Robust estimation in the normal mixture model based on robust clustering †

J.A. Cuesta-Albertos

Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Spain

C. Matrán

Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain

A. Mayo-Iscar

Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain

Summary.
We introduce a robust estimation procedure based on the choice of a representative trimmed subsample through an initial robust clustering procedure, and subsequent improvements based on maximum likelihood. To obtain the initial trimming we resort to the trimmed k-means, a simple procedure designed for finding the core of the clusters under appropriate configurations. By handling the trimmed data as censored, maximum likelihood estimation provides in each step the location and shape of the next trimming. Data-driven restrictions on the parameters, requiring that every distribution in the mixture must be sufficiently represented in the initial clustered region, allow avoiding singularities and guaranteeing the existence of the estimator. Our analysis includes robustness properties and asymptotic results as well as worked examples.

Keywords: Multivariate normal mixture model, identifiability, EM algorithm, censored maximum likelihood, asymptotics, trimmed k-means, breakdown point, influence function.

1. Introduction

Estimation in mixture models has caught the interest of many researchers due to their multiple statistical applications. Although a lot of research has been produced since its publication, McLachlan and Peel [17] gives a general summary of the state of the art in this topic. In this paper, we assume the so-called multivariate normal mixture model (MNMM) given by $\{P_\theta : \theta \in \Theta\}$, with densities

$$f_\theta := \sum_{i=1}^{k} \pi_i g_{\phi_i}, \tag{1}$$

where k is known, g_{ϕ_i}, $\phi_i = (\mu_i, \Sigma_i)$, denotes the density function on $I\!R^d$ of the Gaussian distribution with mean μ_i and covariance matrix Σ_i, and π_i is the mixing proportion of g_{ϕ_i} in the mixture. The multi-parameter θ, indexing the model, varies in the set

$$\Theta := \left\{ \theta = (\pi_1, ..., \pi_k, \phi_1, ..., \phi_k) : \pi_i > 0, \sum_{i=1}^{k} \pi_i = 1, \phi_i \in \Phi, \quad \phi_i \neq \phi_j \text{ if } \ i \neq j \right\},$$

†Research partially supported by the Spanish Ministerio de Ciencia y Tecnología and FEDER, grant BFM2005-04430-C02-01 and 02 and by the Consejería de Educación y Cultura de la Junta de Castilla y León, grant PAPIJCL VA102/06

where $\Phi := I\!R^d \times \mathcal{M}^+_{d\times d}$ and $\mathcal{M}^+_{d\times d}$ is the set of (strictly) positive definite $d \times d$ matrices.

Although the interest in these mixture models has a long history, the implementation of stochastic algorithms solving the enormously complex computational problems involved in the usual statistical approaches is recent. On the one hand, from its introduction, the EM algorithm found one of its principal applications in the maximum likelihood (ML) treatment of mixture models. On the other hand, the Markov chain Monte Carlo (MCMC) algorithms have given the possibility of treating the problem from a Bayesian point of view (an overview on mixtures in this Bayesian setting is Marin, Mengersen and Robert [15]).

In any case the solution of the estimation problem under the mentioned approaches inherits difficulties which can be resumed through its qualification as *an inverse ill-posed problem* (see Section 1.3.3 in [15]). A fact that translates in the unboundedness of the likelihood function, the existence of multiple local maxima, the practical impossibility of using improper priors in this setting and a great instability of the estimators and of the available algorithms computing them.

The main computable approaches in the multivariate setup with a robust motivation seem to be reduced to the one by Fraley and Raftery [6] through the addition of a mixture component accounting for noise modeled as a uniform distribution, and the t mixture model by McLachlan and Peel (see e.g. [18] and Section 7 in [17] for other references) which replaces (1) by a mixture of t distributions. However, as noted by Hennig in [11], *"while a clear gain of stability can be demonstrated for these methods in various examples ..., there is a lack of theoretical justification of their robustness."* In this work, we introduce a new methodology for robust estimation in the MNMM supported by a sound theoretical analysis.

We propose a two-step procedure beginning with a robust estimator, whose efficiency is improved with a maximum likelihood (ML) step. Several iterations of the ML step, leading to an m-step procedure, would increase the efficiency. This method is a natural generalization of that analyzed in Cuesta-Albertos, Matrán and Mayo [4] in the multivariate elliptical model. A similar approach was adopted by Marazzi and Yohai [14] in the univariate regression model. Also Markatou [16] considered a related weighted likelihood method for mixtures, but based on a preliminary nonparametric density estimation step, which could make the procedure undesirable for the multivariate setting due to the curse of the dimensionality.

The procedure searches initially a small (purportedly) uncontaminated core of the data, consisting of k clusters, each one associated with one distribution in the mixture. Then, ML estimation (obtained through a variant of the EM algorithm) of θ based on this trimmed data subset, treating the removed data as censored, produces the estimation. This process can be repeated by updating the trimmed sample in accordance with the present estimation in such a way that in every step the information in the current trimmed sample is used to produce a larger and better-shaped trimmed set. This would be repeated until some suitable stopping rule is met, based on which the final estimate would be obtained.

The troublesome existence of multiple local maxima and the unboundedness of the likelihood function have often been handled with ad hoc procedures. The use of restrictions on the parameter space to circumvent this difficulty was pioneered by Hathaway [10] for mixtures of univariate normal distributions. Here, we introduce a data-driven restriction on the parameters whose meaning is that each distribution in the mixture must have a sufficient representation in the initial trimmed sample.

Features of the procedure are given in Section 2. In Section 3 we compare our method with other procedures and show its performance in several examples. Section 4 is devoted to the theoretical justification of the procedure, including asymptotic results and robustness properties. The main ideas of the proofs can be found in the Appendix (some extra details, as well as some additional examples can be found in the Technical Report [5]). A last section is devoted to discussion on the proposed methodology.

2. Assumptions and description of the procedure

Throughout we will handle the sample space $I\!\!R^d, d \geq 1$, with its usual norm $\| - \|$, and Borel sets β^d. With $\overline{B}(m,r)$ we denote the closed ball centered at m with radius r. For a set $A \subset I\!\!R^d$, A^c denotes its complement and I_A the associated indicator function. We will use two families of sets: For every $\gamma = (m_1, ..., m_k, r) \in \Gamma := \left(I\!\!R^d\right)^k \times I\!\!R^+$, $\mathcal{B}(\gamma)$ will denote the union of closed balls $\cup_{i=1}^k \overline{B}(m_i, r)$, while, for every $\eta = (m_1, ..., m_k, \Sigma_1, ..., \Sigma_k, r_1, ..., r_k) \in \tilde{\Gamma} := \left(I\!\!R^d\right)^k \times \left(\mathcal{M}^+_{d\times d}\right)^k \times \left(I\!\!R^+\right)^k$, $\mathcal{E}(\eta)$ will be the union of ellipsoids $\bigcup_{i=1}^k \{x \in I\!\!R^d : (x-m_i)^T \Sigma_i^{-1}(x-m_i) \leq r_i\}$.

The notation Pf will denote integration of the random variable f with respect to the probability distribution P. Given a random sample $\{X_n\}_n$ of a distribution P, $\{P_n\}_n$ will denote the associated sequence of empirical distributions.

To circumvent the well-known problem of identifiability, we will assume as equivalent $(\pi_1, ..., \pi_k, \phi_1, ..., \phi_k) \in \Theta$ and every $(\pi_{i_1}, ..., \pi_{i_k}, \phi_{i_1}, ..., \phi_{i_k})$ for permutations $(i_1, i_2, ..., i_k)$ of $(1, 2, ..., k)$. Notice that in contrast with the ML approach, this assumption causes a serious "label switching" problem in the Bayesian methodology (see Section 1.3.4 in [15]).

2.1. Initial estimator

We begin by choosing a trimming set $\hat{A}$ through a robust clustering criterion. For this, we use the trimmed k-means introduced in Cuesta-Albertos, Gordaliza and Matrán [3] (the TRIMCLUSTER package includes an R-code, [19], to compute them). This procedure trims a given proportion α of the sample and splits the remaining data into k groups in order to minimize the within groups sums of squares of the distances to the centers of the groups.

It is shown in [3] that, if P is any probability on $I\!\!R^d$, there exists $\gamma_P := (m_1^P, ..., m_k^P, r_P) \in \Gamma$ such that the set $\mathcal{B}(\gamma_P)$ $(= \cup_{i=1}^k \overline{B}(m_i^P, r_P))$ fulfills $P(\mathcal{B}(\gamma_P)) \geq 1-\alpha$ and for every union of closed balls $A := \cup_{i=1}^k \overline{B}(m_i, r_i)$, verifying $P(A) \geq 1-\alpha$, it holds

$$\frac{1}{P(\mathcal{B}(\gamma_P))} \int_{\mathcal{B}(\gamma_P)} \inf_{i=1,\ldots,k} \|x - m_i^P\|^2 P(dx) \leq \frac{1}{P(A)} \int_A \inf_{i=1,\ldots,k} \|x - m_i\|^2 P(dx). \qquad (2)$$

The vector $(m_1^P, ..., m_k^P) \in \left(I\!\!R^d\right)^k$ is called an *α-trimmed k-mean of* P, and we will refer to $\mathcal{B}(\gamma_P)$ as its associated region. When $P = P_\theta$ or P_{θ_0}, instead of γ_P we will use γ_θ or γ_0.

Note that the right hand side term in (2) includes every union of k balls in $I\!\!R^d$, but the minimum is attained by an union of balls with the same radius. This is a peculiarity of the regions associated to the trimmed k-means.

The solution provided by the trimmed k-means is quite simple and it is well suited to our goals when the data set is composed by k approximately spherical distributions with similar dispersions. This can be considered as a limitation of the procedure but through one or several additional iterations it is generally possible to cover a wider framework.

Very often, specially if the components of the mixture do not overlap too much, this non-parametric clustering method allows us to detect the components in a mixture (see [3]).

Moreover, the theory developed in [4] shows that ML estimates based on any trimming of the sample space suffice to successfully estimate the parameters of a Gaussian distribution. Our approach combines both facts. First, the trimmed k-means procedure initializes the search looking for so many zones as components in the mixture in such a way that the data in every zone is likely to be highly representative of just one component. Thus, the (μ_i, Σ_i) parameters in the global maximization of the censored likelihood are approximately the solutions of the k maximizations corresponding to the likelihoods of their normal components. Hopefully, this will avoid the EM algorithm to start in the domains of attraction of spurious solutions. Running the EM we propose initializing the means of the distributions composing the mixture with the trimmed k-means, while the initial values for the covariance matrices and the weights of the distributions are those based on the data in the clusters corresponding to each one of the k balls obtained in the trimming process.

Notice that more reliable initial estimators for particular situations (see e.g. our Example 3.2) are possible and covered by our asymptotic results as soon as they consistently estimate a region in the space. Often these choices can lead to improvements but they should be carefully handled because in some situations they could produce badly behaved solutions.

For a better understanding of the procedure we will apply it to an example, similar to that included in Section 2.12.4 of [17], attributed to Ueda and Nakano.

Example 2.1. Let us consider a random sample of size 600 of the mixture given by

$$\pi_i = 1/3, i = 1, 2, 3; \mu_1^T = (-2, 0), \mu_2^T = (0, 0), \mu_3^T = (2, 0); \Sigma_i = \begin{pmatrix} 0.2 & 0 \\ 0 & 2 \end{pmatrix}, i = 1, 2, 3.$$

To analyze the behavior of the procedure in the presence of contaminated data we added 20 data simulated from the uniform distribution on the set

$$\{(x, y) \in [-5, 5] \times [-8, 8] : x < -4 \text{ or } x > 4 \text{ or } y < -5 \text{ or } y > 5\}.$$

The graph on the left in Figure 1 shows the trimmed region associated to the trimmed 3-means for a trimming level of 0.5 (the union of the three yellow balls), as well as the (non-trimmed) 3-means (marked as bold squares). Here we only stress the scarce influence that the contaminated data have on the trimmed 3-means. On the contrary, the 3-means are greatly influenced by the contamination. The remaining features of the graphs in Figure 1 are explained later. •

2.2. Trimmed sets and the censored likelihood function

In [4] we consider several likelihood functions associated to a subsample constituted by the points of the sample $\{x_1, ..., x_n\}$ belonging to a bounded set $A \in \beta^d$. As stated in the final discussion there, under the hypothesized model the censored point of view is the best choice.

The (artificial) censoring leads to consider the *censored log-likelihood function*:

$$L_{\theta/A}(x) := I_A(x) \log f_\theta(x) + I_{A^c}(x) \log P_\theta(A^c), \; x \in \mathbb{R}^d.$$

Thus, the empirical censored log-likelihood based on a sample of size n is

$$P_n L_{\theta/A} = P_n \left(I_A \log f_\theta \right) + P_n(A^c) \log P_\theta(A^c).$$

In this way we use the full information corresponding to the sample points belonging to A, but also the number of points in A^c. Moreover, as soon as we guarantee the identifiability

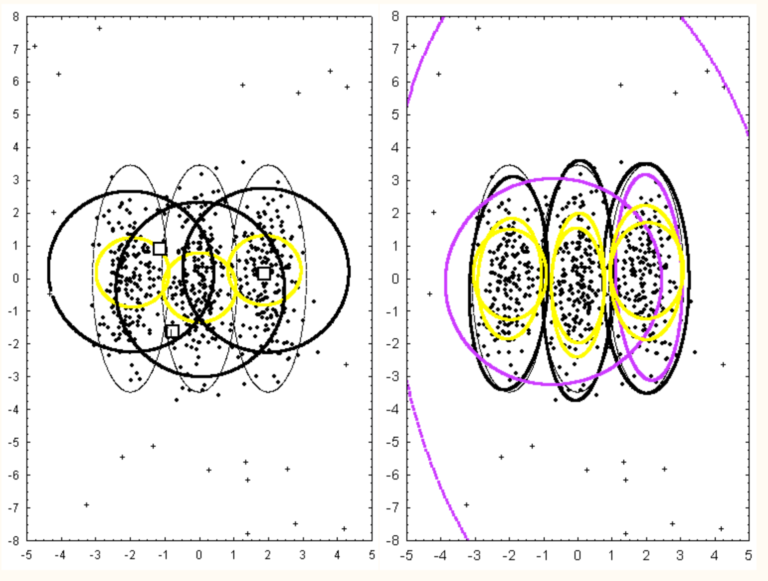

Fig. 1. Realizations of 3-means, 50%-trimmed 3-means, associated (and iterated) regions and different estimations in Example 2.1 (represented by the 95% level curves of the weighted estimated normal distributions in the mixture). The main features are explained there, just before Subsection 2.2.1, and at the end of Subsection 2.2.1.

of the distribution on A (see Theorem 4.1 below) we can guarantee the uniqueness of the maximization of the likelihood under the model (see Proposition 4.2).

To avoid degenerated solutions in the sample optimization problem, we consider restrictions on the parameters based on the presence of k populations. Assuming that the initial procedure is successful in discarding the contaminated data and in searching for a representative subset of $\{x_1, ..., x_n\}$, the selected set $\hat{A}$ should contain sufficient evidence of every population. Therefore, once a threshold value, $u \in (0, 1)$, has been chosen, we consider

$$\Theta_u^n := \left\{ \theta \in \Theta : \frac{1}{\sharp \left\{ t : x_t \in \hat{A} \right\}} \sum_{x_t \in \hat{A}} P_\theta(i/x_t) \geq u, \text{ for every } i = 1, ..., k \right\}, \tag{3}$$

as the restricted parameter set. Defining

$$P_\theta(i/x) := \frac{\pi_i g_{\phi_i}(x)}{f_\theta(x)} = \frac{\pi_i g_{\phi_i}(x)}{\sum_{j=1}^k \pi_j g_{\phi_j}(x)}$$

which is the 'a posteriori' probability of a point x arising from density g_{ϕ_i}, then, the whole quotient in (3) is the sample conditional mean: $P_n[P_\theta(i/\cdot)/\hat{A}]$. Thus the value u limits our search to those parameters which would produce in mean at least a presence of $100u\%$ points of every component in our censored sample. The set Θ_u^n is mostly data-driven and we will call it an *impartially restricted parameter set.*

Now we are in a position to define the (two steps) estimator of θ through

$$\hat{\theta}_n := \arg \max_{\theta \in \Theta_u^n} P_n L_{\theta/\hat{A}}. \tag{4}$$

We use the EM algorithm to solve (4) (the Monte Carlo EM if the involved integrals make the EM infeasible); see Remark 4.3 on the convergence of the EM algorithm in this setup.

Back to Example 2.1. It is not actually necessary to fix a very accurate value for u in order to define the restricted parameter space. Let us consider a light one, $u = 0.1$, the trimmed set $\hat{A}$ already obtained in Example 2.1 and apply the EM algorithm to solve (4).

In the graph on the left in Figure 1 the thin (resp. thick) ellipses show the 95%-level curves of the weighted true (resp. estimated) normal distributions in the mixture. The solution given by the EM algorithm to the classic MLE starting from the 3-means is shown in purple in the graph on the right. As already shown in [17] the poor choice of initial value leads to a very bad solution, but, even with good initial solutions, the EM algorithm for the classic MLE would exhibit a bad behavior in this case due to the contamination. •

2.2.1. Iterations: Improving the trimming regions

As we will see in Theorem 4.1, our model 1 is identifiable even if we use only the information corresponding to any (fixed) open set. Thus, it is theoretically possible to estimate the parameters just by handling the data lying in that set. However, it is intuitively sound (see also our analysis on the Influence Function in Section 4.2) that in order to produce a good estimation for all the parameters through the estimation based on a trimmed sample, we should attempt to obtain a trimming region adapted to the features of the mixture. As stated before, the trimmed k-means are not intended for this. Thus, to better reflect the structure of the mixture, the trimmed set $\hat{A}$ should be substituted by a union of k ellipsoids with appropriate shapes and sizes. We should also improve the use of the information incorporating into the active data set as many (good) data as possible. These facts are patent in the yet unsatisfactory solution provided for Example 2.1.

To pursue in the aforementioned directions, we mimic the EM algorithm in the following way. At this time, we have a value of the trimming parameter $\alpha_1 = \alpha$, an estimated active trimming set $\hat{A}^1 = \hat{A}$ and an estimate $\hat{\theta}^1 = \hat{\theta}_n$ of the parameter, with components $\hat{\theta}^1 = (\hat{\pi}_1^1, ..., \hat{\pi}_k^1, \hat{\phi}_1^1, ..., \hat{\phi}_k^1)$.

Let $M \in I\!N$ and $\alpha_2, ..., \alpha_M \in (0,1), \alpha_1 \geq \alpha_2 \geq ... \geq \alpha_M$ be given (as usually happens with iterative procedures, a smooth enlargement will contribute to avoid brusque changes in the behavior of the procedure). Step 3 consists in replacing the trimming set $\hat{A}^1$ by the set $\hat{A}^2$ composed of the union of the ellipsoids given by the $1-\alpha_2$ level curves of the density functions $g_{\hat{\phi}_1^1}, ..., g_{\hat{\phi}_k^1}$. Now, we can obtain $\hat{\theta}^2$, the MLE associated to the censored likelihood function $L_{\theta/\hat{A}^2}$ with the same impartial restrictions as those used in step 2. We repeat the process using the trimming sizes $\alpha_3, ..., \alpha_M$ and, for every α_i, the last estimation $\hat{\theta}^{i-1}$ as the initial value for the EM algorithm, the active trimming set $\hat{A}^i$ constructed as in step 3 from $\hat{\theta}^{i-1}$ and the new trimming level.

The above process assumes fixed values of M and $\alpha_i, i = 1, ..., M$, although it could be adaptive (by resorting to a similar idea to that in [14] or to a stopping criterion based in penalizing the censored likelihood). We shall not pursue this task here and in all the examples we will use $\alpha_1 = 0.5$ and $\alpha_M = 0.05$, with small changes of size 0.05 (thus M=10).

We keep the initial restrictions, based on the trimmed k-means, to avoid the possibility of a slow, step by step, degeneration of the estimated parameters of some distribution in the mixture. The good performance of the trimmed k-means to select representative zones of the clusters (see [3]) and the nature of our restrictions justify to keep this choice.

Back to Example 2.1. In both graphs in Figure 1 the thin ellipses show the 95%-level curves of the weighted true normal distributions in the mixture; while the thick ellipses show the estimated 95%-level curves of the normal distributions in the mixture given by the initial estimate (graph on the left), where $\alpha_1 = 0.5$, and by the iterated estimate with $M = 10$ steps and $\alpha_M = 0.05$ (graph on the right).

The three inner (resp. outer) yellow ellipses show the intermediate adaptive region corresponding to 35% (resp. 20%) trimming size. •

3. The method in action

In this section we present some features of the method through its behavior on several examples and discuss on the performance of the available alternatives in our framework.

To our best knowledge there is not an explicit robust proposal into the Bayesian methodology. However it is natural to consider the Bayesian analysis of mixtures allowing the contaminating data to be modeled by additional components within the category of robust alternatives. In our context, the main drawbacks of this methodology are the great difficulties of handling multivariate data (the more elaborated and recent proposals focus on, at most, $I\!R^2$), and how to handle isolated outliers as individual components of the mixture.

The first drawback is directly concerned with the specification of manageable and suitable priors for the involved hyper-parameters, which at best would be possible at price of computational challenges provided that strong constraints on the covariance matrices be imposed *"requiring them all to be equal or all to be diagonal for example"* (Stephens [21] p. 64). In this sense, the approach in Bensmail et al. [1], looking at possible factorizations in some aspects of the covariance matrices, could be also of some aid. Regarding the second, when there are contaminating data that are not well explained by a few additional components (as it happens e.g. when there are several isolated outliers or small groups of outliers), allowing for the inclusion of new components to explain very small groups and preventing convergence difficulties for the algorithms by the phenomenon of the "absorbing component" (see Section 9.3 in Robert and Casella [20]) are opposite tasks. In any case, as argued in [11] in a frequentist setting, in practice the maximum number of fitted components will often be fixed and much smaller than the maximum number of outliers thus the lack of robustness of the method for fixed k would remain relevant. Alternatively, a proper definition of the outlying model should be provided and suitably introduced in the model.

In the non-Bayesian framework, the last comment is equally pertinent for the procedures designed for an unknown number of components. For known k, Fraley and Raftery [6] proposed the addition of a component in the mixture accounting for noise. For this task they introduced a uniform distribution on the convex hull of the data. A variation through an improper distribution has been proposed in Hennig [11] in the univariate setting. The t-mixture model (see Section 7.3 in [17]) is based on the use of a variant of the EM algorithm (the ECM algorithm) to estimate the parameters assuming a mixture of multivariate t distributions, including the estimation of the degrees of freedom (which often is assumed to be the same for every distribution in the mixture) and scale matrices.

As noted in [11], these proposals often improve the stability, but they break down in any case under the addition of just one (far enough) outlier. Only the proposal of Hennig circumvents this difficulty, although it is limited to the univariate case and no theoretical justification for its behavior (excepting the breakdown point analysis) has been yet provided.

The examples and table that follow give a comparative perspective of our method. Even when the theoretical results in [11] and those presented in Subsection 4.2 would suffice to show its superiority from a robust point of view, we will show some comparisons of its behavior with that of the classic MLE as well as with the probably better known robust alternative, the t-mixture model.

For the analysis involving the mixture of t distributions method we carry out the computations with the EMMIX algorithm of McLachlan, Peel, Basford and Adams, using the k-means as an initial solution, but also starting from 100 initial randomly chosen solutions, choosing between the results the one that provides the maximum of the likelihood function associated with the mixture of t distributions.

As a general background for the presented graphics, the different colors or symbols show the assignment of the points to the clusters given by the procedure used to produce the initial solution (i.e. the k-means or the trimmed k-means procedure). The cross symbol is always assigned to the trimmed data. The thin (resp. thick) ellipses shows the 95%-level curves of the weighted true (resp. estimated) normal distributions in the mixture.

With respect to the estimations produced through our proposal, in order to show the scarce influence of using very accurate values in the separation threshold to define the constrained parameter space, we have used $u = 0.1$, a light one in comparison with the real proportions of the components considered in the mixtures of the worked examples. Moreover, as stated in Subsection 2.2.1 we start with an initial trim of 50% of the data and the final region, obtained in $M = 10$ steps, contains 95% of the points in the sample.

We begin by noting that the solution provided to Example 2.1 by the t mixture model is similar to that obtained with our method. This often happens for symmetrical contamination, where both methods generally show a good performance. If we now substitute the 20 contaminating data there for another 20 points that constitute a well concentrated contamination arising e.g. from a uniform distribution on the square $[0.5, 1.5] \times [-8, -7]$, then the (bad) behavior of EM for the t mixture model is similar to that EM for the normal mixture model. In fact, it is the classic MLE procedure which is unable to handle the problems arising from the presence of some concentration of outliers.

Example 3.1. In this example we analyze the behavior of the methods in a 10-dimensional problem. The mixture is composed of the product measure of a 8-variate normal distribution with zero mean and covariance matrix equal to 8 times the identity matrix on $\mathbb{R}^8$ with a mixture of three bivariate normal distributions with parameters

$$\pi_i = 1/3, i = 1, 2, 3; \mu_1^T = (-9, 0), \mu_2^T = (1, 5), \mu_3^T = (3.5, -3.5);$$

$$\Sigma_1 = \begin{pmatrix} 16 & 0 \\ 0 & 16 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 8.5 & -7.5 \\ -7.5 & 8.5 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The analysis has been carried out over a sample of size 600, slightly contaminated by 10 additional data obtained from a uniform distribution on the parallelepiped $[-4, 4]^8 \times [6, 10] \times [11, 19]$. The graphs in Figure 2 show the plots of the last two dimensions of the solutions. The graph on the left corresponds to the solution given by the EM algorithm, for the classic MLE, starting from the 3-means as the initial solution (violet) and to the solution provided by the t mixture model (yellow). The violet solution is nearly equivalent to a local maximum found by the EMMIX algorithm for the t mixture model. The graph on the right shows the solution obtained with our method. •

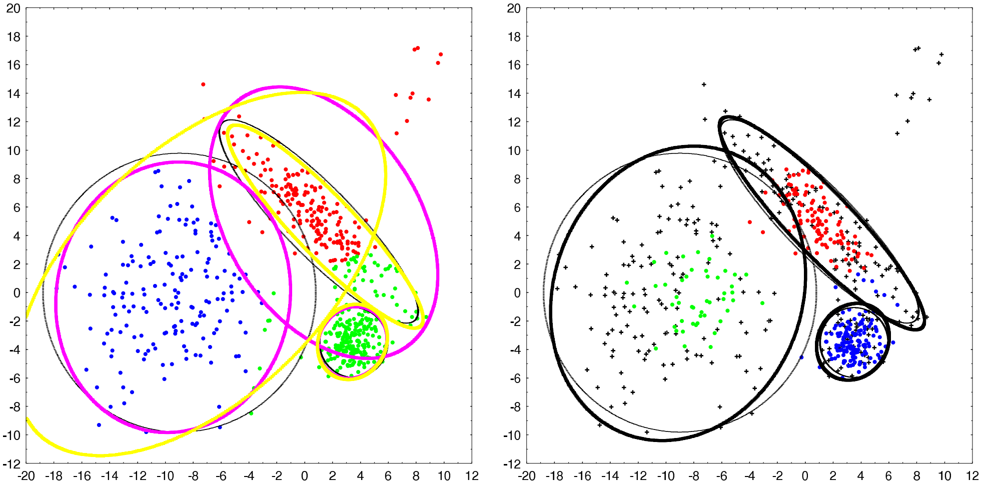

Fig. 2. Plots showing the last two components of the 95% level ellipsoids of the true distribution (thin ones) and the estimated distributions of Example 3.1.

Example 3.2. Here we use the subset of the crabs data set in Campbell and Mahon [2] corresponding to the blue crab species, which includes 50 males and 50 females. The fit, by a mixture of two normal distributions, to the bivariate data provided by the RW and CL variates is studied in Peel and McLachlan [18] and in [17]. That analysis mainly addresses the fitness and robustness of the t mixture model in the classification framework. It includes detailed comments on the influence of the equal covariance matrices hypothesis on the estimation, showing a better performance of the estimation without such restriction. In fact this constraint produces an unnecessary overlapping of the estimated distributions. By introducing one outlier into the original data set, they also show how the normal mixture model fitting can give an outright solution resulting in degenerating one component to explain just the outlier. The t mixture model is robust against the analyzed perturbations.

We will show that a similar case can happen even for the t mixture solution in presence of a small cluster of anomalous observations. For this, we have added three outliers in the left upper corner in the plot of Figure 3 which cause the t mixture model approach to break down. It could be argued that it is also legitimate to consider this set of outliers as a true cluster. If so, we would be in a situation in which the addition of 3 anomalous observations would radically change our model because it forces us to step from $k = 2$ to $k = 3$ clusters. Thus, the very principles of robustness would lead to consider these observations as clustered contamination. In this sense, the restrictions given by (3) and the choice of our trimming level avoid the consideration of non-enoughly representative components in the mixture.

The breakdown solutions provided by the EM algorithm, for the classic MLE, starting from the 2-means as initial solution (violet) and the solution provided by the t mixture model (yellow) coincide. In this case we have used as initial region for our procedure the solution provided by Gallegos and Ritter [7] (covering the 50% of the data by a union of two identical ellipses suitably located). The use of the initial solution given by this procedure is coherent with the hypothesis of homocedasticity discussed in [17], while the iterative ML steps improve the solution without such constraint. A similar solution would be obtained

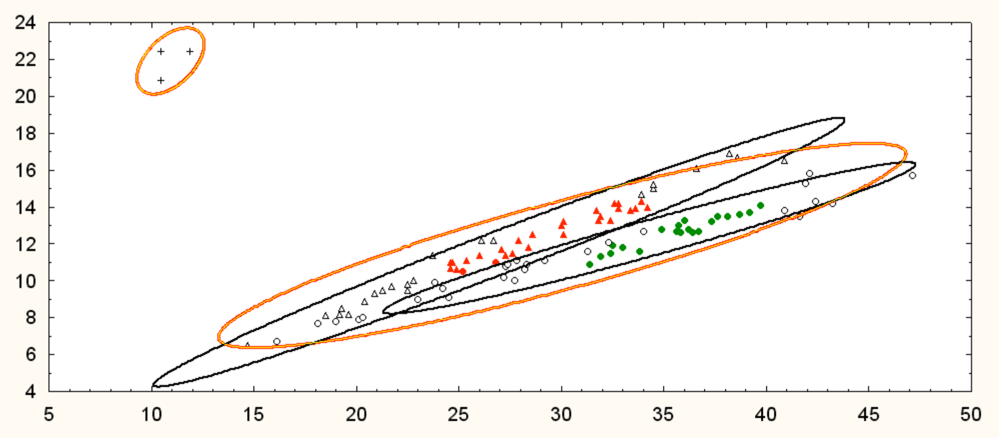

Fig. 3. Plot showing the 95% ellipses for the estimated distributions of the Blue Crab Data set of Example 3.2. Small circles and triangles distinguish the sex of the individuals in the data.

using the 0.7-trimmed 2-means as the initial region, and a final trimming size of 5% (see [5]). Note that in this example, the very elongated shapes of the groups could make impossible to choose a representative region of both populations based on two balls of the same radius unless for a very high trimming level. •

Table 3.1 shows the results of a simulation study to analyze the performance of the method under different types of contamination and dimensions of the sample space. The table shows the proportions of successes obtained by the natural classification rule provided through the estimation procedures under consideration (x is classified in the j-th class, even for trimmed data, iff $\hat{\pi}_j g_{\hat{\phi}_j}(x) > \hat{\pi}_i g_{\hat{\phi}_i}(x)$ for $i \neq j$). The displayed values are the proportions of well classified data. Obviously the outliers are not considered here. To ease the analysis we include the classic MLE solution provided by the EM algorithm, starting from the true theoretical solution, for the non-contaminated data sets. The columns labeled tk-m (resp. G-R) show the results obtained when the procedure starts with the 0.5-trimmed 3-means (resp. with the solution provided by Gallegos and Ritter as a union of 3 equal ellipses suitably located containing 50% of the data). Regarding this table it becomes apparent the scarce influence that the initial solution has over the final estimation in our simulations.

We consider three kinds of the problem. Each kind is handled in three different dimensions (2, 5 and 10) and three different sample sizes. The mixtures will be composed of three populations of equal size (250, 500 or 1000 each) obtained from three normal distributions. In order to ease comparisons between the performances of our procedure when the dimension grows and additionally to give a quick picture of the nature of the problem under consideration, we begin with 2-dimensional problems. Then, we immerse them in dimensions 5 and 10 by completing the last components with independent N(0,1) random variables.

Notice that the particular representation we have chosen does not affect to the real dimension of the problem and, in fact, in dimension 5 we are estimating 62 parameters: 2 for the proportions in the mixture, 3×5 for the means vectors and 3×15 for the covariance matrices. In dimension 10 the number of estimated parameters is $167 = 2 + (3 \times 10) + (3 \times 45)$.

In the first kind of problems (labeled as A) the first two coordinates are obtained from the distributions with parameters

$$\mu_1^T=(0,-4),\mu_2^T=(0,0),\mu_3^T=(0,4);\Sigma_1=\Sigma_2=\Sigma_3=\begin{pmatrix}4\ 0\\0\ 1\end{pmatrix}.$$

In the second kind (labeled as B) the involved distributions are defined by

$$\mu_1^T=(0,-3),\mu_2^T=(0,2),\mu_3^T=(-2,0);\Sigma_1=\begin{pmatrix}4\ 0\\0\ 1\end{pmatrix},\Sigma_2=\begin{pmatrix}6 & -5\\-5 & 6\end{pmatrix},\Sigma_3=\begin{pmatrix}0.4 & 0\\0 & 0.4\end{pmatrix};$$

while in the third kind (Problems C) they are determined by

$$\mu_1^T=(0,-4),\mu_2^T=(-4,0),\mu_3^T=(0,0);\Sigma_1=\begin{pmatrix}4\ 0\\0\ 1\end{pmatrix},\Sigma_2=\begin{pmatrix}1\ 0\\0\ 4\end{pmatrix},\Sigma_3=\begin{pmatrix}6 & -5\\-5 & 6\end{pmatrix}.$$

TABLE 3.1. *Proportions of correctly classified data in our simulation study with the shown sample sizes and dimensions (d). The employed estimation procedures were the usual M.L.E. and the proposed procedure starting from two different trimmed sets: the one obtained with the Gallegos and Ritter methodology (labeled as G-R) and from the trimmed k-means (labeled as tk-m). In both cases the initial trimming level was $\alpha=.5$. Regular samples were generated from a standard Gaussian distribution. Contaminations are described in the text.*

Problem	d	Sizes	No contamination			Contamin. 1		Contamin. 2		Contamin. 3	
			MLE	tk-m	G-R	tk-m	G-R	tk-m	G-R	tk-m	G-R
A	2	3 × 250	.967	.967	.967	.966	.966	.967	.967	.967	.967
		3 × 500	.968	.968	.968	.968	.968	.968	.968	.968	.968
		3 × 1000	.969	.969	.969	.969	.970	.969	.969	.969	.969
	5	3 × 250	.967	.966	.966	.966	.966	.965	.965	.966	.966
		3 × 500	.967	.967	.967	.968	.968	.967	.967	.968	.968
		3 × 1000	.969	.969	.969	.969	.969	.969	.969	.969	.969
	10	3 × 250	.962	.959	.958	.960	.958	.959	.959	.960	.959
		3 × 500	.966	.965	.965	.966	.966	.966	.966	.966	.966
		3 × 1000	.968	.968	.968	.968	.967	.968	.968	.968	.968
B	2	3 × 250	.946	.887	.800	.867	.827	.870	.825	.867	.832
		3 × 500	.946	.889	.799	.884	.842	.907	.852	.909	.823
		3 × 1000	.947	.916	.789	.896	.867	.914	.841	.903	.872
	5	3 × 250	.942	.938	.919	.939	.910	.938	.926	.935	.929
		3 × 500	.945	.944	.942	.944	.940	.944	.944	.945	.941
		3 × 1000	.947	.946	.945	.945	.945	.946	.943	.944	.944
	10	3 × 250	.936	.927	.825	.933	.827	.933	.827	.931	.818
		3 × 500	.942	.940	.918	.940	.907	.941	.926	.942	.909
		3 × 1000	.945	.944	.939	.944	.937	.944	.944	.943	.934
C	2	3 × 250	.860	.839	.835	.822	.810	.830	.823	.828	.822
		3 × 500	.864	.852	.850	.840	.829	.841	.834	.844	.835
		3 × 1000	.867	.853	.854	.842	.836	.845	.841	.845	.835
	5	3 × 250	.850	.831	.802	.830	.783	.836	.794	.835	.785
		3 × 500	.858	.850	.841	.850	.829	.852	.838	.850	.835
		3 × 1000	.863	.857	.854	.857	.853	.857	.849	.857	.851
	10	3 × 250	.833	.782	.688	.770	.695	.786	.697	.778	.715
		3 × 500	.851	.840	.810	.835	.800	.841	.807	.838	.803
		3 × 1000	.858	.853	.850	.854	.848	.854	.848	.853	.851

Each problem is considered without contamination and with three kinds of contamination, where we enlarge the whole sample by adding 5% of outliers. Any one of the already available procedures would produce outright solutions with at least one of these contaminations. In the first kind of contamination we add far away outliers in the sphere centered at the origin with radius 100. In the other cases the contamination is situated outside the zone defining the 99% covering of the populations. In the second kind of contamination the outliers constitute an additional cluster obtained from a normal centered at (8,8,0,...,0) with identity covariance matrix. In the last case we consider sparse outliers obtained from a distribution with independent components N(5,8^2), N(5,8^2), N(0,1),..., N(0,1).

4. Theoretical framework

The following theorem gives the identifiability result that justifies the proposed methodology. The proof, based in Proposition 6.1 in the Appendix, is an easy consequence of the characterizations of identifiability in Yakowitz and Spragins [24].

Theorem 4.1. *Let $\theta_1, \theta_2 \in \Theta$ and let A be a d-dimensional open set. If $f_{\theta_1}(x) = f_{\theta_2}(x)$, for every $x \in A$, then θ_1 and a permutation of θ_2 coincide.*

From this it is straightforward to justify the use of MLE in this framework. The classical proof, based on the use of Jensen's (strict) inequality, works here.

Proposition 4.2. *Let $\theta_0, \theta \in \Theta$ with $\theta \neq \theta_0$. If $A \in \beta^d$ has a nonempty interior, then*

$$P_{\theta_0} L_{\theta_0/A} > P_{\theta_0} L_{\theta/A}.$$

The restrictions considered in Section 2.2 are the sample version of the following general framework: Given a bounded set A in β^d, $u \in (0,1)$, and a distribution P, let

$$\Theta_{A,u} := \left\{ \theta \in \Theta : \frac{1}{P(A)} P[I_A P_\theta(i/\cdot)] \geq u, \text{ for every } i = 1, ..., k \right\}. \tag{5}$$

Thus, $\Theta_{A,u}$ is the family of parameters which give an expected probability to every population, conditioned by A, greater or equal than u. In particular, if $P = P_{\theta_0}$, it is enough that $u \leq \inf_i \pi_i^0 G_{\phi_i^0}(A)$ to guarantee that $\theta_0 \in \Theta_{A,u}$, where G_{ϕ_i} denotes the Gaussian distribution with parameters $\phi_i = (\mu_i, \Sigma_i)$.

Once we fix an $\alpha \in (0,1)$ we are also fixing the trimmed k-means of the subjacent distribution, say P, thus the corresponding trimming set $\mathcal{B}(\gamma_P)$. The choice of the threshold value, u, in turn determines the restricted set $\Theta_P \equiv \Theta_{\mathcal{B}(\gamma_P),u}$, in which we maximize the censored likelihood function obtaining $\theta_P := \arg\max_{\theta \in \Theta_P} PL_{\theta/\mathcal{B}(\gamma_P)}$. From now on we will assume that the value u is fixed and we will omit it in the notation.

Particularizing for the theoretical, P_{θ_0} (resp. sample, P_n), distribution we will use the notation $\mathcal{B}(\gamma_0)$ (resp. $\mathcal{B}(\gamma_n)$) for the trimming set, Θ_{γ_0} (resp. Θ^n) for the restricted sets, and $\hat{\theta}_n := \arg\max_{\theta \in \Theta^n} P_n L_{\theta/\mathcal{B}(\gamma_n)}$ for the resulting estimator (already defined in (4)). Also note that, as soon as θ_0 belongs to Θ_{γ_0}, it fulfills $\theta_0 = \arg\max_{\theta \in \Theta_{\gamma_0}} P_{\theta_0} L_{\theta/\mathcal{B}(\gamma_0)}$.

For our asymptotic results, a technical assumption in relation with the theoretical underlying distribution, P_{θ_0}, is that θ_0 is an interior point of Θ_{γ_0}. It is realistic, for moderately well-separated components in the mixture, when we use the theoretical trimmed k-means to get $\mathcal{B}(\gamma_0)$, if e.g. u is taken as $\frac{1}{k} \inf_{i=1,...,k} v_i$, where v_i is the proportion of non-trimmed sample points in the i-th ball that compose $\mathcal{B}(\gamma_n)$.

Remark 4.3. The impartial restrictions allow to assure the existence of the estimator because (see below) the sets $S_{\theta'} := \left\{\theta \in \Theta^n : P_n L_{\theta/\mathcal{B}(\gamma_n)} \geq P_n L_{\theta'/\mathcal{B}(\gamma_n)}\right\}$ are compact for every $\theta' \in \Theta^n$. Moreover they assure the convergence of the EM algorithm to stationary points of the likelihood function. This is a consequence of Theorem 2 in [23], taking into account that in this setup the likelihood corresponding to the complete data belongs to the curved exponential family and the compactness of $S_{\theta'}$ already noted.

In the model of complete data we assume also known the vector $(z_{i,1},...,z_{i,k})$ of explanatory variables, where $z_{i,j} = 1$ or 0 respectively means that x_i arises or not from the j-th distribution in the mixture. Therefore, the corresponding likelihood function

$$\prod_{i=1}^{n} \exp\left(\sum_{j=1}^{k} z_{i,j}\left(\log(\pi_j) - \frac{1}{2}\log(|\Sigma_j|) - \frac{1}{2}\mu_j\Sigma_j^{-1}\mu_j - \frac{1}{2}x_i\Sigma_j^{-1}x_i + \mu_j\Sigma_j^{-1}x_i\right)\right),$$

belongs to the curved exponential family. On the other hand, notice that

- with probability one, no sample of size $n > d$ of an absolutely continuous distribution on $I\!R^d$ contains more than d points in the same hyperplane,
- the sets used to determine the sample-based restrictions are $\mathcal{B}(\gamma_n)$, which contain at least $[\alpha \cdot n]$ points.

From here, it is possible, by slightly modifying the proof of Proposition 4.5 below, to prove the following compactness property:

Proposition 4.4. *Let $\alpha, u \in (0,1)$, $\gamma \in \Gamma$ and P_n be the sample distribution based on a sample $X_1,...,X_n$ of an absolutely continuous distribution. Assume that $n > 2(d+1)/(u(1-\alpha))$, and that $P_n(\mathcal{B}(\gamma)) \geq 1-\alpha$.*

Let $\theta^*_m = (\pi_1^m,...,\pi_k^m,\phi_1^m,...,\phi_k^m) \in \Theta^n_{\mathcal{B}(\gamma)}$ (the restricted set defined in (5) for the set $\mathcal{B}(\gamma)$, $\gamma \in \Gamma$, taking $P = P_n$), $m \in N$, be a sequence such that if we denote by λ_i^m the smallest eigenvalue of Σ_j^m (where $\phi_j^m = (\mu_j^m, \Sigma_j^m)$), then there exists $i \in \{1,...,k\}$ such that one of the following conditions is satisfied

(a) $\lim_m \lambda_i^m = 0$.
(b) $\lim_m \|\phi_i^m\| = \infty$ *and* $\liminf_m \lambda_j^m > 0,\ j = 1,...,d$.
(c) $\lim_m \pi_i^m = 0$ *and* $\liminf_m \lambda_j^m > 0,\ j = 1,...,d$.

Then $\lim_m P_n L_{\theta^*_m/\mathcal{B}(\gamma)} = -\infty$ a.s. holds.

Thus if, for some $\theta' \in \Theta^n$, the set $S_{\theta'}$ is not compact then there would exist a sequence $\{\theta_m\} \subset S_{\theta'}$ without accumulation points in $S_{\theta'}$. But, $L_{\theta/\mathcal{B}(\gamma_n)}$ being continuous in θ, $S_{\theta'}$ should be a closed subset of Θ^n, so the sequence should satisfy any of the conditions (a), (b) or (c), leading to $P_n L_{\eta/\mathcal{B}(\gamma_n)} \leq \lim_m P_n L_{\theta^*_m/\mathcal{B}(\gamma_n)} = -\infty$. •

4.1. Asymptotics

The proof of the consistency of our procedure is mainly based on an usual compactness argument stated in Proposition 4.5 (see the proof in the Appendix). A chain of statements related to this result and facts involving the behavior of the trimmed k-means (enumerated in Proposition 6.2) allow to obtain the consistency. We want to emphasize the interest of Proposition 4.5 for providing arguments such as those in Remark 4.3 for analyzing the robustness of the estimator. We stress the fact that uniqueness of the trimmed k-means for the parent distribution is not required for our results (see Remark 4.7 in [5]). Notice also that the argument developed in this proposition allows us to prove easily the existence of θ_P for every absolutely continuous distribution or even the continuity with respect to the convergence in distribution, leading to Corollary 4.7.

Proposition 4.5. *Let $\theta_n = (\pi_1^n, ..., \pi_k^n, \phi_1^n, ..., \phi_k^n) \in \Theta^n$, $n \in I\!N$, where $\phi_i^n = (\mu_i^n, \Sigma_i^n)$. Let us denote by λ_i^n the smallest eigenvalue of Σ_i^n and assume that there exists $i \in \{1, ..., k\}$ and a subsequence $\{j_n\}_n$ which satisfy one of the following conditions*

(a) $\lim_n \lambda_i^{j_n} = 0$.
(b) $\lim_n \|\phi_i^{j_n}\| = \infty$ *and* $\liminf_n \lambda_j^{j_n} > 0,\ j = 1, ..., d$.
(c) $\lim_n \pi_i^{j_n} = 0$ *and* $\liminf_n \lambda_j^{j_n} > 0,\ j = 1, ..., d$.

If the random sample was generated from an absolutely continuous distribution, then $\lim_n P_{j_n} L_{\theta_{j_n}/\mathcal{B}(\gamma_{j_n})} = -\infty$ *a.s.*

Theorem 4.6 (Consistency). *Let $\{X_n\}$ be a random sample taken from P_{θ_0}. If θ_0 is an inner point of Θ_{γ_0}, then* $\lim_n \hat{\theta}_n = \theta_0$ *a.s.*

Corollary 4.7 (Qualitative Robustness). *Assume that θ_0 is an inner point of Θ_{γ_0}. If $\{Q_n\}$ is a sequence of probability measures that converges in distribution to P_{θ_0}, then* $\lim_n \theta_{Q_n} = \theta_0$.

To obtain the asymptotic law of the estimator we resort to the Empirical Processes Theory, as developed in van der Vaart and Wellner [22]. For this recall the parameterization by $\tilde{\Gamma}$, indexing the sets $\mathcal{E}(\eta)$, $\eta \in \tilde{\Gamma}$, constituted by the union of k ellipsoids.

We can use arguments of the Empirical Process Theory for the family of functions

$$\mathcal{G}_\Lambda := \left\{ m_{\theta,\eta} := I_{\mathcal{E}(\eta)} \log(f_\theta) + I_{\mathcal{E}(\eta)^c} \log\left(P_\theta\left(\mathcal{E}(\eta)^c\right)\right) : (\theta, \eta) \in \Lambda \right\}, \tag{6}$$

and their derivatives with respect to θ:

$$h_{\theta,\eta} := I_{\mathcal{E}(\eta)} \left(\frac{\partial}{\partial\theta} \log(f_\theta) \right) + I_{\mathcal{E}(\eta)^c} \left(\frac{\partial}{\partial\theta} \log\left(P_\theta\left(\mathcal{E}(\eta)^c\right)\right) \right),$$

where Λ is a suitable subset of $\Theta \times \tilde{\Gamma}$.

As noted in [4] the extension of the arg-max arguments of the Empirical Processes Theory to this semi-parametric model (the η-parameter acts as a nuisance parameter in the model) is an easy fact through the extensions of the results of Section 3.2.4 in [22] given by Theorem 5.2 and Lemma 5.3 in [4]. From these extended statements the results will arise from that work after some algebra on Donsker classes based on the theory included in [22]. A sketch of the proofs is available in Lemma 6.5 in the Appendix.

The estimators $\hat{\theta}_n$ were defined in (4) on the basis of general trimming sets $\hat{A}$. Thus we can consider here the ones based on sets $\mathcal{E}(\eta_n)$ (possibly random).

Theorem 4.8 (Asymptotic distribution). *Let $\{X_n\}$ be a random sample taken from P_{θ_0} and $\eta^* \in \tilde{\Gamma}$. If θ_0 is an inner point of $\Theta_{\mathcal{E}(\eta^*)}$ and $\{\eta_n\}_n$ is a sequence in $\tilde{\Gamma}$ such that $\eta_n \to \eta^*$ a.s. then the sequence $\left\{\hat{\theta}_n\right\}_n$ of estimators based on the sets $\mathcal{E}(\eta_n)$ satisfies*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to_w N\left(0, \left(\left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta_0} P_{\theta_0} h_{\theta,\eta^*}\right)^{-1}\right).$$

The asymptotic covariance matrix can also be expressed as

$$\left(P_{\theta_0}\left((h_{\theta_0,\eta^*})\,(h_{\theta_0,\eta^*})^T\right)\right)^{-1}.$$

We want to remark an important (and somehow surprising) fact already reported in [4]:

Corollary 4.9. *Under the hypotheses in Theorem 4.8, the rate of convergence of $\hat{\theta}_n$ to θ_0 is $n^{1/2}$ and does not depend on the rate of convergence of η_n to η^*.*

The proof of Theorem 4.6 in [5] can easily be modified to cover the m-step estimator, as described in Subsection 2.2.1. Once we have the consistency for the two steps estimator we automatically have the consistency of the trimming sets involved in the next step and so on up to those involved in the step m. Hence, we will have the a.s. consistency of the final estimator as well as its asymptotic law, given in Theorem 4.8, but with η being the element in $\tilde{\Gamma}$ whose components are the parameters which determine the $(1-\alpha_m)$-level curves of the k normal laws involved in the mixture defined by θ_0.

4.2. Measures of robustness

The influence function (IF) and the breakdown point (BP) are central concepts of Hampel's approach to robustness. However, as far as we know, the available proposals for robust estimation in mixtures did not include this kind of analysis until Hennig's work on the BP, [11], and a general approach in Kharin [13].

The IF of the trimmed k-means method was obtained in [8], including a graphical analysis showing its behavior for some variants of a mixture of normal univariate distributions. We resort to a similar explanation that allows us to get some conclusions from the visualization of the involved graphics.

In order to get the IF we will first assume that we have a fixed set $\mathcal{E} \equiv \mathcal{E}(\eta)$, $\eta \in \tilde{\Gamma}$. In this case, the IF of $\hat{\theta}_n$, $\mathrm{IF}(x, \hat{\theta}_n, \theta_0)$, can be obtained as the IF of a MLE, thus

$$\mathrm{IF}(x, \hat{\theta}_n, \theta_0) = -\left(P_{\theta_0}\left(\left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta_0} h_{\theta,\eta}\right)\right)^{-1} h_{\theta_0,\eta}. \tag{7}$$

If $\{\eta\}_n \subset \tilde{\Gamma}$ and $\eta_n \to \eta \in \tilde{\Gamma}$, from the continuity of the estimator with respect to η, it is easy to see that the IF for the estimator $\hat{\theta}_n(\eta_n)$ coincides with that of $\hat{\theta}_n(\eta)$ for the points that do not belong to the boundary of $\mathcal{E}$ (see the proof of Theorem B.1 in [8]). Therefore, the IF for the two steps estimator based on the α-trimmed k-means will be the one given by (7) with $\mathcal{E}(\eta)$ being the union of the k balls associated to the α-trimmed k-means of P_{θ_0}. On the other hand, for the m-step estimator, $m > 1$, the IF will be also (7) with $\mathcal{E}(\eta)$ being the union of the ellipsoids defined by the $1-\alpha_m$ level curves of the k normal laws involved in the mixture determined by θ_0.

The use of this last region, better adapted to the underlying mixture, is not important if the parent distribution is symmetrical, but it becomes very useful in non-symmetrical situations. This arises from the expressions in (8) and is made apparent in the graphs in Figure 4. The lower row in Figure 4 shows an asymmetric case for the one-dimensional mixture $(N(-3,1.5)+N(0,1.5)+2N(3,1.5))/4$. The graph on the left shows the IF when the k-means are used and the one on the right when employing the ellipsoids. In the upper row in Figure 4, we analyze the symmetric mixture $(N(-5,1)+N(0,1)+N(5,1))/3$. Since in this case there is no difference between both regions, to ease the understanding of the figure, we show on the left the IF for the means and on the right the IF for the variances. To avoid excessive noise in the images we excluded the IF for the weights of the component distributions. In all graphs the black curves represent the corresponding density functions augmented 40 times.

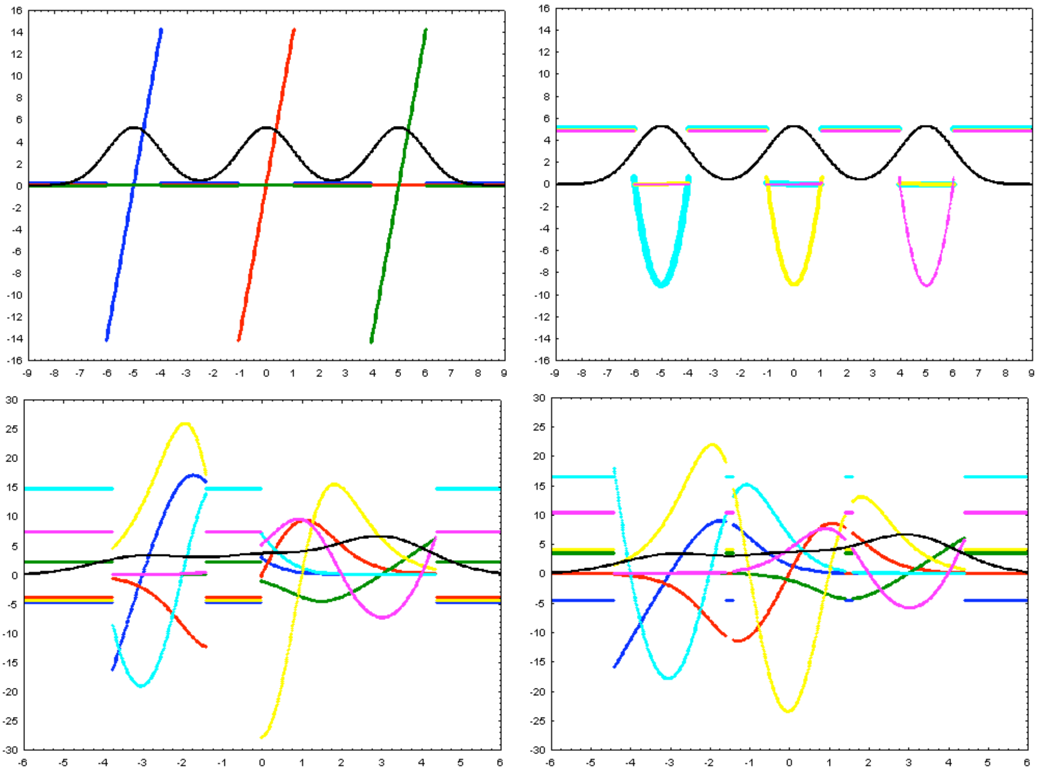

Fig. 4. IF's for the means (blue, green and red) and the variances (cyan, yellow and magenta) of the distributions making up a mixture of three normal distributions. The upper graphs correspond to the two-steps estimator for the mixture $(N(-5,1)+N(0,1)+N(5,1))/3$. The graph on the lower left (resp. lower right) presents the IF for the two-step (resp. m-step) estimator for the one-dimensional mixture $(N(-3,1.5)+N(0,1.5)+2N(3,1.5))/4$. The black curves represent the corresponding density functions augmented 40 times.

To get a more accurate idea of the IF, we include the expression of the components (in π_i, for $i = 1, ..., k-1$, and μ_i and Σ_i, for $i = 1, ..., k$) of $h_{\theta,\eta}(x)$ as a function of $\theta = (\pi_1, ..., \pi_{k-1}, \mu_1, ..., \mu_k, \Sigma_1, ...\Sigma_k)$

$$\begin{aligned}
\frac{\partial}{\partial \pi_i} L_{\theta/\mathcal{E}}(x) &= \left(\frac{P_\theta(i/x)}{\pi_i} - \frac{P_\theta(k/x)}{\pi_k} \right) I_{\mathcal{E}}(x) + P_\theta \left[\frac{P_\theta(i/x)}{\pi_i} - \frac{P_\theta(k/x)}{\pi_k} \Big/ \mathcal{E}^c \right] I_{\mathcal{E}^c}(x) \\
\frac{\partial}{\partial \mu_i} L_{\theta/\mathcal{E}}(x) &= \Sigma_i^{-1}(x-\mu_i) P_\theta(i/x)\ I_{\mathcal{E}}(x) + P_\theta \left[\Sigma_i^{-1}(x-\mu_i) P_\theta(i/x)/\mathcal{E}^c \right] I_{\mathcal{E}^c}(x), \\
\frac{\partial}{\partial \Sigma_i} L_{\theta/\mathcal{E}}(x) &= \frac{1}{2} \left(\Sigma_i^{-1}(x-\mu_i)(x-\mu_i)^T \Sigma_i^{-1} - \Sigma_i^{-1} \right) P_\theta(i/x)\ I_{\mathcal{E}}(x) \\
&\quad + \frac{1}{2} P_\theta \left[\left(\Sigma_i^{-1}(x-\mu_i)(x-\mu_i)^T \Sigma_i^{-1} - \Sigma_i^{-1} \right) P_\theta(i/x)/\mathcal{E}^c \right]\ I_{\mathcal{E}^c}(x).
\end{aligned} \tag{8}$$

The study of the BP of the method is not as simple as that of the IF. Pathological constellations of data may break down even the trimmed k-means procedure by the substitution of only one point by another. However, more favorable configurations can exhibit a BP equal to the trimming level, thus the BP of a procedure in this framework must be considered as data dependent. In any case, the impartial restrictions link the estimations to the procedure used to obtain the initial clustered region. Then, the BP of our, 2 or m-step, estimators for the location parameters are very related to that of the trimmed k-means.

In Donoho and Huber's replacement sample version, some data are replaced by unfortunate data points and the optimistic upper bound $\min\{(\lceil \alpha n \rceil + 1)/n, \min_{i=1,...,k} n_i/n\}$, where n_i is the size of the i-th cluster, is realistic for the location parameters in most well-clustered data sets (see [8]).

Alternatively, the BP may be analyzed under a general assumption of well-clustered data in an idealized situation that permits comparisons between procedures under controlled assumptions in a kind of 'in vitro' analysis. Hennig (Section 4 in [11]), introduces such an ideal model and shows the bad behavior of several estimators for mixtures through an *addition r-components BP*; which is defined as $l/(n+l)$ when l is the minimum number of points to be added to the sample to break down r parameters in the estimation.

Let $k \geq 2$ and $n_1, ..., n_k \in I\!N$ be fixed and such that $n_1 < \ldots < n_k$. Let us consider k sequences of sets indexed by $m \in I\!N$: $A_m^1 = \{x_{1,m}, ..., x_{n_1,m}\}$, $A_m^2 = \{x_{(n_1+1),m}, ..., x_{n_2,m}\}$, ..., $A_m^k = \{x_{(n_{k-1}+1),m}, ..., x_{n_k,m}\}$, and let $\mathcal{X}_m = \cup_i A_m^i$.

Following the ideas in Section 4.1 in [11] we consider this sequence $\mathcal{X}_m$ as an ideal array of well k-clustered data sets whenever there exists $b < \infty$ such that for every $m \in I\!N$,

$$\max_{1 \leq i \leq k} \max\{\|x_{j,m} - x_{l,m}\| : x_{j,m}, x_{l,m} \in A_m^i\} \quad < \quad b \ \text{ and} \tag{9}$$

$$\lim_{m \to \infty} \min\{\|x_{j,m} - x_{l,m}\| : x_{j,m} \in A_m^h,\ x_{l,m} \in A_m^i;\ i \neq h\} \quad = \quad \infty. \tag{10}$$

Under this model the addition of r outliers assumes the existence of a sequence $\mathcal{Y}_m = \{y_{1,m}, ...y_{r,m}\}$ added to $\mathcal{X}_m$ to constitute new data sets $\mathcal{X}_m \cup \mathcal{Y}_m$ verifying

$$\lim_{m \to \infty} \min\{\|y_{j,m} - x_{l,m}\| :\ y_{j,m} \in \mathcal{Y}_m,\ x_{l,m} \in \mathcal{X}_m\} \quad = \quad \infty, \text{ and} \tag{11}$$

$$\lim_{m \to \infty} \min\{\|y_{j,m} - y_{l,m}\| :\ y_{j,m}, y_{l,m} \in \mathcal{Y}_m, j \neq l\} \quad = \quad \infty. \tag{12}$$

The breakdown of an estimator E_n must be understood here in a relative fashion, relating the behavior of the estimator acting over $\mathcal{X}_m$ and over $\mathcal{X}_m \cup \mathcal{Y}_m$ for large values of m.

For estimators related to location (as the k-means) breakdown happens if for every rearrangement of the components of the estimator $\|E_n(\mathcal{X}_m) - E_n(\mathcal{X}_m \cup \mathcal{Y}_m)\| \to \infty$, as $m \to \infty$, holds. For the estimator of the weights, the components breakdown would happen if the minimum weight estimation under $\mathcal{X}_m$ converges to zero while under $\mathcal{X}_m \cup \mathcal{Y}_m$ it remains bounded away from zero, or vice-versa. For the covariance estimators, breakdown would happen, if the smallest eigenvalue of the estimated matrix under $\mathcal{X}_m$ converges to zero while under $\mathcal{X}_m \cup \mathcal{Y}_m$ it remains bounded away from zero, or vice-versa, but also if that is not the case but one of the sequence of matrices is bounded while the other is unbounded.

Hennig handles this ideal model of data sets to show (Theorem 4.4 in [11]) that $r < k$ added outliers break down (in any case!) the estimation of r parameters through the ML estimation, as well as through robustified versions like the t-mixture model of McLachlan and Peel or the Fraley and Raftery proposal (also considered in [17]). In particular, the addition of only one outlier breaks down the estimation of at least one parameter.

Note that this idealized model guarantees, in Hennig's words, that if enough mixture components are fitted, *"eventually there exists a mixture component corresponding to each group, all mixture components correspond to one of the groups and the maximum of the log-likelihood can be obtained from the maxima considering the groups alone; that is, all groups are fitted separately".* Therefore it is easy to show that the α-trimmed k-means do not break down unless we add more than $\lceil \alpha n \rceil$ outliers. The link constituted by the impartial restrictions (3) and an argument similar to that arising from Lemmas 4.1 and 4.2 in [11] (Proposition 4.5 plays here an analogous role) guarantee that our m-step procedure does not break down, if we add $r \leq \lceil \alpha n \rceil$ outliers, if the number of points of every cluster A_m^i is greater than $\lceil \alpha n \rceil + d$ and they are in *general position.* Then every affine hyperplane $H \subset I\!R^d$ contains, at most, d points of A_m^i and there is not possibility of degeneracy of some component into a lower dimension. This leads to the following, even pessimistic, result on the BP of our procedure assuring the lower bound $\lceil \alpha n \rceil/(n + \lceil \alpha n \rceil)$ for the addition BP of 1-component in Hennig's model.

Theorem 4.10. *Let $\mathcal{X}_m, m \in I\!N$, be an ideal array of data sets in $I\!R^d$ well clustered in $k \geq 2$ groups A_m^i, $i = 1, ..., k$, verifying (9) and (10), such that the points in every group A_m^i are in general position and fulfill $n_i - n_{i-1} > \lceil \alpha n \rceil + d$, $i = 1, ..., k$ where we take $n_0 = 0$.*

If $r \leq \lceil \alpha n \rceil$, then the m-step estimator of the parameter $\theta \in \Theta$, determining the mixture of k multivariate normal distributions, does not break down by the addition of r outliers through a sequence $\mathcal{Y}_m = \{y_{1,m}, ..., y_{r,m}\}$ verifying (11) and (12).

A nice complement of our previous analysis on the breakdown point of the trimmed k-means and other clustering methods appears in Hennig [12].

5. Discussion

The connection between mixture and clustering modelings is often used to get a cluster configuration from an estimation of the parameters in a mixture. Here we exploit such connection just in the opposite way. Our estimation procedure starts with a clustering process to estimate the parameters in the mixture. This point of view allows us to take advantage of robust clustering methods to produce robust estimators in the MNMM estimation setup.

We assume the knowledge of the number of components in the mixture. In some situations this assumption can hinder the model, but it is realistic in many problems which involve 'a priori' information of the existence of a determined number of groups in a physical sense (corresponding to say sex, species, kind of illness,...) although the information on the classification of the data might have been lost or, simply, non-recorded. Moreover, in a natural way, our approach allows us to consider situations related to finding a fixed number of main components in the mixtures considering the remaining ones (if any) as contamination.

The introduced procedure is based on making the estimation from a highly representative subset of the data. The choice of such a set begins with a preliminary selection of a core of the data through a clustering-based trimmed procedure. Subsequent improvements are based on ML estimations over increasing sub-sets of representative data obtained in each step by trimming according to the estimated model in the previous step. The additional tools for the estimation process are the EM algorithm, for the involved computations, and impartial restrictions on the parameters, which aid to avoid singularities and spurious solutions. These data-driven restrictions require that the sub-populations which constitute the mixture must be sufficiently represented in the initial trimmed sample.

The proposed method shows a good performance not only under symmetrical contamination but also in the presence of concentration of outliers which often cause other proposals to break down. The estimators obtained are asymptotically Gaussian with $n^{1/2}$ convergence rate and qualitatively robust. The analysis of the BP under Hennig's idealized model shows that the procedure greatly improves those of the available procedures for a fixed number of components. The IF shows finite gross error sensitivity for the estimators. Also, as usually happens for the methods involving data trimming, the IF is discontinuous in the boundary of the region used to trim. The influence of non-trimmed points on the estimation of the parameters of one distribution are modulated by their 'a posteriori' probability of arising from that distribution.

The initial active data set can be selected through the trimmed k-means. In practice, even with this simple method, through the improvement steps based on ML we shall often detect adequate shape and location parameters for the groups as to try the final joint estimation in a successful way. This choice is computationally feasible and can be modulated through the initial trimming level to obtain our goal in well clustered data sets. Moreover most of the asymptotic mathematical analysis of the estimators is valid for other more elaborated clustering-based trimming procedures, as soon as they are consistent.

To conclude, we want to point out that the estimation in the mixture model inherits so many difficulties as to make reliable no method when facing specifically designed unsuited problems. Our proposal shows a nice behavior under the analyzed conditions, where other methods show a poor one. Variations of the presented method, adapted to more involved problems, can be also considered handling other initial robust clustering methods. We think that the results obtained have been encouraging enough to merit the inclusion of our methodology in the toolbox of applied statisticians for estimation of mixtures.

References

[1] Bensmail, H. Celeux, G., Raftery, A.E. and Robert, C. P. (1997) Inference in model-based cluster analysis. *Statist. and Computing* **7**: 1-10.

[2] Campbell, N.A. and Mahon, R.J. (1974). A multivariate study of variation in two species of rock crab of genus Leptograpsus. *Australian J. Zoology* **22**: 417-425.

[3] Cuesta-Albertos, J.A.; Gordaliza, A. and Matrán, C. (1997). Trimmed k-means: An attempt to robustify quantizers, *Ann. Statist.* **25**: 553-576.

[4] Cuesta-Albertos, J.A.; Matrán, C. and Mayo-Iscar, A. (2007). Trimming and likelihood: Robust location and dispersion estimation in the multivariate model. To appear in *Ann. Statist.*

[5] Cuesta-Albertos, J.A.; Matrán, C. and Mayo-Iscar, A. (2007). Estimators based in adaptively trimming cells in the mixture model. *Technical Report.*

[6] Fraley, C. and Raftery, A. E . (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer J.* **41**: 578-588.

[7] Gallegos, M.T. and Ritter, G. (2005). A robust method for cluster analysis. *Ann. Statist.*, **33**: 347-380.

[8] García-Escudero, L.A. and Gordaliza, A. (1999). Robustness properties of k-means and trimmed k-means. *J. Amer. Statist. Assoc.* **94**: 956-969.

[9] García-Escudero, L.A. and Gordaliza, A. (2007). The importance of the scales in heterogeneous robust clustering. *Comput. Statist. Data Anal.* **51**: 4403-4412.

[10] Hathaway, R.J. (1985). A constrained formulation of Maximum Likelihood Estimation for Normal Mixture Distributions. *Ann. Statist.* **13**: 795-800.

[11] Hennig, C. (2004). Breakdown point for maximum likelihood estimators of location-scale mixtures. *Ann. Statist.* **32**: 1313-1340.

[12] Hennig, C. (2007). Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. To appear in *J. Multivariate Anal.*

[13] Kharin, Y. (1996). *Robustness in Statistical Pattern Recognition.* Kluwer, Dordrecht.

[14] Marazzi, A. and Yohai, V.J. (2004). Adaptively truncated maximum likelihood regression with asymmetric errors. *J. Statist. Plann. Inference,* **122**: 271-291.

[15] Marin, J. M., Mengersen, K. and Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions, in *Bayesian Thinking, Modeling and Computation*, eds. Dipak Dey, C.R. Rao. Handbook of Statistics 25, 459-507. Elsevier.

[16] Markatou, M. (2000). Mixture models, Robustness, and the Weighted Likelihood Methodology. *Biometrics* **56**: 483-486.

[17] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

[18] Peel, D. and McLachlan, G.J. (2000). Robust mixture modelling using the t distribution. *Statist. and Computing* **10**: 339-348.

[19] R Development Core Team (2006). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. http://www.R-project.org.

[20] Robert, C.P. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer-Verlag, New York.

[21] Stephens, M. (2000). Bayesian Analysis of Mixture Models with an unknown number of components-An alternative to reversible jump methods. *Ann. Statist.* **28**: 40-74.

[22] Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer-Verlag. New York.

[23] Wu, C. F. (1983). On the Convergence Properties of the EM algorithm. *Ann. Statist.* **11**: 95-103.

[24] Yakowitz, S.J. and Spragins, J.D. (1968). On the identifiability of finite mixtures. *Ann. Math. Statist.* **39**: 209-214.

6. Appendix

PROPOSITION 6.1. *Let $\mathcal{Y} = \{g_\phi : \phi \in \Phi\}$. Let $A \subset I\!R^d$ be a non-empty open set and Ψ be the function defined by $\Psi(f) = fI_A$ on the set $\langle \mathcal{Y} \rangle$ of the linear combinations of elements of $\mathcal{Y}$. Then Ψ is a linear isomorphism of $\langle \mathcal{Y} \rangle$ on the image space.*

PROOF.- Obviously Ψ is linear. To show that Ψ is an injective map, let $\phi_1 \neq \phi_2$ and assume that $g_{\phi_1}(x) = g_{\phi_2}(x)$ for every $x \in A$. Then, if $x \in A$,

$$(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) - (x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2)) = 2\log\left(\frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|^{\frac{1}{2}}}\right).$$

Since the expression on the left hand side can be expanded in a power series, it must also be constant on $I\!R^d$, thus $(\mu_1, \Sigma_1) = (\mu_2, \Sigma_2)$, and both distributions are the same. •

Proposition 6.2 contains some basic properties of the trimmed k-means useful to obtain the results on consistency. They are taken (or are easily deduced) from [8]. In this proposition we employ that, under our model, from the Glivenko-Cantelli theorem, the sequence $\{P_n\}_n$ (a.s.) converges in distribution to P_{θ_0}. Thus, (from Skorohod's Representation Theorem for the weak convergence) we can (and will) assume that $\{P_{\theta_0}, P_1, ...\}$ are the distributions of some random vectors $\{Y_0, Y_1, ...\}$ such that $Y_n \to Y_0$ ν-a.s.

PROPOSITION 6.2. *If P is absolutely continuous, then the sequence of trimmed k-means and associated trimmed regions of the empirical measures P_n fulfills:*

(a) $\lim_n \|\gamma_n - \gamma_P\| = 0$.
(b) $\lim_n I_{\mathcal{B}(\gamma_n)}(Y_n) = I_{\mathcal{B}(\gamma_P)}(Y_0)$, *$\nu$-a.s.*
(c) $\lim_n P_n[\mathcal{B}(\gamma_n)] = P[\mathcal{B}(\gamma_P)] = 1-\alpha$.
(d) $\lim_n P_n\left[I_{\mathcal{B}(\gamma_n)} \log f_{\theta_P}\right] = P\left[I_{\mathcal{B}(\gamma_P)} \log f_{\theta_P}\right]$.
(e) $\lim_n P_n L_{\theta_P/\mathcal{B}(\gamma_n)} = PL_{\theta_P/\mathcal{B}(\gamma_P)}$.

PROOF OF PROPOSITION 4.5.- Let $\phi \in \Phi$, denote $M(\phi) := \sup\{g_\phi(x) : x \in I\!R^d\}$, and assume that (a) holds. By resorting to a subsequence argument and a relabeling, let us assume that $1 \in I \subset \{1, ..., k\}$, where $\lim_n \lambda_i^n = 0$ if $i \in I$, $\liminf_n \lambda_i^n > 0$ if $i \notin I$, and $M(\phi_1^n) = \sup\{M(\phi_i^n) : i \in I\}$, $n \in I\!N$.

Note that

$$K_1 := \sup_{i \notin I} \sup_n M(\phi_i^n) < \infty. \tag{13}$$

Given $r > 0$, let $H_r := \left\{ x \in I\!\!R^d : \langle x - \mu_1^n, v_n \rangle^2 \leq r^2 \right\}$, where v_n is the eigenvector associated to λ_1^n, and set $r_n := \inf \{ r > 0 : P_n[H_r / \mathcal{B}(\gamma_n)] > u/2 \}$.

From the continuity of P, we obtain that $\lim_n P_n[H_{r_n}/\mathcal{B}(\gamma_n)] = u/2$. Thus $\liminf_n r_n > 0$ because, otherwise $\liminf_n P_n[H_{r_n}/\mathcal{B}(\gamma_n)] = 0$ would hold. Let

$$C_n := \left\{ x \in H_{r_n}^c \cap \mathcal{B}(\gamma_n) : P_{\theta_n}(1/x) \geq \frac{u}{4} \right\}.$$

We have that,

$$\begin{aligned} u &\leq \liminf_n \frac{1}{P_n[\mathcal{B}(\gamma_n)]} P_n \left[I_{\mathcal{B}(\gamma_n)} P_{\theta_n}(1/\cdot) \right] \\ &\leq \lim_n P_n[H_{r_n}/\mathcal{B}(\gamma_n)] + \liminf_n \frac{1}{P_n[\mathcal{B}(\gamma_n)]} P_n \left[I_{\mathcal{B}(\gamma_n) \cap H_{r_n}^c} P_{\theta_n}(1/\cdot) \right] \\ &\leq \frac{u}{2} + \frac{u}{4} + \liminf_n P_n[C_n/\mathcal{B}(\gamma_n)], \end{aligned}$$

and, as a consequence, $\liminf_n P_n[C_n/\mathcal{B}(\gamma_n)] \geq u/4$. From here and (c) in Proposition 6.2,

$$\liminf_n P_n[C_n] \geq u(1-\alpha)/4 > 0. \tag{14}$$

On the other hand, if $i \in \{2, ..., k\}$ and $x \in C_n$, we have that

$$\frac{u}{4} \leq P_{\theta_n}(1/x) \leq \frac{\pi_1^n g_{\phi_1^n}(x)}{\pi_i^n g_{\phi_i^n}(x)}.$$

Therefore, if $x \in C_n$,

$$\sup_{i=1,\ldots,k} \pi_i^n g_{\phi_i^n}(x) \leq \frac{4}{u} g_{\phi_1^n}(x) \leq \frac{4}{u} \beta_1^n, \tag{15}$$

where $\beta_1^n = \sup_{x \notin H_{r_n}} g_{\phi_1^n}(x)$. From here and (13), from an index onward, we have that

$$\begin{aligned} P_n L_{\theta_n/\mathcal{B}(\gamma_n)} &\leq P_n \left[I_{\mathcal{B}(\gamma_n) \cap C_n^c} \log f_{\theta_n} \right] + P_n \left[I_{C_n} \log f_{\theta_n} \right] \\ &\leq P_n[\mathcal{B}(\gamma_n) \cap C_n^c] \log \left[\sup(K_1, M(\phi_1^n)) \right] + P_n[C_n] \log \left[k 4 \beta_1^n / u \right] \\ &\leq \log(k4/u) + \log^+(K_1) + \log \left[(\beta_1^n)^{P_n[C_n]} M(\phi_1^n) \right], \end{aligned}$$

which converges to $-\infty$ due to (14), to $\beta_1^n = (2\pi\lambda^n)^{-d/2} \exp\left(-r_n^2/(2\lambda^n)\right))$ and to $M(\phi^n) \leq (2\pi\lambda^n)^{-d/2}$.

Now, let us suppose that (b) or (c) hold. By repeating the subsequence argument and the notation simplifications, we assume that for every $i \in \{1, ..., k\}$, $\liminf_n \lambda_i^n > 0$ and $\lim_n \|\phi_i^n\| = \infty$ or $\lim_n \phi_i^n = \phi_i \in \Phi$, and that $\lim_n \|\phi_1^n\| = \infty$, or $\lim_n \pi_1^n = 0$.

Define $D_n := \{ x \in \mathcal{B}(\gamma_n) : P_{\theta_n}[1/x] > u/2 \}$. Then $P_n[D_n/\mathcal{B}(\gamma_n)] > u/2$, and arguing as in (14) and (15), we have that $u(1-\alpha)/2 \leq \liminf_n P_n(D_n)$ and, if $x \in D_n$, that

$$f_{\theta_n}(x) \leq k 2 \delta_1^n g_{\phi_1^n}(x)/u, \tag{16}$$

and $K_2 := \sup_n \sup_i M(\phi_i^n) < \infty$. Therefore:

$$\begin{aligned} P_n L_{\theta_n/\mathcal{B}(\gamma_n)} &\leq P_n \left[I_{\mathcal{B}(\gamma_n) \cap D_n^c} \log f_{\theta_n} \right] + P_n \left[I_{D_n} \log f_{\theta_n} \right] \\ &\leq \log^+(K_2) + \log(k2/u) + P_n \left[I_{D_n} \log(\pi_1^n g_{\phi_1^n}) \right], \end{aligned}$$

which converges to $-\infty$. The proof ends as in the previous case. •

Lemma 6.3. *If θ_0 is an inner point of Θ_{γ_0}, then there exists $N_0 \in I\!N$ such that if $n \geq N_0$, then $\theta_0 \in \Theta^n$.*

Proof.- The continuity of the map $x :\to P_{\theta_0}(i/x)$ and (b) in Proposition 6.2, give that

$$P_{\theta_0}(i/Y_n)I_{\mathcal{B}(\gamma_n)}(Y_n) \to \text{a.s. } P_{\theta_P}(i/Y_0)I_{\mathcal{B}(\gamma_0)}(Y_0). \tag{17}$$

Now, taking into account that $P_{\theta_0}(i/\cdot) \in [0,1]$, we obtain that, for some $\delta > 0$:

$$\begin{aligned} P_n\left[P_{\theta_0}(i/\cdot)I_{\mathcal{B}(\gamma_n)}\right] &= \nu\left[P_{\theta_0}(i/Y_n)I_{\mathcal{B}(\gamma_n)}(Y_n)\right] \\ &\to \nu\left[P_{\theta_0}(i/Y_0)I_{\mathcal{B}(\gamma_0)}(Y_0)\right] \\ &= P\left[P_{\theta_0}(i/\cdot)I_{\mathcal{B}(\gamma_0)}\right] \geq (u+\delta)P\left[\mathcal{B}(\gamma_0)\right], \end{aligned}$$

and the proof ends by applying (c) in Proposition 6.2. •

Corollary 6.4 follows from Lemma 6.3 and *(e)* in Proposition 6.2, taking into account that in Proposition 4.5 we can take the vectors θ_n as close as desired to the optimum parameters.

Corollary 6.4. *If θ_0 is an inner point of Θ_{γ_0}, then, from an index onward, the sequence $\{\hat{\theta}_n\}_n$ belongs to a compact set contained in Θ.*

Proof of Theorem 4.6: It is straightforward by resorting to standard techniques and to the properties enumerated in Proposition 6.2. •

Proof of Theorem 4.8: After our consistency results, for the analysis of the asymptotic distribution, we can assume that the η-parameters belong to a compact subset K of $\tilde{\Gamma}$, as well as that the θ-parameters fulfill the restrictions given by Θ^n and belong to the set $\{\theta : \|\theta - \theta_0\| < \delta\}$ for some small enough $\delta > 0$ and large enough n.

Now the proof parallels that given in [4] based on extended versions of the results in Section 3.2.4 in [22] to this semiparametric framework. In our case, for $m_{\theta,\eta}$ defined as in (6) the components of $\dot{m}_{\theta,\eta} := h_{\theta,\eta}$ are those given in (8) with $\mathcal{E}(\eta)$ as $\mathcal{E}$. The result is then the consequence of Lemma 6.5 below, similar to Lemma 3.12 in [4]. From here, taking into account Proposition 4.2 and some easy computations, obtaining the asymptotic distribution given in the theorem as well as its different expressions is straightforward. •

Lemma 6.5. *There exist $\delta > 0$ and a compact neighborhood K of η^* such that*

$$\left\{\frac{m_{\theta\eta} - m_{\theta_0\eta} - (\theta-\theta_0)^T\,\dot{m}_{\theta_0\eta}}{\|\theta - \theta_0\|} : \|\theta-\theta_0\| \leq \delta, \eta \in K\right\} \tag{18}$$

is P-Donsker and

$$P\left(m_{\theta\eta} - m_{\theta_0\eta} - (\theta-\theta_0)^T\,\dot{m}_{\theta_0\eta}\right)^2 = o\left(\|\theta-\theta_0\|\right)^2, \tag{19}$$

uniformly in $\eta \in K$.

Proof.- Let δ small enough to assure that the parameters in $\Theta_\delta := \{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta\}$ do not lead to degeneration of the mixture, and let K be any compact neighborhood of η^*. If we choose a compact ball, B_0, in $I\!R^p$ containing all the ellipsoids composing the sets $\mathcal{E}(\eta), \eta \in K$, the continuity of $m_{\theta,\eta}$ and $\dot{m}_{\theta,\eta}$ with respect to the argument and with respect to the parameters guarantee that the functions in the family (18) are uniformly bounded by a constant over the set B_0. This implies the uniform L_2-Frechet derivability (19).

The first statement is then consequence of a chain of arguments beginning with:

- The class $\mathcal{M}_\delta$ of density functions of mixtures of normal distributions with parameters in V_δ fulfills the uniform entropy condition (see Section 2.5.1 in [22]).

The class of functions given by

$$\mathcal{I} := \left\{ \log\left((2\pi)^{-\frac{p}{2}} \left(\det\left(\Sigma\right)\right)^{-\frac{1}{2}} \right) - \frac{1}{2}\left(x-\mu\right)' \Sigma^{-1}\left(x-\mu\right) : \ \mu \in I\!\!R^p, \ \Sigma \in \mathcal{M}^+_{p\times p} \right\}$$

defines a linear space of finite dimension, thus it is a VC-class of functions (see Lemma 2.6.15 in [22]). The density functions of normal distributions are obtained by composing a function in the class $\mathcal{I}$ with the exponential function, $\exp\left(\mathcal{I}\right)$, hence it is also a VC-class of functions (see Lemma 2.6.18 in [22]). Now, we can assure that the finite mixtures of normal distributions are a VC-hull class, and from Corollary 2.6.12 and the previous arguments in [22], a such class fulfills the uniform entropy condition.

- The class of functions $\log(\mathcal{M}_\delta)I_{B_0} := \{\log(f)I_{B_0} : \ f \in \mathcal{M}_\delta\}$ fulfills the uniform entropy condition.

The class of functions $\mathcal{M}_\delta$ fulfills the condition, so we can apply Theorem 2.10.20 in [22] to assure that the transformed class $\log\left(\mathcal{M}_\delta\right) I_{B_0}$ also fulfills that condition. We only need to show that there exists a constant, A, such that

$$\left(\log\left(f\left(x\right)\right) I_{B_0}\left(x\right) - \log\left(g\left(x\right)\right) I_{B_0}\left(x\right)\right)^2 \leq A^2 \left(f\left(x\right) - g\left(x\right)\right)^2, \ \ x \in I\!\!R^p, \quad f, g \in \mathcal{M}_\delta,$$

but this is an easy consequence of the mean value theorem and the fact that we can obtain two constants $0 < c < C$ such that $c < f(x) < C$, for all $f \in \mathcal{M}_\delta$ and $x \in B_0$.

- The class of indicator functions of unions of k ellipsoids and the class of indicator functions of complementary of unions of k ellipsoids fulfill the condition of uniform entropy

- $\inf\left\{P_\theta\left(\mathcal{E}(\eta)^c\right) : \ \ \theta \in \Theta_\delta \text{ and } \eta \in K\right\} > 0.$

- The family $\left\{I_{\mathcal{E}(\eta)} \log\left(f_\theta\right) + I_{\mathcal{E}(\eta)^c} \log\left(P_\theta\left(\mathcal{E}(\eta)^c\right)\right) : \ \ (\theta, \gamma) \in \Theta_\delta \times K\right\}$ fulfills the uniform entropy condition.

Theorem 2.10.20 in [22] leads to this statement, because this class of functions is constituted by sums of functions verifying the uniform entropy condition.

- The class of the functions $I_{\mathcal{E}(\eta)}\left(x\right) \frac{\partial}{\partial\theta} \log\left(f_\theta\left(x\right)\right) + I_{\mathcal{E}(\eta)^c}\left(x\right) \frac{\partial}{\partial\theta} \log\left(P_\theta\left(\mathcal{E}(\eta)^c\right)\right)$, where $\theta \in \Theta_\delta$ and $\eta \in K\}$ is a Donsker class.

This statement can be proved by a chain of arguments similar to the above, beginning with the fact that the class of functions $\{I_{B_0}\left(x\right) \frac{\partial}{\partial\theta} \log\left(f_\theta\left(x\right)\right) : \theta \in \Theta_\delta\}$ is a Donsker class of functions. But this follows from the fact that the components of these functions are products of $P_\theta(i/x)I_{B_0}$ with functions of the types $\frac{1}{\pi_i}, \Sigma_i^{-1}\left(x-\mu_i\right)$ and $-\frac{1}{2}\Sigma_i^{-1} + \frac{1}{2}\Sigma_i^{-1}\left(x-\mu_i\right)\left(x-\mu_i\right)' \Sigma_i^{-1}$. •