neq 0$ and $-\infty < \theta_z < \infty$ (see e.g. [24] p. 116). In the first model GMM, the copula is correctly specified but the marginals are not and in the third model SPF, it is the opposite: the marginals are correctly specified but the copula is not. Note also that GMM corresponds to a copula-based mixture model (2) in which the copulas c_z are Gaussian copulas and the marginals h_{jz} are Gaussian distributions.

We used `R` (`https://www.r-project.org/`) and the package `Mclust` [8] for estimation in the GMM model. For the semiparametric models, we implemented the algorithm of Section 3.2. Kernel density estimation was performed with the package `ks` [6] (we kept the bandwidth provided by default) and the starting parameters were obtained with `stats::kmeans`, the default nearest neighbors algorithm [12] of `R`. The raw simulated data are shown in Figure 1, while Figure 2 displays the expected complete log-likelihood conditionally on the data at each step of the EM-like algorithm as a check of convergence. A quick stabilization indicates that the algorithms have converged.

Figure 3 shows the estimated groups under the three tested models GMM, SPG and SPF. The estimated contour lines of level 95% were added for comparison. These are defined as $\{(x_1, x_2) \in \mathcal{R}^2 : h_z(x_1, x_2; \hat{\boldsymbol{\phi}}) = c\}$, where $\hat{\boldsymbol{\phi}}$ is the estimated parameter vector and c is a constant such that $P(h_z(X_1, X_2; \hat{\boldsymbol{\phi}}) \leq c) \approx 0.05$ (the approximation was carried out by drawing a bootstrap sample of size 10000). Note that the group symbols are interchanged from one picture to another due to label switching, but the Figure is still easily readable. While, of course, the shapes of the contour lines are elliptical under the GMM model, it is not so for the models SPG and SPF: these are able to capture wider shapes. In particular, one sees clearly the effect of nonparametric estimation. Note that

the group assignments are different with respect to the point marked by the symbol "P". Neither the standard model GMM nor the misspecified model SPF were able to correctly classify this point. But the correct model SPG did so, which demonstrates the reliability of our estimation procedure.

Regarding the marginal distributions, the results, displayed in Figure 4, are interesting, too. The univariate data are displayed along with the estimated densities under the three tested models, plus the true underlying distribution. For convenience, Table 1, which contains L_2 rescaled distances between the estimated densities and the true density, precisely

$$10000 * \int_{-\infty}^{\infty} \left[h_z(x_j; \hat{\phi}) - h_z(x_j) \right]^2 dx_j, \tag{11}$$

$j = 1, 2$, $z = 1, 2, 3$, summarizes the results.

One can see that the distances for the SPG and SPF models are much lower than for the GMM model. In the first marginal, the error is reduced by about 22—68%. The gain in the second marginal is even more striking: the error is reduced by about 81—90%. As the model GMM has misspecified marginals, it comes as no surprise that it is unable to capture the shape of the true marginals. It is not so for SPF and SPG. Interestingly enough, we found that the estimated marginals for SPG and SPF were indistinguishable (and so were represented by the same line on the graphs) and closer to the true marginal. Note that one of these models, SPF, was wrong — as the copulas were misspecified —, but this did not deteriorate marginal estimation. This suggests that, in the model (5), marginal estimation is robust in regard to copula misspecification. Thus, thanks to the nonparametric part of the semiparametric and location-shift mixture model, marginals can be estimated correctly irrespective of copula modeling.

The results of this section emphasize the following points. In the case of heterogeneous data — in the sense that the marginal distributions may be different, as was the case in our illustration —, standard mixture models such as the Gaussian or Student mixture models and their variants will likely fail. Copula-based mixture models may come as a remedy, but a difficulty stands up in front of the practitioner: which parametric families must be chosen for the copulas and the marginals? The results of the model GMM showed that even if the copula is correctly specified, marginal misspecifications will lead to poor estimation. In this context, we think that the model proposed in this paper brought a first answer to these questions at least for the marginals as they are estimated nonparametrically.

4.2 A second illustration

Consider the following data generating process. All the copulas c_z are Gaussian copulas with correlation parameters $\theta_z = 0.5$ for all z; the symmetric distributions G_1 and G_2 are Gaussian distributions with variances 2 and 0.5 respectively; the location parameters and cluster weights are the same as in Section 4.1. We simulated data according to the above data generating process for different noise ratios, namely, for noise ratios of 0%, 10% and 20%. A noise ratio of 10% means that 10% of the simulated vectors are removed and replaced by vectors coming from a uniform distribution with independent coordinates

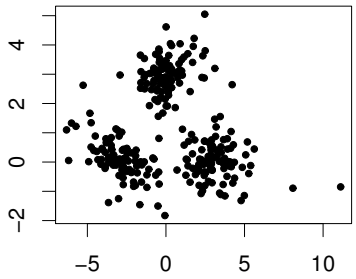
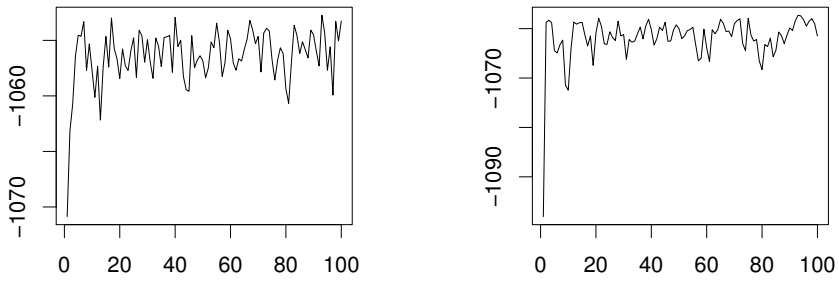


Figure 1: The raw simulated data.

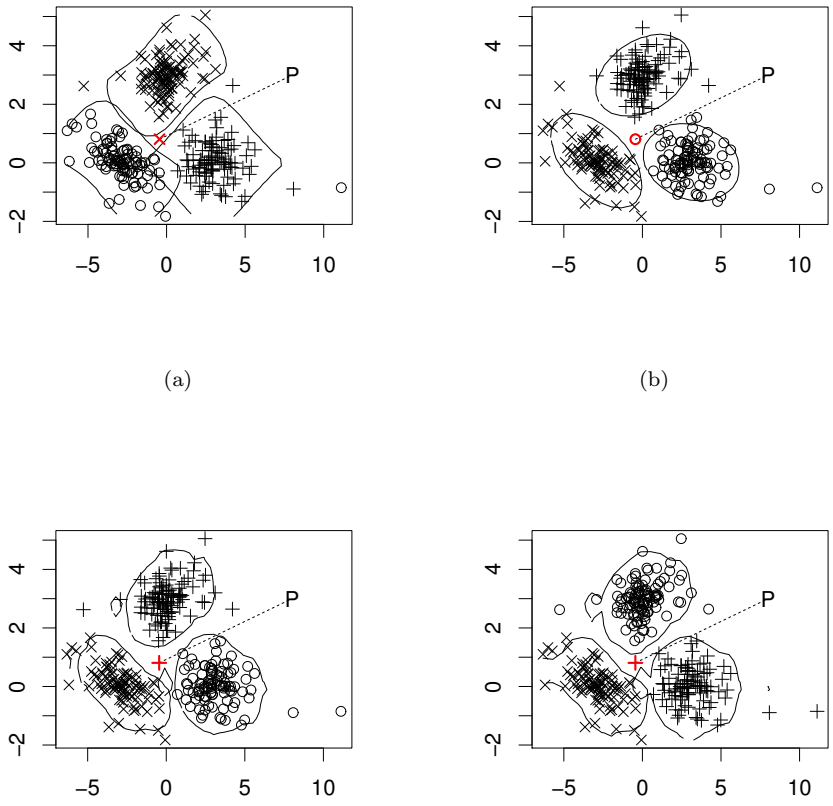


(a) (b)

Figure 2: Expected log-likelihood values conditionally on the data for SPG (a) and SPF (b).

		first marginal	second marginal
group 1	GMM	62	248
	SPG	35	60
	SPF	30	59
group 2	GMM	76	210
	SPG	27	35
	SPF	24	39
group 3	GMM	45	442
	SPG	28	43
	SPF	35	48

Table 1: Rescaled L_2 distances, see (11), between the estimated marginal distributions and the true distribution, for each group and each marginal.



(a) (b)

(c) (d)

Figure 3: Estimated group assignments for the model GMM (b), SPG (c) and SPF (d). The true group assignments corresponds to (a). The contour lines are such that, for each group, the probability of falling inside them is 95%. The group symbols are represented up to label switching.

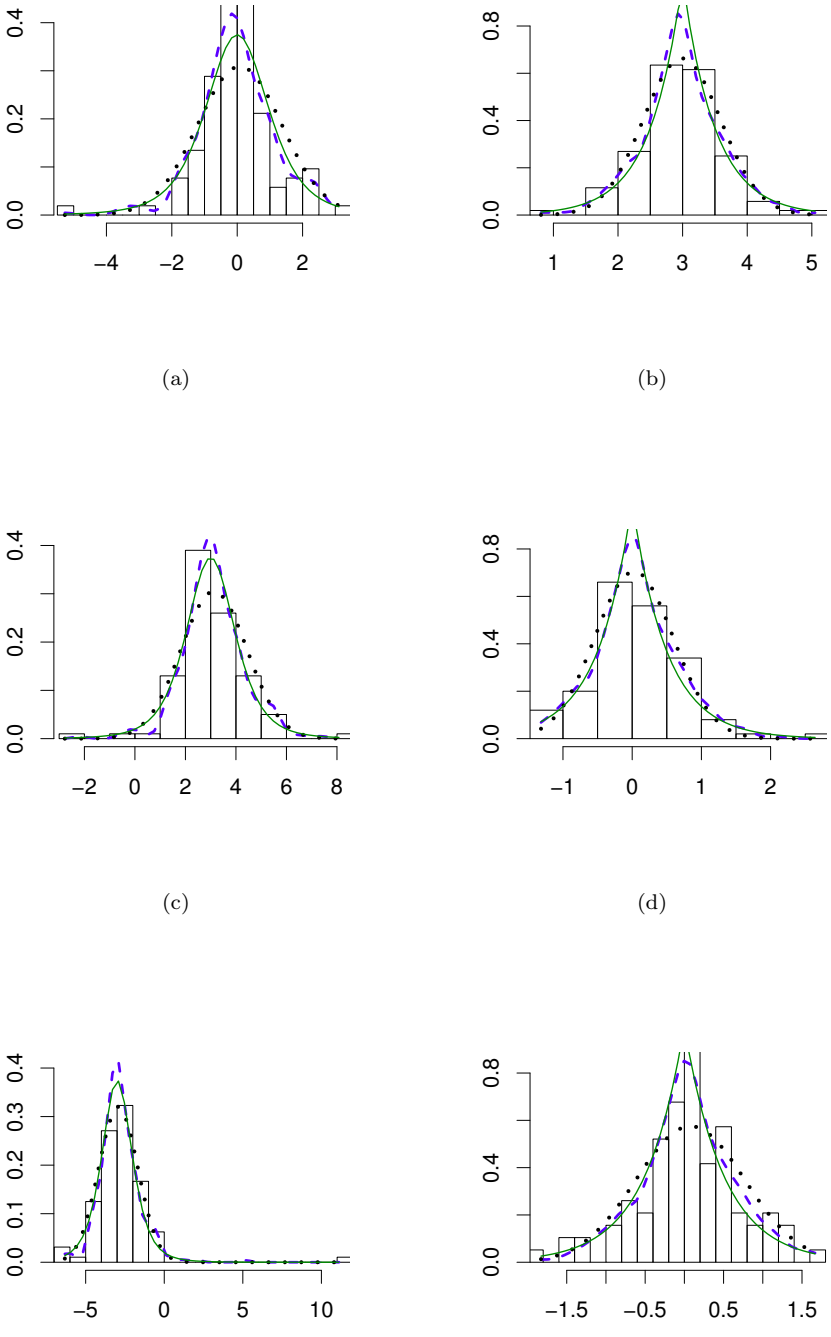


Figure 4: Histograms for the univariate data along with the estimated densities under the three tested models (GMM: black dotted line; SPG and SPF: blue dashed line), as well as the true density (plain green line). The first column from the left is the first marginal, and each row represents one group: the first row from top is the first group, etc.

over a grid covering the range of the data. Estimation in the standard Gaussian mixture model (GMM) and the semiparametric and location-shift mixture model with Gaussian copulas (SPG) was performed with the standard EM algorithm of Section 3.1 and the semiparametric and stochastic EM-like algorithm presented in Section 3.2 respectively. In both cases we assumed the knowledge that all groups have a common copula.

The results of our experiment are summarized in Figure 5, Figure 6 and Figure 7. Figure 5 displays minus the objective function $-Q(\boldsymbol{\phi}^t|\boldsymbol{\phi}^{t-1})$, $t = 0, 1, \ldots$, evolving through the iterations of the considered algorithms. First, we observe that GMM and SPG perform quite similarly in terms of convergence speed but also in terms of the magnitude of the objective values. This suggests that SPG can perform as well as GMM even in situations where the parametric assumptions of the later are met. Second, we observe that increasing the level of noise will increase the fitting error, as expected. For the noise ratio of 20%, the convergence of SPG is slower.

Figure 6 displays the quantities ($t = 0, 1, \ldots$)

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\left(\sum_z \pi_z^t g_j^t(x_j^{(i)} - \mu_{jz}^t) - \pi_z g_j(x_j^{(i)} - \mu_{jz})\right)^2}{\sum_z \pi_z g_j(x_j^{(i)} - \mu_{jz})}, \qquad j = 1, 2, \tag{12}$$

which are approximately the L_2 distances between the estimated marginals at the t-th step and the true marginals. Here π_z, μ_{jz} and g_j denote the truth. Without noise, SPG performs better than GMM for the first marginal but poorer for the second. (Note that the scales in the y-lines of the pictures are different.) The fit deteriorates as the noise level increases; an exception being the noise level of 20% for the first marginal. Again, the convergence speed of GMM and SPG are similar. This observation is a point in favor of SPG since it makes no parametric assumptions and yet performed as good as GMM. Figure 7 displays the mean square error (MSE) values for the copula parameter $|\theta^t - \theta|$ and the cluster weights $(\sum_{z=1}^{K}(\pi_z^t - \pi_z)^2/K)^{1/2}$. Quite interestingly, for a noise ratio less than 20%, the copula parameter was better estimated with SPG. As already noticed before, the convergence speed of both SPG and GMM algorithms are similar. For a noise ratio of 20% however, SPG did not converge yet after 50 iterations. The cluster weights were estimated similarly for both models at all noise levels.

4.3 A third illustration

In this third illustration, data were simulated under the same generating process as in the second Illustration of Section 4.2 but without noise only. We performed estimation under 20 different semiparametric location-shift copula-based models: the number of clusters were assumed to be 2,3 or 4 and the copulas were assumed to belong to the Gauss, Student, Frank family or assumed to be the independence copulas. For the Gauss family, both the free model and the restricted model were tried. The free model is that in which the different clusters are allowed to have different copulas, and the restricted model is that in which all clusters have a common copula. For the other models, all clusters were assumed to have a common copula. Our goal is to compare the 20 pseudo-AIC values and see wether this criterion is able to select the true model, that is, the

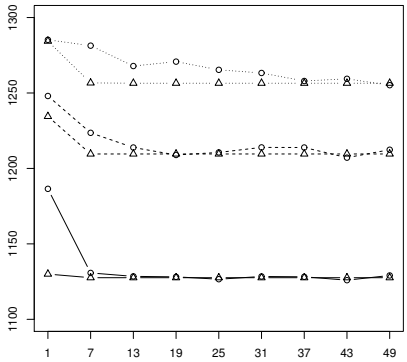
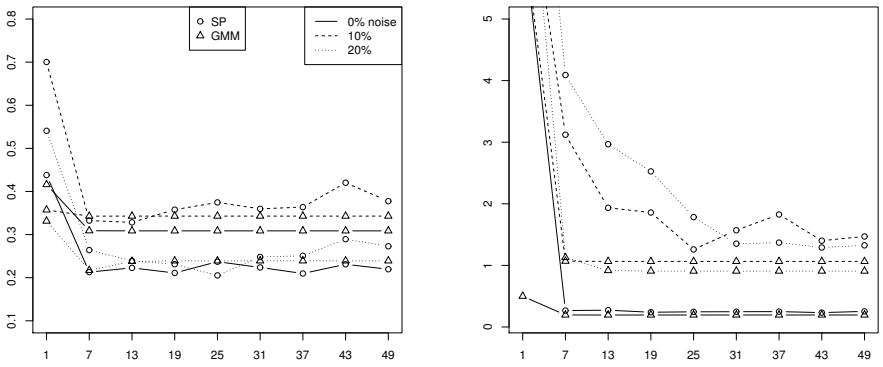
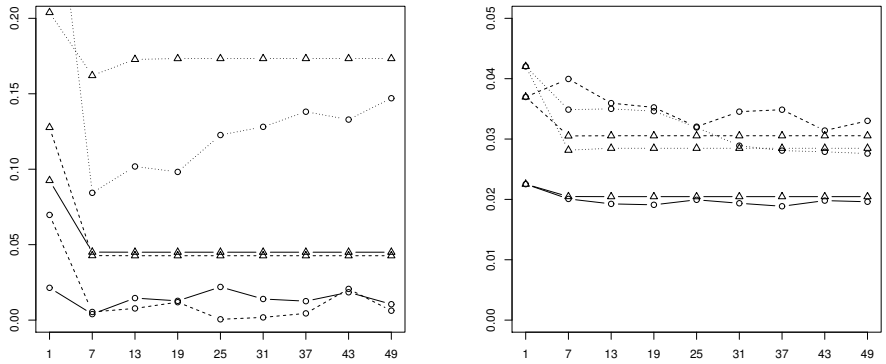


Figure 5: Minus objective function values $-Q(\boldsymbol{\phi}^{t+1}|\boldsymbol{\phi}^t)$ for $t = 0, 1, \ldots$. There are three different levels of noise and two models. The plain, dashed and dotted lines correspond to a noise level of 0%, 10% and 20% respectively. The triangles and the circles correspond to the standard Gaussian mixture model and the semiparametric model, respectively.



(a) (b)

Figure 6: Approximated marginal L_2 distances (12) for $t = 0, 1, \ldots$. There are three different levels of noise and two models. The plain, dashed and dotted lines correspond to a noise level of 0%, 10% and 20% respectively. The triangles and the circles correspond to the standard Gaussian mixture model and the semiparametric model, respectively. (a) is the first marginal and (b) is the second.



(a) (b)

Figure 7: Mean Square Error values for the copula parameter (a) and the cluster weight (b) through the iterations of the estimation algorithms. There are three different levels of noise and two models. The plain, dashed and dotted lines correspond to a noise level of 0%, 10% and 20% respectively. The triangles and the circles correspond to the standard Gaussian mixture model and the semiparametric model, respectively.

model in which the copulas are Gaussian and the number of clusters is $K = 3$. The computation of the maximum log-likelihood was approximated by the objective function Q. Indeed, heuristically, $Q(\boldsymbol{\phi}^{t+1}|\boldsymbol{\phi}^t)$ tends to the maximum log-likelihood as $t \to \infty$. The results are reported in Table 2.

K	Gauss$^{\text{res}}$	Independence	Student	Frank	Gauss$^{\text{free}}$
2	-1306	-1349	-1211	-1299	-1191
3	-1126	-1159	-1138	-1131	-1127
4	-1202	-1188	-1211	-1203	-1162

Table 2: Approximated pseudo-AIC values (10) for different parametric families and assuming a different number of groups. Gauss$^{\text{res}}$ stands for the model where all the K clusters have a common Gaussian copula. Gauss$^{\text{free}}$ stands for the model where the copulas of the K different clusters are allowed to have different parameters (but all being Gaussian).

One can see that for each column of the table, that is, for each given parametric family, the pseudo AIC criterion is able to select the correct number of clusters. Moreover, for the third row of the table, that, is, assuming the correct number of clusters, we see that the pseudo-AIC criterion is able to select the correct parametric family for the copulas. These results are encouraging and suggest that the pseudo-AIC is a reasonable criterion for model selection within the class of location-shift semiparametric copula-based mixture models.

5 Beyond location-shift models: hints

While the location-shift hypothesis (4) may be in order in some applications, it is nevertheless a strong hypothesis. Relaxing this assumption in the context of semiparametric copula-based mixture modeling is still an open research problem. In this section, we give some hints that may permit to get rid of this hypothesis.

In order to go beyond the location-shift hypothesis, a natural extension of the model (5) may be to assume, in place of (4), a hypothesis of the form

$$H_{jz}(x_j) = G_j\left(\frac{x_j - \mu_{jz}}{\sigma_{jz}}\right), \qquad \sigma_{jz} > 0,$$

where, in addition to the symmetry properties, the G_j should verify

$$\int x^2 g_j(x)\, dx = 1, \qquad j = 1, \ldots, d.$$

The semiparametric and stochastic EM-like algorithm of Section 3.2 could be adapted by modifying the S step. First, and obviously, we would replace S step (c) by

$$\tilde{x}_j^{(i),t+1} = \frac{x_j^{(i)} - \mu^t_{j,\tilde{Z}^{t+1}(\mathbf{x}^{(i)})}}{\sigma^t_{j,\tilde{Z}^{t+1}(\mathbf{x}^{(i)})}}.$$

Then, we would need to find a nonparametric estimation procedure which permits to compute an estimate g_j^{t+1} based on the sample $\tilde{x}_j^{(i),t+1}$, $i = 1, \ldots, n$. Based on numerical experiments that we made, we believe that finding an estimator which ensures that

$$\int x g_j^{t+1}(x)\, dx = 0, \qquad \int x^2 g_j^{t+1}(x)\, dx = 1 \tag{13}$$

is crucial as, if the estimator fails to satisfy these properties, then the variance of the distribution represented by g_j and the scaling parameters σ_{jz} could be cofounded. The standard kernel estimator used so far does not satisfy the properties (13).

Thus, it appears that a key issue in relaxing the location-shift hypothesis, is to construct a nonparametric density estimator which satisfies the properties (13). To the best of our knowledge, such an estimator is not directly available in the literature, and one would have to adapt existing methods. Maximum penalized likelihood methods, see e.g. [11] or [27] Chapter 5.4, might be adapted to satisfy our requirements but one would have to take care of computational convenience as these methods will have to be embedded within a EM-like algorithm. Another direction of research would be to get rid of any structure whatsoever for the marginals and try to estimate them with weighted kernel estimators, as in [3, 19].

Finally, it is to be stressed that relaxing the location-shift only hypothesis may raise identifiability issues. The advantage of the location-shift only hypothesis is that these issues are conjectured to be limited, as was shown in [13] for univariate semiparametric location-shift mixture models.

6 Discussion

In this paper, we built a new copula-based mixture model. Novelty lies in the fact that it is nonparametric, and therefore allows reduction of the modeling assumptions. The nonparametric part of the model resides in its marginal distributions. For each dimension, an observation is assumed to come from a nonparametric and symmetric distribution whose location shifts according to group assignment. As a result, nonparametric estimation of this invariant distribution is made possible, and thus one can estimate many different types of marginals without any modeling effort. We saw that marginal estimation is robust with respect to copula misspecification. Moreover, we saw on a simulation experiment that the stochastic and semiparametric EM-like algorithm could converge as fast as the standard EM algorithm. We also saw in this experiment that the copula parameter was better estimated with the semiparametric approach. Finally, in order to relax the assumption that the number of clusters is known or that the parametric copula families are known, we proposed a version of the Akaike Information Criterion (AIC) for model selection within the class of semiparametric location-shift copula-based mixture models. This criterion was shown to constantly select the true model in our simulations.

We showed that location-shift semiparametric copula-based mixture models have several good properties, but they rely on the assumption that the clusters have the same scale. Relaxing this assumption is an obvious and important line of research for the future. A possible way one might want to follow for addressing this open problem was given in a devoted section.

Acknowledgment. The author thanks three anonymous referees and the editor in chief who made suggestions that helped to improve this paper. The research by G. Mazo was funded by a "Projet de Recherche" of the "Fonds de la Recherche Scientifique — FNRS" (Belgium).

A Appendix

Sklar's theorem states that any distribution function H with continuous marginals $H_1, \ldots, H_d$ can be decomposed as

$$H(x_1, \ldots, x_d) = C(H_1(x_1), \ldots, H_d(x_d)), \tag{14}$$

for any $(x_1, \ldots, x_d)$ in the domain of H. The function $C : [0,1]^d \to [0,1]$ is unique and is called the *copula* and can be viewed as the dependence structure of the random vector $(X_1, \ldots, X_d) \sim H$. In particular, C is itself a distribution function with standard uniform marginals. It can be seen that it is the distribution of $(H_1(X_1), \ldots, H_d(X_d))$. Differentiating with respect to $x_1, \ldots, x_d$ in (14), we get

$$h(x_1, \ldots, x_d) = c(H_1(x_1), \ldots, H_d(x_d)) \prod_{j=1}^{d} h_j(x_j),$$

where h is the probability density function of H, h_j is the probability density function of H_j, and c is the the probability density function of C. If we apply the above formula within each cluster, we get our formula (2).

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd Inter. Symp. on Information Theory*. Petrov, B. N. and Csaki, F., 1973.

[2] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[3] T. Benaglia, D. Chauveau, and D. R. Hunter. An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526, 2009.

[4] L. Bordes, D. Chauveau, and P. Vandekerkhove. A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics & Data Analysis*, 51, 2007.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39:1–38, 1977.

[6] T. Duong. ks: Kernel Smoothing, 2015. R package.

[7] F. Forbes and D. Wraith. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing*, 24(6):971–984, 2014.

[8] C. Fraley, A. E. Raftery, T. M. Brendan, and L. Scrucca. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*, 2012.

[9] C. Genest and A.-C. Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368, 2007.

[10] C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

[11] I. J. Good and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58:255–277, 1971.

[12] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C*, 28:100–108, 1979.

[13] D. R. Hunter, S. Wang, and T. P. Hettmansperger. Inference for mixtures of symmetric distributions. *The Annals of Statistics*, 35:224–251, 2007.

[14] H. Joe. *Multivariate models and dependence concepts*. Chapman & Hall/CRC, Boca Raton, FL, 2001.

[15] H. Joe. *Dependence Modeling with Copulas*. Chapman & Hall, 2014.

[16] M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996.

[17] I. Kosmidis and D. Karlis. Model-based clustering using copulas with applications. *Statistics and Computing*, pages 1–21, 2015.

[18] S. Lee and G. J. McLachlan. Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24(2):181–202, 2014.

[19] M. Levine, D. R. Hunter, and D. Chauveau. Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, 98(2):403–416, 2011.

[20] M. Marbac, C. Biernacki, and V. Vandewalle. Model-based clustering of gaussian copulas for mixed data. *arXiv preprint arXiv:1405.1299*, 2014.

[21] M. Marbac, C. Biernacki, and V. Vandewalle. Model-based clustering for conditionally correlated categorical data. *Journal of Classification*, 32(2):145–175, 2015.

[22] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[23] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[24] R. B. Nelsen. *An introduction to copulas*. Springer, New York, 2006.

[25] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

[26] Y. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics*. KTK Scientific Publishers, 1986.

[27] B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall, 1998.

[28] A. Sklar. Fonction de répartition dont les marges sont données. *Inst. Stat. Univ. Paris*, 8:229–231, 1959.

[29] M. Vrac, L. Billard, E. Diday, and A. Chédin. Copula analysis of mixture models. *Computational Statistics*, 27(3):427–457, 2012.

[30] M. P. Wand and M. C. Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116, 1994.