Nonparametric mixture models with conditionally independent multivariate component densities

Didier Chauveau [1] & Vy Thuy Lynh Hoang [2]

[1] [2] *Univ. Orléans, CNRS, MAPMO, UMR 7349, Orléans, France*
[1] *didier.chauveau@univ-orleans.fr*
[2] *vy-thuy-lynh.hoang@etu.univ-orleans.fr*

Résumé. Les mélanges non-paramétriques font l'objet de nombreux travaux récents, portants sur la détermination de modèles identifiables ainsi que de méthodes d'estimation souvent fondées sur le principe de l'algorithme EM. Ces modèles sont plus flexibles que les mélanges paramétriques car les densités des composantes y sont semi- ou totalement non-paramétriques. Dans le cas d'observations multivariées, l'hypothèse communément posée afin d'assurer l'identifiabilité consiste à admettre que les coordonnées sont indépendantes, conditionnellement à la sous-population de provenance des individus. Or dans de nombreux cas cette hypothèse n'est pas raisonnable. Nous proposons ici un nouveau modèle de mélange multivarié, dans lequel les densités des composantes sont composées de blocs indépendants conditionnellement à la sous-population, mais eux-mêmes multivariés et non-paramétriques. Ce modèle est identifiable, et nous définissons un algorithme de type "EM non paramétrique" incluant une stratégie de choix de fenêtres, afin d'en estimer les paramètres. Les performances de ce modèle et cet algorithme sont illustrés au travers de simulations et d'une étude sur un jeu de données réel pour un objectif de classification.

Mots-clés. Algorithme EM, Estimation non paramétrique de densité multivariées, Mélanges non-paramétriques multivariés.

Abstract. Recent works in the literature have proposed models and algorithms for nonparametric estimation of finite multivariate mixtures. In these works, the model assumes independent coordinates, conditional on the subpopulation from which each observation is drawn, so that the dependence structure comes only from the mixture. Here, we relax this assumption, allowing in the multivariate observations independent multivariate blocks of coordinates conditional upon knowing which mixture component from which they come. Otherwise their density functions are completely multivariate and nonparametric. We check that this new model is identifiable, and propose an EM-like algorithm for the statistical estimation of its parameters. We then derive some strategies for selecting the bandwidth matrix involved in the nonparametric estimation step of it. The performance of this algorithm is illustrated through several numerical simulations. We also experiment this new model and algorithm on an actual dataset from the model based, unsupervised clustering perspective, to illustrate its potential.

Keywords. EM algorithm, multivariate kernel density estimation, multivariate mixture, nonparametric mixture.

1 Introduction

The most general model for nonparametric multivariate mixtures is as follows: suppose the vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are a simple random sample from a finite mixture of $m > 1$ arbitrary distributions. The density of each $\boldsymbol{X}_i$ may be written

$$g_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \sum_{j=1}^{m} \lambda_j f_j(\boldsymbol{x}_i), \tag{1}$$

where $\boldsymbol{x}_i \in \mathbb{R}^r$, and $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{f}) = (\lambda_1, \ldots, \lambda_m, f_1, \ldots, f_m)$ denotes the parameters of the statistical model. In this model λ_j denotes the proportion (weight) of component j in the population; the λ_j's are thus positive and $\sum_{j=1}^{m} \lambda_j = 1$. The f_j's are the component densities, drawn from some family of multivariate density functions $\mathcal{F}$ absolutely continuous with respect to Lebesgue measure, and the term "nonparametric" means that no parametric assumptions are made about the form of the f_j's.

Model (1) is not identifiable if no restrictions are placed on $\mathcal{F}$, where "identifiable" means that $g_{\boldsymbol{\theta}}$ has a *unique* representation of the form (1) and also that we do not consider that "label-switching" — i.e., reordering the m pairs $(\lambda_1, f_1), \ldots, (\lambda_m, f_m)$ — produces a distinct representation. The common restriction placed on $\mathcal{F}$ in a number of recent theoretical and algorithmic developments in the statistical literature, since its introduction by Hall and Zhou (2003), is that each joint density $f_j(\cdot)$ is equal to the product of its marginal densities. In other words, the coordinates of the $\boldsymbol{X}_i$ vector are independent, conditional on the subpopulation or component (f_1 through f_m) from which $\boldsymbol{X}_i$ is drawn. Therefore, model (1) becomes

$$g_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_{ik}). \tag{2}$$

For the multivariate model (2), an empirical "EM-like" (`npEM`) algorithm for statistical estimation of its parameter has been introduced in Benaglia et al. (2009).

In this work, we relax the assumption underlying model (2) by assuming that each joint density f_j is equal to the product of $B < r$ multivariate densities that will correspond to independent multivariate *blocks*, conditional on the subpopulation from which each observation is drawn. Let the set of indices $\{1, ..., r\}$ be partitioned into B disjoint subsets s_l, i.e. $\{1, ..., r\} = \bigcup_{l=1}^{B} s_l$, where B is the total number of such blocks, and d_l is the number of coordinates in lth block, i.e. lth block dimension. Here, the indices i, j, k and l denote a generic individual, component, coordinate, and block, $1 \leq i \leq n, 1 \leq j \leq m, 1 \leq k \leq r$ and $1 \leq l \leq B$ (m, r, B and n stand for the number of mixture components, repeated measurements, blocks, and the sample size). Then model (1) becomes

$$g_{\theta}(\boldsymbol{x}_i) = \sum_{j=1}^{m} \lambda_j \prod_{l=1}^{B} f_{jl}(x_{is_l}), \tag{3}$$

where $x_{is_l} = \{x_{ik}, k \in s_l\}$ is the multivariate variable which have its coordinates in lth block and multivariate density function f_{jl}. This is a main difference in comparison to model (2): here the dependence structure does not come only from the mixture structure; some additional within-block dependence is allowed. This model thus brings more flexibility with respect to the conditional independence assumption. From a modelisation perspective, the way to choose these blocks depends on the structure of the data, see the example in Section 3. In view of Allman et al. (2009), the fundamental result of identifiability is established for model (2) if $r \geq 3$, regardless of m. We check that this result is indeed generalized to model (3) where at least three multivariate blocks are independent, conditioned on the latent structure.

2 Estimating the parameters

The algorithm we propose is an extension of the original `npEM` algorithm that was designed for estimation in the multivariate mixture model (2). The EM principle is first applied in the E-step, i.e. computation of the posterior probabilities given the current value $\boldsymbol{\theta}^{(t)}$ of the whole parameter. The EM machinery is also applied straightforwardly for the M-step of the Euclidean part that are only the weights $\boldsymbol{\lambda}$. Then a nonparametric Weighted Kernel Density Estimation (WKDE) is applied to update the component densities per blocks. The main difference is that in this model, we need multivariate density estimates. This is also where this algorithm becomes "EM-like", since kernel density estimation is not a genuine maximization step.

In finite mixture models, the *complete data* associated with the actually observed sample $\boldsymbol{x}$ is $(\boldsymbol{x}, \boldsymbol{Z})$, where to each individual (multivariate) observation $\boldsymbol{x}_i$ is associated an indicator variable Z_i denoting its component of origin. It is common to define $Z_i = (Z_{i1}, \ldots, Z_{im})$, the indicator variables $Z_{ij} = \mathbb{I}\{\text{observation } i \text{ comes from component } j\}$, $\sum_{j=1}^{m} Z_{ij} = 1$. From (1), this means that $\mathbb{P}_{\boldsymbol{\theta}}(Z_{ij} = 1) = \lambda_j$, and $(\boldsymbol{X}_i | Z_{ij} = 1) \sim f_j$, $j = 1, ..., m$.

The `mvnpEM` algorithm. Given initial values $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\lambda}^{(0)}, \boldsymbol{f}^{(0)})$, the algorithm consists in iterating the following steps:

1. **E-step:** Calculate the posterior probabilities (conditional on the data and $\boldsymbol{\theta}^{(t)}$), for each $i = 1, \ldots, n$, $j = 1, \ldots, m$, and where $f_j^{(t)}(\boldsymbol{x}_i) = \prod_{l=1}^{B} f_{jl}^{(t)}(x_{is_l})$,

$$p_{ij}^{(t)} := \mathbb{P}_{\boldsymbol{\theta}^{(t)}}(Z_{ij} = 1 | \boldsymbol{x}_i) = \frac{\lambda_j^{(t)} f_j^{(t)}(\boldsymbol{x}_i)}{\sum_{j'=1}^{m} \lambda_{j'}^{(t)} f_{j'}^{(t)}(\boldsymbol{x}_i)}. \tag{4}$$

2. **M-step for λ**:

$$\lambda_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(t)}, \quad j = 1, \ldots, m. \tag{5}$$

3. **Nonparametric kernel density estimation step:** For any $\boldsymbol{u}$ in $\mathbb{R}^{d_l}$, define for each component $j \in \{1, \ldots, m\}$ and block $l \in \{1, \ldots, B\}$,

$$f_{jl}^{(t+1)}(\boldsymbol{u}) = \frac{1}{n\lambda_j^{(t+1)}} \sum_{i=1}^{n} p_{ij}^{(t)} K_{H_{jl}}(\boldsymbol{u} - x_{is_l}), \tag{6}$$

where, for $\boldsymbol{u} \in \mathbb{R}^{d_l}$, $K_{H_{jl}}(\boldsymbol{u}) = |H_{jl}|^{-1/2} K(H_{jl}^{-1/2}.\boldsymbol{u})$, K is a multivariate kernel function typically Gaussian, H_{jl} is a symmetric positive definite $d_l \times d_l$ "bandwidth matrix", that may depend on the lth block and jth component, and even on the tth iteration, as briefly precise below.

Bandwidth selection in multivariate KDE. The central decision in the nonparametric density estimation step of both the `npEM` and `mvnpEM` algorithm is the selection of an appropriate value for the (scalar or matrix) bandwidth or smoothing parameter. We restrict ourselves to diagonal bandwidth matrices in this work. Firstly, as in Benaglia et al. (2009) it is possible to simply use a single fixed bandwidth for all components per coordinate within each block, selected by default according to a rule of thumb from Silverman (1986). Secondly, we investigate a often more appropriate strategy defining iterative and per component and coordinate bandwidths by adapting Silverman's rule of thumb as in Benaglia et al. (2011). There, the scalar bandwidths for each coordinates in the block depend also on component j and current algorithm iteration t through the posterior probabilities $p_{ij}^{(t)}$'s, that are used to compute weighted interquartile range and standard deviations involved in Silverman's rule.

3 Implementation and examples

Initialization of the mvnpEM algorithm. To initialize the algorithm, the first E-step requires initial values for the $f_j^{(0)}$'s which themselves require an initial $n \times m$ matrix of posteriors $(p_{ij}^{(0)})$. To obtain this matrix, we apply, like in other EM-strategies, a k-means algorithm to get a first clustering of the data.

Monte-Carlo Experiments. We first performed experiments on simulated data and computed the errors in terms of the square root of the Mean Integrated Squared Error (MISE) for the densities as in Hall et al.(2005): $MISE_{jl} = \frac{1}{S}\sum_{s=1}^{S} \int (\hat{f}_{jl}^{(s)}(\boldsymbol{u}) - f_{jl}(\boldsymbol{u}))^2 d\boldsymbol{u}$, where $\hat{f}_{jl}^{(s)}$ is the density estimate at replication s, computed from (6) but using the final

values $\hat{p}_{ij}$'s of the posterior probabilities after numerical convergence of the algorithm, and where the integral is computed numerically. We computed also the mean squared error (MSE) for the $m-1$ proportions that are the only scalar parameters in these models. For instance the MSE for the proportion of component 1 is $MSE_{\lambda_1} = \frac{1}{S}\sum_{s=1}^{S}(\hat{\lambda}_1^{(s)} - \lambda_1)^2$, where $\hat{\lambda}_1^{(s)}$ is computed using (5) together with the final posterior probabilities $\hat{p}_{ij}$'s. Note that we computed and provided as well MSE's for other scalar measures of precision (means, variances,...) that are not genuine parameters of the model.

Several models have been tested, and we just give here brief results from one model with $r = 6$ variables, $m = 2$ components with λ_1 =30 %, and 3 blocks of bivariate ($d_l = 2$) Gaussian densities with some covariance structure. The adaptive bandwidth strategy proved its superiority over the fixed bandwidth strategy in this particular model. In addition, all the MSE's and MISE's decrease when the sample size n increases, as expected. For brevity, a single output concerning one situation is displayed in Fig.1.

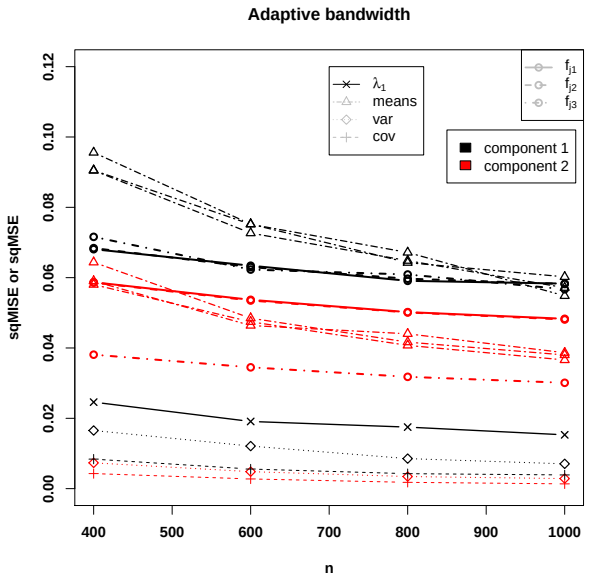

Figure 1: Square roots of MISE's for the densities and square roots of MSE's for the parameters as a function of the sample size n, $S = 300$ replications, $k-$means initialization, and the adaptive bandwidths settings.. Lines types in grey correspond to same types but per-component colored, as indicated in the plots.

An example on actual data We consider a real dataset from an experiment involving $n = 569$ instances of Wisconsin Diagnostic Breast Cancer (WDBC). This database is available through the UW CS ftp server[1]. The details of the attributes found in WDBC

[1] `ftp.cs.wisc.edu`, see `math-prog/cpo-dataset/machine-learn/WDBC/`

dataset are: ID number, Diagnosis (M = malignant, B = benign) and ten groups of three real-valued features that are computed for each cell nucleus. The total number of attributes is 32 (ID, diagnosis, 30 real-valued input features). In these data, some features are related to others, so that we can expect dependences apart from any mixture structure. Some scatterplots also confirm this. We thus applied the `mvnpEM` algorithm to model (3) with $B = 8$ blocks (1 block of size 9 and 7 blocks of size 3), $m = 2$ components. Then we used the posteriors after convergence of the algorithm to obtain the correct classification p and the distribution of Maximum A Posteriori (MAP) strategy given by each subpopulation; we compared them with a k-means classification where we can see that the solution of `mvnpEM` using MAP is better than the k-means strategy. This gave sensible results in the perspective of unsupervised, model-based clustering.

4 Perspectives

Using a non-diagonal bandwidth matrix is an interesting perspective for future work, to better recover multivariate and strongly correlated densities. The smoothed EM idea in Levine et al. (2012) where introduced a smoothed loglikelihood objective function and developed an iterative algorithm, is also the subject of an ongoing work.

References

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist*, 37(6A):3099–3132.

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

Benaglia, T., Chauveau, D., and Hunter, D. R. (2011). *Bandwidth Selection in an EM-like algorithm for nonparametric multivariate mixtures*, pages 15–27. Number Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger. World Scientific Publishing Co.

Hall, P. and Zhou, X. H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, 31:201–224.

Levine, M., Hunter, D. R., and Chauveau, D. (2012). Maximum smoothed likelihood for multivariate maximum. *Biometrika*, 98(2):403–416.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.