EXPLORING THE NUMBER OF GROUPS IN ROBUST MODEL-BASED CLUSTERING *

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

Abstract

Two key questions in Clustering problems are how to determine the number of groups properly and measure the strength of group-assignments. These questions are specially involved when the presence of certain fraction of outlying data is also expected.

Any answer to these two key questions should depend on the assumed probabilistic-model, the allowed group scatters and what we understand by noise. With this in mind, some exploratory "trimming-based" tools are presented in this work together with their justifications. The monitoring of optimal values reached when solving a robust clustering criteria and the use of some "discriminant" factors are the basis for these exploratory tools.

Keywords: Heterogeneous clusters, number of groups, strength of group-assignments, trimming.

1 Introduction

Two key questions in Clustering problems are how to choose the number of groups properly and measure the strength of group-assignments. These two questions are specially involved when we also expect the presence of noise or outliers in our data set, as occurs when robust

*Research partially supported by the Spanish Ministerio de Ciencia e Innovación, grant MTM2008-06067-C02-01, and 02 and by Consejería de Educación y Cultura de la Junta de Castilla y León, GR150.

†Departamento de Estadística e Investigación Operativa. Facultad de Ciencias. Universidad de Valladolid. 47002, Valladolid. Spain.

clustering methods are applied. We think that it is difficult to provide sensible answers to these two important questions without stating formal assumptions concerning the following issues:

a) *State clearly which one is the probabilistic model assumed.* Without making that decision, previous questions are not well-posed ones. This formal statement avoids the view of Cluster Analysis as "model-free" or merely "heuristically" motivated techniques (see, e.g., Flury 1997). A review of adequate probabilistic models to be applied in Clustering can be found in Bock (1996).

b) *Put constraints on the allowed group structures.* For instance, we should clearly specify in advance if we assume spherical, elliptical or other different shapes. Moreover, we can force similar group scatters or we can allow for very different ones. Banfield and Raftery (1993) and Celeux and Govaert (1995) presented several ways of forcing these constraints on the group structures.

c) *State clearly what we understand by noise or outliers.* A clear distinction should be made about what a "proper group" means compared to what could be merely a (small) fraction of outlying data. This establishes a subtle link between Clustering and Robust Estimation methods (see, e.g., Hawkins and Olive 2002, Rocke and Woodruff 2002, Hennig and Chrislieb 2002, García-Escudero et al. 2003, Hardin and Rocke 2004 or Woodruff and Reiners 2004).

Several "mixture modeling" and "crisp clustering" approaches to model-based Clustering can be found in the literature. Mixture modeling approaches assume that data at hand $x_1, ..., x_n$ in $\mathbb{R}^p$ come from a probability distribution with density $\sum_{j=1}^{k} \pi_j \phi(\cdot; \theta_j)$ with $\phi(\cdot; \theta_j)$ being p-variate densities with parameters θ_j for $j = 1, ..., k$. This leads to likelihoods of the form

$$\prod_{i=1}^{n} \left[\sum_{j=1}^{k} \pi_j \phi(x_i; \theta_j) \right]. \tag{1.1}$$

On the other hand, "crisp" (0-1) clustering approaches assume classification likelihoods

$$\prod_{j=1}^{k} \prod_{i \in R_j} \phi(x_i; \theta_j), \tag{1.2}$$

with R_j containing the indexes of the observations that are assigned to group j.

In this work, we adopt a crisp clustering approach and therefore we try only to provide answers to the initial questions within this framework. Moreover, we put special emphasis in robustness aspects due to the (aforementioned) connections between Clustering and Robust Estimation as well as the real experience of practitioners that reveals that the presence of

outlying data in Clustering is the rule rather than the exception. We also assume that $\phi(\cdot, \theta_j)$ stands for the multivariate normal distribution with parameters $\theta_j = (\mu_j, \Sigma_j)$ where μ_j is the population mean and Σ_j is the variance-covariance matrix. This is our choice of underlying probability model in a) which, indeed, it is a frequent choice in many Clustering procedures.

A general framework for trying to simultaneously handle issues a), b) and c) in the crisp clustering framework was introduced in Gallegos (2002) and Gallegos and Ritter (2005). The so-called "spurious-outliers model" assumes the presence of a fraction α of the data generated by an extraneous mechanism that may be trimmed off or discarded. Within this framework, the TCLUST methodology presented in García-Escudero et al. (2008) is able to handle different types of constraints for the group scatter matrices which allows for addressing point b) through a restriction on the group scatter matrix eigenvalues. A different way of controlling the group scatters has been recently introduced in Gallegos and Ritter (2009). Alternatively, constraining the minimum group size is proposed in Gallegos and Ritter (2010) also within this "spurious-outliers model" framework.

The TCLUST has a theoretically well-founded background and a feasible algorithm for its implementation was given in García-Escudero et al. (2008) once parameters k (number of groups) and α (trimming proportion) are fixed in advance. However, procedures for choosing k and α when applying TCLUST have not been proposed yet. We will see in this work how suitable choices for k and α can be obtained throughout the careful examination of some Classification Trimmed Likelihoods curves which are based on "trimmed" modifications of (1.2). The presented methodology can be surely adapted to other types of group covariance constraints within the "spurious-outliers model" framework.

The choice of k in clustering problems has a long history in Statistics (see, e.g., Milligan and Cooper 1988). When trimming is allowed, some recent proposals for choosing k and α are based on modified BIC notions (Neykov et al 2007 and Gallegos and Ritter 2009 and 2010). Gallegos and Ritter's proposals also include the consideration of normality goodness-of-fit tests and outlier identification tools (see, e.g., Becker and Gather 1999 and references therein).

The choice of k in the mixture modeling has also received a lot of attention in the literature (e.g., Wolfe 1970, Titterington et al. 1985, McLachlan 1987, Fraley and Raftery 1998 and Keribin 2000). Robust mixture modeling approaches try to fit noisy data through MLE with additional components for the noise (Fraley and Raftery 1998 and Dasgupta and Raftery 1998) or fitting mixtures of t distributions (McLachlan and Peel 2000). Unfortunately, it is easy to see that these approaches do not always work correctly when noisy data depart notably from the assumptions (Hennig 2004a). The here presented "trimming" approach differs from "robust mixture modeling" ones in that noisy data are completely avoided and no attempt to fit them is tried. A "trimming" approach to mixture modeling can be found in Neykov et al (2004 and 2007) and in Cuesta-Albertos et al. (2008). Take also into account

that "mixture modeling" and "crisp clustering" approaches pursue different goals and so answers to our initial questions may be completely different (see, Biernacki, Celeux and Govaert 2000). There are other proposals for choosing k resorting to validity measures being functionals of the group partitions and quantifying concepts like clusters "separation" or "compactness". These ideas will not be explored here.

The outline of the work is as follows. Section 2 briefly reviews the TCLUST methodology while the Classification Trimmed Likelihood curves are introduced in Section 3. Section 4 provides graphical displays based on "discriminant" factors which provide confirmatory graphical tools to see the appropriateness of the final choice of k and α. Some simulated and real data examples are presented in Section 5. The Appendix includes some theoretical results justifying the use of the presented tools.

An `R` package called `tclust` for implementing the TCLUST and the proposed graphical tools is available at the CRAN (`http://cran.r-project.org`) repository.

2 The TCLUST methodology

Starting from the "classification likelihood" in (1.2), Gallegos (2002) and Gallegos and Ritter (2005) proposed the "spurious-outlier" model by considering a "likelihood" like:

$$\prod_{j=1}^{k} \prod_{i \in R_j} \phi(x_i; \theta_j) \prod_{i \in R_0} g_i(x_i), \tag{2.1}$$

with $\{R_0, R_1, ..., R_k\}$ being a partition of indexes $\{1, ..., n\}$ in $\mathcal{R}_\alpha$ defined as

$$\mathcal{R}_\alpha = \big\{ \{R_0, R_1, ..., R_k\} : \cup_{j=0}^{k} R_j = \{1, ..., n\}, R_r \cap R_s = \emptyset \text{ for } r \neq s \text{ and } \#R_0 = [n\alpha] \big\}.$$

The densities $\phi(\cdot; \theta_j)$ are p-variate normal densities with parameters $\theta_j = (\mu_j, \Sigma_j)$ and the $g_i(\cdot)$'s are assumed to be probability density functions in $\mathbb{R}^p$ satisfying

$$\arg\max_{\mathcal{R}_\alpha} \max_{\{\theta_j\}_{j=1}^{k}} \prod_{j=1}^{k} \prod_{i \in R_j} \phi(x_i; \theta_j) \subseteq \arg\max_{\mathcal{R}_\alpha} \prod_{i \notin R_1 \cup ... \cup R_k} g_i(x_i).$$

This last assumption is reasonable whenever a fraction α of non regular data points (with indexes in R_0) may be just considered as noisy observations. Precise examples for these $g_i(\cdot)$'s are given in Gallegos and Ritter' papers. We can see there that the maximization of (2.1) simplifies into the maximization of

$$\prod_{j=1}^{k} \prod_{i \in R_j} \phi(x_i; \theta_j).$$

Additionally, the TCLUST method introduced in García-Escudero et al. (2008) assumes the presence of some unknown weights $\pi_j \in [0,1]$ for the regular part of data proposing the maximization of

$$\prod_{j=1}^{k} \prod_{i \in R_j} \pi_j \phi(x_i; \theta_j),$$

with $\{R_0, R_1, ..., R_k\}$ ranging in $\mathcal{R}_\alpha$.

Notice that this likelihood is not a "mixture likelihood" and these weights are intended to take into account the different sizes of the groups when making the final group assignments.

The TCLUST method also considers a group scatters similarity constraint in terms of the eigenvalues of the group covariance matrices. If $\lambda_l(\Sigma_j)$'s are the eigenvalues of the group covariance matrix Σ_j and

$$M_n = \max_{j=1,...,k} \max_{l=1,...,p} \lambda_l(\Sigma_j) \text{ and } m_n = \min_{j=1,...,k} \min_{l=1,...,p} \lambda_l(\Sigma_j)$$

are the maximum and minimum eigenvalues, the constraint $M_n/m_n \leq c$ for a constant $c \geq 1$ is imposed. Closely related constraints were first applied in the mixture modeling framework by Hathaway (1985).

Hence, the TCLUST problem is defined throughout the maximization:

$$\mathcal{L}_c^{\Pi}(\alpha, k) := \max_{\{R_j\}_{j=0}^k \in \mathcal{R}_\alpha, \{\theta_j\}_{j=1}^k \in \Theta_c, \{\pi_j\}_{j=1}^k \in [0,1]^k} \sum_{j=1}^{k} \sum_{i \in R_j} \log(\pi_j \phi(x_i; \theta_j)), \tag{2.2}$$

with $\Theta_c = \{(\theta_1, ..., \theta_c)$ such that θ_j's satisfy the constraint $M_n/m_n \leq c\}$. The values where the maximum of (2.2) is attained are denoted by $\{\widehat{R_0}, \widehat{R_1}, ..., \widehat{R_k}\}$, $(\widehat{\theta_1}, ..., \widehat{\theta_k})$ and $(\widehat{\pi_1}, ..., \widehat{\pi_k})$.

An "unweighed" problem can be also defined throughout the maximization:

$$\mathcal{L}_c(\alpha, k) := \max_{\{R_j\}_{j=0}^k \in \mathcal{R}_\alpha, \{\theta_j\}_{j=1}^k \in \Theta_c} \sum_{j=1}^{k} \sum_{i \in R_j} \log \phi(x_i; \theta_j). \tag{2.3}$$

These problems are closely connected with some trimmed likelihood proposals in Neykov et al (2004 and 2007). However, instead of using explicit constraints on the group scatter matrices to avoid undesired singularities, no explicit model-based rules for choosing/eliminating spurious solutions are provided in those works (apart from some sensible protections in the algorithms aimed at "maximizing" the trimmed likelihood).

The maximization of (2.2) or (2.3) has obviously a very high computational complexity due to the "combinatorial" nature of the problem. A feasible algorithm aimed at approximately solving (2.2) was given in García-Escudero et al. (2008). This algorithm may be easily adapted to solve (2.3) just assuming equal weights $\pi_j = 1/k$. These algorithms belong to the family of Classification EM algorithms (Celeux and Govaert 1992) but, in order to perform the data-driven trimming, some "concentration" steps (as those behind the fast-MCD

algorithm in Rousseeuw and van Driessen 1999) are also applied. A quadratic programming problem is solved to force the solutions to satisfy the eigenvalue ratio constraints.

Solving problems like those in (2.2) and (2.3) in the untrimmed case with different group scatter constraints has a long history in Clustering. For instance, classical k-means (McQueen 1967) are just the solution of problem (2.3) with $\alpha = 0$ and assuming $\Sigma_1 = ... = \Sigma_k = \sigma^2 I$. The determinant criterium (Friedman and Rubin 1967) is based on the assumption $\Sigma_1 = ... = \Sigma_k = \Sigma$. Other constraints can be found in Banfield and Raftery (1998) and Celeux and Govaert (1992). The consideration of weights π_j's in classification likelihoods like in (2.2) goes back to Symons (1981). This criterium also appears in Bryant (1991) under the name of "penalized classification likelihood".

Many procedures for choosing k in Clustering are based on monitoring the size of "likelihoods" like (2.3) depending on k. For instance, see Engelman and Hartigan (1974) or Calinski and Harabasz (1974) when assuming $\Sigma_1 = ... = \Sigma_k = \sigma^2 I$ and when $\alpha = 0$ or Marriot (1991) when $\Sigma_1 = ... = \Sigma_k = \Sigma$. However, there is a clear drawback in directly monitoring the size of (2.3) since it *strictly* increases when k is increased. A formal proof of this claim is given in Proposition 2 in the Appendix. This fact could lead us to overestimate k if no remedy is adopted. Therefore, searching for an "elbow" in summarizing graphs or considering non-linear transformations (Sugar and James 2003) are proposed in the literature.

Since weights $\pi_j = 0$ are possible, (2.2) does not necessarily increase strictly when increasing k. This fact was already noticed in Bryant (1991) in the untrimmed case $\alpha = 0$. He also mentioned the possible merit of it in order to provide helpful guidance for choosing the number of groups in Clustering. See also the discussion of the CL2 method in Biernacki and Govaert (1997).

As we commented in the Introduction, the choice of k should depend on the assumptions made for issues b) and c). As a simple example, let us consider a data set with $n = 1000$ observations simulated through a bivariate three-component Gaussian mixture with mixing parameters $\pi_1 = .36$, $\pi_2 = .54$ and $\pi_3 = .1$. The assumed means are $\mu_1 = (0, 0)'$, $\mu_2 = (5, 5)'$ and $\mu_3 = (2.5, 2.5)'$, and the covariance matrices are

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}, \text{ and, } \Sigma_3 = \begin{pmatrix} 50 & 0 \\ 0 & 50 \end{pmatrix}.$$

The result of applying the TCLUST to this data set appears in Figure 1,(a) when $k = 3$, $\alpha = 0$ and a large value for the group scatters constraint constant $c = 50$ are chosen. This value of c allows for finding a "main" group containing all the most scattered data points. Alternatively, if we are not willing to accept these scattered observations as being a proper group, a more sensible cluster solution can be found in Figure 1,(b). There, we can see $k = 2$ groups for a $c = 5$ value and a trimming proportion $\alpha = .1$ (that now serves to trim all the

most scattered data points). This example shows that the proper determination of α and k are two closely related tasks.

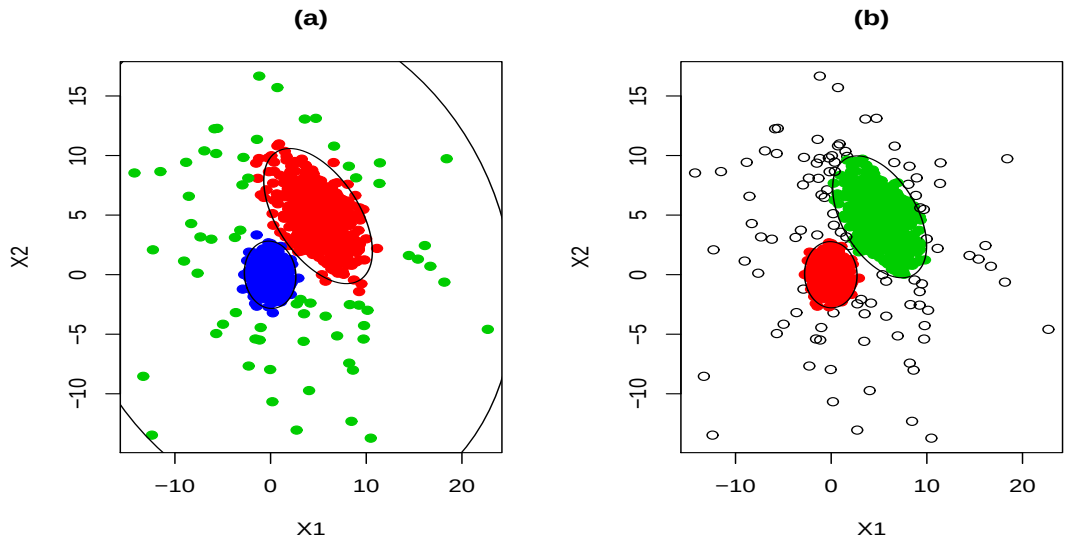

Figure 1: TCLUST results for a simulated data set: $k = 3$, $\alpha = 0$ and $c = 50$ are fixed in (a) and $k = 2$, $\alpha = .1$ and $c = 5$ in (b). Symbols "$\circ$" stand for the trimmed observations and shown ellipsoids are 97.5% tolerance based on the estimated group means and scatters.

3 Classification Trimmed Likelihood Curves

In these section, we introduce some "classification trimmed likelihood curves" as useful tools for choosing the number of groups k in Clustering. The k-th trimmed likelihood curve is defined through the function:

$$\alpha \mapsto \mathcal{L}_c^{\Pi}(\alpha, k) \text{ for } \alpha \in [0, 1),$$

with $\mathcal{L}_c^{\Pi}(\alpha, k)$ as defined in (2.2). This definition explicitly depends on the constant c stating the allowed differences between group scatter matrices eigenvalues.

These curves are used to measure $\Delta_c^{\Pi}(\alpha, k) = \mathcal{L}_c^{\Pi}(\alpha, k+1) - \mathcal{L}_c^{\Pi}(\alpha, k)$, which quantifies the "gain" achieved by allowing increasing the number of groups from k to $k+1$ for a given trimming level α.

Notice that we might have also considered $\Delta_c(\alpha, k) = \mathcal{L}_c(\alpha, k+1) - \mathcal{L}_c(\alpha, k)$, but $\Delta_c(\alpha, k)$ would always be strictly greater than 0. This property is proved in Proposition 2 when excluding some (non-interesting) pathological cases which happen when the data set is concentrated on k points after trimming a proportion α.

On the other hand, we can see that $\Delta_c^{\Pi}(\alpha, k)$ may be equal to 0 for values of k greater or equal than the "appropriate" number of groups. For instance, Figure 2 is based on a

simulated data set obtained as in Figure 1 but now taking $\pi_1 = .4$, $\pi_2 = .6$ and $\pi_3 = 0$. Figure 2,(a) shows the clustering result associated to the maximization of (2.3) when $k = 3$, $c = 5$ and $\alpha = 0$. We can see in Figure 2,(b) the values of $\mathcal{L}_5(0, k)$ for $k = 1, 2, 3$. By examining this graph, we might be tempted to think that increasing k from 2 to 3 is needed. Figures 2,(c) shows the clustering results obtained when maximizing (2.2) for the same values of k, c and α. Notice that one of the estimated group weights $\widehat{\pi}_j$ is now set to zero (with arbitrary values for $\widehat{\mu}_j$ and $\widehat{\Sigma}_j$ satisfying the group scatters constraint). This fact leads to $\Delta_5^{\Pi}(0, 3) = 0$, as we can see in the plot of values of $\mathcal{L}_5^{\Pi}(0, k)$ for $k = 1, 2, 3$ that appear in Figure 2,(d).

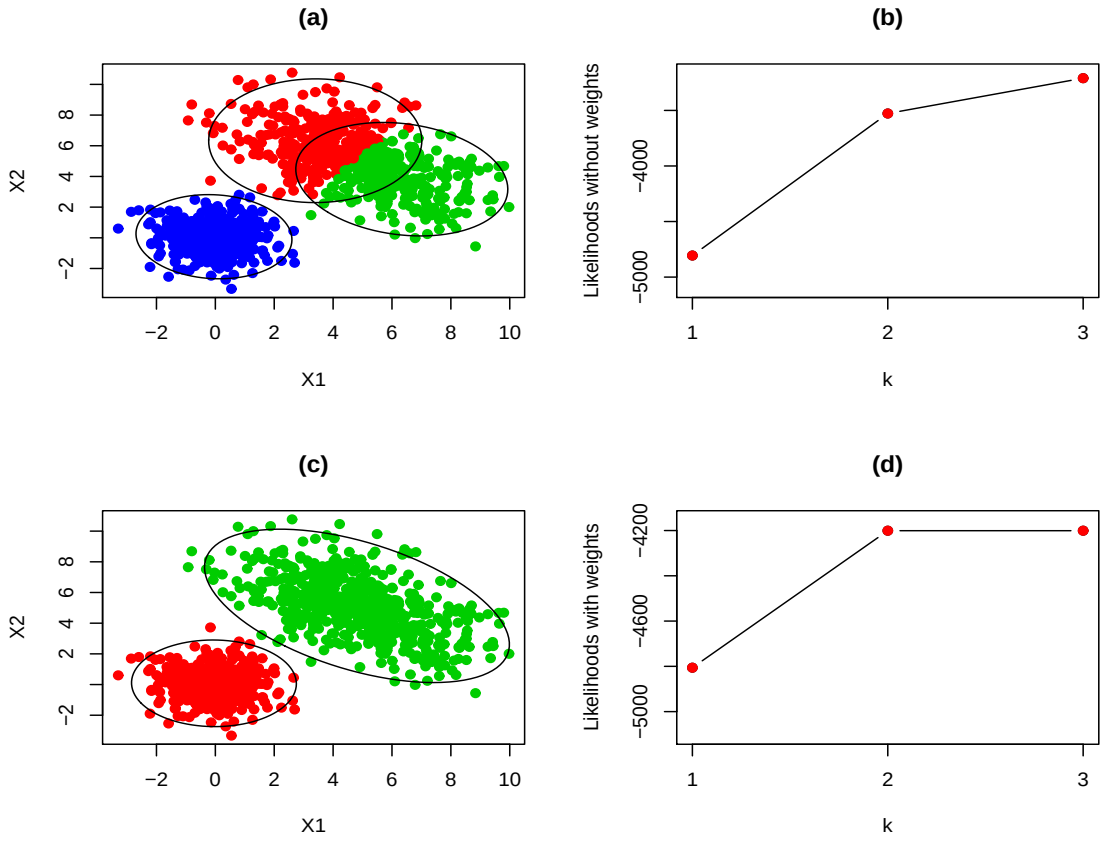

Figure 2: Clustering results when solving the problem (2.3) when $k = 3$, $c = 5$ and $\alpha = 0$ is shown in (a). The results when solving (2.2) appear in (c) for the same values of k, c and α. Values of $\mathcal{L}_5(0, k)$ for $k = 1, 2, 3$ are shown in (b) and those of $\mathcal{L}_5^{\Pi}(0, k)$ in (d).

By using these classification trimmed likelihood curves, we recommend choosing the number of groups as the smallest value of k such that $\Delta_c^{\Pi}(\alpha, k)$ is always (close to) 0 except for small values of α. Once k is fixed, the first trimming size α_0 such that $\Delta_c^{\Pi}(\alpha, k)$ is (close to) 0 for every $\alpha \geq \alpha_0$ is a good choice for the trimming level.

Figure 3 shows the classification trimmed likelihood curves $\mathcal{L}_5^{\Pi}(\alpha, k)$ when $k = 1, 2, 3, 4$ and α ranges in $[0, .3]$ and $c = 50$. We can see that no significant improvement happens

when increasing k from 3 to 4 and $\alpha = 0$. Moreover, no significant improvement is detected when k increases from 2 to 3 once we discard the proportion $\alpha_0 = .1$ of the most scattered data points. This figure so suggests $k = 3$ and $\alpha = 0$ or $k = 2$ and $\alpha = .1$ as two possible sensible choices for k and α for this data set when $c = 50$.

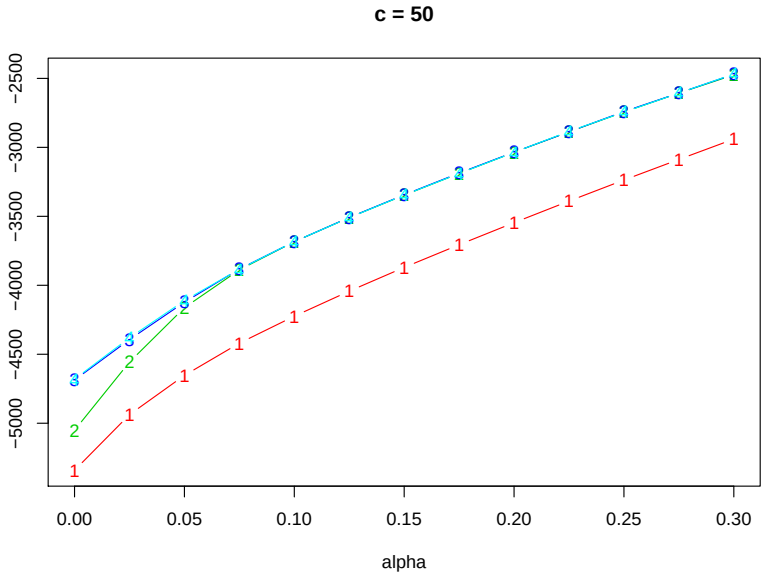

Figure 3: Classification trimmed likelihood curves $\mathcal{L}_5^{\Pi}(\alpha, k)$ when $k = 1, 2, 3, 4$, α ranges in $[0, .3]$ and $c = 50$ for the data set in Figure 1.

The presented approach just provides some graphical "exploratory" tools. Although these curves are very informative, it would also be interesting to develop formal statistical tools in order to numerically quantify when $\Delta_c^{\Pi}(\alpha, k)$ is (close to) 0. This is an ongoing work.

As the TCLUST is a computationally demanding procedure, the Classification Trimmed Likelihood curves are only evaluated over a grid of α values. Moreover, as it was pointed out in García-Escudero et al. (2003), a precise resolution is indeed only needed when $\alpha \in [0, 1/(k+1)]$ for the k-th curve. Notice that García-Escudero et al. (2003) had already suggested the interest of monitoring (trimmed k-means based) likelihoods like (2.2) when choosing k and α in a more constrained case.

4 Strength of Cluster Assignments

For a given TCLUST clustering solution, we now introduce some "confirmatory" graphical tools that will help us to evaluate the quality of the cluster assignments and the strength of the trimming decisions.

Let us considerer an optimal solution $\widehat{R} = \{\widehat{R_0}, \widehat{R_1}, ..., \widehat{R_k}\}$, $\widehat{\theta} = (\widehat{\theta_1}, ..., \widehat{\theta_k})$ and $\widehat{\pi} = (\widehat{\pi_1}, ..., \widehat{\pi_k})$ returned by the TCLUST for some k, α and c values. Given an observation x_i,

let us define

$$D_j(x_i;\widehat{\theta},\widehat{\pi}) = \widehat{\pi}_j\phi(x_i,\widehat{\theta_j}) \text{ for } j = 1,...,k. \tag{4.1}$$

The values in (4.1) can be sorted as $D_{(1)}(x_i;\widehat{\theta},\widehat{\pi}) \leq ... \leq D_{(k)}(x_i;\widehat{\theta},\widehat{\pi})$. A nontrimmed observation x_i would be assigned to group j if $D_j(x_i;\widehat{\theta},\widehat{\pi}) = D_{(k)}(x_i;\widehat{\theta},\widehat{\pi})$ (García-Escudero et al. 2008). Therefore, we can measure the strength of the assignment of x_i to group j by analyzing the size of $D_{(k)}(x_i;\widehat{\theta},\widehat{\pi})$ with respect to the second largest value $D_{(k-1)}(x_i;\widehat{\theta},\widehat{\pi})$. We, thus, define the discriminant factors $\mathrm{DF}(i)$'s as

$$\mathrm{DF}(i) = \log\left(D_{(k-1)}(x_i;\widehat{\theta},\widehat{\pi})/D_{(k)}(x_i;\widehat{\theta},\widehat{\pi})\right).$$

(throughout this section and in the Appendix section, we will omit the dependence on the c value in the notation).

The idea of using "posterior probabilities" like (4.1) to measure assignment strengths is not new in Clustering. The use of these $\mathrm{DF}(i)$'s was already suggested in Van Aelst et al. (2006). The main novelty here will be the definition of discriminant factors also for trimmed observations. Let us consider $d_i = D_{(k)}(x_i;\widehat{\theta},\widehat{\pi})$ for all the observations in the sample and sort them in $d_{(1)} \leq ... \leq d_{(n)}$. The TCLUST trims off a proportion α of observations with smallest assignment strengths. In other words, the trimmed observations are $\widehat{R_0} = \{i \in \{1,...,n\} : d_{(i)} \leq d_{[n\alpha]}\}$. Therefore, we can quantify the certainty of the "trimming" decision for the trimmed observation x_i through

$$\mathrm{DF}(i) = \log\left(d_{([n\alpha]+1)}/D_{(k)}(x_i;\widehat{\theta},\widehat{\pi})\right).$$

Large values of $\mathrm{DF}(i)$ (e.g., $\mathrm{DF}(i) > \log(1/8)$) indicate doubtful assignments or trimming decisions. Of course, this $\log(1/8)$ threshold value is a subjective choice. With this in mind, different summaries of the discriminant factors may be obtained. For instance, "silhouette" plots (Rousseeuw 1987) like those in Figure 4,(b) and Figure 5,(b) can be made. The presence of groups in the silhouette plot containing many large $\mathrm{DF}(i)$ values or, equivalently, having a large $\mathrm{DF}(i)$ mean-value, would indicate that the obtained solution includes some groups having not enough strength. Moreover, we can also plot observations having large $\mathrm{DF}(i)$ values (Figure 4,(c) and Figure 5,(c)) and these observations correspond to doubtful assignment or trimming decisions. For instance, the observations in the frontier between the two clusters that appear when (artificially) splitting one of the main groups in Figure 4,(a) are labeled as doubtful assignments. Some trimmed observations in the boundaries of the main groups may be considered as "doubtfully trimmed" ones. Notice that appropriate lower dimensional representations of the data are needed when $p > 2$ (see, e.g., Hennig 2004b).

By examining these plots, we can see that many doubtful assignments are made when $k = 3$ (Figure 4) and less are made when $k = 2$ (Figure 5).

The here presented discriminant factors are not directly connected with the well-known "Bayes Rule" which constitutes the benchmark for Discriminant Analysis techniques. This

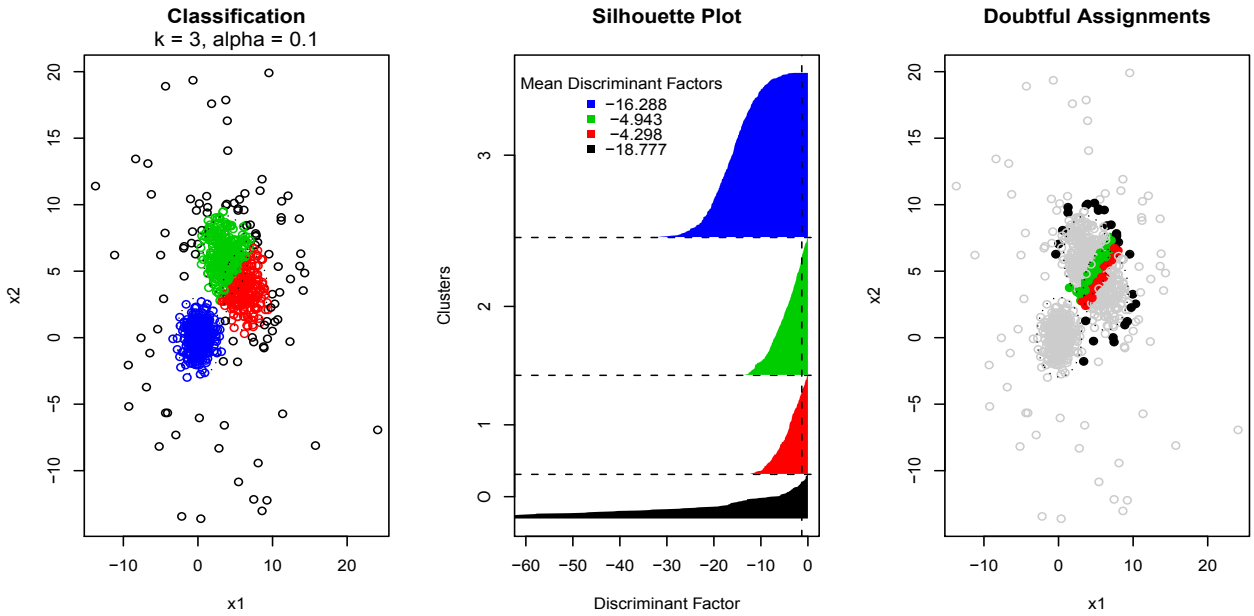

Figure 4: (a) TCLUST solution for $k = 3$, $\alpha = .1$ and $c = 1$. (b) Associated silhouette plot. (c) Observations with discriminant factors larger than $\log(1/8)$ (dotted vertical line in (b)).

connection is only possible when clusters are very separated. Recall also that $\widehat{\theta}$ and $\widehat{\pi}$ are biased "clustering" estimators for the "mixture" parameters θ and π that appear when assuming likelihoods like (1.1). Anyway, Proposition 3 in the Appendix shows that the here presented discriminant factors consistently estimate some population discriminant factors defined for the theoretical (unknown) distribution that generates our data set.

5 Examples

5.1 Simulated Examples

5.1.1 "Noise" and "proper groups"

Let us go back to the simulated data in Figure 1. We had seen there (through the examination of Figure 3) that we can obtain a cluster including all the most scattered data points when we fix $k = 3$, $\alpha = 0$ and a large enough value for the constant c (for instance, $c = 50$). We can see in Figure 6, with a smaller value $c = 5$, that increasing k from 3 to 4 when $\alpha = 0$ seems to be still needed since the most scattered data points can not be fitted within a single group. We can also see in Figure 6 that $k = 2$ and α around .1 are still reasonable choices for k and α when $c = 5$.

This example clearly shows how the answer to the question about whether we have a background noise or a proper main group should depend on the choice of constant c made by the researcher.

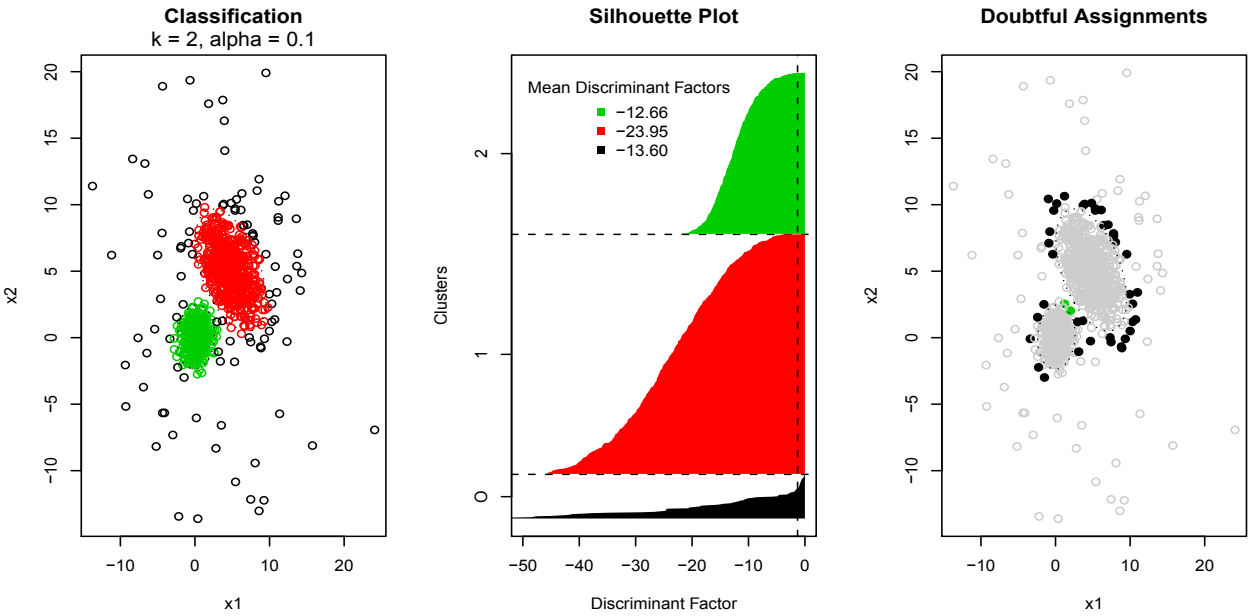

Figure 5: (a) TCLUST solution for $k = 2$, $\alpha = .1$ and $c = 5$. (b) Associated silhouette plot. (c) Observations with discriminant factors larger than $\log(1/8)$ (dotted vertical line in (b)).

It is also important to note that the presented approach does not pose any strong distributional assumption on the contaminating part of the data. This fact makes it very flexible. Other approaches can be very sensitive to deviations from the requirements posed on the contamination. For instance, we can see that Fraley and Raftery's MCLUST (1998) applied to a data set like that in Figure 1 would correctly find 3 groups or 2 groups plus a 10% noise proportion, depending on whether we allow for a uniform background noise component or not. In this case, the non-overlapping part of the third more scattered group is properly fitted through the addition of a uniform component. However, the behavior of the MCLUST is not so satisfactory when handling more "structured" types of noise like those shown in Figure 7. In this figure, we have simulated a data set like that in Figure 1 but replacing the 10% proportion of data corresponding to the more scattered group by more structured noise patterns. With this in mind, a 10% proportion of data is placed in two circumference segments and in a straight line segment.

MCLUST finds a new extra group in Figure 7(1.a) because this "outlying" data is far from being a uniform background noise. Moreover, the clustering solutions may be notably altered as we can see in Figure 7(2.a) and Figure 7(3.a). In all these examples, MCLUST was implemented in such a way that a 10% contamination level was declared as the expected number of outlying data when initializing the procedure. In a similar fashion, these types of noise would also affect McLachlan and Peel's (2000) approach based on fitting mixtures of t distributions. On the other hand, the TCLUST approach has no problem to address these types of noise as we can see in Figure 7(1.b), 7(2.b) and 7(3.b). Notice also that

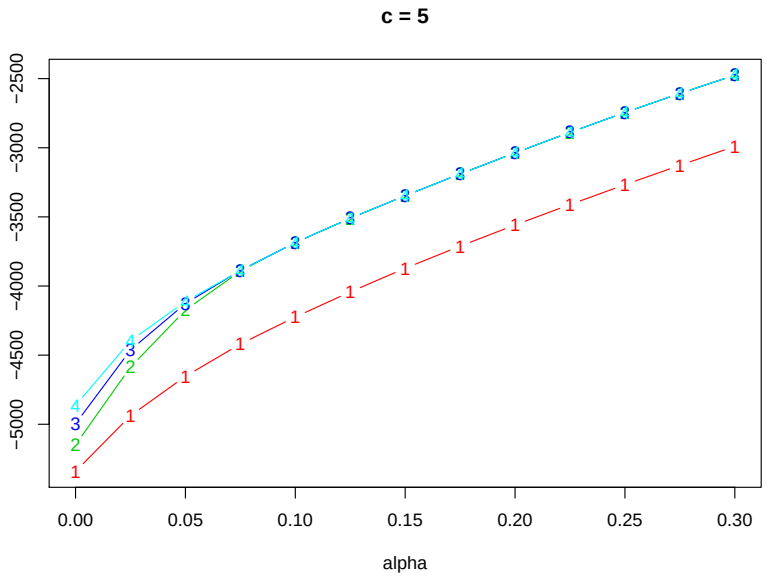

Figure 6: Classification trimmed likelihood curves $\mathcal{L}_5^{\Pi}(\alpha, k)$ when $k = 1, 2, 3, 4$ and α ranges in $[0, .3]$ for the data set in Figure 1 with $c = 5$.

the consideration of eigenvalue ratio constraints has also been very important to achieve the clustering results shown in Figure 7(3.a) and 7(3.b). Otherwise, the subset made of collinear data would introduce a singularity in the likelihood to be maximized. This fact could severely affect clustering methods that do not consider any kind of protection against this problem.

Figure 8 shows the associated classification trimmed likelihood curves for the data sets in Figure 7. The examination of these curves clearly suggests the choice of parameters $k = 2$ and $\alpha \approx .1$ for applying the TCLUST method.

It may be argued that the three contaminations added in Figure 7 could indeed constitute further main groups, but these decisions should again be dependent on the type of clusters we are searching for. For instance, these added points should clearly be outlying data whenever non-degenerated (elliptical) normally distributed clusters are expected.

5.1.2 Clustering and mixture approaches

In this example, we consider two types of three-component Gaussian mixtures presented in Biernacki et al (2000). $n = 400$ data points are randomly drawn from mixtures with mixing proportions $\pi_1 = \pi_2 = .25$ and $\pi_3 = .5$, means given by $\mu_1 = \mu_2 = (0, 0)'$ and $\mu_3 = (8, 0)'$, and, covariance matrices given by

$$\Sigma_1 = \Sigma_3 = \begin{pmatrix} .11 & 0 \\ 0 & 9 \end{pmatrix} \text{ and } \Sigma_2.$$

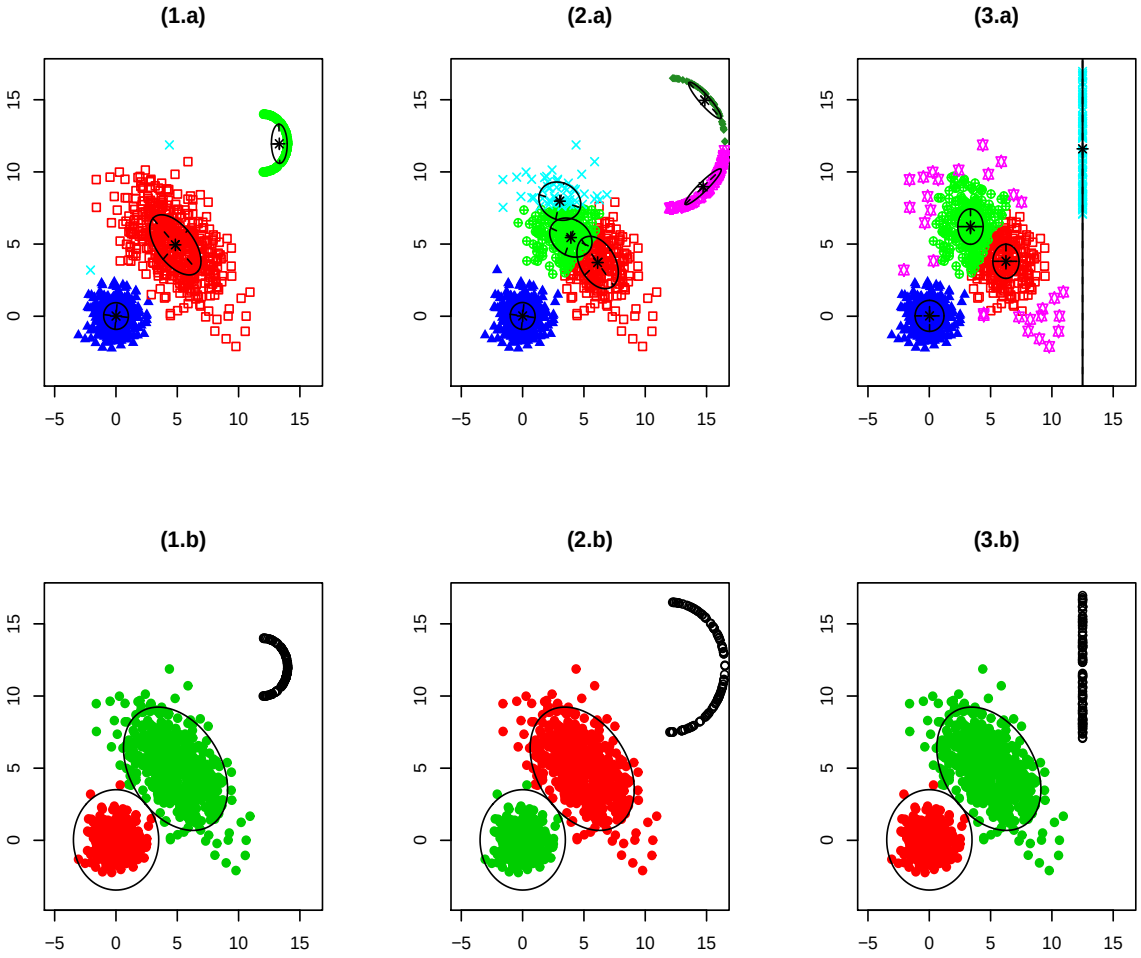

Figure 7: Clustering results for a simulated data set with 3 different sources of structured outlying data. MCLUST results are shown in the top row while TCLUST results with $k = 2$ and $\alpha = .1$ are shown in the bottom row.

Depending on the matrix Σ_2, we have two different mixtures. As a first case, Biernacki et al (2000) set

$$\Sigma_2 = \begin{pmatrix} .96 & 2.61 \\ 2.61 & 8.15 \end{pmatrix},$$

obtaining data sets like that in Figure 9,(a) with an angle between the first eigenvector of Σ_1 and Σ_2 equal to 18 degrees. However, with

$$\Sigma_2 = \begin{pmatrix} 7.33 & 2.64 \\ 2.64 & 1.67 \end{pmatrix},$$

we get an angle equal to 68.5 degrees as it appears in Figure 9,(b). Figure 9 also shows the associated TCLUST cluster solutions for these two simulated data sets when $k = 3$, $c = 90$ and $\alpha = 0$. We can see that 3 groups can be found when the two mixture components centered at $(0, 0)'$ are not "excessively" overlapped in Figure 9,(b). However, when they are

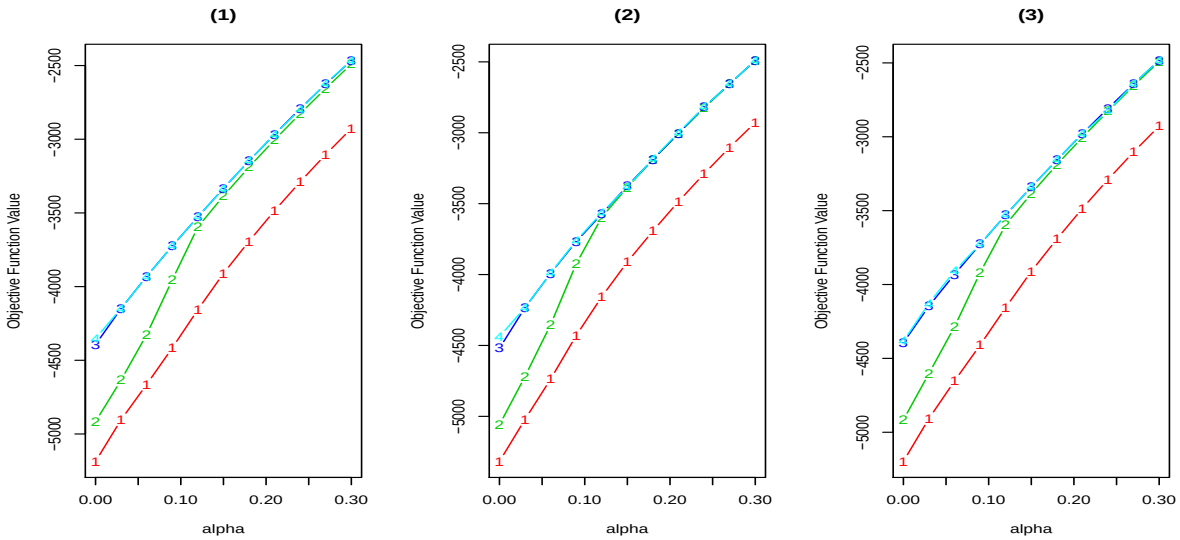

Figure 8: Classification trimmed likelihood curves with $c = 50$ for the data sets in Figure 7 ((1) corresponds to the data in (1.a), (2) to the data in (2.a) and (3) to the data in (3.a)).

very overlapped, we found only two main clusters in Figure 9,(a) and a smaller "spurious" one. The simulated data sets are made of three Gaussian components in both cases from a "mixture modeling" viewpoint. But, from a pure "clustering" viewpoint, we can perhaps recognize only two clusters in Figure 9,(a) while we recognize three in Figure 9,(b). This coincides with the feeling expressed in Biernacki et al (2000).

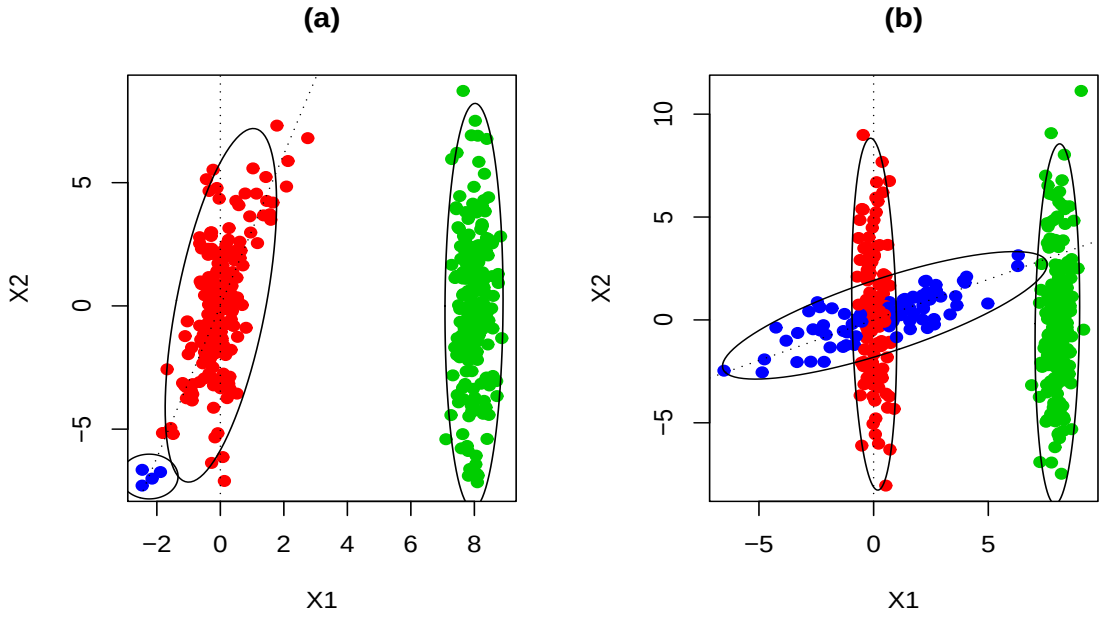

Figure 9: Two simulated data sets following Biernacki et al (2000) simulation scheme and the associated TCLUST cluster solutions when $k = 3$, $c = 90$ and $\alpha = 0$.

Figure 10 shows the classification trimmed likelihood curves for the two simulated data

sets that appear in Figure 9 when $c = 90$. Notice that this value of c satisfies the eigenvalue ratio constraints for the theoretical underlying covariance matrices. Figure 10,(a) suggests a choice of $k = 2$ for data set in Figure 9,(a). Unless a very high trimming value was chosen, a value $k = 3$ is suggested by Figure 10,(b) for the data set in Figure 9,(b).

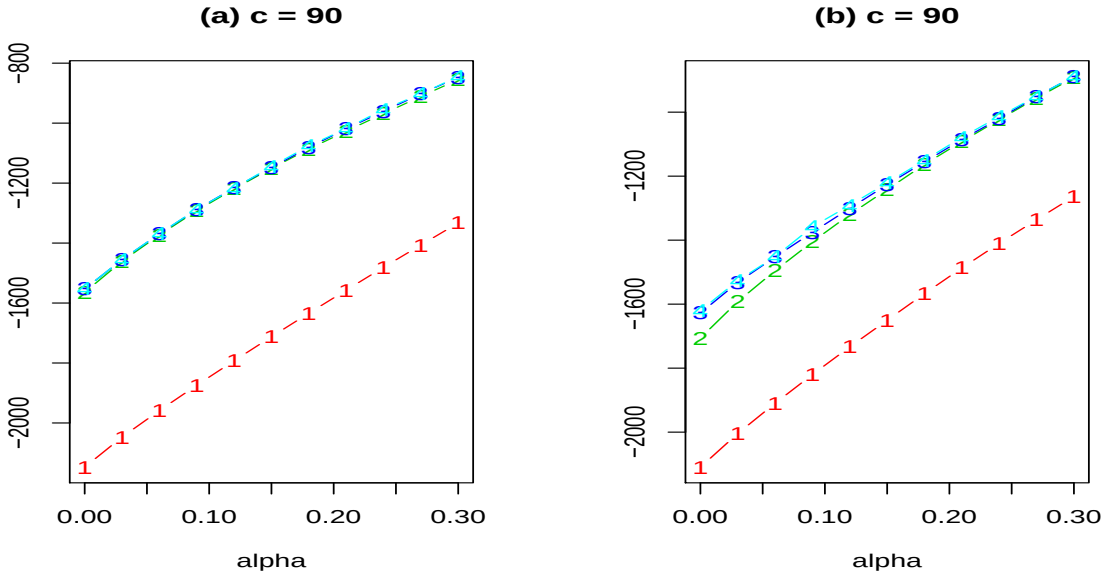

Figure 10: Classification trimmed likelihood curves for the data sets appearing in Figure 9 for $c = 90$. The curves for Figure 9,(a) is given in (a) and for Figure 9,(b) in (b).

5.2 Two Real Data Examples

5.2.1 Old Faithful Geyser data

The "Old Faithful" Geyser data set here considered contains 272 observations on eruption lengths of this geyser. A bivariate data set is obtained by considering these eruption lengths and their associated previous eruption lengths, so, having $n = 271$ data points that are plotted in Figure 12. Three clear main groups together with the presence of six (rare) "short followed by short" eruptions may be seen there.

When computing the classification trimmed likelihoods for this data set, we obtain those appearing in Figure 11. By examining them, it seems sensible to consider at least $k = 3$ and that we only should consider $k = 4$ when we are willing to accept a fraction of approximately 8% of the data joined together as being a proper group.

The cluster solution for $k = 3$ and $\alpha = .08$ appears in Figure 12,(a) and that for $k = 4$ and $\alpha = .02$ in Figure 12,(b).

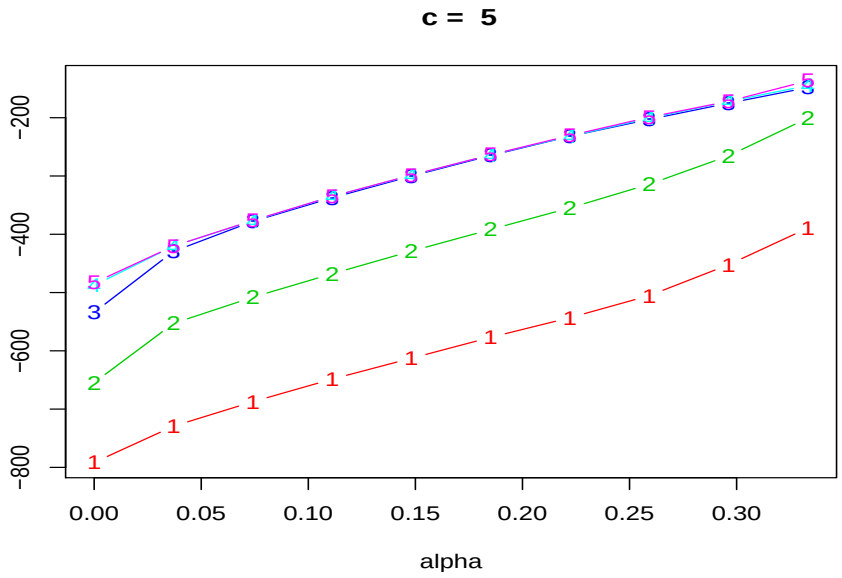

Figure 11: Classification trimmed likelihoods curves for the Old Faithful geyser data.

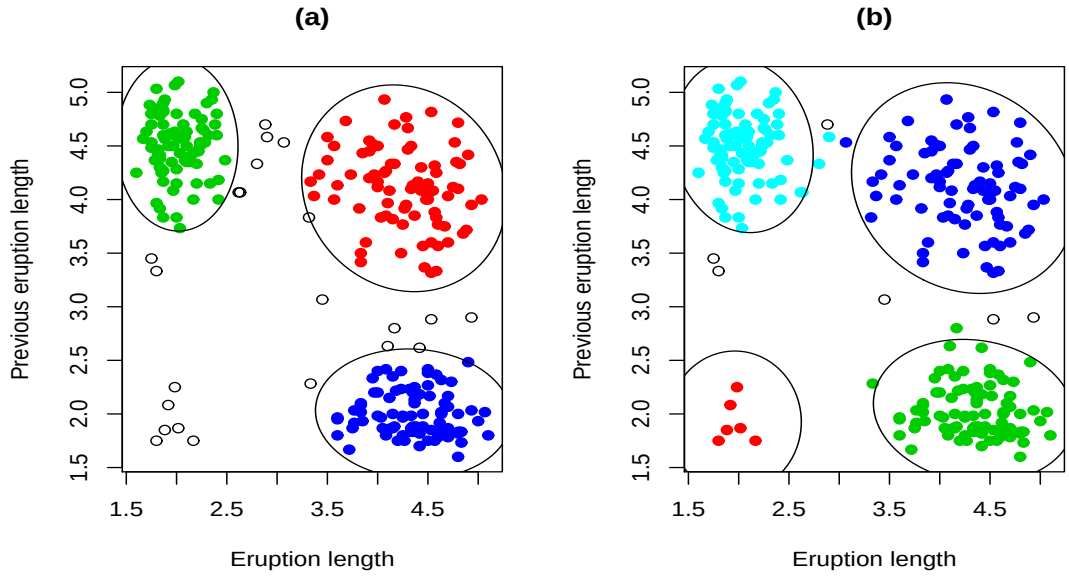

Figure 12: TCLUST-based clustering results for the Old Faithful geyser data: $k = 3$, $\alpha = .08$ and $c = 3$ are used in (a) while $k = 4$, $\alpha = .02$ and $c = 3$ are used in (b).

5.2.2 Swiss Bank Notes data

In this well-known data set (Flury and Riedwyl 1988), $p = 6$ variables are measured on 100 genuine and 100 counterfeit old Swiss 1000-franc bank notes roughly quantifying the size and position of certain features in the printed image. When analyzing the associated classification trimmed likelihoods curves in Figure 13, we can easily guess a two groups structure corresponding due to the presence of "genuine" and "forged" bills. Moreover, when comparing "2 vs. 3 groups", we can observe certain improvement considering $k = 3$

group unless we use a trimming level approximately equal to $\alpha = .1$. In fact, this is essentially motivated by the non-homogeneity of the group of forgeries. For instance, it is quite well-known (see, e.g., Flury and Riedwyl 1988 or Cook 1999) the existence of a further subgroup containing 15 data points (perhaps due to the presence of other forger at work).

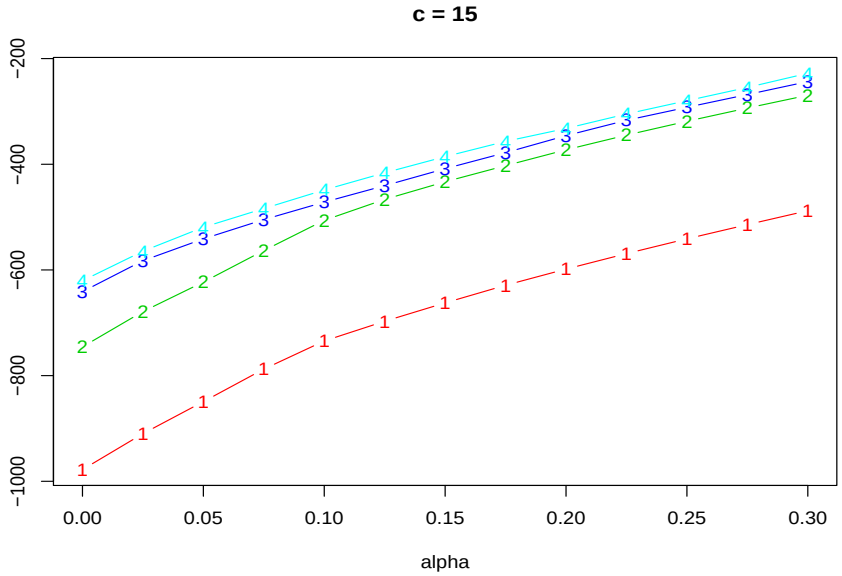

Figure 13: Classification trimmed likelihoods curves for the Swiss Bank Notes data when $c = 15$.

Figure 14 shows a scatterplot of the fourth variable against the sixth one and showing the resultant cluster assignment after the application of the TCLUST with $k = 2$, $\alpha = .08$ and $c = 15$. Trimmed points essentially coincide with the previously commented subset of 15 forged bills with a different forgery pattern.

Appendix: Technical Results

In this section, $I_A(\cdot)$ stands for the indicator function for the set A, A^c for the complementary set of A and $B(m, r)$ denotes the ball centered at m with radius r.

The problem (2.2) introduced in Section 2 admits a population version. Given a probability measure P, we will search for $\widehat{\theta} = (\widehat{\theta_1}, ..., \widehat{\theta_k})$, $\widehat{\pi} = (\widehat{\pi_1}, ..., \widehat{\pi_k})$, and, some *zones* $\{\widehat{Z_0}, \widehat{Z_1}, ..., \widehat{Z_k}\} \subset \mathbb{R}^p$ such that

$$\cup_{j=0}^{k} \widehat{Z_j} = \mathbb{R}^p, \widehat{Z_r} \cap \widehat{Z_s} = \emptyset \text{ for } r \neq s \text{ and } P\big[\cup_{j=1}^{k} \widehat{Z_j}\big] = 1 - \alpha. \tag{5.1}$$

Let $\mathcal{Z}_\alpha$ denote the set of all possible partitions of $\mathbb{R}^p$ satisfying (5.1).

We consider the following population problem:

$$\mathcal{L}^{\Pi}_{c,\alpha,k}(P) := \max_{\{Z_j\}_{j=0}^{k} \in \mathcal{Z}_\alpha, \{\theta_j\}_{j=1}^{k} \in \Theta_c, \{\pi_j\}_{j=1}^{k} \in [0,1]^k} P\left[\sum_{j=1}^{k} I_{Z_j}(\cdot) \log(\pi_j \phi(\cdot; \theta_j))\right]. \tag{5.2}$$

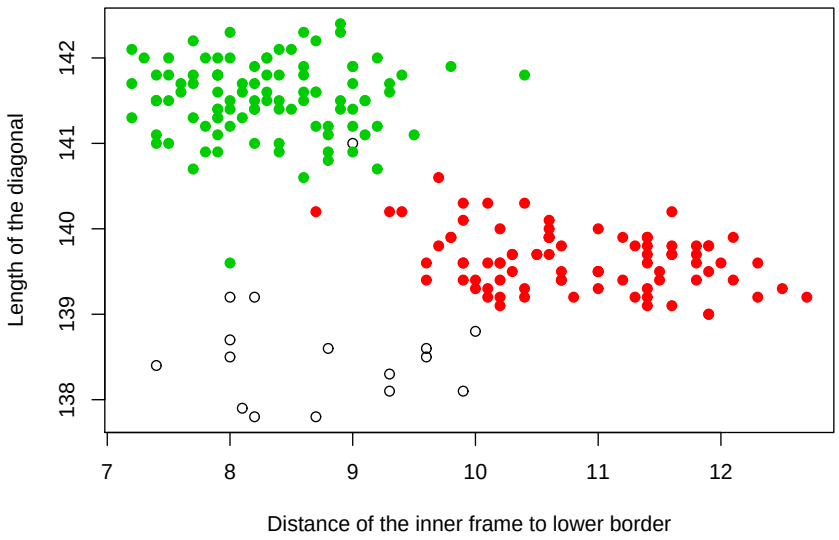

Figure 14: TCLUST-based clustering results for the Swiss Bank Notes data. $k = 2$, $\alpha = .08$ and $c = 15$ are used and only the fourth and sixth variables are plotted.

Notice that $\mathcal{L}_c^{\Pi}(\alpha, k)$ coincides with $\mathcal{L}_{c,\alpha,k}^{\Pi}(P_n)$ in this wider framework, where P_n denotes the empirical probability measure based on the sample $x_1, ..., x_n$.

Analogously, we can extend the problem (2.3) to general probability measures through the problem:

$$\mathcal{L}_{c,\alpha,k}(P) := \max_{\{Z_j\}_{j=0}^k \in \mathcal{Z}_\alpha, \{\theta_j\}_{j=1}^k \in \Theta_c} P\left[\sum_{j=1}^{k} I_{Z_j}(\cdot) \log(\phi(\cdot; \theta_j))\right]. \tag{5.3}$$

Again, we have $\mathcal{L}_{c,\alpha,k}(P_n) = \mathcal{L}_c(\alpha, k)$.

We will start showing that the optimal values of the sample problems converges toward the optimal values of the theoretical ones.

Proposition 1 *When ϕ is the p.d.f. of a multivariate standard normal distribution, we have (for every α, c and k) that*

i) $\mathcal{L}_{c,\alpha,k}^{\Pi}(P_n) \to \mathcal{L}_{c,\alpha,k}^{\Pi}(P)$ *almost surely as* $n \to \infty$, *and,*

ii) $\mathcal{L}_{c,\alpha,k}(P_n) \to \mathcal{L}_{c,\alpha,k}(P)$ *almost surely as* $n \to \infty$.

Proof: These convergences easily follow from standard Glivenko-Cantelli results similar to those applied in Section A.2 in García-Escudero et al. (2008). □

We continue analyzing how $\mathcal{L}_{c,\alpha,k}^{\Pi}(P)$ and $\mathcal{L}_{c,\alpha,k}(P)$ changes when increasing k. As it was already commented in García-Escudero et al. (2008), we could have $\mathcal{L}_{c,\alpha,k}^{\Pi}(P) = \mathcal{L}_{c,\alpha,k+1}^{\Pi}(P)$ since we can set an optimal zone as being $\widehat{Z_j} = \emptyset$ and take $\widehat{\pi}_j = 0$. In this work, we have proposed to take advantage of this fact when trying to choose a suitable k in Clustering

problems. On the other hand, we now prove in Proposition 2 that $\mathcal{L}_{c,\alpha,k}(P)$ strictly increases when increasing k to $k+1$ whenever P is not concentrated on k points after trimming a proportion α.

Proposition 2 *If ϕ is the p.d.f. of a multivariate standard normal distribution and if*

$$\text{there is not a set with } k \text{ points } M = \{m_1, ..., m_k\} \text{ such that } P[M] \geq 1 - \alpha, \tag{5.4}$$

then we have

$$\mathcal{L}_{c,\alpha,k}(P) < \mathcal{L}_{c,\alpha,k+1}(P). \tag{5.5}$$

Proof: Given $\theta = (\theta_1, ..., \theta_k)$, let us define

$$\varphi_j(x;\theta) := (x-\mu_j)'\Sigma_j^{-1}(x-\mu_j) + \log(|\Sigma_j|) := d^2_{\Sigma_j}(x,\mu_j) + \log(|\Sigma_j|).$$

It is easy to see that solving the problem in (5.3) when ϕ is the p.d.f. of a multivariate standard normal distribution reduces to the minimization of

$$\mathcal{V}_{c,\alpha,k}(P) := \min_{\{Z_j\}_{j=0}^k \in \mathcal{Z}_\alpha, \{\theta_j\}_{j=1}^k \in \Theta_c} P\Big[\sum_{j=1}^k I_{Z_j}(\cdot)\varphi_j(\cdot;\theta)\Big]. \tag{5.6}$$

Consequently, the inequality (5.5) would be proven if we prove that $\mathcal{V}_{c,\alpha,k}(P) > \mathcal{V}_{c,\alpha,k+1}(P)$ whenever we have (5.4).

Let $\widehat{\theta} = (\widehat{\theta}_1, ..., \widehat{\theta}_k)$ and $\{\widehat{Z}_0, \widehat{Z}_1, ..., \widehat{Z}_k\}$ be the solutions of problem (5.6) (and consequently (5.3)). It may be seen that the optimal zones $\{\widehat{Z}_0, \widehat{Z}_1, ..., \widehat{Z}_k\}$ may be determined through $\widehat{\theta}$ and the probability measure P. In order to do that, let us consider $\varphi(x;\theta) = \min_{j=1,...,k} \varphi_j(x;\theta)$,

$$R(\theta, P) = \inf_r \big\{P[x : \varphi(x;\theta) \leq r] \geq 1 - \alpha\big\}, \tag{5.7}$$

and, $Z_j(r,\theta) = \big\{x : \{\varphi(x;\theta) = \varphi_j(x;\theta)\} \cap \{\varphi(x;\theta) \leq r\}\big\}$. We have that the optimal zones can be obtained as $\widehat{Z}_j = Z_j\big(R(\widehat{\theta}, P), \widehat{\theta}\big)$, for $j = 1, ..., k$, and $\widehat{Z}_0 = \mathbb{R}^p \setminus \cup_{j=1}^k \widehat{Z}_j$.

Since P satisfy (5.4), we have that $P[\cup_{j=1}^k Z_j(r, \widehat{\theta})] < 1 - \alpha$ for every $r < R(\widehat{\theta}, P)$ (recall definition (5.7)). Therefore, we can see that there exists $m_0 \in \cup_{j=1}^k \widehat{Z}_j$ and j satisfying that:

a) $d^2_{\Sigma_j}(m_0, \widehat{\mu}_j) \geq \frac{2}{3}\big(R(\widehat{\theta}, P) - \log(|\Sigma_j|)\big) > 0$, and,

b) $P[B(m_0, \varepsilon)] > 0$ for every $\varepsilon > 0$.

Without loss of generality, let us assume that this j is equal to k and, thus, $m_0 \in \widehat{Z}_k$. Take now a $\delta > 0$ such that $\delta \leq \frac{1}{3}\big(R(\hat{\theta}, P) - \log(|\widehat{\Sigma}_k|)\big)$. If we now choose a set $B_0 \subset B(m_0, \delta)$,

we have

$$\begin{aligned}\mathcal{V}_{c,\alpha,k}(P) &= P\Big[\sum_{j=1}^{k-1} I_{\widehat{Z}_j}(\cdot)\varphi_j(\cdot;\hat{\theta})\Big] + P[\widehat{Z}_k]\log(|\widehat{\Sigma}_k|) \\ &\quad + P\Big[I_{B_0\cap\widehat{Z}_k}(\cdot)d^2_{\widehat{Z}_k}(\cdot,\widehat{\mu}_k)\Big] + P\Big[I_{B_0^c\cap\widehat{Z}_k}(\cdot)d^2_{\widehat{Z}_k}(\cdot,\widehat{\mu}_k)\Big] \\ &> P\Big[\sum_{j=1}^{k-1} I_{\widehat{Z}_j}(\cdot)\varphi_j(\cdot;\hat{\theta})\Big] + P[B_0\cap\widehat{Z}_k]\log(|\widehat{\Sigma}_k|) + P[B_0^c\cap\widehat{Z}_k]\log(|\widehat{\Sigma}_k|) \\ &\quad + P\Big[I_{B_0\cap\widehat{Z}_k}(\cdot)d^2_{\widehat{Z}_k}(\cdot,m_0)\Big] + P\Big[I_{B_0^c\cap\widehat{Z}_k}(\cdot)d^2_{\widehat{Z}_k}(\cdot,\widehat{\mu}_k)\Big].\end{aligned}$$

Therefore, considering $\widehat{\theta}_1 = (\widehat{\mu}_1, \widehat{\Sigma}_1)$,..., $\widehat{\theta}_{k-1} = (\widehat{\mu}_{k-1}, \widehat{\Sigma}_{k-1})$, $\widehat{\theta}_k = (\widehat{\mu}_k, \widehat{\Sigma}_k)$ and $\widehat{\theta}_{k+1} = (m_0, \widehat{\Sigma}_k)$, we get a possible solution for problem (5.6), when k is increased to $k+1$, with a target value strictly smaller than $\mathcal{V}_{c,\alpha,k}(P)$. Consequently, resorting to the optimality of $\mathcal{V}_{c,\alpha,k+1}(P)$, we have that $\mathcal{V}_{c,\alpha,k}(P) > \mathcal{V}_{c,\alpha,k+1}(P)$ and the result is proved. □

Let us consider $D_j(x;\theta,\pi) = \pi_j\phi(x,\theta_j)$ defined for every $x \in \mathbb{R}^p$ and $j = 1,...,k$ and sort them to get $D_{(1)}(x;\theta,\pi) \leq ... \leq D_{(k)}(x;\theta,\pi)$. We define

$$R(\theta,\pi,P) = \inf_r \big\{P[x : D_{(k)}(x;\theta,\pi) \geq r] \geq 1-\alpha\big\}.$$

and $Z_j(r,\theta,\pi) = \big\{x : \{D_{(k)}(x;\theta,\pi) = D_j(x;\theta,\pi)\} \cap \{D_{(k)}(x;\theta,\pi) \geq r\}\big\}$. If $\hat{\theta}$, $\hat{\pi}$ and $\{\widehat{Z}_0, \widehat{Z}_1, ..., \widehat{Z}_k\}$ are the optimal solution of the problem (5.2), it can be proved that $\widehat{Z}_j = Z_j\big(R(\widehat{\theta},\widehat{\pi},P),\widehat{\theta},\widehat{\pi}\big)$, for $j = 1,...,k$, and, $\widehat{Z}_0 = \mathbb{R}^p \setminus \cup_{j=1}^k \widehat{Z}_j$.

Given a probability measure P, we can define (by using previous notation) a discriminant factor value for $x \in \mathbb{R}^p$ as:

$$\mathrm{DF}(x;P) = \log\big(D_{(k)}(x;\widehat{\theta},\widehat{\pi})/D_{(k-1)}(x;\widehat{\theta},\widehat{\pi})\big) \text{ for } x \in \cup_{j=1}^k \widehat{Z}_j,$$

and

$$\mathrm{DF}(x;P) = \log\big(D_{(k)}(x;\widehat{\theta},\widehat{\pi})/R(\hat{\theta},\hat{\pi},P)\big) \text{ for } x \in \widehat{Z}_0.$$

Notice that $\mathrm{DF}(x_i;P_n) = \mathrm{DF}(i)$ when P_n is the empirical measure.

The following result states the consistency of the empirical discriminant factors toward the theoretical ones.

Proposition 3 *If ϕ is the p.d.f. of a multivariate standard distribution, P has a strictly positive density function and the solution of the problem* (5.2) *for that P is unique, then* $\mathrm{DF}(x;P_n) \to \mathrm{DF}(x;P)$ *almost surely as $n \to \infty$ for every $x \in \mathbb{R}^p$.*

Proof: Let us denote by $\widehat{\theta^n}$, $\widehat{\pi^n}$ and $\{\widehat{Z_0^n}, \widehat{Z_1^n}, ..., \widehat{Z_k^n}\}$ to the solutions of the problem (5.2) when P is the empirical measure P_n. Let us also denote by $\widehat{\theta^0}$, $\widehat{\pi^0}$ and $\{\widehat{Z_0^0}, \widehat{Z_1^0}, ..., \widehat{Z_k^0}\}$ to

the solutions of (5.2) for the true underlying distribution P. As shown in Proposition 3 in García-Escudero et al. (2008), we have $\widehat{\theta^n} \to \widehat{\theta^0}$ and $\widehat{\theta^n} \to \widehat{\theta^0}$ almost surely. Moreover, Lemma A.7 in García-Escudero et al. (2008) guaranteed $R(\widehat{\theta^n}, \widehat{\pi^n}, P_n) \to R(\widehat{\theta^0}, \widehat{\pi^0}, P)$ almost surely. All these consistencies easily entail the consistency result we want to prove. □

Remark 1 An uniqueness condition for the solution of the population problem is needed in Proposition 3. This is not surprising because otherwise the solution of the sample problem would be highly unstable and, thus, it is not logical to expect consistency for the empirical discriminant factors. Moreover, the uniqueness condition many times has to do with the lack of appropriateness of the choice of parameter k.

Acknowledgements:

We thank the associate editor and two anonymous referees for their careful reading and for their valuable comments. The `tclust` package has been elaborated with the collaboration of Heinrich Fritz.

References

[1] Banfield, J.D. and Raftery, A.E. (1993), "Model-based Gaussian and non-Gaussian clustering," *Biometrics,* **49**, 803-821.

[2] Becker, C. and Gather, U. (1999), "The masking breakdown point for outlier identification rules," *J. Amer. Statist. Assoc.,* **94**, 947-955.

[3] Biernacki, C. and Govaert, G. (1997), "Using the Classification Likelihood to Choose the Number of Clusters," *Computing Science and Statistics*, **29**, 451-457

[4] Biernacki, C., Celeux, G. and Govaert, G. (2000), "Assesing a mixture model for clustering with the integrated completed likelihood," *IEEE Trans. on Pattern Analysis and Machine Learning,* **22**, 719-725.

[5] Bryant, P.G. (1991), "Large-sample results for optimization-based clustering methods," *J. Classific.*, **8**, 31-44.

[6] Bock, H.-H. (1996), "Probabistic models in cluster analysis," *Comput. Statit. Data Anal.*, **23**, 5-28.

[7] Calinski, R.B. and Harabasz, J (1974), "A dendrite method for cluster analysis" *Communications in Statistics*, **3**, 1-27.

[8] Celeux, G. and Govaert, A. (1992), "Classification EM algorithm for clustering and two stochastic versions", *Comput. Statit. Data Anal.,* **13**, 315-332

[9] Celeux, G. and Govaert, A. (1992), "Gaussian parsimonious clustering models", *Pattern Recognition.,* **28**, 781-793.

[10] Cook, D. (1999), "Graphical detection of regression outliers and mixtures," *Procedings ISI, 99. Helsinki.*

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

[12] Cuesta-Albertos, J.A., Matran, C. and Mayo-Iscar, A. (2008), "Robust estimation in the normal mixture model based on robust clustering", *Journal of the Royal Statistical Society. Ser. B.* , **70**, 779-802.

[13] Dasgupta, A. and Raftery, A.E. (1998) "Detecting features in spatial point processes with clutter via model-based clustering," *J. Amer. Statist. Assoc.*, **93**, 294-302.

[14] Engelman, L. and Hartigan, J.A. (1969), "Percentage Points of a Test for Clusters," *J. Amer. Statist. Assoc.* **64**, 1647-1648.

[15] Flury, B. (1997), *A first course in Multivariate Statistics,* Springer-Verlag New York.

[16] Flury, B. and Riedwyl, H. (1988), *Multivariate Statistics, A Practical Approach,* Cambridge University Press.

[17] Friedman, H.P. and Rubin, J. (1967), "On some invariant criterion for grouping data" *J. Amer. Statist. Assoc.*, **63**, 1159-1178.

[18] 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.

[19] Gallegos, M.T. (2002), "Maximum likelihood clustering with outliers", in *Classification, Clustering and Data Analysis: Recent advances and applications*, K. Jajuga, A. Sokolowski, and H.-H. Bock eds., 247-255, Springer-Verlag.

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

[21] Gallegos, M.T. and Ritter, G. (2009), "Trimming algorithms for clustering contaminated grouped data and their robustness," *Adv. Data Analysis and Classification*, **3**, 135-167.

[22] Gallegos, M.T. and Ritter, G. (2010), "Using combinatorial optimization in model-based trimmed clustering with cardinality constraints," *Comput. Statit. Data Anal.* **54**, 637-654.

[23] García-Escudero, L.A., Gordaliza, A. and Matrán, C. (2003), "Trimming tools in exploratory data analysis," *J. Comput. Graph. Statist.,* **12**, 434-449.

[24] García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008), "A general trimming approach to robust cluster analysis", *Ann. Statist.*, **36**, 1324-1345.

[25] Hardin, J. and Rocke, D. (2004), "Outlier Detection in the Multiple Cluster Setting Using the Minimum Covariance Determinant Estimator", *Comput. Statist. Data Anal.*, **44**, 625-638.

[26] Hathaway, R.J. (1985), "A constrained formulation of maximum likelihood estimation for normal mixture distributions," *Ann. Statist*, **13**, 795-800.

[27] Hawkins, D.M. and Olive, D.J. (2002), "Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm", with discussion *J. Amer. Statist. Assoc.*, **97**, 136-159.

[28] Hennig, C. and Christlieb, N. (2002), "Validating visual clusters in large datasets: fixed point clusters of spectral features", *Comput. Statit. Data Anal.*, **40**, 723-739.

Hennig, C. (2004a), "Breakdown points for maximum likelihood-estimators of location-scale mixtures", *Ann. Statist* **32**, 1313-1340.

[29] Hennig, C. (2004b), "Asymmetric linear dimension reduction for classification", *J. Comput. Graph. Statist.,* **13**, 930-945 .

[30] Keribin, C. (2000), "Consistent estimation of the order of mixture models", *Sankhyā Ser. A*, **62**, 49-62.

[31] McLachlan, G. (1987), "On Bootstrapping The Likelihood Ratio Test Statistic For The Number Of Components In A Normal Mixture," *Applied Statistics*, **37**, 318-324.

[32] McLachlan, G. and Peel, D. (2000), *Finite Mixture Models,* John Wiley Sons, Ltd., New York.

[33] McQueen, J. (1967), "Some methods for classification and analysis of multivariate observations", 5^{th} *Berkeley Symposium on Mathematics, Statistics, and Probability.* Vol **1**, 281-298.

[34] Milligan, G.W. and Cooper, M.C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, **50**, 159-179.

[35] Neykov, N.M., Filzmoser, P., Dimova, R. and Neytchev, P.N. (2004), "Mixture of Generalized Linear Models and the Trimmed Likelihood Methodology" in *Proc. in Computational Statistics*, J. Antoch (ed.), Physica-Verlag, 1585-1592.

[36] Neykov, N.M., Filzmoser, P., Dimova, R., and Neytchev, P. (2007), "Robust fitting of mixtures using the trimmed likelihood estimator", *Comput. Statit. Data Anal.*, **52**, 299-308.

[37] Rocke, D.M. and Woodruff, D.M. (2002), "Computational Connections Between Robust Multivariate Analysis and Clustering", in COMPSTAT 2002 Proceedings in Computational Statistics, W. Härdle and B. Rönz eds., 255-260, Heidelberg:Physica-Verlag.

[38] Rousseeuw, P.J. (1987), "Silhouettes: a graphical aid to the interpretation and validation of cluster analyisis," *J. Comput. Appl. Math.*, **20**, 53-65.

[39] Rousseeuw, P.J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, **41**, 212-223.

[40] Scott, A.J. and Symons, M.J. (1971), "Clustering based on likelihood ratio criteria," *Biometrics,* **27**, 387-397.

[41] Symons, M.J. (1981), "Clustering criteria and Multivariate Normal Mixtures," *Biometrics,* **37**, 35-43.

[42] Titterington, D.M., Smith A.F. and Makov, U.E. (1985), *Statistical analysis of finite mixture distributions*, Wiley, New York.

[43] Van Aelst, S., Wang, X., Zamar, R.H. and Zhu, R. (2006), Linear grouping using orthogonal regression, *Comput. Statit. Data Anal.*, **50**, 1287-1312.

[44] Woodruff, D.L. and Reiners, T. (2004), "Experiments with, and on, algorithms for maximum likelihood clustering", *Comput. Statist. Data Anal.*, **47**, 237-253.

[45] Wolfe, J.H. (1970), "Pattern clustering by multivariate analysis", *Multivariate Behavioral Research*, **5**, 329-350.