A semiparametric and location-shift copula-based mixture model

Gildas Mazo

▶ **To cite this version:**

Gildas Mazo. A semiparametric and location-shift copula-based mixture model. 2016. hal-01263382v1

HAL Id: hal-01263382

https://hal.science/hal-01263382v1

Preprint submitted on 27 Jan 2016 (v1), last revised 28 Jan 2017 (v3)

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

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

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 researchers have started to construct and study broader generic models without appealing to these 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, it increases much the number of choices the practitioner has to make, 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, which allows for data adaptation without any modeling efforts.

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

1 Introduction

The modeling of a mixture of distributions has long rested upon Gaussian distributions [14] and it is only recently that researchers have started to construct and study broader generic models [4, 12, 13, 11, 14].

Among these extensions, models featuring copulas are still rare, but certainly promising [13, 11, 18]. Indeed, copulas allow for building very flexible models, as they permit to handle the marginal distributions and the dependence separately. More precisely, denoting by h_z the conditional density within the z-th population group, the copula c_z associated to h_z can be defined as

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

where, for $j = 1, \ldots, d$, H_{jz} and h_{jz} denote the distribution function and the density of the j-th variable of interest in the z-th group. This decomposition is sometimes called the copula decomposition or the Sklar's decomposition, in view of Sklar's theorem [17]. For more details about copulas in general, see e.g. [15, 10, 7].

So, in order to build a parametric model for the z-th group out of (1), $d+1$ parametric families have to be chosen: one for the copula c_z, and one for each marginal distribution h_{zj}, $j = 1, \ldots, d$. This situation can be embarrassing in practice as testing all the existing parametric families of the literature is unfeasible. Often, the practitioner simply assumes that all the margins come from the same parametric family, that can be too restrictive in applications. This issue, in particular, was pointed out in [18].

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 margins. In this respect it echos the common semiparametric copula models of the "nonmixture" literature, see e.g. [8]. 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 efforts.

This paper is organized as follows. Section 2 presents our new location-shift semiparametric copula-based mixture model. Section 3 deals with estimation. Identifiability issues are also discussed. Section 4 illustrates the model's features and a Discussion closes this 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 assumed to be known and 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 $\pi_z = P(Z = z)$. As Sklar's decomposition (1) can be applied to each $z = 1, \ldots, K$, the density of any copula-based mixture model

writes [11]

$$h(x_1,\ldots,x_d)=\sum_{z=1}^{K}\pi_z c_z\{H_{1z}(x_1),\ldots,H_{dz}(x_d)\}\prod_{j=1}^{d}h_{jz}(x_j). \tag{2}$$

Now, for all $j=1,\ldots,d$, we assume

$$H_{jz}(x_j)=G_j\left(x_j-\mu_{jz}\right), \qquad x_j\in\mathcal{R} \tag{3}$$

so that

$$h(x_1,\ldots,x_d)=\sum_{z=1}^{K}\pi_z c_z\{G_1(x_1-\mu_{1z}),\ldots,G_d(x_d-\mu_{dz})\}\prod_{j=1}^{d}g_j(x_j-\mu_{jz}), \tag{4}$$

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 $-\infty<x_j<+\infty$. This location shift hypothesis induces 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 (3) means that the marginal distributions in a cluster z and a cluster z' differ only by a shift. 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.

Several arguments can be made in favor of Hypothesis (3). 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, popular mixture models satisfy a location-shift hypothesis, as the standard Gaussian mixture model with equal variances, in which (3) holds with

$$g_j(x_j)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x_j-\mu_{jz}}{\sigma}\right)^2\right\}, \qquad x_j\in\mathcal{R}$$

where μ_{jz} and σ are the location and standard deviation of the j-th component in the z-th cluster respectively. To play the role of g_j, other symmetric distributions can be called for, as for instance, the student distribution

$$g_j(x_j)=\frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma\nu/2}(1+\frac{x_j^2}{\nu})^{-(\nu+1)/2}, \qquad \nu>2,\, x_j\in\mathcal{R}, \tag{5}$$

or the Laplace distribution

$$g_j(x_j)=\frac{1}{2b}\exp\left(-\frac{|x_j|}{b}\right), \qquad b>0,\, x_j\in\mathcal{R}. \tag{6}$$

Last but not least, hypothesis (3) allow for nonparametric estimation, as explained next.

3 Estimation

This section presents semiparametric estimation in the model (4) through a semiparametric Expectation-Maximization (EM) algorithm. Recall that the number of clusters is assumed to be known and fixed. But first, identifiability results are given.

3.1 Identifiability results

In statistics, for estimation to make sense, models have to be identifiable, that is, one distribution in the model can be identified with one parameter vector. But mixture models are particularly prone to identifiability issues.

Suppose that a parametric family has been chosen for c_z, $z = 1, \ldots, K$, in (4), so that $c_z(u_1, \ldots, u_d) = c_z(u_1, \ldots, u_d; \theta_z)$ for all $u_1, \ldots, u_d \in [0, 1]$, where θ_z is the parameter vector in the z-th group. Denote the z-th parameter set by Θ_z, so that $\theta_z \in \Theta_z$, and denote by $\Theta = \Theta_1 \times \cdots \times \Theta_K$ the set of all copula parameters. Also, let

$$\Lambda = \left\{ \pi = (\pi_1, \ldots, \pi_K) \in [0, 1]^K : \sum_{z=1}^{K} \pi_z = 1 \right\}$$

be the set containing the weigh parameters and write $\mu = (\mu_{11}, \ldots, \mu_{d1}, \ldots, \mu_{dK}) \in \mathcal{R}^{dK}$ for the location parameters. The set of all absolutely continuous symmetric distributions is denoted by $\mathcal{G}$ and we write $G = (G_1, \ldots, G_d) \in \mathcal{G}^d$ or equivalently $g = (g_1, \ldots, g_d) \in \mathcal{G}^d$.

The parametrized semiparametric model corresponding to (4) writes

$$\mathcal{H} = \left\{ h(\cdot; \pi, \mu, \theta, G) : \pi \in \Lambda, \mu \in R^{dK}, \theta \in \Theta, G \in \mathcal{G}^d \right\},$$

where $h(x_1, \ldots, x_d; \pi, \mu, \theta, G) = h(x_1, \ldots, x_d)$ in (4). The set of parameters corresponding to $\mathcal{H}$ is given by $\Lambda \times \mathcal{R}^{Kd} \times \Theta \times \mathcal{G}^d$ and is called the *admissible* set of parameters. In (4), not all distributions h are identifiable. As a counter example, suppose that $K = 2$ and that the copula parametric families of both groups are the same. The parameter vector is then $(\pi_1, \mu_{11}, \mu_{21}, \mu_{12}, \mu_{22}, \theta_1, \theta_2)$. Since both $(1, \mu_{11}, \mu_{21}, \mu_{12}, \mu_{22}, \theta_1, \theta_2)$ and $(0, \mu_{12}, \mu_{22}, \mu_{11}, \mu_{21}, \theta_2, \theta_1)$ lead to the same distribution h, it is not identifiable.

Thus, the set of all identifiable distributions h is a subset, even a proper subset, of $\mathcal{H}$. To this set of identifiable h corresponds a set of parameters, called the *effective* parameter set. Naturally, the more the admissible set agree with the effective set, the less we expect to have identifiability issues. The sense of "agree" can be loose, but was given a precise meaning in [9].

Consider the j-th margin of our semiparametric location-shift mixture model (4),

$$\mathcal{H}_j = \left\{ h_j(\cdot; \pi, \mu_j, g_j) : h_j(x_j; \pi, \mu_j, g_j) = \sum_{z=1}^{K} g_j(x_j - \mu_{jz})\pi_z, \, \forall x_j \in \mathcal{R} \right\},$$

where $\mu_j = (\mu_{j1}, \ldots, \mu_{jK})$. The admissible parameter set is then $\Lambda \times \mathcal{R}^K \times \mathcal{G}$. Define the mapping $\varphi : \Lambda \times \mathcal{R}^K \times \mathcal{G} \to \mathcal{H}_j$ such that for each $\pi \in \Lambda$, $\mu_j \in \mathcal{R}^K$ and $g_j \in \mathcal{G}$, the element $\varphi(\pi, \mu_j, g_j) \in \mathcal{H}_j$ is a function, defined by $\varphi(\pi, \mu_j, g_j)(x_j) = h_j(x_j; \pi, \mu_j, g_j)$ for all $x_j \in \mathcal{R}$. A distribution h_j in $\mathcal{H}_j$ is *identifiable* if $\varphi^{-1}(h_j)$ is a singleton in $\Lambda \times \mathcal{R}^K \times \mathcal{G}$. Let $(\Lambda \times \mathcal{R}^K)^*$ be the biggest subset of $\Lambda \times \mathcal{R}^K$ such that for all (π, μ_j) in $(\Lambda \times \mathcal{R}^K)^*$, $\varphi(\pi, \mu_j, g_j)$ is identifiable for all g_j in $\mathcal{G}$. Let us call $(\Lambda \times \mathcal{R}^K)^* \times \mathcal{G}$ the *effective parameter set* in the model $\mathcal{H}_j$.

In [9], the authors established that for $K = 2$ and $K = 3$, the difference between the admissible parameter set and the effective parameter set, that is, the difference between $\Lambda \times \mathcal{R}^K$ and $(\Lambda \times \mathcal{R}^K)^*$, is small, meaning that their difference, in the sense of the set theory terminology, has Lebesgue measure 0.

Proposition 1 (Hunter, Wang and Hettmansperger [9]). *The set* $\Lambda \times \mathcal{R}^K \times \mathcal{G} \setminus (\Lambda \times \mathcal{R}^K)^* \times \mathcal{G}$ *is of Lebesgue measure zero for* $K = 2$ *and* $K = 3$.

Proof. It was established in [9] Theorem 2 that $(\Lambda \times \mathcal{R}^K)^* = \{\pi \in \Lambda : \pi_1 \notin \{0, 1/2, 1\}\} \times \mathcal{R}^K$ for $K = 2$. For $K = 3$, Corollary 1 in the same reference established that $(\Lambda \times \mathcal{R}^K)^* \supset \{\pi \in \Lambda : \pi_1\pi_2\pi_3 \neq 0\} \times \{\mu \in \mathcal{R}^K : (\mu_2 - \mu_1)/(\mu_3 - \mu_2) \neq 1\}$. The proof is complete since a countable set has Lebesgue measure zero.

For $K \geq 4$, the authors left a conjecture.

Conjecture 1 (Hunter, Wang and Hettmansperger [9]). *Proposition 1 carries over for all* K.

Proposition 1 and Conjecture 1 state that the distributions in $\mathcal{H}_j$, that is, the marginal distributions of $\mathcal{H}$, are almost all identifiable, as the difference between the admissible parameter set and the effective parameter set is of Lebesgue measure zero. Now, if we substitute "of Lebesgue measure zero" by "at most countable", Proposition 1 still holds, as, actually, the difference *is* at most countable for $K = 2$ and $K = 3$. For larger K, this result is still unknown, but, since it is also unknown in regard to the Lebesgue measure, it is logically equivalent to slightly strengthen a bit Conjecture 1.

Conjecture 2. *The set* $\Lambda \times \mathcal{R}^K \times \mathcal{G} \setminus (\Lambda \times \mathcal{R}^K)^* \times \mathcal{G}$ *is at most countable for* $K \geq 4$.

The remaining of this section aims at extending these results to the multivariate case. One must take care that the joint model involves taking d times the set $\mathcal{R}^K$ but only one time the set Λ. Let $\psi : \Lambda \times \mathcal{R}^{dK} \times \Theta \times \mathcal{G}^d \mapsto \mathcal{H}$. Moreover, let Λ^* be the set Λ from which we have removed all π such that there exists a $\mu \in \mathcal{R}^K$ satisfying $(\pi, \mu) \notin (\Lambda \times \mathcal{R}^K)^*$. The following proposition, whose proof is given is the Appendix, identifies a set which is contained is the effective parameter set of the joint model $\mathcal{H}$.

Proposition 2. *Assume the parametric families chosen for the copulas* c_z *in* (4) *are identifiable. Then, for* $K = 2$ *and* $K = 3$, *we have*

$$\Lambda^* \times \mathcal{R}^{dK} \times \Theta \subset (\Lambda \times \mathcal{R}^{dK} \times \Theta)^* \subset \Lambda \times \mathcal{R}^{dK} \times \Theta. \tag{7}$$

Moreover, if Conjecture 2 holds, then the above statement holds for all K. *Finally, we have that*

$$(\Lambda \times \mathcal{R}^{dK} \times \Theta) \setminus (\Lambda^* \times \mathcal{R}^{dK} \times \Theta)$$

has Lebesgue measure zero.

Proposition 2 trivially implies that the difference between the admissible parameter set and the effective parameter set has Lebesgue measure zero. But this means that the distributions in the model $\mathcal{H}$ are almost all identifiable, in the Lebesgue sense.

3.2 Estimation

This section presents an EM algorithm to estimate the unknown parameters $\pi_z, \mu_{jz}, \theta_z, G_j$, $j = 1, \ldots, d$, $z = 1, \ldots, K$ of the model (4). The parameters G_j, $j = 1, \ldots, d$, are infinite-dimensional and therefore standard EM algorithms [14, 16] are unfeasible. There are however refinements on which we can rest in order to perform estimation. One such refinement is given in [1], where a stochastic and semiparametric EM algorithm is applied to a univariate semiparametric location-shift model.

Capitalizing on their results, we also present a stochastic and semiparametric EM algorithm allowing for semiparametric estimation in our model (4). But first, the ideas underlying the standard EM algorithm are recalled. Generally, consider a model of the form

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

where $\pi = (\pi_1, \ldots, \pi_K)$ and $\phi = (\phi_1, \ldots, \phi_K)$ are the parameters to be estimated. Let $X^{(i)}$, $i = 1, \ldots, n$, with $X^{(i)} = (X_1^{(i)}, \ldots, X_d^{(i)})$, be a random sample (independent and identically distributed variables) and $x^{(i)}$, $i = 1, \ldots, n$, with $x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})$, 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 $(x^{(i)}, z^{(i)})$, $i = 1, \ldots, n$, were observed, we could maximize the log-likelihood

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

where $X = (X^{(1)}, \ldots, X^{(n)})$, $Z = (Z^{(1)}, \ldots, Z^{(n)})$ and $h(X^{(i)}, Z^{(i)}; \pi, \phi)$ denotes the density of $(X^{(1)}, Z^{(1)})$ at $(X^{(i)}, Z^{(i)})$. Since we do not have the data for Z, the idea is to replace the log-likelihood $L(X, Z; \pi, \phi)$ by its expectation given the data, $E[L(x, Z; \pi, \phi)|X = x]$, as the objective function to maximize over π and ϕ. Thus, given an initial estimate (π^t, ϕ^t), one wishes to solve

$$\arg\max_{\pi,\phi} Q(\pi, \phi|\pi^t, \phi^t) \equiv \arg\max_{\pi,\phi} E_{\pi^t,\phi^t}\left[L(x, Z; \pi, \phi)|X = x\right],$$

where the subscripts attached to the expectation symbol mean that the expectation is taken with respect to these estimates. Remarkably, for convergence properties of the EM algorithm to hold, finding the exact maximizer is not needed. Only required is to find a parameter (π^*, ϕ^*) such that

$$Q(\pi^*, \phi^*|\pi^t, \phi^t) \geq Q(\pi^t, \phi^t|\pi^t, \phi^t).$$

One then proceeds iteratively, that is, sets $(\pi^{t+1}, \phi^{t+1}) = (\pi^*, \phi^*)$ and repeats the process until the obtained estimates become stable. For more details, see for instance [16, 14].

Our semiparametric and stochastic EM algorithm for doing estimation in the model (4) is given next. Let $\phi^t = (\pi^t, \mu^t, \theta^t, G^t)$ be the parameter vector of interest, where $\mu^t = (\mu_{11}^t, \ldots, \mu_{d1}^t, \mu_{12}^t, \ldots, \mu_{dK}^t)$, $\pi^t = (\pi_1^t, \ldots, \pi_K^t)$, $\theta^t =$

$(\theta_1^t, \ldots, \theta_K^t)$, and $G^t = (G_1^t, \ldots, G_d^t)$. The density of $Z^{(1)}$ given $X^{(1)} = x^{(i)}$ at z under the parameter ϕ^t is written as

$$h(z|x^{(i)}; \pi^t, \phi^t) = \frac{\pi^t c_z(G_1^t(x_1^{(i)} - \mu_{1z}^t), \ldots, G_d^t(x_d^{(i)} - \mu_{dz}^t); \theta_z^t) \prod_{j=1}^d g_j^t(x_j^{(i)} - \mu_{jz}^t)}{h(x_1^{(i)}, \ldots, x_d^{(i)}; \pi^t, \phi^t)},$$

where the function h was given in (4). The function to maximize over (π, ϕ) is given by

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

The maximization of the above function can be achieved with the following steps.

1. Update the cluster weights

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

2. Update the location parameters

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

3. Update the symmetric distribution
   (a) For each $i = 1, \ldots, n$, define $\tilde{Z}^{(i)}(u)$, $u \in \mathcal{R}$, so that it follows a multinomial distribution with probabilities $P(Z^{(1)} = z|X^{(1)} = u)$, $z = 1, \ldots, K$.
   (b) Given the data $x^{(i)}$, $i = 1, \ldots, n$, generate a sample $\tilde{Z}^{(i)}(x^{(i)})$, $i = 1, \ldots, n$.
   (c) Put $\tilde{x}_j^{(i),t+1} = x_j^{(i)} - \mu_{j,\tilde{Z}^{(i)}(x_j^{(i)})}^t$ for $i = 1, \ldots, n$ (this constitutes a sample of G_j, see the Appendix), $j = 1, \ldots, d$
   (d) Compute 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 (see e.g. [2] for more details about nonparametric estimation).
   (e) Symmetrize $g_j^{t+1}(u) \equiv \{\hat{g}_j(u) + \hat{g}_j(-u)\}/2$

4. Update the copula parameters; for $z = 1, \ldots, K$,

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

The update of the weights is the same as in the standard EM algorithm, see e.g. [16, 12]. The update of the location parameters comes from the following idea [1]. If the variables $Z^{(i)}$ were observed, maximizing the objective function would lead to

$$\mu_{jz}^{t+1} = \frac{\sum_{i=1}^n x_j^{(i)} \mathbf{1}(Z^{(i)} = z)}{\sum_{i=1}^n \mathbf{1}(Z^{(i)} = z)},$$

where $\mathbf{1}(\cdot)$ is the indicator function. But since they are unobserved, $\mathbf{1}(Z^{(i)} = z)$ in the above expression is replaced by $E[\mathbf{1}(Z^{(i)} = z)|X = x] = h(z|x^{(i)}; \pi^t, \phi^t)$, its expectation given the data, similarly as in [1]. Let us note that one could have chosen to maximize $\sum_{i=1}^n \log g_j^{t+1}(x_j^{(i)} - \mu_{jz}) h(z|x^{(i)}; \pi^t, \phi^t)$ over μ_{jz}; in fact these two possibilities coincide when the distribution g_j comes from an exponential family, see e.g. [16]. The update of the symmetric distributions are as in [1]. The update of the copula parameters can be made more explicit for some copula families. The updates of the weights, location and copula parameters and the invariant distributions are independent of each other.

4 Illustrations

This section provides with an illustration of our model (4) at work. We simulated $n = 300$ observations of dimension $d = 2$ under the model (4) 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 (5) and a Laplace distribution (6) such that its variance equals $1/2$. Finally, the copulas were Gaussian copulas, that is,

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

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 our EM algorithm of Section 3 under three different models: the standard Gaussian mixture model (hereafter denoted by S, for Standard), our model with Gaussian copulas (LSG for Location Shift Gauss) and our model with Frank copulas (LSF for Location Shift 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. [15] p. 116). In the first model S, the copula is correctly specified but the margins are not and in the third model

LSF, it is the opposite: the margins are correctly specified but the copula is not. Note also that the standard Gaussian mixture model corresponds to the model (1) with the copulas c_z being Gaussian copulas and the margins h_{jz} being Gaussian distributions.

We used `R` (`https://www.r-project.org/`) and the package `Mclust` for estimation in the standard S model [5, 6] but implemented our own version of the algorithm of Section 3, in which kernel density estimation was performed with the package [3] (we kept the bandwidth provided by default). 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 algorithm as a check of convergence. A quick stabilization indicates that our algorithms have converged.

Figure 3 shows the estimated groups under the three tested models S, LSG and LSF. The estimated contour lines of level 95% were added for comparison. These are defined as $\{(x_1, x_2) \in \mathcal{R}^2 : h_z(x_1, x_2; \hat{\pi}, \hat{\phi}) = c\}$, where $\hat{\pi}, \hat{\phi}$ are the estimated parameter vectors and c is a constant such that $P(h_z(X_1, X_2; \hat{\pi}, \hat{\phi}) \leq c) \approx 0.05$ (the approximation was carried out by drawing a bootstrap sample of size 10000). Note that the group symbols were 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 S model, it is not so for the models LSG and LSF: 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 S nor the misspecified model LSF were able to correctly classify this point. But the correct model LSG 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

$$10000 * \int_{-\infty}^{\infty} \left[h_z(x_j; \hat{\pi}, \hat{\phi}) - h_z(x_j) \right]^2 dx_j, \tag{8}$$

$j = 1, 2$, $z = 1, 2, 3$, was made.

One can see that the distances for the LSG and LSF models are much lower than for the S model. In the first margin, one gains an error reduction of about 22—68%. The gain in the second margin is even more striking: it ranges from 81—90%. This was expected because the second margin, a Laplace distribution, differs more from the Gaussian distribution than the first margin, a Student distribution. As the model S has misspecified margins, it comes as no surprise that it is unable to capture the shape of the true margins. It is not so for LSF and LSG. Interestingly enough, we found that the estimated margins for LSG and LSF were indistinguishable (and so were represented by the same line on the graphs) and closer to the true margin. Interestingly enough, note that one of these models, LSF, was wrong — as the copulas were misspecified —, but this did not deteriorate marginal estimation. This suggests that, in the model (4), marginal estimation is robust in regard to copula misspecification. Thus, thanks to the nonparametric part of our model, margins can be estimated correctly irrespective of copula modeling.

		first margin	second margin
group 1	S	62	248
	LSG	35	60
	LSF	30	59
group 2	S	76	210
	LSG	27	35
	LSF	24	39
group 3	S	45	442
	LSG	28	43
	LSF	35	48

Table 1: Rescaled L_2 distances, see (8), of the estimated marginal distributions to the true distribution, for each group and each margin.

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 to choose for the copulas and the margins? The results of the model S 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 margins as they are estimated nonparametrically.

5 Discussion

In this paper, we built a new copula-based mixture model. Novelty lies in the fact that it is nonparametric, and therefore allows to reduce 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 margins without any modeling effort. Moreover, we saw that marginal estimation is robust with respect to copula misspecification. Finally, theoretical results showed that identifiability issues are limited, if one accepts a slightly stronger conjecture than that of Hunter, Wang and Hettmansperger [9].

Nevertheless, research questions still remain. First, sensibility of estimation performance to the choice of the bandwidth in kernel density estimation would need to be assessed. Second, the location shift hypothesis could be replaced, in practice, by an affine transformation of the form $h_{jz}(t) = g_j((t - \mu_{jz})/\sigma_{jz})$, but then identifiability results would not hold anymore. Last, the possibility of carrying over our model to higher dimension is still unclear.

Acknowledgment. The research by G. Mazo was funded by a "Projet de Recherche" of the "Fonds de la Recherche Scientifique — FNRS" (Belgium).

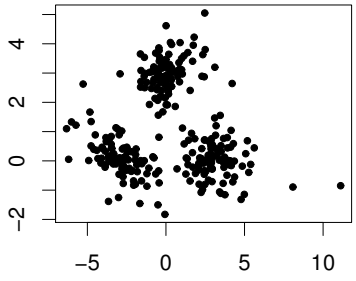
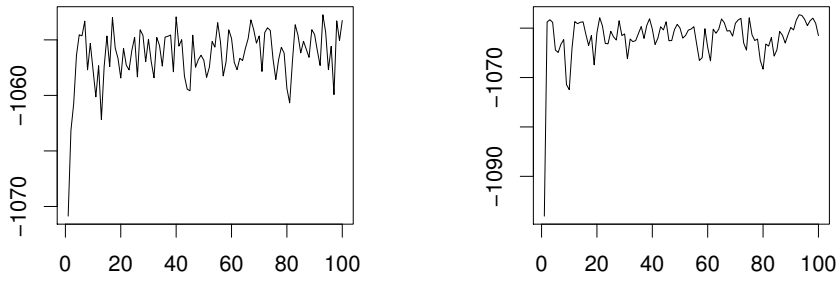
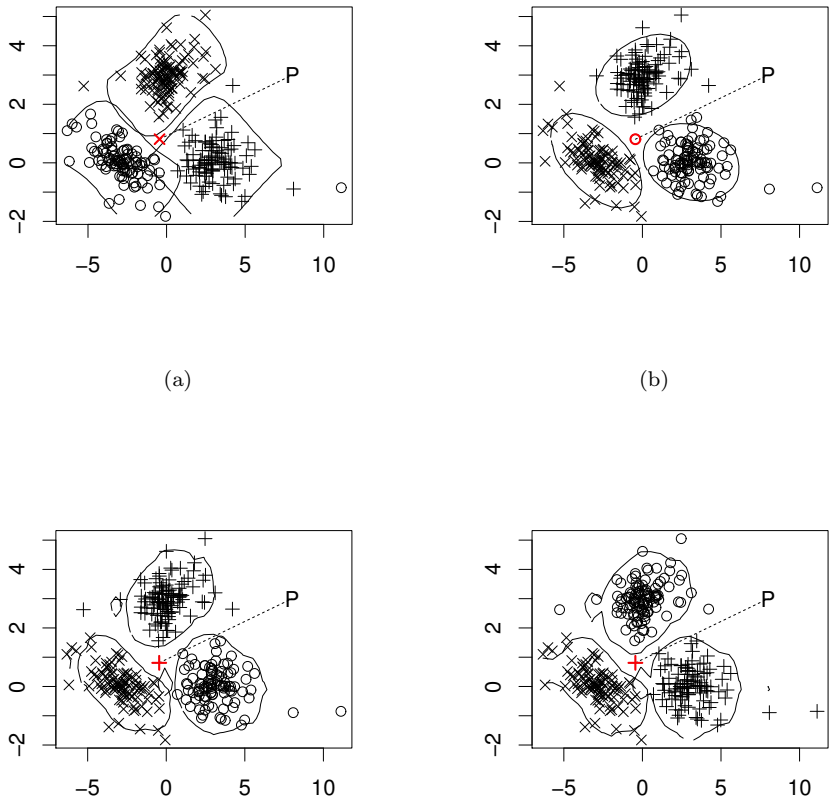


Figure 1: The raw simulated data.



(a) (b)

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



(a) (b)

(c) (d)

Figure 3: Estimated group assignments for the model S (b), LSG (c) and LSF (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.

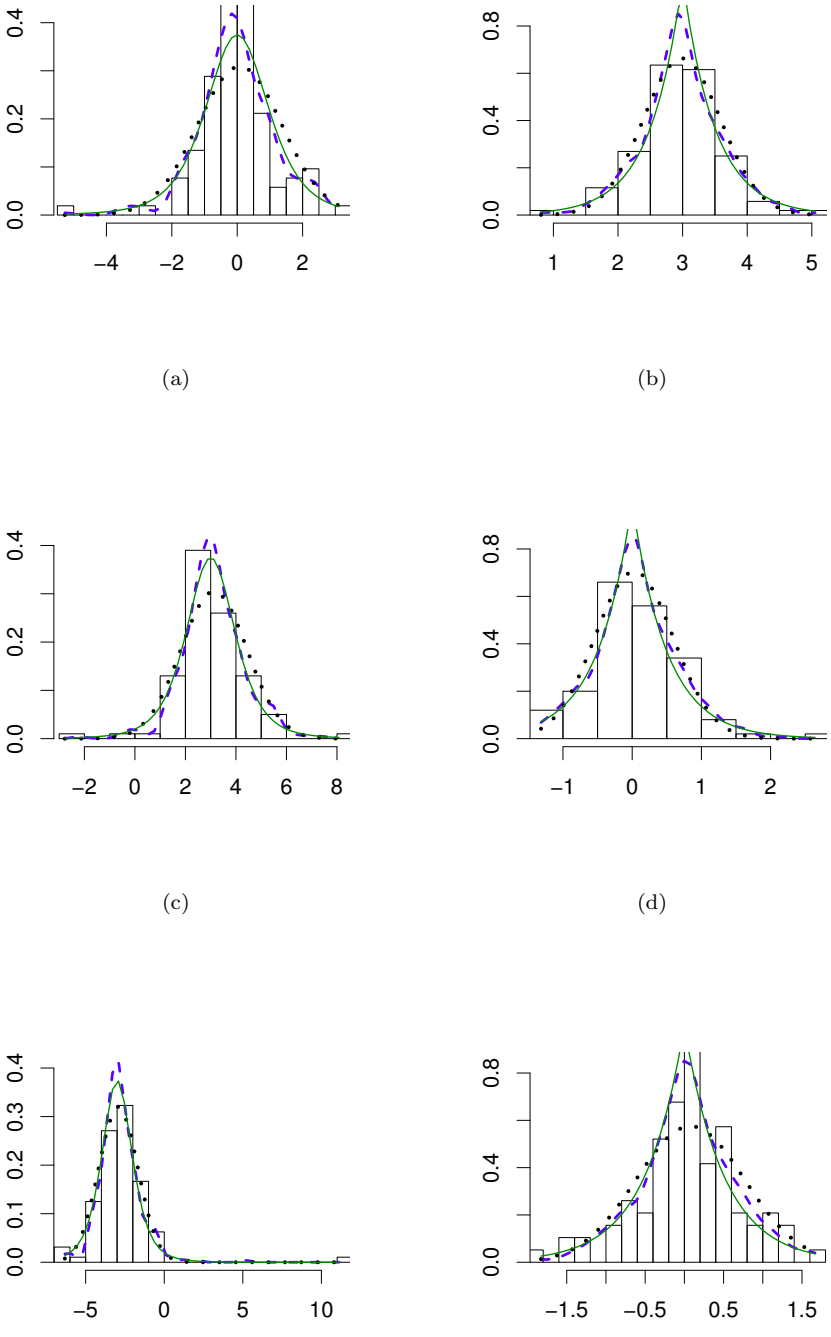


Figure 4: Histograms for the univariate data along with the estimated densities under the three tested models (S: black dotted line; LSG and LSF: blue dashed line), as well as the true density (plain green line). The first column from the left is the first margin, and each row represents one group: the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

6 Appendix

Proof of Proposition 2

Let $(\pi, \mu, \theta, g), (\tilde{\pi}, \tilde{\mu}, \tilde{\theta}, \tilde{g}) \in \Lambda^* \times \mathcal{R}^{dK} \times \Theta \times \mathcal{G}^d$, with $g = (g_1, \ldots, g_d)$, and suppose that, for all $x_1, \ldots, x_d \in \mathcal{R}$,

$$h(x_1, \ldots, x_d; \pi, \mu, \theta, g) = h(x_1, \ldots, x_d; \tilde{\pi}, \tilde{\mu}, \tilde{\theta}, \tilde{g}) \tag{9}$$

so that

$$\begin{aligned} &\sum_{z=1}^{K} \pi_z c_z\{G_1(x_1 - \mu_{1z}), \ldots, G_d(x_d - \mu_{dz}); \theta_z\} \prod_{j=1}^{d} g_j(x_j - \mu_{jz}) \\ =&\sum_{z=1}^{K} \tilde{\pi}_z c_z\{\tilde{G}_1(x_1 - \tilde{\mu}_{1z}), \ldots, \tilde{G}_d(x_d - \tilde{\mu}_{dz}); \tilde{\theta}_z\} \prod_{j=1}^{d} \tilde{g}_j(x_j - \tilde{\mu}_{jz}). \end{aligned} \tag{10}$$

By marginalizing, we obtain

$$\sum_{z=1}^{K} \pi_z g_j(x_j - \mu_{jz}) = \sum_{z=1}^{K} \tilde{\pi}_z \tilde{g}_j(x_j - \tilde{\mu}_{jz}).$$

But since $\pi \in \Lambda^*$, we have $(\pi, \mu_j) \in (\Lambda \times \mathcal{R}^K)^*$ for all $\mu_j \in \mathcal{R}^K$. This means, by definition of $(\Lambda \times \mathcal{R}^K)^*$, that $\sum_{z=1}^{K} \pi_z g_j(x_j - \mu_{jz})$ is identifiable, and therefore $\pi_z = \tilde{\pi}_z$, $\mu_{jz} = \tilde{\mu}_{jz}$ and $g_j = \tilde{g}_j$, for all $j = 1, \ldots, d$ and $z = 1, \ldots, K$. Now, since, by (10), $h(\cdot; \pi, \mu, \theta, g) = h(\cdot; \tilde{\pi}, \tilde{\mu}, \tilde{\theta}, \tilde{g})$ are the densities of the same distribution, we have

$$\begin{aligned} &c_z\{G_1(x_1 - \mu_{1z}), \ldots, G_d(x_d - \mu_{dz}); \theta_z\} \prod_{j=1}^{d} g_j(x_j - \mu_{jz}) \\ =&c_z\{\tilde{G}_1(x_1 - \tilde{\mu}_{1z}), \ldots, \tilde{G}_d(x_d - \tilde{\mu}_{dz}); \tilde{\theta}_z\} \prod_{j=1}^{d} \tilde{g}_j(x_j - \tilde{\mu}_{jz}), \end{aligned}$$

hence $c_z(u_1, \ldots, u_d; \theta_z) = c_z(u_1, \ldots, u_d; \tilde{\theta}_z)$ for all $u_1, \ldots, u_d \in [0, 1]$ which implies $\theta_z = \tilde{\theta}_z$ for all $z = 1, \ldots, K$. Therefore, $h(\cdot; \pi, \mu, \theta, g)$ in (9) is identifiable. But since $(\Lambda \times \mathcal{R}^{dK} \times \Theta)^*$ is the biggest subset of $\Lambda \times \mathcal{R}^{dK} \times \Theta$ such that identifiability holds, we conclude $\Lambda^* \times \mathcal{R}^{dK} \times \Theta \subset (\Lambda \times \mathcal{R}^{dK} \times \Theta)^*$.

Now, note that $\Lambda \setminus \Lambda^*$ is countable, hence of Lebesgue measure zero. Therefore, since

$$\Lambda^* \times \mathcal{R}^{dK} \times \Theta \subset (\Lambda \times \mathcal{R}^{dK} \times \Theta)^* \subset \Lambda \times \mathcal{R}^{dK} \times \Theta,$$

and because the Lebesgue difference between the set on the right and the set on the left is zero, the claim follows.

Proof that $\tilde{X}_j^{(i)}$, $i = 1, \ldots, n$ is a sample of G_j

The following proof is that in [1]. Let us show that $\tilde{x}_j^{(i)}$, $i = 1, \ldots, n$, is a sample of G_j. Let $x = (x_1, \ldots, x_d) \in \mathcal{R}^d$. The following calculations prove the result:

$$
\begin{aligned}
& P(\tilde{X}_j^{(i)} \leq t_j) = P(X_j^{(i)} - \mu_{j,\tilde{Z}^{(i)}(X_j^{(i)})} \leq t_j) \\
&= \int \cdots \int \sum_z P(x_j - \mu_{j,z} \leq t_j | X^{(i)} = x, \tilde{Z}^{(i)}(x) = z) P(\tilde{Z}^{(i)}(x) = z | X^{(i)} = x) h(x) dx \\
&= \int \cdots \int \sum_z P(x_j \leq t_j + \mu_{j,z} | X^{(i)} = x, \tilde{Z}^{(i)}(x) = z) P(Z^{(i)}(x) = z | X^{(i)} = x) h(x) dx \\
&= \sum_z \int \cdots \int \mathbf{1}_{(-\infty, \mu_{j,z} + t_j]}(x_j) P(Z^{(i)} = z | X^{(i)} = x) h(x) dx \\
&= \sum_z \int \cdots \int \mathbf{1}_{(-\infty, \mu_{j,z} + t_1]}(x_j) h_{X^{(i)} | Z^{(i)} = z}(x|z) \pi_z dx \\
&= \sum_z \pi_z \int_{-\infty}^{\mu_{j,z} + t_j} h_{jz}(x_j) dx_j \\
&= \sum_z \pi_z \int_{-\infty}^{\mu_{j,z} + t_j} g_j \left(x - \mu_{jz} \right) dx = \sum_z \pi_z G_j(t_j) = G_j(t_j).
\end{aligned}
$$

References

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

[2] A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, 1997.

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

[4] 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.

[5] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] I. Kosmidis and D. Karlis. Model-based clustering using copulas with applications, 2015. arXiv.

[12] 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.

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

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

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

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

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

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