A semiparametric and location-shift copula-based mixture model

Gildas Mazo

Inria Grenoble Rhône-Alpes, gildas.mazo@inria.fr

Résumé. La modélisation des modèles de mélange se sont longtemps appuyés sur les lois gaussiennes et/ou l'hypothèse d'indépendance conditionnelle. Ce n'est que (relativement) récemment que les chercheurs ont construit des modèles plus généraux sans faire appel à de telles hypothèses. Certaines de ces constructions utilisent les copules qui permettent de séparer l'analyse des effets marginaux de la structure de dépendance. Mais cette approche a aussi des inconvénients. D'abord, l'utilisateur doit faire plus de choix arbitraires, et ensuite, des problèmes de spécifications peuvent apparaitre. Cette communication a pour but de limiter ces problèmes en proposant un modèle de mélange basé sur les copules et qui a l'avantage d'être semi-paramétrique. Grâce à une hypothèse de translation, l'estimation semi-paramétrique est également réalisable, permettant l'adaptation aux données sans effort de modélisation.

Mots-clés. algorithme EM, modèle de mélange, non-parametrique, semi-parametrique, copule, classification non-supervisée

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 communication 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. EM algorithm, mixture model, non-parametric, semi-parametric, copula, clustering

1 Introduction

The modeling of a mixture of distributions [4] has long rested upon Gaussian distributions [6] and it is only recently that researchers have started to construct and study broader generic models. Among these extensions, models featuring copulas [3] are still

rare, but certainly promising. Indeed, copulas allow for building very flexible models, as they permit to handle the marginal distributions and the dependence separately [5].

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.

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.

In this work, 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. 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.

2 The method

2.1 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

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

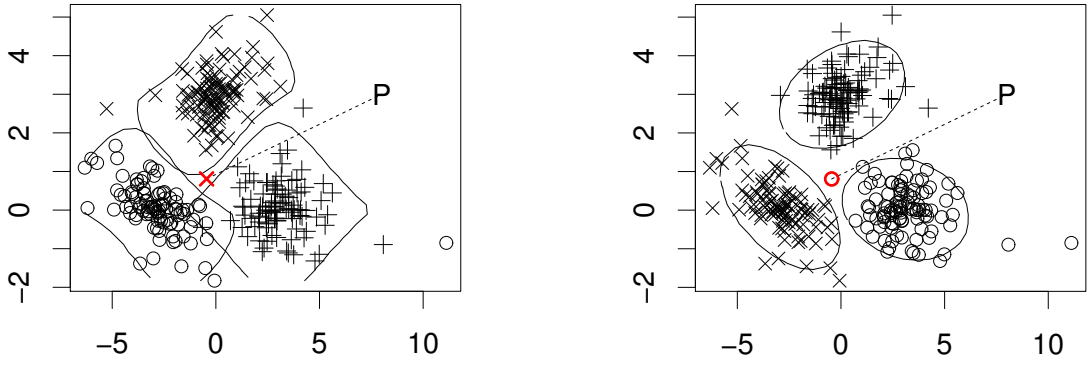
where G_j and g_j are respectively the distribution function and the density of a symmetric distribution, that is, $G_j(x_j) = 1 - G_j(-x_j)$ for all continuity points x_j. This location-shift hypothesis (4) implies that X_j, $j = 1, \ldots, d$, is assumed to have support $(-\infty, +\infty)$. If it 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).

Examples of data generated through the above model are presented in Figure 1. Notice the rough contour lines of the estimated model in Figure 1 (c) and (d), typical of non-parametric estimation procedures.

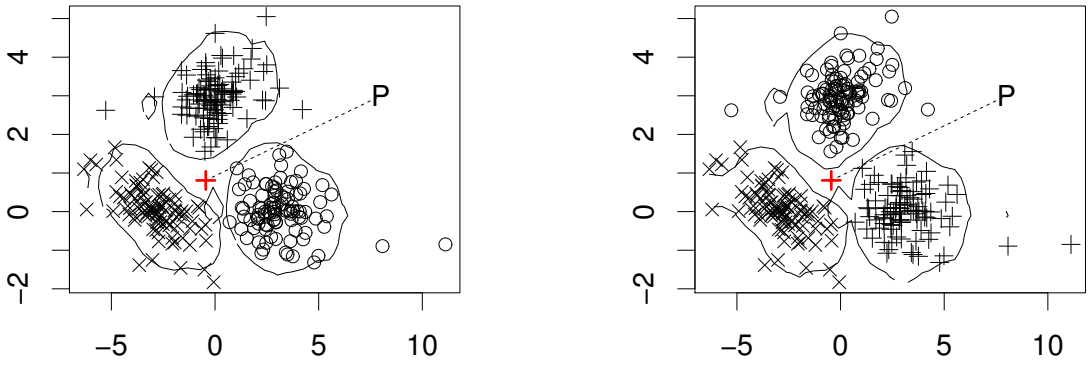
2.2 The estimation procedure

Estimation is performed through a EM-like algorithm in the same spirit as in [2]. Let $\mathbf{X}^{(i)}$, $i = 1, \ldots, n$ be the given data, where $\mathbf{X}^{(i)} = (X_1^{(i)}, \ldots, X_d^{(i)})$. 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

$$h(z|\mathbf{x}^{(i)}; \boldsymbol{\phi}^t) = \frac{\pi_z^t c_z \left(G_1^t\left(x_1^{(i)} - \mu_{1z}^t\right), \ldots, G_d^t\left(x_d^{(i)} - \mu_{dz}^t\right); \boldsymbol{\theta}_z^t\right) \prod_{j=1}^{d} g_j^t\left(x_j^{(i)} - \mu_{jz}^t\right)}{h(x_1^{(i)}, \ldots, x_d^{(i)}; \boldsymbol{\phi}^t)}, \tag{6}$$



(a) (b)



(c) (d)

Figure 1: Data generated according to (5) and estimated distributions for different copulas. Details are in [1].

where the function h in the denominator was given in (5). We should maximize over $\boldsymbol{\phi}$ the objective function

$$\begin{aligned} Q(\boldsymbol{\phi}|\boldsymbol{\phi}^t) =& E\left[\sum_{i=1}^{n}\left\{\log h(\mathbf{x}^{(i)}|Z^{(i)};\boldsymbol{\phi}) + \log P(Z^{(i)}=z)\right\}\Big|\mathbf{X}=\mathbf{x}\right] \\ =& \sum_{z,i}\Big[\log c_z\left\{G_1(x_1^{(i)}-\mu_{1z}),\ldots,G_d(x_d^{(i)}-\mu_{dz});\boldsymbol{\theta}_z\right\} \\ &+\sum_{j=1}^{d}\log g_j(x_j^{(i)}-\mu_{jz}) + \log\pi_z\Big]h(z|\mathbf{x}^{(i)};\boldsymbol{\phi}^t), \end{aligned} \tag{7}$$

which involves an infinite dimensional parameter. A solution of this problem being unknown, we instead adapt 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 (6).
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 $\tilde{x}_j^{(i),t+1} = x_j^{(i)} - \mu^t_{j,\tilde{Z}^{t+1}(\mathbf{x}^{(i)})}$ 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 $z = 1, \ldots, K$,

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

This algorithm permitted to obtain Figure 1 (c) and (d). The details are available in [1].

Bibliographie

[1] Mazo, G. (2017), A semiparametric and location-shift copula-based mixture model, accepted for publication, Journal of Classification, `https://hal.archives-ouvertes.fr/hal-01263382v3`.

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

[3] Kosmidis, I. and Karlis, D. (2015), Model-based clustering using copulas with applications, Statistics and Computing.

[4] McLachlan, G. and Peel, D. (2004), Finite mixture models, John Wiley & Sons.

[5] Genest, C. and Favre, A.-C. (2007), Everything you always wanted to know about copula modeling but were afraid to ask, Journal of Hydrologic Engineering.

[6] Fraley, C. and Raftery, A. E. (2002), Model-based Clustering, Discriminant Analysis and Density Estimation, Journal of the American Statistical Association.