Estimators based in adaptively trimming cells in the mixture model *

J.A. Cuesta-Albertos
Departamento de Matemáticas, Estadística y Computación,
Universidad de Cantabria, Spain

C. Matrán† and A. Mayo-Iscar Departamento de Estadística e Investigación Operativa,
Universidad de Valladolid, Spain

Abstract

The paper introduces a robust estimation procedure, in the multivariate normal mixture model, 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. Maximum likelihood estimation, handling the trimmed data as censored, 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 clusterized region, allow to avoid singularities and to guarantee the existence of the estimator. Our analisis 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.

A.M.S. 2000 subject classification: Primary: 62F35, secondary: 62H12, 62P99

1 Introduction

Estimation in mixture models has caught the interest of many researchers due to their multiple statistical applications. McLachlan and Peel [17], for example, offers an updated account of the development and state of the art in this topic. In this paper we will assume the so called multivariate normal mixture model (MNMM) which is given by $\{I\!P_\theta : \theta \in \Theta\}$, where $I\!P_\theta$ is the distribution on $I\!R^d$ with density

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

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 Q_{ϕ_i} with mean μ_i and covariance matrix Σ_i, and π_i is the mixing proportion of Q_{ϕ_i} in the mixture. θ includes all the parameters in the model. Thus, θ varies in the set

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

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.

The necessity to employ robust estimation procedures is patent. On the first hand, the reported difficulties in the clustering setting (see e.g. Cuesta-Albertos, Gordaliza and Matrán [3], Banfield and Raftery [1], Hardin and Rocke [10], Gallegos [6] or Hennig [12]) are here in force. On the other hand, we

*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

shall also face a great instability of the classical estimators and of the available algorithms to compute them.

The main computable approaches to the problem in this multivariate setting with a robust motivation seem to be reduced to the one proposed by Fraley and Raftery [5] through the addition of a mixture component accounting for noise modeled as a uniform distribution, and the based on a mixture of t distributions by McLachlan and Peel (see e.g. [18] and Section 7 in [17] for other references). However, as noted by Hennig in [12], *"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 to obtain a robust estimator in the MNMM and we include a sound theoretical analysis of its properties.

Our approach is based on a m-step procedure beginning with a robust initial estimator whose efficiency is improved through several iterations of a suitable maximum likelihood (ML) step. This method constitutes 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 [15] in the univariate regression model. Also Markatou [16] considered a related weighted likelihood procedure for mixtures, but based on a preliminary nonparametric density estimation procedure, which could made the procedure undesirable for the multivariate setting due to the curse of the dimensionality.

The procedure initially searches a small (purportedly) uncontaminated core of the data, consisting of k well separated clusters, each one associated to 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 first estimation. This process is 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. Hence the procedure gradually increases the representative clusters, decreasing the trimming size, and re-estimating θ in each step. The process is repeated until some maximal uncontaminated trimmed set is reached, based on which the final estimate is obtained.

The scheme is similar to that employed in the elliptical model in [4] although here it involves more steps. Practice shows that a gradual enlargement of the regions is useful to avoid masking effects between the populations. However there are other additional problems associated to this scheme in our framework.

The choice of the initial clustering method can be carried through different nonparametric procedures (e.g. the Hardin and Rocke [10], Gallegos [6] or García-Escudero and Gordaliza [8] proposals). To simplify the exposition focusing in the adaptiveness of the methodology, our choice is based on the perhaps simplest one, the impartial trimmed k-means introduced by Cuesta-Albertos, Gordaliza and Matrán in [3]. This procedure involves a trimming of the data and it is intended to look for spherical clusters with similar weights. However, we will see (in Subsection 2.2.1) that, in practice, if in the iteration steps we gradually reduce the trimming sizes, then the proposed procedure often produces a good estimation of the parameters in general mixtures.

On the other hand, as it is well known, in this setup the likelihood function is usually unbounded and can present multiple local maxima, leading to ad hoc procedures to eliminate spurious solutions. The use of restrictions on the parameter space to circumvent this difficulty was pioneered by Hathaway [11], in the setting of the mixture of univariate normal distributions. Here, we introduce a data-driven proposal,

the impartial restriction on the parameter which only requires that each distribution in the mixture has enough representation in the initial trimmed sample.

Although we will develop the method in the general framework of (possibly) different covariance matrices and unclassified data, we want to point out that the method can easily be adapted to work with additional information concerning, say, partially classified data or the same covariance structure for the distributions in the mixture.

The details of the procedure are given in Section 2. In Section 3 we show the performance of the procedure in several examples where our method is compared with other robust procedures. Section 4 is devoted to the theoretical justification of the procedure, including asymptotic results and robustness properties. The paper ends with a section containing a discussion of the proposed procedure and an appendix containing the proofs of our results.

2 Assumptions and description of the procedure

Throughout we will handle probability measures defined on β^d, the Borel sets in the euclidean space $\mathbb{R}^d$, $d \geq 1$, whose usual norm will be denoted by $\|-\|$. Given $m \in \mathbb{R}^d$ and $r > 0$, $B(m, r)$ will denote the open ball centered at m with radius r. Given $A \subset \mathbb{R}^d$, $\overline{A}$ will denote the topological closure of A, A^c its complement and if $h > 0$, then $A^h := \{x : \inf_{a \in A} \|x - a\| \leq h\}$.

Unless otherwise stated, we will assume that all random vectors (r.v.'s) are defined on the same rich enough probability space (Ω, σ, ν). The notation $\mathbb{P}f$ will denote integration of the random variable f with respect to the probability $\mathbb{P}$. $\{\mathbb{P}_n\}_n$ will denote the associated sequence of empirical distributions associated to a random sample $\{X_n\}_n$ of a distribution $\mathbb{P}$ which generally will be assumed to be absolutely continuous.

To circumvent the problem of the identifiability of the mixture models due to possible permutations of the weights and the distributions in the mixture, we will assume as equivalent $\theta := (\pi_1, ..., \pi_k, \phi_1, ..., \phi_k) \in \Theta$ and every $(\pi_{i_1}, ..., \pi_{i_k}, \phi_{i_1}, ..., \phi_{i_k})$ such that $(i_1, i_2, ..., i_k)$ is a permutation of $(1, 2, ..., k)$.

2.1 Initial estimator

Given a random sample, we begin by choosing a set $\hat{A}$ based on our data, using a robust clustering criterion. Roughly speaking, we are assuming that the population is composed of k clusters and that some contamination could be present in the sample as outliers or, also, as *bridge points* between clusters. Then, $\hat{A}$ should be the union of k bounded sets which are located well inside the data set. This choice is directed at having a specialized region for each sub-population in the mixture with low influence from any other.

To choose $\hat{A}$, we follow the impartial trimmed k-means (from now on trimmed k-means) approach. This solution is quite simple and it is designed to accomplish our goals when the data set is composed by k approximately spherical groups with similar weights and sizes. This can be considered as a limitation of the procedure but in Subsection 2.2.1 we will analyze the improvement of this initial estimator through one or several additional steps to cover a wider framework, and, in any case we emphasize that other alternative initial estimators (more reliable for a particular situation) are possible and covered by our theoretical results.

The trimmed k-means were introduced in [3] (the TRIMCLUSTER package includes an R-code to compute them). Broadly speaking, this procedure first trims the sample and then splits the remaining data into k spherical components in order to minimize the within sums of squares of the distances to the center of the group:

For every $\gamma := (r, m_1, ..., m_k) \in \Gamma := I\!R^+ \times \left(I\!R^d\right)^k$, let $A_\gamma = \cup_{i=1}^k \overline{B}(m_i, r)$. Let us consider a fixed value $\alpha \in (0,1)$ and let $I\!P$ be any probability on $I\!R^d$. It is shown in [3] that there exists $\gamma_P := (r_P, m_1^P, ..., m_k^P) \in \Gamma$ such that the set $A_{\gamma_P} = \cup_{i=1}^k \overline{B}(m_i^P, r_P)$ verifies $I\!P[A_{\gamma_P}] \geq 1-\alpha$ and for every union of closed balls $A := \cup_{i=1}^k \overline{B}(m_i, r_i)$ verifying $I\!P[A] \geq 1-\alpha$

$$\frac{1}{I\!P[A_{\gamma_P}]} \int_{A_{\gamma_P}} \inf_{i=1,...,k} \|x - m_i^P\|^2 I\!P(dx) \leq \frac{1}{I\!P[A]} \int_A \inf_{i=1,...,k} \|x - m_i\|^2 I\!P(dx). \tag{2}$$

The vector $(m_1^P, ..., m_k^P) \in \left(I\!R^d\right)^k$ is called an *α-trimmed k-mean of $I\!P$*, and the associated region is A_{γ_P}. 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 a union of balls with the same radius. This is a peculiarity of the trimmed k-means because the obtained region A_{γ_P} (and the sample version $\hat{A} = A_{\gamma_{P_n}}$) gives identical treatment to each distribution composing the mixture. As already announced, the improvement of this region will be analyzed in Subsection 2.2.1.

The choice of this initial trimming set has two main consequences in our setup. First, if we choose an high trimming value, this choice should eliminate possible outliers and masking effects. On the other hand, once a sub-sample has been chosen we try to estimate θ using the EM algorithm and it is very well known that this algorithm is very sensitive to wrong selections of the initial value. Our proposal consists of initialize 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.

Obviously, other choices (for instance, using some additional information) could lead to improvements but, by the same reason, it should not be very difficult to show situations where this particular choice produces bad behaved solutions. This observation can be extended also to principles that are considered as unquestionable in other settings. For example, a largely used principle in robust estimation is that affine equivariance. However (as pointed out by Hampel in [9]), it is still far from being obvious that contamination and right data must share the same behavior under affine transformations, and this consideration is much more patent in presence of several groups of right data.

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

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

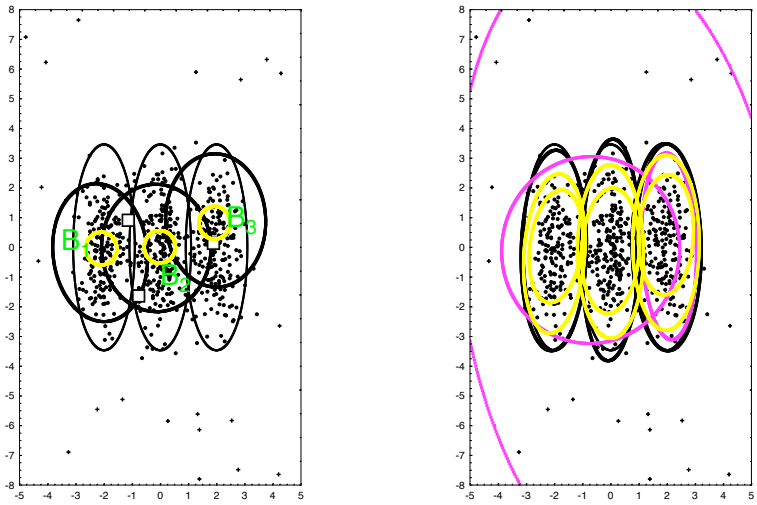

Figure 1: Realizations of 3-means, trimmed 3-means, associated (and iterated) regions and different estimations (represented by the 95% level curves of the weighted estimated normal distributions in the mixture).

The graph on the left in Figure 1 shows the trimmed region, $\hat{A} = \hat{B}_1 \cup \hat{B}_2 \cup \hat{B}_3$, associated to the trimmed 3-means for a trimming level of 0.7 (the union of the three yellow balls), as well as the (non-trimmed) 3-means (marked as bold squares). The only thing to be stressed here is the scarce influence that the contaminated data have on the trimmed 3-mean. On the contrary, the 3-means are badly located within the clusters because they are greatly influenced by the contamination. The remaining features of this graphic and the graphic in the right hand side will be 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 belonging to a bounded set $A \in \beta^d$. As stated in the final discussion there, between the natural possibilities associated to this situation, the censored point of view improves the convergence of the algorithms and under the hypothesized model is the best choice. Therefore we will assume here this setup which we briefly explain now.

Given a fixed set $A \in \beta^d$ the (artificial) censoring leads to consider the *censored log-likelihood function*:

$$L_{\theta/A}(x) := I_A(x) \log f_\theta(x) + I_{A^c}(x) \log I\!P_\theta(A^c), \ x \in I\!R^d, \tag{3}$$

where $0 \times \infty$ is taken as 0 and I_A denotes the indicator function of the set A.

The empirical censored log-likelihood based on a sample of size n is

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

We recall that, under this model, the information we handle is not only the sample points belonging to A, but we are also assuming as known the number of points in A^c. Also, as soon as we guarantee the identifiability of the model on A (see Theorem 4.1 below) we have the uniqueness of the maximization of the likelihood under the model: If $A \in \beta^d$ has a nonempty interior, then

$$I\!P_{\theta_0} L_{\theta_0/A} > I\!P_{\theta_0} L_{\theta/A}, \quad \text{if} \quad \theta \neq \theta_0. \tag{4}$$

However, the sample optimization problem has multiple local maxima and singularity points even in the non-censored case. Therefore, to avoid degenerated or undesirable solutions, we need to impose some kind of restriction to guarantee the existence of the estimator arising from the sample maximization. We propose restrictions based on our assumption on the presence of k populations. So, assuming that the initial procedure is successful in deleting the contaminated data and searching for a representative subset of the sample data, the set $\hat{A}$ should contain sufficient evidence of every population. Therefore, once a threshold value, $u \in (0,1)$, has been chosen, we consider the restricted set

$$\hat{\Theta}_u^n := \left\{ \theta \in \Theta : \frac{1}{\sharp\left\{ r : x_r \in \hat{A} \right\}} \sum_{x_r \in \hat{A}} I\!P_\theta(i/x_r) \geq u, \text{ for every } i = 1, ..., k \right\}, \tag{5}$$

where

$$I\!P_\theta(i/x) = \frac{\pi_i g_{\phi_i}(x)}{f_\theta(x)} = \frac{\pi_i g_{\phi_i}(x)}{\sum_{j=1}^k \pi_j g_{\phi_j}(x)} \tag{6}$$

denotes the a posteriori probability of a point x arising from density g_{ϕ_i}. Thus, the whole quotient in (5) is the sample conditional mean: $I\!P_n[I\!P_\theta(i/\cdot)/\hat{A}]$.

In defining the set $\hat{\Theta}_u^n$ the only man-made selection is that of u. Thus, this set is mostly data-driven and we call it *impartial restricted parameter set*.

Now we are in position to define the (one step) estimator of θ

$$\hat{\theta}_n := \arg\max_{\theta \in \hat{\Theta}_u^n} I\!P_n L_{\theta/\hat{A}}. \tag{7}$$

At this level our proposal to solve (7) is based on the EM algorithm (or the Monte Carlo EM when the involved integrals make the EM infeasible).

Back to Example 2.1. It is not actually necessary to fix very accurate values in the threshold value u in order to define the constrained parameter space. Let us consider a light one, $u = 0.1$, and the trimmed set $\hat{A}$ already obtained in Example 2.1.

We applied the EM algorithm to solve (7) with $\hat{\Theta}_u^n$ given by (5), for the set $\hat{A}$, and $u = 0.1$. 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 starting from the 3-means initial solution is shown in the graph on the right in violet. As already shown in [17] the poor choice of initial value leads to a very bad solution. However, even with good initial solutions, the EM algorithm would exhibit a bad behavior in this case due to contamination. •

2.2.1 Iterations: Improving the trimming regions

We still have some problems which will be eased with the iterative procedure developed in this subsection. The noted drawback of the k-means should be corrected to better reflect the structure of the mixture through a union of k ellipsoids with different 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 our example (as to be expected taking into account the relative size of the set on which we based the estimation).

On the other hand, it would be desirable to recover the principle of affine equivariance of the estimators, once we discarded the possible contaminating data.

In order to advance in the aforementioned directions we mimic the EM-algorithm in the following way. As a result of the already described step 1, we have a value of the trimming parameter $\alpha_1 = \alpha$, an estimated active trimming set $A_n^1 = \hat{A}$ and an estimate $\theta_n^1 = \theta_n$ of the parameter. We will denote $\theta_n^1 = (\pi_1^{n,1}, ..., \pi_k^{n,1}, \phi_1^{n,1}, ..., \phi_k^{n,1})$.

Let $m \in I\!N$ and let $\alpha_2, ..., \alpha_m \in (0,1), \alpha_1 > \alpha_2 > ... > \alpha_m$. Step 2 consists in replacing the trimming set A_n^1 by the set A_n^2 composed of the union of the ellipsoids given by the $1 - \alpha_2$ level curves of the density functions $g_{\phi_1^{n,1}}, ..., g_{\phi_k^{n,1}}$.

Obviously, $I\!P_{\theta_n^1}[A_n^2] \geq 1 - \alpha_2$ and the equality is not satisfied in general. In order to achieve this, we should have taken the α_2-level curve of $f_{\theta_n^1}$. The chosen procedure leads to every distribution in the mixture to be equally well represented, thus improving the estimation of the ϕ's parameters. If we were mostly interested in the estimation of the π's parameters, the α_2-level curve of $f_{\theta_n^1}$ would be our proposal. This precise selection is not too important, and we have chosen the proposed one just to fix ideas.

Now, we can obtain θ_n^2, the MLE associated to the censored likelihood function L_{θ/A_n^2} with the same impartial restrictions as those used in step 1. Then we repeat the process using the trimming sizes $\alpha_3, ..., \alpha_m$ and, for every α_i, the last estimation θ_n^{i-1} as the initial value for the EM algorithm, the active trimming set A_n^i constructed as in step 2 from θ_n^{i-1} and the new trimming level. The process continues until a stopping criteria is met (maybe penalizing the censored likelihood) or until the value m is reached.

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, with a high trimming level to select representative zones of the clusters, and the nature of our restrictions justify the adequacy of keeping this choice as fixed.

We wish to remark on the different kinds of initial solutions used to start the EM algorithm in step 1 and successive steps. From step 2 on, the initial values are the previous estimations given by the procedure for the whole parameter. This should provide a more representative subsample, with a proportional presence of sample points of every distribution making up the mixture.

In relation with the equivariance properties of the procedure we must point out that

- The procedure based on the trimmed k-means is equivariant with respect to isometries up to constants (i.e. transformations T verifying that for some λ, λT is an isometry).
- The iterations allow much of the affine equivariance of ML to be recovered, because the effect of our initial choice gradually disappears through the sequence of estimations.

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 method when based on one step (graph on the left) and on several steps (graph on the right).

The three inside (resp. outside) yellow ellipses show the intermediate adaptive regions corresponding to 50% (resp. 25%) trimming size. The thick ellipses show the 95%-level curves corresponding to the final estimation obtained on the basis of the adaptive region with a 5% trimming size. •

3 The method in action

In this section we present the method acting in several examples. Examples 3.1 and 3.2 are based on simulated data, while Example 3.3 is based on a real data set analyzed in [17]. In Example 3.4 we present a real data set which has been already analyzed by Jorgensen in [14] and later by Markatou in [16] and permits to illustrate the performance of our method from the data analysis point of view. Example 3.5 exhibits the behaviour of the process involved in the obtention of the estimator paying special attention to the behaviour of the cells. It is also based on simulated data. The graphs are sufficiently eloquent so we will usually not provide the explicit values for the data or for the estimations.

We will compare our method with two approaches: with the classical MLE and with the estimation based on a mixture of multivariate t distributions. The MLE will be obtained with the EM algorithm employing the k-means as the initial solution.

The mixture of t distributions was proposed as a robust alternative for estimation in the mixture normal model (see Section 7.3 in [17]). It 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 we assume to be the same for every distribution in the mixture) and scale matrices. We carry out he 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.

Note that for large values of the degrees of freedom there are no practical differences between the multivariate t and the normal distributions. Therefore, the multivariate normal mixture could be considered as a particular case of the t mixture and often the estimations obtained from both models are very similar.

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). Since we use multiple initial solutions to compute the multivariate t mixture solutions, when provided, these solutions are considered as the best ones that the method can achieve, but we do not give the initial solutions from which they arose. 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 a light one ($u = 0.1$) in the examples in this section. Moreover, unless otherwise stated, we start with an initial trim of 50% of the data and the final region contains 95% of the points in the sample.

We want to remark 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.

Example 3.1 The first example is a variation of Example 2.1 obtained by changing the 20 contaminating data for another 20 points which constitute a well concentrated contamination arising from a uniform distribution on the square $[0.5, 1.5] \times [-8, -7]$.

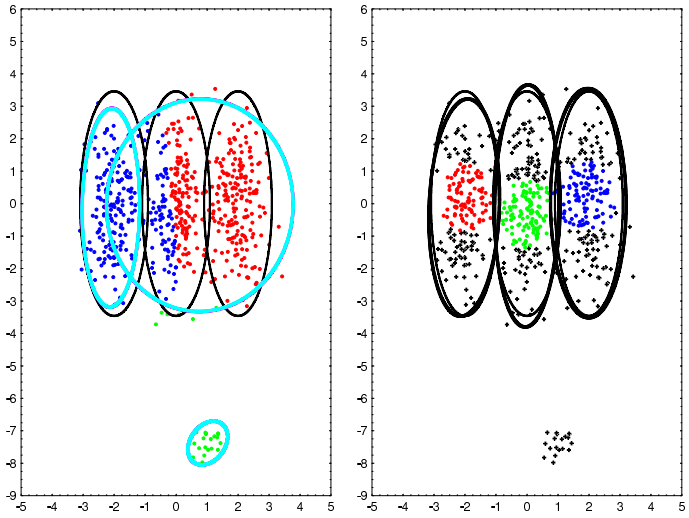

Figure 2: Plots of the 95% ellipses of the true distribution (thin ones) and the estimated distributions for Example 3.1.

Figure 2 shows the behavior of the different methods applied to the new contaminated data set. The graph on the right provides the solution obtained through our method, while the plot on the left side shows the behavior of the estimations provided by the EM algorithm for the normal mixture and t mixture models (which in fact almost coincide).

In this example, 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 MLE procedure which is unable to handle the problems arising from the presence of some concentration of outliers. •

Example 3.2 Here 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 $I\!R^8$ and a mixture of three bivariate normal distributions with parameters

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

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

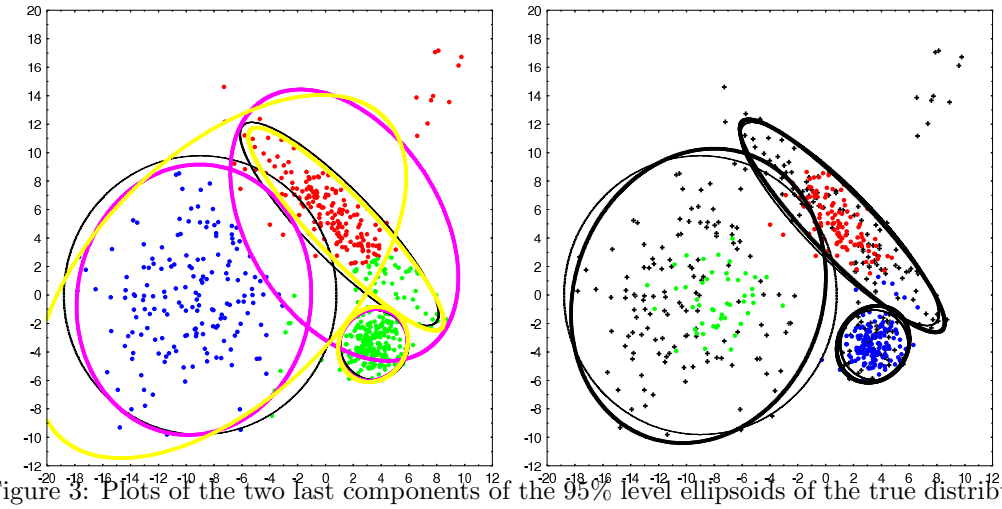

Figure 3: Plots of the two last components of the 95% level ellipsoids of the true distribution (thin ones) and the estimated distributions of Example 3.2.

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 3 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 starting from the 3-means as the initial solution (violet) and to the solution provided by the t mixture model (yellow). We should note that the violet solution is nearly equivalent to a local maximum found by the EMMIX algorithm for the t mixture model. Also note the bad behavior of both methods in this problem due to the curse of dimensionality. The graph on the right shows the solution obtained with our method. •

Example 3.3 This example uses the crab data set in Campbell and Mahon [2]. We analyze the subset 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. The analysis includes detailed comments on the influence of the homocedasticity hypothesis on the estimation, showing a better performance of the estimator without such restriction. In fact this constraint produces an unnecessary overlapping of the estimated distributions, so we do not consider it in our analysis.

Here we give a comparative solution for these data with a contamination, provided by three outliers included in the left upper corner in the plot of Figure 4 which even cause the t mixture model approach to break down. To ease the comparisons the graph shows the true sex group of the crabs through different marks.

The solutions provided by the EM algorithm starting from the 2-means as initial solution (violet) and the solution provided by the t mixture model (yellow) coincide. Our solution is obtained using the 0.7-

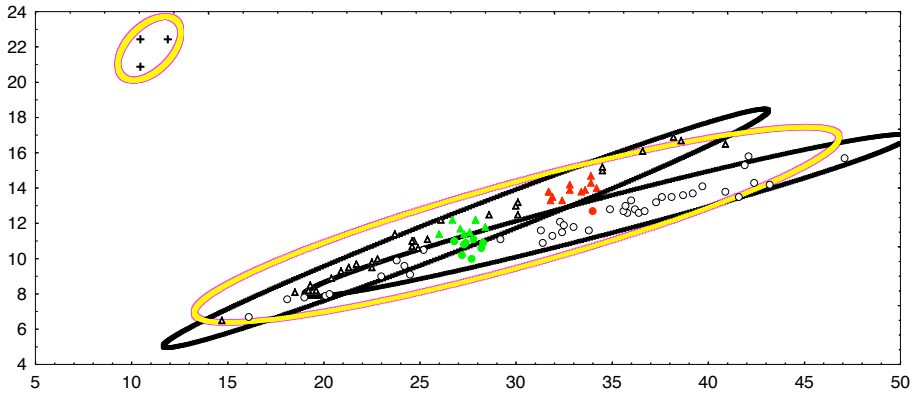

Figure 4: Plots showing the 95% ellipses for the estimated distributions and sex of the Blue Crab Data set of Example 3.3.

trimmed 2-means as the initial region (leading to the colored points), a separation threshold of $u = 0.1$, and a final trimming size of 5%. This time we have chosen a higher initial trimming level than usually because, if not, 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. •

Example 3.4 This example differs from the precedent ones by three main facts: It is based on a real data set, the data are univariate and the number of components in the mixture is unknown. Moreover in previous analysis it has been assumed as a simplification of the model that the subyacent populations have the same variance, a fact that enormously simplifies the problem by avoiding singular solutions. In consequence, under such assumption, the use of restrictions is not necessary to obtain the estimators. The data, represented by the bars in Figure 5, are constituted by the lenght of 222 scallops caught in an area of 79 m^2 in Mercury Bay, New Zealand, and present a clear component of smaller shellfish containing the bulk of the observations and a more confused and spread out tail of larger animals. An additional trouble for the analysis is the scarce precission of the measurements, leading to many repetitions.

This data set was first treated by Jorgensen, in [14], to introduce some diagnostic statistics that permit the analysis of the influence of data in the estimation of a finite mixture. Jorgensen, arguing from the fact that individual scallops have a diminishing rate of linear growht, considers unlikely a great heterogeneity within a cohort. Thus, if the data within the range 62-82 mm constitutes one component, the remaining data must constitute at least two. In the absence of other information, Jorgensen opted by fitting a mixture of three normal distributions with the same variance $\pi_1 N(\mu_1, \sigma^2) + \pi_2 N(\mu_2, \sigma^2) + (1 - \pi_1 - \pi_2) N(\mu_3, \sigma^2)$, noting the excesive influence of the value 126 mm as well as the very scarce representation of the third (in importance) distribution.

Later, in [16], Markatou introduced the weighted likelihood approach in the framework of the estimation of finite mixtures, and worked two examples of mixtures of normal distributions in the real line with same variance. The weighted likelihood approach is based on weighting the terms in the likelihood equation according with a measure of discrepancy of the data with respect to the model. This procedure

involves the use of a previous estimation based on kernel density estimation, the choice of an adequate discrepancy measure and the use of the method of moment estimates on multiple bootstrap subsamples to identify the important data substructures. The paper also provides a stopping rule to terminate the weighted likelihood algorithm. In the analysis, with two normal components in the mixture, the weighted likelihood method produced the values (0.804, 71.643, 95.287, 4.856) for $(\pi, \mu_1, \mu_2, \sigma)$ while the non-robust maximum likelihood method (through the EM algorithm) gave (0.799, 72.243, 100.163, 6.295). Moreover in two cases from the 100 initial bootstrap root searches, a root corresponding to the third component was identified. By fitting a mixture based on three normal distributions the maximum likelihood estimator gave the estimate (0.764, 0.163, 71.557, 92.830, 110.514, 4.540), while the weighted likelihood procedure gave the values (0.768, 0.156, 71.533, 92.265, 108.554, 4.38). Markatou also reported that, in agreement with Jorgensen, the only value which obtained a remarkable low weight in the weighted likelihood equation was that of the scallop of lenght 126 mm.

With respect to our method we must remark that the scarce presence of some cohortes in the data avoids the use of high initial trimming levels. This would completely eliminate the presence of some sub-populations in the sample. In fact the first analysis should be to answer what is the minimum size of a subsample to be considered as a subjacent population in the data or simply as contamination. Taking into account the observation of Jorgensen about the heterogeneity within a cohort in our opinion there are two possibilities for the analysis:

- To fit a mixture with three components and same standard deviation and a final small trimming size. In this way we want to elliminate only the troublesome data that could be not well explained by the mixture. This final trimming size permits to mantain the assumption on equal variances for the components that, in other case, could not be justified taking into account that the spread tail of larger animals could be constituted by three or more cohorts of scallops.

- To fit a mixture with two components and same standard deviation with a final moderate trimming size. This would reduce the contamination effects of a third (and may be more) component in the estimation, giving more precission to the estimates of the parameters of the two main components in the sample.

The graphics in Figure 5 show in a comparative way the estimated weighted densities corresponding to every component in the mixture, according to the estimation method employed and the number of components assumed.

For the estimation in the three components setup we used an initial trimming size of 2%, that produced an initial 3-cell given by the interval (61.9526, 113), thus the trimming data were the four data with greater lenght. The final trimming size was 0.5% and the only discarded value was 126, which agrees with its more influential character in Jorgensen and Markatou reports. Our final estimation was (0.7672, 0.1573, 71.4723, 92.3579, 108.7987, 4.4640).

In the two components setup we used 20% as the initial trimming size, leading to the 2-cell given by [65, 77.3221)$\cup$(85.0113, 97.3335). The final trimming size was 10%, which produced the final 2-cell: (62.1197, 80,5475)$\cup$(83.9271, 102.3549) and the estimation (0.8229, 71.3183, 93.1549, 5.2690).

As a remarkable (and distinguishible) fact we want to point out that under both models our method produced coherent estimations for the two main components in the mixture. This would be also coherent

with the observations of Jorgensen and could explain the excesive spread of the tail of larger animals by the presence of a third not well defined cohort, that could be also considered as a noise effect.

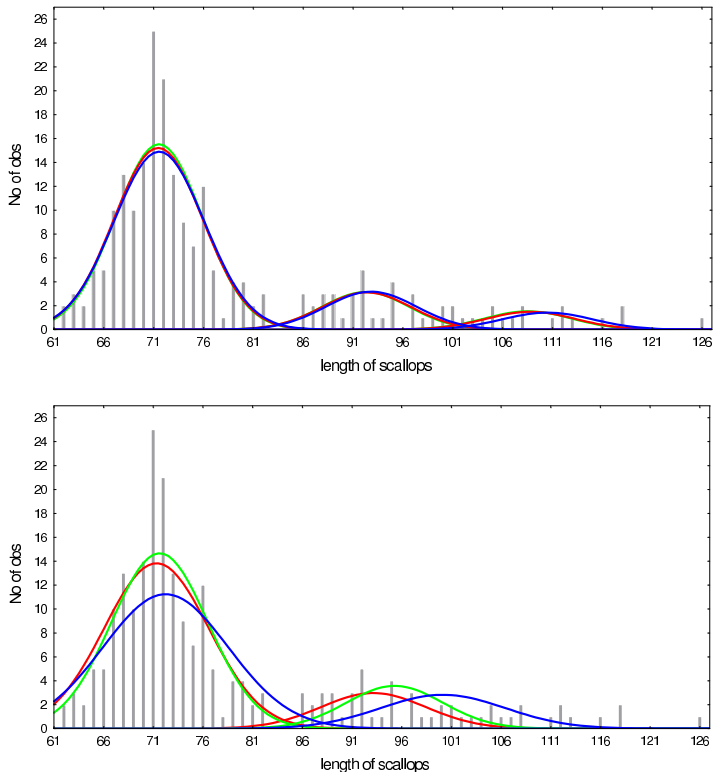

Figure 5: The bars represent the number of scallops with same lenght corresponding to the abscises axis. In each graph the curves of same colour give the estimated weighted densities for the distributions composing the mixture. To make more apparent the graphics the curves are scaled according to the total number of scallops. In the upper graph the estimations based on the assumption of three components are represented, while the lower graph show those based in two components. Blue curves correspond to the maximum likelihood estimators and green curves are those reported by Markatou. Red curves correspond to our estimations based on initial trimming sizes of 2% (upper) and 20% (lower) and final trimming sizes of 0.5% and 10% respectively.

Example 3.5 This example is included to exhibit the process of enlargement and adaptation of the cells and the corresponding estimations based on these (illustrated by their 95% level curves) through the successive iterations which compose the procedure. The model is a mixture of three bivariate normal distributions with parameters

$$\pi_i = 1/3, i = 1, 2, 3; \mu_1^T = (-8, 0), \mu_2^T = (-4, 10), \mu_3^T = (10, 0);$$

$$\Sigma_1 = \begin{pmatrix} 16 & 0 \\ 0 & 16 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 16 & -10 \\ -10 & 16 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 16 & 10 \\ 10 & 16 \end{pmatrix}.$$

The analysis is based on a sample of size 3000 contaminated by the additional inclusion of 90 data obtained from a uniform distribution on the set

$$\{(x, y) \in [-30, 30] \times [-30, 30] : x < -15 \text{ or } x > 15 \text{ or } y < -10 \text{ or } y > 10\}.$$

The graphics in Figure 6 show the evolution of the cells corresponding to trimming sizes of 80%, 50%, 25% and 5% (marked in red) as well as the corresponding estimated ellipses. As in the other examples the restricted parameter set remained fixed in the process. It is based on the 80%-trimmed 3-means with a separation threshold given by $u = 0.1$.

The final estimates are:

$$\hat{\pi}_1 = 0.3668, \hat{\pi}_2 = 0.2973, \hat{\pi}_3 = 0.3359,$$

$$\hat{\mu}_1^T = (-7.1778, 0.5706), \hat{\mu}_2^T = (-4.5442, 10.5898), \hat{\mu}_3^T = (10.0608, -0.0382);$$

$$\hat{\Sigma}_1 = \begin{pmatrix} 23.9378 & 5.9790 \\ 5.9790 & 23.1718 \end{pmatrix}, \hat{\Sigma}_2 = \begin{pmatrix} 18.7432 & -9.3451 \\ -9.3451 & 15.8172 \end{pmatrix}, \hat{\Sigma}_3 = \begin{pmatrix} 14.1699 & 7.5683 \\ 7.5683 & 17.6247 \end{pmatrix}.$$

4 Theoretical framework

The following theorem contains the property that constitutes the mathematical justification for the applicability of the proposed methodology. The proof is an immediate consequence of the characterizations of identifiability in Yakowitz and Spragins [23] and Proposition 6.1 in the Appendix. Although obvious, it can be appropriate mentioning that this result could be useless in the applications for specially unfortunate choices of A.

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 $\theta_1 = \theta_2$.*

This theorem leads to the validation of the classical argument that justifies the use of MLE in this framework, given in the next proposition. The classical proof, based on the use of Jensen's (strict) inequality, works here.

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

$$I\!P_{\theta_0} L_{\theta_0/A} > I\!P_{\theta_0} L_{\theta/A}, \tag{8}$$

for every $\theta \in \Theta$ such that $\theta \neq \theta_0$.

The restrictions considered is Section 2.2 are the sample version of the following general framework: Given $\gamma \in \Gamma$ and the threshold value $u \in (0, 1)$, take

$$\Theta_{\gamma,u} := \left\{ \theta \in \Theta : \frac{1}{I\!P[A_\gamma]} I\!P[I_{A_\gamma} I\!P_\theta(i/\cdot)] \geq u, \text{ for every } i = 1, ..., k \right\}, \tag{9}$$

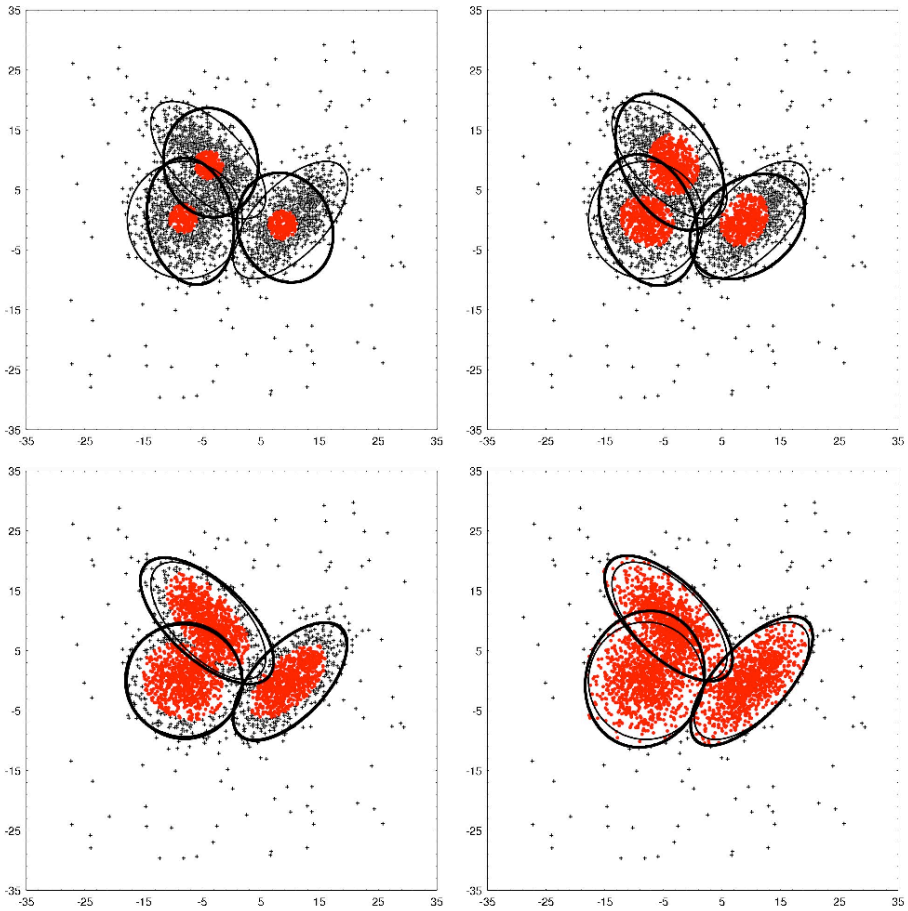

Figure 6: Simulation of 3000 points of a mixture of three 2-dimensional Gaussian distributions plus 90 contaminated observations generated as described in Example 3.3. Thin curves represent the 95% level ellipses of the true distribution. From left to right and from above to below, thick curves represent the evolution of the estimated 95% level ellipses by using the cell device when the trimming size takes the values 80%, 50%, 25% and 5%.

where $I\!P_\theta(i/x)$ was defined in (6).

Thus, $\Theta_{\gamma,u}$ is the family of parameters which give to every population an expected probability, conditioned by A_γ, greater or equal than u. We will assume that there exists $\theta(\gamma,u;I\!P)\in\Theta_{\gamma,u}$ such that $\theta(\gamma,u;I\!P):=\arg\max_{\theta\in\Theta_{\gamma,u}} I\!PL_{\theta/\gamma}$, where $L_{\theta/\gamma}$ denotes L_{θ/A_γ}. Of course, as soon as θ_0 belongs to $\Theta_{\gamma,u}$ we will have $\theta(\gamma,u;I\!P_{\theta_0})=\theta_0$.

If we replace $I\!P$ in (9) by the empirical distribution $I\!P_n$, we obtain the restricted set $\Theta^n_{\gamma,u}$ (which, for a general set $\hat{A}$ was defined in (5)) and the natural estimate of $\theta(\gamma,u;I\!P)$

$$\hat{\theta}_n(\gamma,u;I\!P_n):=\arg\max_{\theta\in\Theta^n_{\gamma,u}} I\!P_nL_{\theta/\gamma}. \tag{10}$$

Then, we construct the estimate, firstly, by selecting $\alpha\in(0,1)$, which leads to the parameters $\gamma_P\equiv\gamma_{_P}(\alpha)$ and $\gamma_{P_n}\equiv\gamma_{_{P_n}}(\alpha),\ n\in I\!N$, which in turn determine the restricted sets $\Theta_{\gamma_P,u}$ and $\Theta^n_{\gamma_{P_n},u}$ in which we maximize the censored likelihood function. In order to keep the notation as simple as possible, we will also write θ_P instead of $\theta(\gamma_{_P}(\alpha),u;I\!P)$ and γ_n and $\hat{\theta}_n$ instead of $\gamma_{_{P_n}}(\alpha)$ and $\hat{\theta}_n(\gamma_n(\alpha),u;I\!P_n)$. We will also often employ the notation $\theta_P=(\pi_1^P,...,\pi_k^P,\phi_1^P,...,\phi_k^P)$ and $\hat{\theta}_n=(\pi_1^n,...,\pi_k^n,\phi_1^n,...,\phi_k^n)$.

From now on we will assume that the value u which determines the restricted parametric set is fixed.

Remark 4.3 The impartial restrictions contribute towards allowing the existence of the estimator and assuring convergence of the EM algorithm to stationary points of the likelihood function. This is a consequence of Theorem 2 in [22], taking into account that in this setup the likelihood corresponding to the complete data belongs to the curved exponential family and that the restrictions make the sets $\left\{\theta\in\Theta^n_{\gamma,u}:\ I\!P_nL_{\theta/\gamma}\geq I\!P_nL_{\eta/\gamma}\right\}$ compact for every $\eta\in\Theta^n_{\gamma,u}$.

For the first statement note that in the model of complete data we assume every value x_i to be known as well as the vector $z_i=(z_{i1},...z_{ik})$ that explains the subpopulation j from which such x_i arises (i.e. $z_{ij}=1$ or 0 respectively means that x_i arises or not from the j-th distribution in the mixture). Therefore, the corresponding likelihood function is

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

Concerning the convergence of the EM algorithm, 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 A_γ used to determine the sample-based restrictions are A_{γ_n}, which contain at least $[\alpha\cdot n]$ points.

From here, it is possible, by slightly modifying the proof of Proposition 4.4 below, to ensure that from a fixed n, depending on α,d and u, the set $\Theta^n_{\gamma,u}$ allows us to define the estimator through (10) and even more:

Let $0<\alpha,u<1$, $\gamma\in\Gamma$ and $I\!P_n$ be the sample distribution based on a sample $X_1,...,X_n$ of an absolutely continuous distribution. Let us denote $\phi_i^m=(\mu_i^m,\Sigma_i^m)$ and let λ_i^m be the smallest eigenvalue of Σ_i^m. Assume that $n>2(d+1)/(u(1-\alpha))$, that $I\!P_n(A_\gamma)\geq 1-\alpha$, and that, for a sequence $\theta_m^*=$

$(\pi_1^m, ..., \pi_k^m, \phi_1^m, ..., \phi_k^m) \in \Theta_{\gamma,u}^n$, $m \in N$ *there exists* $i \in \{1, ..., k\}$ *such that one of the following conditions is satisfied*

1. $\lim_m \lambda_i^m = 0$.
2. $\lim_m \|\phi_i^m\| = \infty$ *and* $\liminf_m \lambda_j^m > 0,\ j = 1, ..., d$.
3. $\lim_m \pi_i^m = 0$ *and* $\liminf_m \lambda_j^m > 0,\ j = 1, ..., d$.

Then $\lim_m I\!P_n L_{\theta_m^*/\gamma} = -\infty$ *a.s. holds.*

Thus if, for some $\eta \in \Theta_{\gamma,u}^n$, the set $S_\eta = \left\{\theta \in \Theta_{\gamma,u}^n : I\!P_n L_{\theta/\gamma} \geq I\!P_n L_{\eta/\gamma}\right\}$ is not compact then there would exist a sequence $\{\theta_m\} \subset S_\eta$ without accumulation points in S_η. But, $L_{\theta/\gamma}$ being continuous in θ, S_η should be a closed subset of $\Theta_{\gamma,u}$, so the sequence should verify any of the conditions (a), (b) or (c), leading to $I\!P_n L_{\eta/\gamma} \leq \lim_m I\!P_n L_{\theta_m^*/\gamma} = -\infty$. •

4.1 Asymptotics

To improve the readability we will not discuss the problem of uniqueness of the trimmed k-means but we will suppose throughout that every theoretical probability has a unique trimmed k-mean (however, see Remark 4.6).

The proof of the consistency of our procedure will be based on an usual compactness argument stated in Proposition 4.4. We want to emphasize the interest of this proposition for providing arguments such as those in Remark 4.3 on the robustness of the estimator.

Proposition 4.4 *Let* $\theta_n^* = (\delta_1^n, ..., \delta_k^n, \psi_1^n, ..., \psi_k^n) \in \Theta_{\gamma_n,u}^n$, $n \in I\!N$, *where* $\psi_i^n = (\mu_i^n, \Sigma_i^n)$. *Let us denote by* λ_i^n *the smallest eigenvalue of* Σ_i^n. *Let us assume that there exist* $i \in \{1, ..., k\}$ *and a subsequence* $\{j_n\}_n$ *which satisfy one of the following conditions*

1. $\lim_n \lambda_i^{j_n} = 0$.
2. $\lim_n \|\psi_i^{j_n}\| = \infty$ *and* $\liminf_n \lambda_j^{j_n} > 0,\ j = 1, ..., d$.
3. $\lim_n \delta_i^{j_n} = 0$ *and* $\liminf_n \lambda_j^{j_n} > 0,\ j = 1, ..., d$.

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

Theorem 4.5 (Consistency) *Let* $\{X_n\}$ *be a random sample of an absolutely continuous distribution* $I\!P$. *If* θ_P *is unique and there exists* $\delta > 0$ *such that* $\theta_P \in \Theta_{\gamma_P,u+\delta}$, *then*

$$\lim_n \hat{\theta}_n = \theta_P \ \text{a.s.} \tag{11}$$

Remark 4.6 Theorem 4.5 holds without the uniqueness of the trimmed k-means assumption, because, without this hypothesis, we should change the statement of Proposition 6.2 to say that there exists a ν-probability one set Ω_0 such that if $\omega \in \Omega_0$, then every subsequence of trimmed k-means of $\{I\!P_n\}$ contains a further subsequence which converges to a trimmed k-mean of $I\!P$.

But, if we keep the uniqueness of θ_P and we apply the argument in Theorem 4.5 to each of the subsequences for which the trimmed k-means converge, we would have that, if $\omega \in \Omega_0$, then every subsequence of $\{\hat{\theta}_n\}$ contains a further subsequence which converges to θ_P, and, in consequence, the whole sequence converges to θ_P. •

The reasoning leading to the consistency of the procedure is only based on the (a.s.) weak convergence of the sample probability measures to $I\!P$. Therefore, the same proof works to prove its continuity with respect to the weak convergence:

Corollary 4.7 (Qualitative Robustness) *Let $I\!P = I\!P_{\theta_0}$ for some $\theta_0 \in \Theta$ and assume that $\theta_0 \in \Theta_{\gamma_P, u+\delta}$, for some $\delta > 0$. Let $\{Q_n\}$ be a sequence of probability measures that converges in distribution to $I\!P$. 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 Waart and Wellner [21]. We will take advantage of the parametric nature of the trimmed sets under consideration. To this end, let $\tilde{\Gamma} := \left(I\!R^d\right)^k \times \left(\mathcal{M}^+_{d\times d}\right)^k \times \left(I\!R^+\right)^k$ indexing the sets constituted by the union of k ellipsoids. For $\gamma = (m_1, ..., m_k, \Sigma_1, ..., \Sigma_k, r_1, ..., r_k) \in \tilde{\Gamma}$, let $A_\gamma := \bigcup_{i=1}^k \{x \in I\!R^d : (x - m_i)^T \Sigma_i^{-1} (x - m_i) \leq r_i\}$.

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

$$\mathcal{G}_\Lambda := \left\{ m_{\theta,\gamma} := I_{A_\gamma} \log(f_\theta) + I_{A_\gamma^c} \log\left(I\!P_\theta\left(A_\gamma^c\right)\right), (\theta, \gamma) \in \Lambda \right\}, \tag{12}$$

and their derivatives with respect to θ:

$$h_{\theta,\gamma} := I_{A_\gamma} \left(\frac{\partial}{\partial \theta} \log(f_\theta) \right) + I_{A_\gamma^c} \left(\frac{\partial}{\partial \theta} \log\left(I\!P_\theta\left(A_\gamma^c\right)\right) \right)$$

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

As noted in [4] the extension of the argmax arguments of the Empirical Processes Theory to this semiparametric model is an easy fact through the extensions of the results of Section 3.2.4 in [21] 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 [21] as we will prove in Lemma 6.5.

Theorem 4.8 (Asymptotic distribution) *Let $I\!P = I\!P_{\theta_0}$, for some $\theta_0 \in \Theta$, and $\gamma_0 \in \tilde{\Gamma}$. If $\theta_0 \in \Theta_{\gamma_0, u+\delta}$ for some $\delta > 0$ and $\{\gamma_n\}_n$ is a sequence (possibly random) in $\tilde{\Gamma}$ such that $\gamma_n \to \gamma_0$ a.s. then the sequence $\left\{\hat{\theta}_n(\gamma_n)\right\}_n$ of estimators based on the sets A_{γ_n} verifies*

$$\sqrt{n}\left(\hat{\theta}_n(\gamma_n) - \theta_0\right) \to_w N\left(0, \left(\left.\frac{\partial}{\partial \theta}\right|_{\theta=\theta_0} I\!P_{\theta_0} h_{\theta,\gamma_0} \right)^{-1}\right).$$

The asymptotic covariance matrix can also be expressed as

$$\left(I\!P_{\theta_0}\left(\left(h_{\theta_0,\gamma_0}\right)\left(h_{\theta_0,\gamma_0}\right)^T\right)\right)^{-1}.$$

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 $\hat{\theta}_n(\gamma_n)$ to θ_0 is $n^{1/2}$ and does not depend on the rate of convergence of γ_n to γ_0.*

On the other hand, our proof of consistency can easily be modified to cover the m-step estimator, obtained through the iteration of the procedure a fixed number of times, $m > 1$, as described in Subsection 2.2.1. Once we have the consistency for the first step we automatically have the consistency of the trimming sets involved in the second 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 γ_0 being 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

Besides the qualitative robustness given in Corollary 4.7, the theoretical study of the robustness properties of a method is usually carried out through the so-called quantitative approach. The influence function (IF) and the breakdown point (BP) are the central concepts of Hampel's infinitesimal 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, [12].

The IF of the trimmed k-means method was obtained in [7], including a graphical analysis showing its behavior for some variants of a mixture of normal univariate distributions. We resort to a similar explanation that permits conclusions to be obtained from the visualization of the involved graphics.

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

$$IF(x, \hat{\theta}(\gamma), \theta_0) = -\left(I\!P_{\theta_0} \left(\left. \frac{\partial}{\partial \theta} \right|_{\theta=\theta_0} h_{\theta,\gamma} \right) \right)^{-1} h_{\theta_0,\gamma}(x). \tag{13}$$

Because of the continuity of the estimator with respect to γ, if we apply the idea in the proof of Theorem B.1 in [7] to the points that do not belong to the boundary of A_γ, it is easy to see that the IF for the estimator $\hat{\theta}_n(\gamma_n)$ coincides with that of $\hat{\theta}_n(\gamma)$, if $\{\gamma_n\}_n \subset \tilde{\Gamma}$ and $\gamma_n \to \gamma \in \tilde{\Gamma}$. Therefore, the IF for the one step estimator based on the α-trimmed k-means will be the one given by (13) with A_γ being the union of the k balls associated to the α-trimmed k-means of $I\!P_{\theta_0}$. On the other hand, for the m-step estimator, $m > 1$, the IF will be also (13) with A_γ 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. However, it becomes very useful in non-symmetrical situations. This can be analyzed through the expressions in (14) and is made apparent in the graphs in Figure 7. This figure shows, in the lower row, an asymmetric case in which the mixture is $\frac{1}{4}(N(-3, 1.5) + N(0, 1.5) + 2N(3, 1.5))$. The left-hand side graph shows the IF when the k-means are used and the right-hand side one when employing the ellipsoids. In the upper row in Figure 7, we analyze the symmetric mixture $\frac{1}{3}(N(-5, 1) + N(0, 1) + N(5, 1))$. Since in this case there is no difference between both regions, to ease the understanding of the figure, we show on the left-hand side the IF for the means and on the right-hand

side 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.

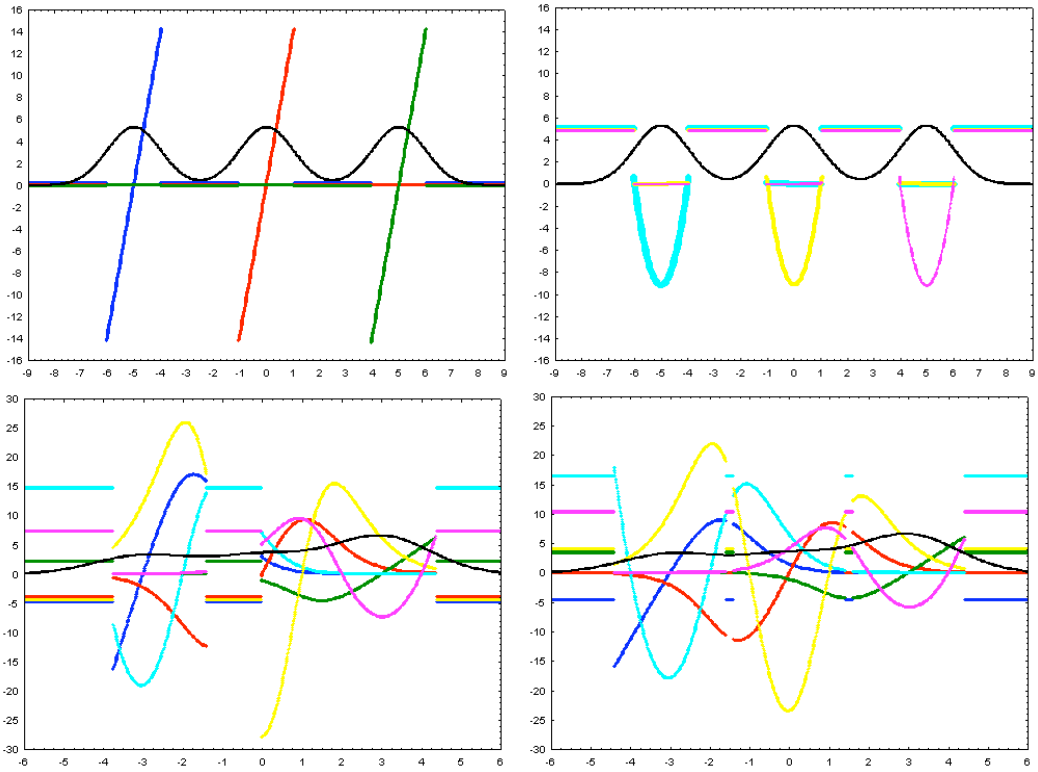

Figure 7: IF's for the means (blue, green and red) and the variances (cyan, yellow and magenta) of the distributions making up a mixture of three normal distributions. The upper graphs correspond to the mixture $\frac{1}{3}(N(-5,1)+N(0,1)+N(5,1))$. The graph on the lower left (resp. lower right) presents the IF for the one-step (resp. m-step) estimator for the mixture $\frac{1}{4}(N(-3,1.5)+N(0,1.5)+2N(3,1.5))$. 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 $h_{\theta,\gamma}(x)$ as a function of $\theta = (\pi_1,...,\pi_{k-1},\mu_1,...,\mu_k,\Sigma_1,...\Sigma_k)$

$$
\begin{aligned}
\frac{\partial}{\partial \pi_i} L_{\theta/A}(x) &= \left(\frac{I\!P_\theta(i/x)}{\pi_i} - \frac{I\!P_\theta(k/x)}{\pi_k}\right) I_A(x) + I\!P_\theta\left[\frac{I\!P_\theta(i/x)}{\pi_i} - \frac{I\!P_\theta(k/x)}{\pi_k} \Big/ A^c\right] I_{A^c}(x) \\
\frac{\partial}{\partial \mu_i} L_{\theta/A}(x) &= \Sigma_i^{-1}(x-\mu_i) I\!P_\theta(i/x)\ I_A(x) + I\!P_\theta\left[\Sigma_i^{-1}(x-\mu_i) I\!P_\theta(i/x)/A^c\right] I_{A^c}(x), \\
\frac{\partial}{\partial \Sigma_i} L_{\theta/A}(x) &= \frac{1}{2}\left(\ \Sigma_i^{-1}(x-\mu_i)(x-\mu_i)^T\Sigma_i^{-1} - \Sigma_i^{-1}\right) I\!P_\theta(i/x)\ I_A(x) \\
&\quad + \frac{1}{2} I\!P_\theta\left[\left(\Sigma_i^{-1}(x-\mu_i)(x-\mu_i)^T\Sigma_i^{-1} - \Sigma_i^{-1}\right) I\!P_\theta(i/x)/A^c\right]\ I_{A^c}(x).
\end{aligned}
\tag{14}
$$

The study of the BP of the method is not as simple as that of the IF. Indeed, we do agree with García-Escudero and Gordaliza [7] and Hennig [12] on the peculiarities of the BP in this setting. Pathological constellations of data may break down even the trimmed k-means procedure through the substitution of only one point by another. But more favorable configurations may have a BP equal to the trimming level. Therefore, the BP of a procedure in this framework must be considered as data dependent. In any case, we must stress the fact that, after the arguments in Remark 4.3, the impartial restrictions link the estimations to the procedure used to obtain the initial clusterized region. Thus, the BP of our, one or m-step, estimators for the location parameters is 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-clusterized data sets (see [7]).

Alternatively the BP may be analyzed under a general assumption of well-clusterized data in an idealized situation which permits comparisons between procedures under controlled assumptions in a kind of 'in vitro' analysis. Hennig, in Section 4 in [12], introduces a such ideal model and analyzes several estimators for mixtures using his addition r-components BP which is defined as follows:

If l is the minimum number of points to be added to the sample to break down r parameters in the estimation, then, the *addition r-components BP* is $l/(n+l)$.

Let us assume that $\mathcal{X}_m = \{x_{1,m}, x_{2,m}, ..., x_{n,m}\}, m \in I\!N$ is a sequence of data sets clusterized in $k \geq 2$ groups $A_m^i,\ i = 1, ..., k;\ m \in I\!N$:

$$A_m^1 = \{x_{1,m}, ..., x_{n_1,m}\}, A_m^2 = \{x_{(n_1+1),m}, ..., x_{n_2,m}\}, ..., A_m^k = \{x_{(n_{k-1}+1),m}, ..., x_{n_k,m}\}.$$

Following the ideas in Section 4.1 in [12] we consider this sequence $\mathcal{X}_m$ as an ideal array of well k-clusterized data sets whenever there exists $b < \infty$ such that for every $m \in I\!N$,

$$\max_{1 \leq i \leq k} \max\{\|x_{jm} - x_{lm}\| : x_{jm}, x_{lm} \in A_m^i\} \quad < \quad b \ \text{ and} \tag{15}$$

$$\lim_{m \to \infty} \min\{\|x_{jm} - x_{lm}\| : x_{jm} \in A_m^h,\ x_{lm} \in A_m^i;\ i \neq h\} \quad = \quad \infty. \tag{16}$$

Under this idealized model the addition of r outliers must be analyzed under the assumption of a sequence $\mathcal{Y}_m = \{y_{1.m}, ...y_{r.m}\}$ added to $\mathcal{X}_m$ to constitute the new data sets $\mathcal{X}_m \cup \mathcal{Y}_m$ verifying also that

$$\lim_{m \to \infty} \min\{\|y_{jm} - x_{lm}\| :\ y_{jm} \in \mathcal{Y}_m,\ x_{lm} \in \mathcal{X}_m\} \quad = \quad \infty, \text{ and} \tag{17}$$

$$\lim_{m \to \infty} \min\{\|y_{jm} - y_{lm}\| :\ y_{jm}, y_{lm} \in \mathcal{Y}_m, j \neq l\} \quad = \quad \infty. \tag{18}$$

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. In particular for estimators related to location (including here the k-means) breakdown holds if for every rearrangement of the components (if more than one) of the estimator it holds $\|E_n(\mathcal{X}_m) - E_n(\mathcal{X}_m \cup \mathcal{Y}_m)\| \to \infty$ as $m \to \infty$. However, for the estimator of the weights, components breakdown would happen if the minimum weight estimation under $\mathcal{X}_m$ converges to zero while under $\mathcal{X}_m \cup \mathcal{Y}_m$ 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 $\mathcal{X}_m \cup \mathcal{Y}_m$ 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 [12]) that $r < k$ added outliers break down the estimation of r parameters through the ML estimation, as well as through robustified versions like the t-procedure of McLachlan and Peel or the Fraley and Raftery proposal (also considered in [17]). In particular, the addition of only 1 outlier breaks down the estimation of at least one parameter.

Note that this idealized model guarantees, in Hennig's words, that *"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. Thus the link constituted by the impartial restrictions (5) and an argument similar to that arising from Lemmas 4.1 and 4.2 in [12] (Proposition 4.4 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, at least whenever the number of points of every cluster A_m^i is greater or equal than $\lceil \alpha n \rceil + d + 1$ and they are in *general position*. This means that every affine hyperplane $H \subset \mathbb{R}^d$ contains, at most, d points of A_m^i and prevents against the degeneracy of some distribution in the mixture into a lower dimension, which could happen if the data points contained in any of the balls associated to the trimmed k-means can live in a space of 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 $\mathcal{X}_m = \{x_{1,m}, x_{2,m}, ...x_{n,m}\}, m \in \mathbb{N}$ be an ideal array of data sets in $\mathbb{R}^d$ well clusterized in $k \geq 2$ groups A_m^i, $i = 1, ..., k$; $m \in \mathbb{N}$, verifying (15) and (16), such that the points in every group A_m^i are in general position and their numbers verify $n_i - n_{i-1} \geq \lceil \alpha n \rceil + d + 1$, $i = 1, ..., k$ ($n_0 = 0$). If $r \leq \lceil \alpha n \rceil$, then the m-step estimator of the parameter $\theta \in \Theta$, corresponding to 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 (17) and (18).*

5 Discussion

It is well known that there is a strong connection between the mixture and the clustering modellings. This connection is often used to obtain a cluster configuration from an estimation of the parameters in a mixture. Here we exploit this connection just in the opposite way. Our estimation objective is understood as the improvement of a clustering process to estimate the parameters of every group as well as in their respective weights in the mixture. This point of view allows to take advantage of robust clustering methods to produce robust estimators in the MNMM estimation setup.

We assume the knowledge of the number of populations in the mixture. Although in some situations this assumption can hinder the model, it is realistic for a very large amount of 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,...).

The introduced procedure is based on making the estimation from a highly representative subset of the data. The choice of such a set is adaptive and 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 sample.

The proposed method shows a good performance not only under symmetrical contamination but also under concentration of outliers or in the presence of bridge points which often cause other proposals to break down. The estimators obtained are asymptotically gaussian and qualitatively robust. Moreover, the analysis of the BP of the procedure under Hennig's idealized model shows that it greatly improves those of the available procedures. 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.

Mention should be made of the relevance of a good choice of the active data set in the initial step as well as a controlled enlargement of the active set in the successive steps. Initial active data set can be successfully 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. Through a high trimming level this robust method can give a good representation of each distribution in the mixture by providing a noise-free data set focused on the uncontaminated cores of the k clusters of the data. When employing a high trimming level, this procedure takes advantage of the fact that if the distributions in the mixture are well separated and the trimmed mixture is composed by k clusters, it can serve to prevent masking effects that hide the nature of the mixture. We recall that in the mixture model context, noise effect on any distribution composing the mixture can arise not only from contaminating data due to an external stream but also from the contiguous distributions living in the mixture. This choice is computationally feasible and can be modulated through the initial trimming level to obtain, in well clusterized data sets, our goal. However 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 unappropriated 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. Therefore we consider that the methodology can be included in the toolbox of applied statisticians as an alternative to other available methods.

6 Appendix

The following proposition leads to a simple proof of the identifiability of our model.

Proposition 6.1 *Let $\mathcal{Y}$ be the set of density functions of non-singular multivariate normal distributions on $\mathbb{R}^d$. Let $A \subset \mathbb{R}^d$ 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 $\mathbb{R}^d$, thus $(\mu_1, \Sigma_1) = (\mu_2, \Sigma_2)$, and both distributions are the same. •

For the proofs of the results on consistency we note that from the Glivenko-Cantelli theorem, the sequence $\{\mathbb{P}_n\}_n$ (a.s.) converges in distribution to the probability measure $\mathbb{P}$. In fact (from Skorohod's Representation Theorem for the weak convergence) we will assume that $\{\mathbb{P}, \mathbb{P}_1, ...\}$ are the distributions of some random vectors $\{Y_0, Y_1, ...\}$ such that $Y_n \to Y_0$ ν-a.s.

In the following proposition we enumerate some basic properties of the trimmed k-means which are taken (or are easily deduced) from [7].

Proposition 6.2 *If* $\mathbb{P}$ *is absolutely continuous, then the sequence of trimmed* k*-means and associated trimmed regions of* $\mathbb{P}_n$ *verify:*

1. $\lim_n \|\gamma_n - \gamma_P\| = 0$.
2. $\lim_n I_{A_{\gamma_n}}(Y_n) = I_{A_{\gamma_P}}(Y_0)$, ν*-a.s.*
3. $\lim_n \mathbb{P}_n \left[A_{\gamma_n} \right] = \mathbb{P} \left[A_{\gamma_P} \right] = 1 - \alpha$.
4. $\lim_n \mathbb{P}_n \left[I_{A_{\gamma_n}} \log f_{\theta_P} \right] = \mathbb{P} \left[I_{A_{\gamma_P}} \log f_{\theta_P} \right]$.
5. $\lim_n \mathbb{P}_n L_{\theta_P/\gamma_n} = \mathbb{P} L_{\theta_P/\gamma_P}$.

PROOF OF PROPOSITION 4.4.- Given $\phi \in \Phi$, we denote $M(\phi) := \sup\{g_\phi(x) : x \in \mathbb{R}^d\}$. Let us assume first, that 1 holds. Let $\{j_n'\}_n$ be a subsequence of $\{j_n\}$. There exists a subsequence $\{j_n^*\}_n$ of $\{j_n'\}_n$ and a non-empty set $I \subset \{1, ..., k\}$ such that

$$\begin{aligned} &\text{if } i \in I, \quad \text{then} \quad \lim_n \lambda_i^{j_n} = 0 \\ &\text{if } i \notin I, \quad \text{then} \quad \liminf_n \lambda_i^{j_n} > 0. \end{aligned}$$

Moreover, we can also assume that there exists $i_0 \in I$ such that

$$M\left(\psi_{i_0}^{j_n^*}\right) = \sup\left\{ M\left(\psi_i^{j_n^*}\right) : i \in I \right\}, \ n \in \mathbb{N}.$$

Obviously,

$$K_1 := \sup_{i \notin I} \sup_n M\left(\psi_i^{j_n^*}\right) < \infty. \tag{19}$$

Without loss of generality, we can assume that $i_0 = 1$ and, to avoid complications in the notation, we will denote $j_n^* = n,\ n \in I\!N$.

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 . Let

$$r_n := \inf\left\{r > 0 : I\!P_n[H_r/A_{\gamma_n}] > u/2\right\},$$

where $I\!P_n[A/B]$ denotes the conditional $I\!P_n$-probability of the set A given B. From the continuity of $I\!P$, we have that $\lim_n I\!P_n[H_{r_n}/A_{\gamma_n}] = u/2$ and, then, $\liminf_n r_n > 0$ because, otherwise, it would be $\liminf_n I\!P_n[H_{r_n}/A_{\gamma_n}] = 0$. Let

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

We have that,

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

and, as a consequence, $\liminf_n I\!P_n[C_n/A_{\gamma_n}] \geq u/4$. From here and 3 in Proposition 6.2,

$$\liminf_n I\!P_n[C_n] \geq u(1-\alpha)/4 > 0. \tag{20}$$

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

$$\frac{u}{4} \leq I\!P_{\theta_n^*}(1/x) \leq \frac{\delta_1^n g_{\psi_1^n}(x)}{\delta_i^n g_{\psi_i^n}(x)}.$$

Therefore, if $x \in C_n$,

$$\sup_{i=1,...,k} \delta_i^n g_{\psi_i^n}(x) \leq \frac{4}{u} g_{\psi_1^n}(x) \leq \frac{4}{u}\beta_1^n, \tag{21}$$

where $\beta_1^n = \sup_{x\notin H_{r_n}} g_{\psi_1^n}(x)$. From here and (19), from an index onward, we have that

$$\begin{aligned} I\!P_n L_{\theta_n^*/\gamma_n} &\leq I\!P_n\left[I_{A_{\gamma_n}\cap C_n^c} \log f_{\theta_n^*}\right] + I\!P_n\left[I_{C_n} \log f_{\theta_n^*}\right] \\ &\leq I\!P_n[A_{\gamma_n} \cap C_n^c] \log\left[\sup(K_1, M(\psi_1^n))\right] + I\!P_n[C_n] \log\left[k4\beta_1^n/u\right] \\ &\leq \log(k4/u) + \log^+(K_1) + \log\left[(\beta_1^n)^{I\!P_n[C_n]} M(\psi_1^n)\right], \end{aligned}$$

which converges to $-\infty$ because of (20), $M(\phi^n) \leq (2\pi\lambda^n)^{-d/2}$ and

$$\beta_1^n = \frac{1}{(2\pi\lambda^n)^{d/2}} \exp\left(-\frac{r_n^2}{2}(\lambda^n)^{-1}\right).$$

Thus, we have shown that every subsequence of $\{j_n\}_n$ admits a new subsequence (which at the beginning we called $\{j_n^*\}_n$) such that

$$\lim_n I\!P_{j_n^*} L_{\theta_{j_n^*}^*/\gamma_{j_n^*}} = -\infty. \tag{22}$$

But this property is only possible if the subsequence $\{j_n\}_n$ satisfies (22).

Now, let us suppose that 2 or 3 hold. Let $\{j_n^{'}\}_n$ be a subsequence of $\{j_n\}_n$. Obviously, there exists a subsequence $\{j_n^*\}_n$ of $\{j_n^{'}\}_n$ which satisfies that for every $i \in \{1, ..., k\}$,

$$\liminf_n \lambda_i^{j_n^*} > 0 \quad \text{and} \quad \lim_n \|\psi_i^{j_n^*}\| = \infty \quad \text{or} \quad \lim_n \psi_i^{j_n^*} = \psi_i \in \Phi,$$

and there also exists i_0 such that $\lim_n \|\psi_{i_0}^{j_n^*}\| = \infty$, or $\lim_n \delta_{i_0}^{j_n^*} = 0$. Without loss of generality we can assume that $i_0 = 1$. As before, we will denote $j_n^* = n$, for $n \in I\!N$. Let

$$D_n := \left\{x \in A_{\gamma_n} : I\!P_{\theta_n^*}[1/x] > u/2\right\}.$$

Then $I\!P_n[D_n/A_{\gamma_n}] > u/2$, and arguing as in (20) and (21), we have that $u(1-\alpha)/2 \leq \liminf_n I\!P_n(D_n)$ and, if $x \in D_n$, that

$$f_{\theta_n^*}(x) \leq k2\delta_1^n g_{\psi_1^n}(x)/u. \tag{23}$$

On the other hand, in this case, we have that $K_2 := \sup_n \sup_i M(\psi_i^n) < \infty$. From here and (23) we have that

$$\begin{aligned} I\!P_n L_{\theta_n^*/\gamma_n} &\leq I\!P_n\left[I_{A_{\gamma_n}\cap D_n^c} \log f_{\theta_n^*}\right] + I\!P_n\left[I_{D_n} \log f_{\theta_n^*}\right] \\ &\leq \log^+(K_2) + \log(k2/u) + I\!P_n\left[I_{D_n} \log(\delta_1^n g_{\psi_1^n})\right], \end{aligned}$$

which converges to $-\infty$ for the subsequence we are considering, and the proof ends as in the previous case. •

Lemma 6.3 *If $I\!P$ is absolutely continuous and there exists $\delta > 0$ such that $\theta_P \in \Theta_{\gamma_P, u+\delta}$, then there exists $N_0 \in I\!N$ such that if $n \geq N_0$, then $\theta_P \in \Theta^n_{\gamma_n, u}$.*

PROOF.- From the continuity of the map $x :\to I\!P_{\theta_P}(i/x)$ and 2 in Proposition 6.2, we have that

$$I\!P_{\theta_P}(i/Y_n) I_{A_{\gamma_n}}(Y_n) \to \text{a.s. } I\!P_{\theta_P}(i/Y_0) I_{A_{\gamma_P}}(Y_0). \tag{24}$$

Now, taking into account that $I\!P_{\theta_P}(i/\cdot) \in [0, 1]$, we obtain that

$$\begin{aligned} I\!P_n\left[I\!P_{\theta_P}(i/\cdot) I_{A_{\gamma_n}}\right] &= \nu\left[I\!P_{\theta_P}(i/Y_n) I_{A_{\gamma_n}}(Y_n)\right] \\ &\to \nu\left[I\!P_{\theta_P}(i/Y_0) I_{A_{\gamma_P}}(Y_0)\right] \\ &= I\!P\left[I\!P_{\theta_P}(i/\cdot) I_{A_{\gamma_P}}\right] \geq (u+\delta) I\!P\left[A_{\gamma_P}\right], \end{aligned}$$

and the proof ends by applying 3 in Proposition 6.2. •

Corollary 6.4 follows from Lemma 6.3 and 5 in Proposition 6.2, taking into account that in Proposition 4.4 we can take the vectors θ_n^* as close as desired to the optimum parameters.

Corollary 6.4 *If $I\!P$ is absolutely continuous and there exists $\delta > 0$ such that $\theta_P \in \Theta_{\gamma_P, u+\delta}$, then, from an index onward, the sequence $\{\hat{\theta}_n\}_n$ belongs to a compact set contained in Θ.*

Now, we are ready to prove the consistency of the procedure.

PROOF OF THEOREM 4.5.- By Lemma 6.3 and the definition of $\hat{\theta}_n$, if $n \geq N_0$, we have

$$I\!P_n \left[L_{\hat{\theta}_n/\gamma_n} \right] \geq I\!P_n \left[L_{\theta_P/\gamma_n} \right] \to I\!P \left[L_{\theta_P/\gamma_P} \right], \tag{25}$$

where the last convergence was stated in 5 in Proposition 6.2.

Let us assume that (11) does not hold. By Corollary 6.4 the sequence $\{\hat{\theta}_n\}_n$ contains a subsequence such that

$$\lim_n \hat{\theta}_{j_n} = \theta^* = (\pi_1^*, ..., \pi_k^*, \phi_1^*, ..., \phi_k^*) \neq \theta_P,$$

and $\theta^* \in \Theta$. From 2 and 3 in Proposition 6.2, we have that

$$L_{\hat{\theta}_{j_n}/\gamma_{j_n}}(Y_{j_n}) \to \text{a.s. } L_{\theta^*/\gamma_P}(Y_0). \tag{26}$$

However,

$$L_{\hat{\theta}_{j_n}/\gamma_{j_n}}(Y_{j_n}) \leq I_{A_{\gamma_{j_n}}}(Y_{j_n}) \log f_{\hat{\theta}_{j_n}}(Y_{j_n}),$$

which is a bounded function, and we can apply Fubini's Theorem to conclude that

$$\limsup_n I\!P_{j_n} L_{\hat{\theta}_{j_n}/\gamma_{j_n}} = \limsup_n \nu \left[L_{\hat{\theta}_{j_n}/\gamma_{j_n}}(Y_{j_n}) \right] \leq \nu \left[L_{\theta^*/\gamma_P}(Y_0) \right] = I\!P L_{\theta^*/\gamma_P}. \tag{27}$$

Then, if we could prove that $\theta^* \in \Theta_{\gamma_P, u}$, we would have a contradiction between (27), (25) and the uniqueness of θ_P. To prove this, first notice that

$$I\!P_{\hat{\theta}_{j_n}}(i/Y_{j_n}) I_{A_{\gamma_{j_n}}}(Y_{j_n}) \to \text{a.s. } I\!P_{\theta^*}(i/Y_0) I_{A_{\gamma_P}}(Y_0). \tag{28}$$

As the functions $I\!P_{\theta_{j_n}}(i/\cdot)$ are bounded by 1, we can apply 3 in Proposition 6.2 to have

$$\begin{aligned} uI\!P \left[A_{\gamma_P} \right] &= \lim_n u I\!P_{j_n} \left[A_{\gamma_{j_n}} \right] \leq \limsup_n I\!P_{j_n} \left[I\!P_{\theta_{j_n}}(i/\cdot) I_{A_{\gamma_{j_n}}} \right] \\ &= \limsup \nu \left[I\!P_{\hat{\theta}_{j_n}}(i/Y_{j_n}) I_{A_{\gamma_{j_n}}}(Y_{j_n}) \right] \\ &\leq \nu \left[I\!P_{\theta^*}(i/Y_0) I_{A_{\gamma_P}}(Y_0) \right] = I\!P \left[I\!P_{\theta^*}(i/\cdot) I_{A_{\gamma_P}} \right], \end{aligned}$$

so the proof is complete because Corollary 6.4 implies that $\pi_i^* > 0$ for every $i = 1, ..., k$. •

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

Now the proof parallels that given in [4] based on extended versions of the results in Section 3.2.4 in [21] to this semiparametric framework. In our case, for $m_{\theta,\gamma}$ defined as in (12) the components of $\dot{m}_{\theta,\gamma} := h_{\theta,\gamma}$ are those given in (14) with A_γ as A. The result is then the consequence of Lemma 6.5 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 γ_0 such that*

$$\left\{ \frac{m_{\theta\gamma} - m_{\theta_0\gamma} - (\theta - \theta_0)^T \dot{m}_{\theta_0\gamma}}{\|\theta - \theta_0\|} : \|\theta - \theta_0\| \leq \delta, \gamma \in K \right\} \tag{29}$$

is $I\!P$-Donsker and

$$I\!P \left(m_{\theta\gamma} - m_{\theta_0\gamma} - (\theta - \theta_0)^T \dot{m}_{\theta_0\gamma} \right)^2 = o\left(\|\theta - \theta_0\| \right)^2, \tag{30}$$

uniformly in $\gamma \in K$.

PROOF.- Let δ small enough to assure that the parameters in $V_\delta := \{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta\}$ do not lead to degeneration of the mixture, and let K be any compact neighborhood of γ_0. If we choose a compact ball, B_0, in $I\!R^p$ containing all the ellipsoids A_γ, $\gamma \in K$, the continuity of $m_{\theta,\gamma}$ and $\dot{m}_{\theta,\gamma}$ with respect to the argument and with respect to the parameters guarantee that the functions in the family (29) are uniformly bounded by a constant over the set B_0. This implies the uniform L_2-Frechet derivability (30).

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 parameter in V_δ verifies the uniform entropy condition (see Section 2.5.1 in [21]).

The class of functions given by

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

defines a linear space of finite dimension, thus it is a VC-class of functions (see Lemma 2.6.15 in [21]). The density functions of normal distributions are obtained by composing a function in the class $\mathcal{I}$ with the exponential function, $\exp(\mathcal{I})$, hence it is also a VC-class of functions (see Lemma 2.6.18 in [21]). 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 [21], a such class verifies the uniform entropy condition.

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

The class of functions $\mathcal{M}_\delta$ verifies the condition, so we can apply Theorem 2.10.20 in [21] to assure that the transformed class $\log(\mathcal{M}_\delta) I_{B_0}$ also verifies that condition. We only need to show that there exists a constant, A, such that

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

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 verify the condition of uniform entropy

- $\inf\left\{ I\!P_\theta\left(A^c_\gamma\right) : \ \theta \in \Theta_\delta \text{ and } \gamma \in K \right\} > 0.$

- The family $\left\{ I_{A_\gamma} \log(f_\theta) + I_{A^c_\gamma} \log\left(I\!P_\theta\left(A^c_\gamma\right) \right) : \ (\theta, \gamma) \in \Theta_\delta \times K \right\}$ verifies the uniform entropy condition.

Theorem 2.10.20 in [21] 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 $I_{A_\gamma}(x)\frac{\partial}{\partial\theta}\log\left(f_\theta(x)\right)+I_{A_\gamma^c}(x)\frac{\partial}{\partial\theta}\log\left(I\!P_\theta\left(A_\gamma^c\right)\right)$ where $\theta\in\Theta_\delta$ and $\gamma\in K$ 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 $\{I_{B_0}(x)\frac{\partial}{\partial\theta}\log\left(f_\theta(x)\right):\theta\in\Theta_\delta\}$ is a Donsker class of functions. But this follows from the fact that the components of these functions are products of $I\!P_\theta(i/x)I_{B_0}$ with functions of the types $\frac{1}{\pi_i},\Sigma_i^{-1}(x-\mu_i)$ and $-\frac{1}{2}\Sigma_i^{-1}+\frac{1}{2}\Sigma_i^{-1}(x-\mu_i)(x-\mu_i)'\Sigma_i^{-1}$. •

References

[1] Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49 803821.

[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. (2006). Trimming and likelihood: Robust location and dispersion estimation in the multivariate model. *Submitted*

[5] Fraley, C. and Raftery, A. E . (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer J.* **41**: 578588.

[6] Gallegos, M.T. (2003). Robust clustering under general normal assumptions. Available at: www.fmi.uni-passau.de/forschung/mip-berichte/MIP-0103.ps.

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

[8] García-Escudero, L.A. and Gordaliza, A. (2006). The importance of the scales in heterogeneous robust clustering. To appear in *Comput. Statist. Data Anal.*

[9] Hampel, F. (2002). Some Thoughts about Classification. In *Classification, Clustering and Data Analysis*, eds. Jajuga, K., Sokolowski, A. and Bock, H.H. Springer, New York.

[10] Hardin, J, and Rocke, D.M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Comput. Statist. and Data Anal.*, **44**: 625-638.

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

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

[13] Hunter, D. R.; Wang, S. and Hettmansperger , T. P. (2006) Inference for mixtures of symmetric distributions. *To appear in Ann. Statist.*

[14] Jorgensen, M. A. (1990). Influence-based diagnostics for finite mixture models. *Biometrics* **46**: 1047-1058.

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

[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] Rousseeuw, P.J.(1985). Multivariate estimation with high Breakdown Point, in *Mathematical Statistics and Applications, Volume B*, eds. W. Grossmann, G. Pflug, I. Vincze and W. Werz, Reidel, Dordrecht.

[20] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

[21] Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak convergence and empirical processes. With applications to statistics.* Springer-Verlag. New York.

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

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