Robust Constrained Fuzzy Clustering

Heinrich Fritz[a], Luis A. García-Escudero[b,*], Agustín Mayo-Iscar[b]

[a]*Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria*
[b]*Department of Statistics and Operations Research and IMUVA, University of Valladolid, Valladolid, Spain*

Abstract

It is well-known that outliers and noisy data can be very harmful when applying clustering methods. Several fuzzy clustering methods which are able to handle the presence of noise have been proposed. In this work, we propose a robust clustering approach called F-TCLUST based on an "impartial" (i.e., self-determined by data) trimming. The proposed approach considers an eigenvalue ratio constraint that makes it a mathematically well-defined problem and serves to control the allowed differences among cluster scatters. A computationally feasible algorithm is proposed for its practical implementation. Some guidelines about how to choose the parameters controlling the performance of the fuzzy clustering procedure are also given.

Keywords: Clustering, fuzzy clustering, noise, outliers, constraints, trimming.

1. Introduction

Hard clustering procedures are aimed at searching for a partition of data into k disjoint clusters, with similar subjects grouped in the same cluster and dissimilar subjects in different ones. On the other hand, fuzzy clustering methods provide nonnegative membership values of observations to clusters, and this generates overlapping clusters where every subject is shared among all clusters [see, e.g., 33, 10, 2, 20].

It is also widely recognized that clustering methods need to be robust if they are to be useful in practice. Notice that, otherwise, clustering ability

*Corresponding author

may deteriorate drastically due to the presence of even a small fraction of outlying data. In fact, historically, the fuzzy clustering community was the first one to face the robustness challenge in Clustering. This is mainly due to the fact that outliers tend to be approximately "equally remote" from all clusters and, thus, they may have similar (but not necessarily small) membership values. For instance, membership values for outlying observations could be close to $1/k$ for all the clusters, with k being the number of groups whenever membership values are assumed to sum up to 1. [7] provide a general review of robust fuzzy clustering methods (see also [15] for a review of robust hard clustering procedures).

One of the methods which is more widely considered in robust fuzzy clustering is the fuzzy C-means method with "noise component" [5] and a plethora of modifications. Unfortunately, this approach inherits its preference for spherical clusters from fuzzy C-means. So, this method is often unable to properly detect clusters with very different shapes. Several procedures have been proposed to address this problem [see, e.g., 18, 17, 34, 31].

In this work, we adapt a hard robust clustering approach called TCLUST [14] to the fuzzy clustering framework. The proposed approach also extends the "Least Trimmed Squares" approach to fuzzy clustering introduced by [21] toward a more general methodology. The proposed methodology is thus based on trimming a fixed fraction α of the "most outlying" observations. We may denote this trimming as "impartial" since the data set itself tells us which are the observations to be trimmed off without the intervention of the user declaring privileged directions or zones for trimming. The fixed trimming level controls the number of observations to be discarded in a different way from other methods that are based on fixing a "noise distance" [see, e.g., 6, 8, 29]. These methods are also cited as "noise clustering" in the literature. Discarding a fixed fraction of data has also been considered in [22].

There exist other interesting fuzzy clustering proposals where robustness is incorporated through the replacement of the Euclidean distance as a measure of the discrepancies between observations and cluster centers [35, 25, 37]. However, they are mainly aimed at searching spherical equally scattered groups as fuzzy C-means methods do.

An important feature of the proposed approach is that it allows for non spherically-shaped clusters, but it also forces the obtained clusters to be "comparable" in terms of cluster scatters. In this way, clusters with arbitrarily very different scatters are not allowed. This is done by imposing an eigenvalue ratio constraint on the cluster scatter matrices. Some type of con-

straint on the scatter matrices is compulsory because, otherwise, the fuzzy clustering problem would become a mathematically ill-posed problem.

The proposed methodology, called F-TCLUST, is presented in Section 2. A feasible algorithm for its practical application is given in Section 3. The algorithm is theoretically justified in Section 4. Section 5 provides some guidance about how to choose the several parameters that F-TCLUST takes into account. Section 6 presents an application to a well-known real data set. Finally, the paper concludes with some closing remarks and future research lines.

2. The F-TCLUST method

Suppose that we have n observations $\{x_1, ..., x_n\}$ in $\mathbb{R}^p$ and we want to group them into k clusters in a fuzzy way. Therefore, our aim is to obtain a collection of nonnegative membership values $u_{ij} \in [0, 1]$ for all $i = 1, ..., n$ and $j = 1, ..., k$. A membership value 1 indicates that object i fully belongs to cluster j while a 0 membership value means that it does not belong at all to this cluster. However, intermediate degrees of membership are allowed when $u_{ij} \in (0, 1)$. We consider that an observation is fully trimmed if $u_{ij} = 0$ for all $j = 1, ..., k$ and, thus, this observation has no membership contribution to any cluster.

Let $\varphi(\cdot; \mu, \Sigma)$ stand for the probability density function of a p-variate normal distribution $N_p(\mu, \Sigma)$ defined as

$$\varphi(x; \mu, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1} \exp\big(-(x-\mu)'\Sigma^{-1}(x-\mu)/2\big).$$

Given a fixed trimming proportion $\alpha \in [0, 1)$, a fixed constant $c \geq 1$ and a fixed value of the fuzzifier parameter value $m > 1$; a robust constrained fuzzy clustering problem can be defined through the maximization of the objective function:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m \log \varphi(x_i; m_j, S_j), \tag{1}$$

where the membership values $u_{ij} \geq 0$ are assumed to satisfy

$$\sum_{j=1}^{k} u_{ij} = 1 \text{ if } i \in \mathcal{I} \text{ and } \sum_{j=1}^{k} u_{ij} = 0 \text{ if } i \notin \mathcal{I},$$

for a subset

$$\mathcal{I} \subset \{1, 2, ..., n\} \text{ with } \#\mathcal{I} = [n(1-\alpha)],$$

where $m_1, ..., m_k$ are vectors in $\mathbb{R}^p$, and, $S_1, ..., S_k$ are positive semidefinite $p \times p$ matrices obeying the following eigenvalue ratio constraint

$$\frac{\max_{j=1}^{k} \max_{l=1}^{p} \lambda_l(S_j)}{\min_{j=1}^{k} \min_{l=1}^{p} \lambda_l(S_j)} \leq c, \tag{2}$$

where $\{\lambda_l(S)\}_{l=1}^{p}$ denote the p eigenvalues of the matrix S.

Notice that $u_{i1} = ... = u_{ik} = 0$ for all $i \notin \mathcal{I}$, so these observations do not contribute to the summation in the target function (1).

Using a maximum likelihood criterium like that in (1) implies fixing a specific underlying statistical model, which indeed allows us to better understand what the fuzzy clustering method is really aimed at. This maximum likelihood approach has already been considered, among others, in [17], [34], [36], [4] and [31].

One of the main features of the proposed methodology is the application of the eigenvalue ratio constraint in (2). It is important to see that some type of constraint in this maximum likelihood approach is compulsory because, otherwise, the objective function (1) would become unbounded, just by taking one of the m_j equal to one of the observations x_i, setting $u_{ij} = 1$, and taking a sequence of scatter matrices S_j such that $|S_j| \to 0$. This problem is recurrent in Cluster Analysis whenever general scatter matrices are allowed. For instance, this trouble was already noticed in fuzzy clustering by [18], where they also proposed constraining the relative volumes $|S_j|$ to be equal to some constants fixed in advance. Other different types of constraint can be found in [34] and [31].

In our approach, the unboundedness problem is addressed by constraining the ratio between the largest and smallest eigenvalues of the scatter matrices. In other words, we are assuming that the square root of the ratio between the lengths of the axes of the tolerance ellipsoids defined through the S_j scatter matrices are smaller than a constant c. This approach can be seen as an extension of [19]. The smaller the constant c, the more similarly scattered the groups are. For instance, the clusters should fall within spheres of the same radius when $c = 1$ and the associated clustering results are close to those obtained when applying fuzzy C-means. Larger values of c lead to an almost unconstrained fuzzy clustering approach. This type of constraints on the eigenvalues was also considered in [3] when updating the scatter matrix

with the aim of controlling the cluster shapes. In this work, this constraint on the eigenvalues is explicitly posed in the maximization of the objective function.

The use of an objective function like that in (1) lends the method a bias toward clusters with similar sizes (more precisely toward clusters with similar values of $\sum_{i=1}^{n} u_{ij}^m$). If this effect is not desired then it is better to replace the objective function (1) by

$$\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m \log\left(p_j \varphi(x_i; m_j, S_k)\right), \tag{3}$$

where $p_j \in [0,1]$ and $\sum_{j=1}^{k} p_j = 1$ are some weights that the objective function also needs to be maximized on. Notice that, once the membership values are known, these weights are optimally determined as $p_j = \sum_{i=1}^{n} u_{ij}^m / \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m$. Thus, this approach implies adding the term

$$\sum_{j=1}^{k}\Big(\sum_{i=1}^{n} u_{ij}^m\Big)\log\Big(\sum_{i=1}^{n} u_{ij}^m \Big/ \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m\Big)$$

to the target function (1). This type of regularization is related to the "entropy regularizations" [27] which have already appeared in the literature. We will explain, through a simulated data set in Section 5, the impact that the consideration of the weights p_j has on the type of clusters we are looking for.

By replacing the target function (1) by (3) in the previously introduced robust constrained fuzzy clustering problem, we obtain the F-TCLUST approach to fuzzy clustering. It is easy to see that it exactly reduces to the TCLUST hard robust clustering method introduced in [14] when the value of the fuzzifier parameter is set at $m = 1$.

3. A feasible algorithm

The maximization of the objective functions (1) and (3) with all these constraints is not an easy task. In this section, we propose a computationally feasible algorithm aimed at solving this complex problem. The proposed algorithm is based on two alternating steps. First, given the values of the parameters in a given iteration, the best possible membership values are obtained. Conversely, given some membership values, the parameters are updated by maximizing expression (3) on these parameters.

Therefore, the algorithm is an "Expectation-Maximization (EM)" type of algorithm [9] as are those often applied when fitting mixtures to data sets [see, e.g., 26]. In any case, we can see that the updating formulas for the membership values and for the parameters are similar to those applied in other fuzzy clustering algorithms when different cluster scatter matrices are allowed (see, e.g., [18] or the algorithms given in [31]).

The algorithm pays special attention to how the constraint on the eigenvalue ratios are imposed when updating the S_j matrices in Step 2.3. Moreover, in Step 2.2., the algorithm incorporates a type of "concentration step" analogous to that applied in many high-breakdown point robust algorithms [32]. Note also that the algorithm may be seen as an extension to fuzzy clustering of the TCLUST algorithm in [14] and [12].

Although its proper justification will be deferred to Section 4, the proposed algorithm may be described as follows:

1. *Initialization:* The procedure is initialized several times by randomly selecting parameters $p_1, ..., p_k$, $m_1, ..., m_k$, and, $S_1, ..., S_k$. For this purpose, we propose to randomly select $k \times (p+1)$ observations and to accordingly compute k cluster centers m_j and scatter matrices S_j based on these chosen data points. If needed, the S_j matrices must be modified properly so that they satisfy the required eigenvalue ratio constraints by following the approach described in Step 2.3. Weights $p_1, ..., p_k$ in the interval $(0, 1)$ and summing up to 1 are also randomly chosen.
2. *Iterative steps:* The following steps are executed until convergence or a maximum number of iterations is reached.

   2.1. *Membership values:* Based on the current parameters, if

   $$\max_{q=1,...,k} p_q \varphi(x_i; m_q, S_q) \geq 1,$$

   then

   $$u_{ij} = I\{p_j \varphi(x_i; m_j, S_j) = \max_{q=1,...,k} p_q \varphi(x_i; m_q, S_q)\} \text{ (hard assignment)},$$

   with $I\{\cdot\}$ being a 0-1 indicator function which takes the value 1 if the expression within the brackets holds. If

   $$\max_{q=1,...,k} p_q \varphi(x_i; m_q, S_q) < 1,$$

then

$$u_{ij} = \Big(\sum_{q=1}^{k} \Big(\frac{\log(p_j \varphi(x_i; m_j, S_j))}{\log(p_q \varphi(x_i; m_q, S_q))} \Big)^{\frac{1}{m-1}} \Big)^{-1} \text{ (fuzzy assignment).}$$

2.2. *Trimmed observations:* Let

$$r_i = \sum_{j=1}^{k} u_{ij}^m \log(p_j \varphi(x_i; m_j, S_j)) \tag{4}$$

and $r_{(1)} \leq r_{(2)} \leq ... \leq r_{(n)}$ be these values sorted. The observations to be trimmed are those with indexes $\{i : r_i < r_{([n\alpha])}\}$. The membership values for those observations are redefined as

$$u_{ij} = 0, \text{ for every } j, \text{ if } r_i < r_{([n\alpha])}.$$

2.3. *Update parameters:* Given the membership values obtained in the previous steps, the parameters are updated as

$$p_j = \frac{\sum_{i=1}^{n} u_{ij}^m}{\sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m}, \tag{5}$$

and,

$$m_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m}. \tag{6}$$

Updating the scatter matrices S_j is more complex, since the matrices that are often used for updating them, defined as

$$T_j = \frac{\sum_{i=1}^{n} u_{ij}^m (x_i - m_j)(x_i - m_j)'}{\sum_{i=1}^{n} u_{ij}^m}, \tag{7}$$

may not satisfy the required eigenvalue ratio constraint. In that case, the singular-value decomposition of $T_j = U_j' D_j U_j$ is considered, with U_j being an orthogonal matrix and $D_j = \text{diag}(d_{j1}, d_{j2}, ..., d_{jp})$ a diagonal matrix. Let us define the truncated eigenvalues as

$$[d_{jl}]_t = \begin{cases} d_{jl} & \text{if } d_{jl} \in [t, ct] \\ t & \text{if } d_{jl} < t \\ ct & \text{if } d_{jl} > ct \end{cases}, \tag{8}$$

with t being a threshold value. The scatter matrices are then updated as

$$S_j = U_j' D_j^{opt} U_j,$$

with $D_j^{opt} = \text{diag}\left([d_{j1}]_{t_{opt}}, [d_{j2}]_{t_{opt}}, ..., [d_{jp}]_{t_{opt}}\right)$ where t_{opt} minimizes the real-valued function:

$$t \mapsto \sum_{j=1}^{k} p_j \sum_{l=1}^{p} \left(\log\left([d_{jl}]_t\right) + \frac{d_{jl}}{[d_{jl}]_t} \right), \tag{9}$$

with p_j as defined in (5).
In fact, there is a closed form to obtain t_{opt} by evaluating the function (9) $2pk+1$ times. In order to do that, let us consider the values $e_1 \leq e_2 \leq ... \leq e_{2kp}$ obtained after ordering the $2kp$ values:

$$d_{11}, d_{12}, ..., d_{jl}, ..., d_{kp}, d_{11}/c, d_{12}/c, ..., d_{jl}/c, ..., d_{kp}/c.$$

Consider any $2pk+1$ values $f_1, ..., f_{2kp+1}$ satisfying:

$$f_1 < e_1 \leq f_2 \leq e_2 \leq ... \leq f_{2kp} \leq e_{2kp} < f_{2kp+1},$$

and, compute

$$t_i = \frac{\sum_{j=1}^{k} p_j (\sum_{l=1}^{p} d_{jl} I\{d_{jl} < f_i\} + \frac{1}{c} \sum_{l=1}^{p} d_{jl} I\{d_{jl} > c f_i\})}{\sum_{j=1}^{k} p_j (\sum_{l=1}^{p} (I\{d_{jl} < f_i\} + I\{d_{jl} > c f_i\}))}, \tag{10}$$

for $i = 1, ..., 2kp+1$. Finally, choose t_{opt} as the value of t_i which yields the minimum value of (9).

3. *Evaluate objective function:* Finally, after this iterative process, the value of the associated target function (3) is computed. The set of parameters and membership values yielding the highest value of this objective function are returned as the algorithm's output.

The algorithm presented here is focused on maximizing the objective function (3) but it can be easily adapted to perform the maximization of (1) just by assuming fixed equal weights $p_j = 1/k$ throughout all the iterations.

The number of random initializations and the maximum number of iterations play a key role in the performance of the algorithm. The larger these two numbers are, the higher the probability that the algorithm ends up finding the global constrained maximum. Of course, these higher numbers also

imply a higher computational cost (as happens with other closely related algorithms). Experience tells us that not many random initializations and iterations are required when the dimension p is not huge and parameters m, c, α and k are chosen in a sensible way. Note that the minimization of function (9) does not notably increase the computational complexity of the algorithm.

4. Justification of the proposed algorithm

As previously commented, the proposed algorithm is based on two alternating steps. In one of them, we search for the membership values maximizing the target function given the current parameters (Steps 2.1 and 2.2), and, in the other, we search for the parameters maximizing the target function under the constraint on the eigenvalues (Steps 2.3) given the current membership values. The algorithm thus increases the value of the target function through this iterative process, which allows to find a local maximum of the target function (3). The iterative process is randomly initialized several times trying to find the global maximum.

Membership values: Let us assume as known the values of the parameters p_j, m_j and S_j. Then, we search for the membership values that make (3) as large as possible. The maximization of (3) is equivalent to the minimization of

$$\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^{m} D_{ij}, \tag{11}$$

with $D_{ij} = -\log\left(p_j \varphi(x_i; m_j, S_k)\right) = \log\left(p_j^{-1}\det(S_j)^{1/2}\exp((x_i - m_j)' S_j^{-1}(x_i - m_j))\right)$.

If we assume that $p_j\varphi(x_i; m_j, S_j) < 1$ for all x_i, then $D_{ij}(> 0)$ can be seen as a measure of the distances of the observation x_i to the center m_j (in fact, $\exp(D_{ij})$ is the "exponential distance measure" introduced in [17]. In this way, the minimization of (11) on the membership values has a similar statement as that considered in fuzzy C-means clustering. Standard Lagrange multiplier arguments lead to optimal membership values as

$$u_{ij} = \left(\sum_{q=1}^{k}\left(\frac{D_{ij}}{D_{iq}}\right)^{\frac{1}{m-1}}\right)^{-1},$$

which indeed coincide with the "fuzzy assignments" proposed in Step 2.1.

On the other hand, let us now assume that there exists any j such that $p_j\varphi(x_i; m_j, S_j) \geq 1$ (that is: $\log(p_j\varphi(x_i; m_j, S_j)) \geq 0$). We can assume w.l.o.g. that $\log(p_1\varphi(x_i; m_1, S_1)) = \max_{j=1,..k} p_j \log(\varphi(x_i; m_j, S_j)) \geq 0$ and, then, we have

$$\sum_{j=1}^{k} u_{ij}^m \log\left(p_j\varphi(x_i; m_j, S_k)\right) \leq \log\left(p_1\varphi(x_i; m_1, S_1)\right) \sum_{j=1}^{k} u_{ij}^m$$
$$\leq \log\left(p_1\varphi(x_i; m_1, S_1)\right) \sum_{j=1}^{k} u_{ij} = \log\left(p_1\varphi(x_i; m_1, S_1)\right).$$

Thus, we just need $u_{i1} = 1$ and $u_{ij} = 0$ for $j \neq 1$ to maximize (3) and the "hard assignments" proposed in Step 2.1 are justified.

Trimmed observations: When a proportion α of observations is allowed to be discarded, it is quite easy to see that the discarded observations are those yielding the smallest values of r_i with r_i as defined in (4) to maximize (3) (recall that (3) is equal to $\sum_{i=1}^{n} r_i$). This type of argument is the basis of the "concentration steps" that some Robust Statistics algorithms apply [32]. The "concentration steps" have been previously used in hard clustering problems [14, 28, 13]. Consequently, the algorithm fixes $u_{ij} = 0$ for all the indexes i such that $r_i < r_{([n\alpha])}$. This idea also underlies the "Least Trimmed Squares" approach to fuzzy clustering in [21].

Update parameters: We now assume that the membership values are known and we want to maximize the objective function (3) on the parameters p_j, m_j and S_j.

As happens with other maximum likelihood approaches to fuzzy clustering, it is not difficult to see that the best choices for p_j and m_j are those given in expressions (5) and (6) [see, e.g., 18, 36]. Plugging these values into (3) and applying the cyclic property of the trace operator, the maximization of (3) can be reduced to the minimization on S_j of

$$\sum_{j=1}^{k} \Big(\sum_{i=1}^{n} u_{ij}^m \Big) \Big(\log(\det(S_j)) + \operatorname{trace}(S_j^{-1} T_j) \Big), \tag{12}$$

with T_j as given in (7).

Let us consider the spectral decomposition of the matrices $T_j = U_j' D_j U_j$ and $S_j = V_j' E_j V_j$, with diagonal matrices $D_j = \text{diag}(d_{j1}, ..., d_{jp})$ and $E_j = \text{diag}(e_{j1}, ..., e_{jp})$, and, orthogonal matrices U_j and V_j. Mimicking the reasoning in [14], it can be shown that the V_j matrix must exactly coincide with U_j. This tells us that the "shapes" of the optimal matrices S_j are uniquely determined by the T_j matrices. Therefore, it is only necessary to choose the optimal eigenvalues e_{ij} properly.

By using $U_j = V_j$ and the fact that these matrices are orthogonal, we can easily see that the minimization of (12) simplifies to the minimization of

$$\sum_{j=1}^{k} \Big(\sum_{i=1}^{n} u_{ij}^m \Big) \sum_{l=1}^{p} \left(\log e_{jl} + \frac{d_{jl}}{e_{jl}} \right), \tag{13}$$

with e_{jl} taking some values that satisfy the constraint $e_{jl}/e_{uv} \leq c$ for every j, l, u and v. This can be done [see details in 12] by truncating the eigenvalues d_{jl} from below by a constant t and from above by $c \cdot t$ (i.e., considering the truncated $[d_{jl}]_t$ given in (8)) and searching for the optimal t which minimizes the real-valued function (9). This function is continuously differentiable and, thus, it attains the minimum value at one of its critical points (with expressions like those in (10)).

5. Choice of parameters

Since the proposed methodology aims to be very general, several parameters are involved in it. In this section, we will explain the different roles that the parameters in the F-TCLUST play through a simulated data set. Thus we consider a very simple example made up of 450 random observations in $\mathbb{R}^2$ from the $N_2(0, I)$ distribution and another 450 observations from the

$$N_2 \left(\begin{pmatrix} 5 \\ 10 \end{pmatrix}, \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix} \right)$$

distribution. We also add 100 uniformly distributed observations in the rectangle $[-10, 15] \times [-10, 15]$, but not considering those observations whose Mahalanobis distances (using the parameters of these two bivariate normal distributions) are smaller than $\chi^2_{2;0.975}$ (where $\chi^2_{2;0.975}$ is the 0.975 quantile of the Chi-squared distribution with 2 degrees of freedom). We thus mitigate the overlapping of the generated noise with the two normal components. In

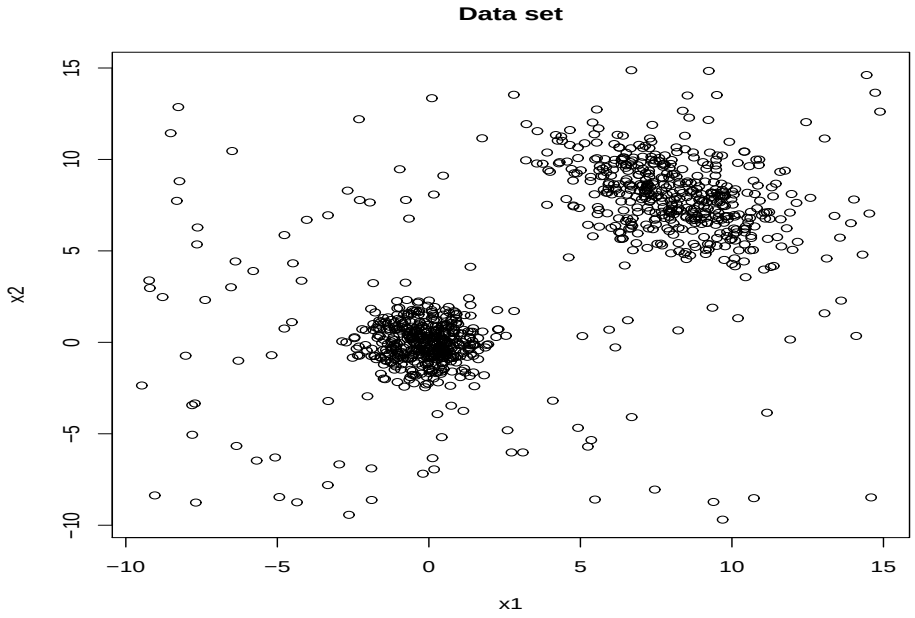

Figure 1: Simulated data set with a 10% noise proportion.

this way, these 100 added observations can actually be considered as a 10% background noise. Figure 1 shows a scatterplot of that simulated data set.

Fuzzifier parameter: We first study the effect of the fuzzifier parameter m in the clustering results when applying the F-TCLUST procedure with $k = 2$, $c = 5$ and $\alpha = 0.1$. Figure 2 shows the obtained cluster membership values by plotting observations with point sizes proportional to these membership values. Trimmed observations are shown in a separate plot.

As previously commented, the $m = 1$ case coincides with the TCLUST method yielding "hard" or "crisp" membership values, where each observation is fully trimmed or fully assigned to a cluster as shown in Figure 2,(b). On the contrary, all non-trimmed observations are "shared" with almost equal membership values in Figure 2,(c) when a large value of m, like $m = 2$, is chosen. Intermediate values of m, like $m = 1.3$, surely yield more interesting membership values, as shown in Figure 2,(a).

The proposed approach is equivariant with respect to location shifts and rotations but it is non-affine equivariant due to the lack of equivariance of the eigenvalue ratio constraint. However, a large value of c yields an almost affine equivariant procedure, but it also increases the risk of finding spurious

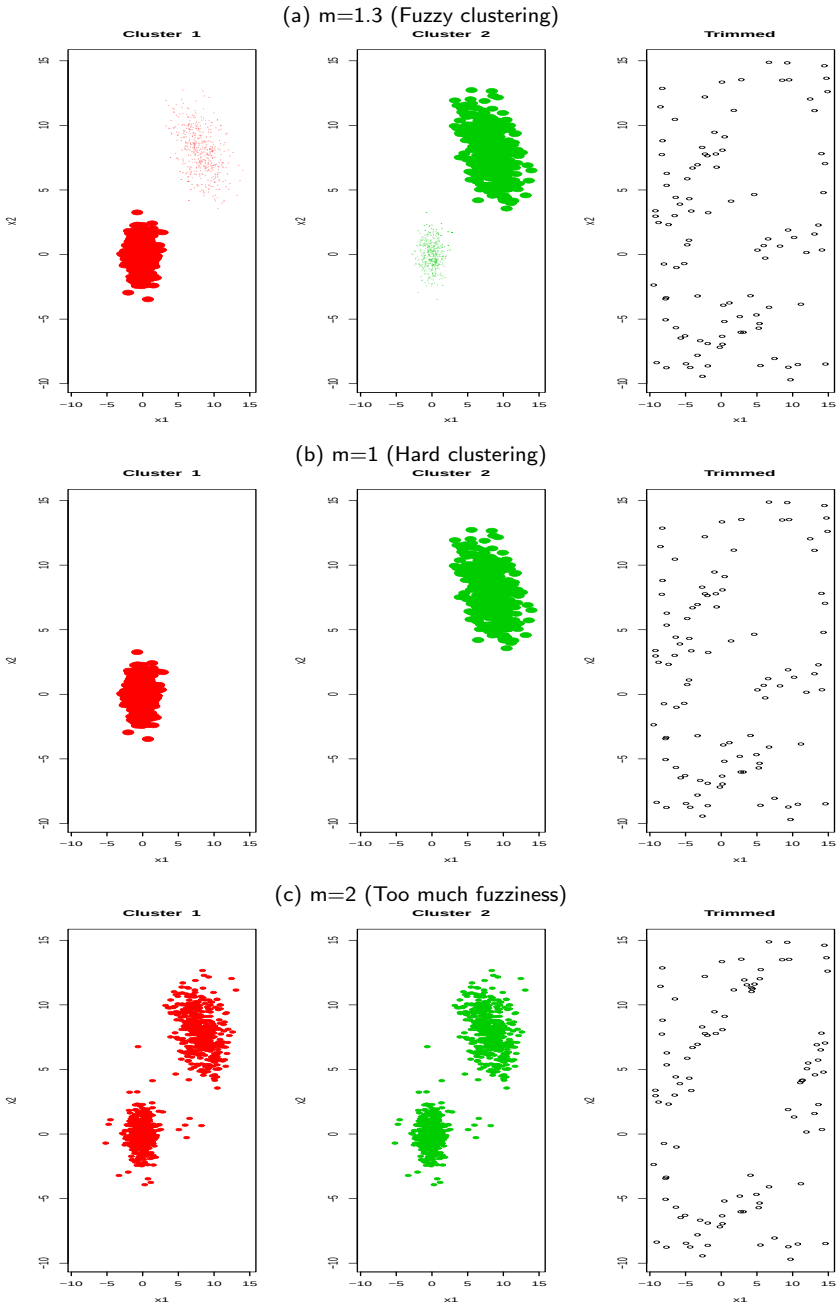

Figure 2: Effect of the value of the fuzzifier parameter m on the cluster membership values (proportional to the point sizes).

clusters.

It is also important to pay special attention to an inherent problem of fuzzy clustering approaches based on the maximum likelihood principle. To see this, let us assume that the variables in our example are scaled by a constant factor S. That is, we change the variables as follows: $X_1 \leftarrow X_1/S$ and $X_2 \leftarrow X_2/S$. Figure 3,(a) shows that $m = 5$ yields a very high degree of fuzzification (in fact, $m = 2$ already did so). However, a logical degree

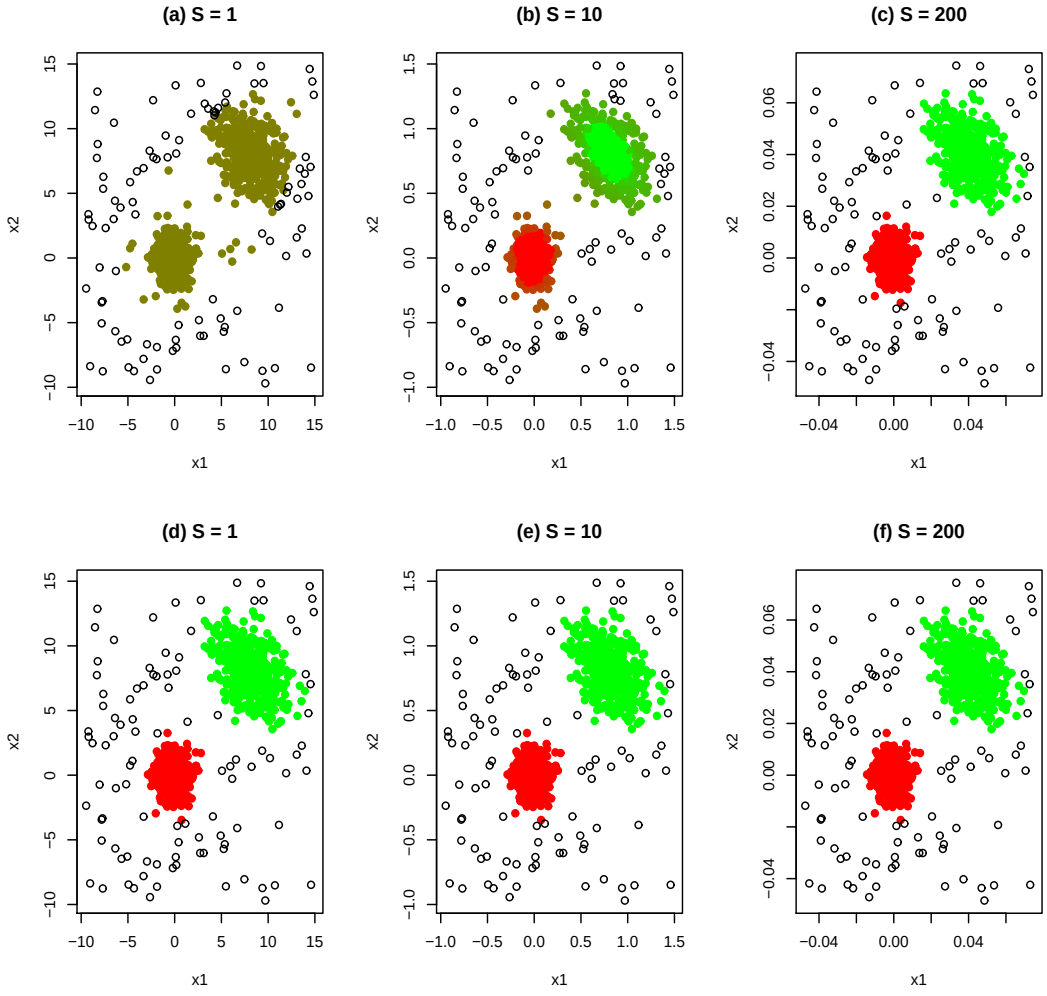

Figure 3: Clustering solutions when applying F-TCLUST with $k = 2$, $\alpha = 0.1$, $c = 5$ and $m = 5$ in (a), (b) and (c) and $m = 1$ in (d), (e) and (f). Different scaling factors S are considered (observe the x_1-axes in these figures).

of fuzzification is obtained in Figure 3,(b) with $m = 5$ when the variables are scaled with an $S = 10$ factor. Finally, we obtain a hard clustering partition of the data set with the same value $m = 5$ when $S = 100$ in Figure 3,(c). In these figures, we use a mixture of "red" and "green" colors with intensities proportional to the membership values to summarize the fuzzy clustering results, while trimmed points are always represented by "○" symbols. This dependence on the scale factor S no longer appears when using a hard clustering approach (i.e. when $m = 1$), as can be seen in Figures 3,(d), (e) and (f).

The clustering results shown in Figure 3,(b) (augmented in Figure 4) are particularly interesting. We can see that "hard" assignment decisions are made in the "core" of the clusters, with observations that are undoubtedly assigned. "Fuzzy" assignments are made for the observations that are more difficult to be classified in the tails of the two normal components. These two different types of assignment decisions follow from the application of Step 2.1 in the proposed algorithm. This naturally leads to a fuzzy clustering method with "high contrast" [30]; that is, it may be seen as a compromise between "hard" and "fuzzy" clustering methods. If "high contrast" partitions are specially interesting for the user, then this provides an appropriate way to

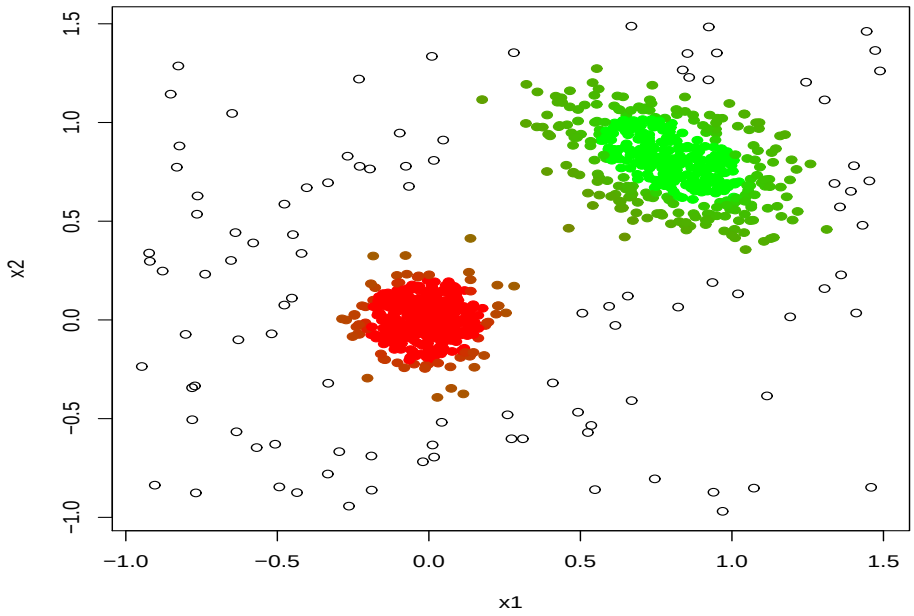

Figure 4: Clustering results in Figure 3,(b) augmented.

scale the data by determining S in such a way that a fixed proportion of "hard" assignments are done.

Weights and number of clusters: The addition of weights p_j in expression (3) (instead of considering (1)) is very important. We are "ideally" assuming that the clusters have similar sizes when weights p_j do not appear in the target function to be maximized. This does not necessarily imply that the resulting clusters satisfy $\sum_{i=1}^{n} u_{i1}^m = ... = \sum_{i=1}^{n} u_{ik}^m$, but we are "ideally" searching for this type of clustering solutions, and not very interested in clustering solutions too far from that case.

We can see in Figure 5,(a) the clustering results for $k = 3$, $c = 5$, $\alpha = 0.1$ and $m = 1.3$ when maximizing (1) (that is: "equal weights") and when maximizing (3) in Figure 5,(b) (that is: "unequal weights"). The clustering solution in Figure 5,(b) is essentially made up of just two clusters, while the third cluster has small membership values for all observations. In this example, $k = 2$ is clearly a good choice for the number of clusters (once 10% of the outlying data points are trimmed). Thus, the fact of allowing for weights p_j in the target function could provide interesting information about

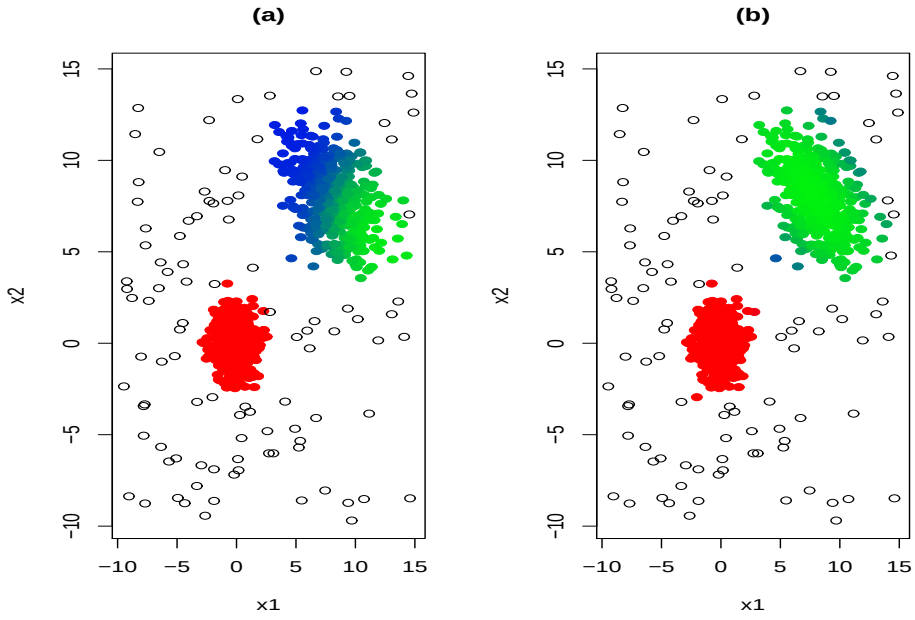

Figure 5: Fuzzy clustering results for $k = 3$, $m = 1.3$, $c = 5$ and $\alpha = 0.1$ depending on whether "equal weights", i.e. maximizing (1), are assumed in (a) or "unequal weights", i.e. maximizing (3), in (b).

how to make sensible choices for k. This possibility was already considered in hard clustering problems in [16].

Note also that Figure 5,(a) shows many intermediate membership values due to the clear overlap between the two clusters that share one of the two normal components. This "soft" transition between clusters is only feasible when applying fuzzy clustering techniques.

Eigenvalue ratio restriction constant: One of the most distinctive features of the proposed F-TCLUST approach is the consideration of constraints on the cluster scatters following from the control of relative sizes of the scatter matrix eigenvalues through constant c. For instance, we allow for clusters with very different scatters in Figure 6,(a) when fixing a large c value (like $c = 50$). This has allowed for the detection of a very scattered group (shown in "blue" color). On the other hand, the clusters are forced to have very similar scatters when $c = 1$, as shown in Figure 6,(b). In this second case, the more scattered cluster is no longer possible. Moreover, when c is close to 1, the clusters are forced to be almost spherical (and with the same scatter among them) and, thus, F-TCLUST may be seen as an extension of the fuzzy C-means method.

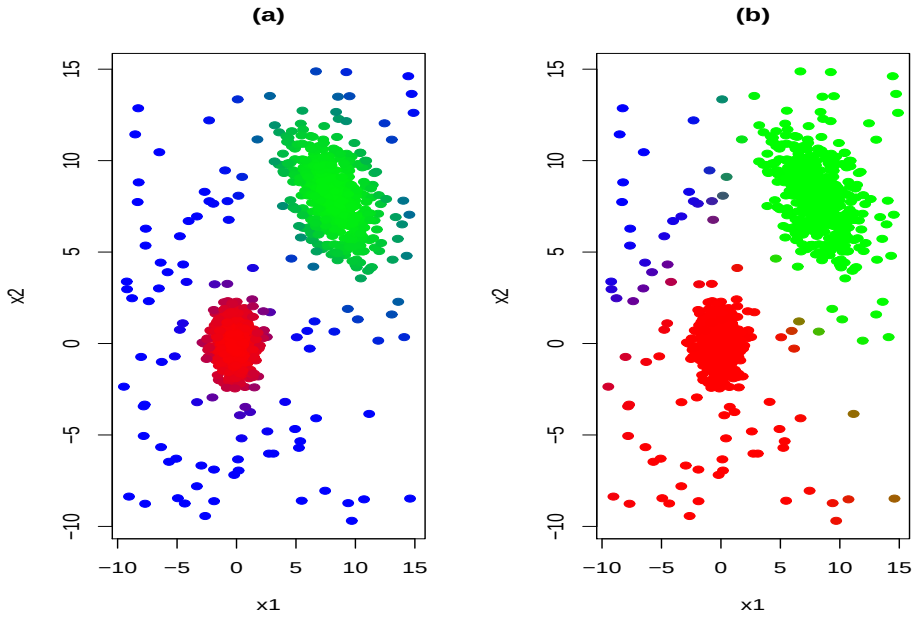

Figure 6: Fuzzy clustering results for $k = 3$ and $\alpha = 0.1$ depending on whether a large value for the restriction factor $c = 50$ is chosen in (a) or a smaller one $c = 1$ in (b).

With respect to the choice of parameter c in a specific clustering problem, the researcher sometimes has an initial idea of the differences between group scatters that he/she is willing to accept, depending on the application in mind for the clustering results. When such information is not available, we propose to monitor the sequence of clustering solutions obtained when moving c. In our experience, not too many essentially different clustering solutions need to be evaluated. The interpretation of c in terms of the lengths of the axes of the ellipsoids defined by the S_j matrices will be very important in this process. It is also informative to check whether the resulting S_j matrices satisfy $\lambda_l(S_j) = \lambda_v(S_u)$ for some $(l, j) \neq (v, u)$ or not, to see if the algorithm is "artificially" forcing the fulfillment of the required constraints for a given value of constant c.

Trimming level: Another parameter that plays a key role in the F-TCLUST methodology is the trimming level α. Recall that observations are fully declared as noise when they are trimmed and no "intermediate" noise assignments are allowed (as, for instance, the methods based on the "noise clustering" approach do).

Sometimes, the researcher has an approximate initial idea of the underlying "contamination level" in the data set, but at other times this contamination level is completely unknown. In the case where this underlying contamination level is unknown, we can see that monitoring the r_i values introduced in (4) provides valuable information to see whether the choice made for α was sensible or not. Recall that these values are used to determine the trimmed observations as long as outlying observations take small r_i values. Thus, for a tentative trimming level α, we propose plotting the points $\{(i/n, r_{(i)})\}_{i=1,2,3,...}$ obtained with F-TCLUST for this value of α. Recall that $r_{(1)} \leq ... \leq r_{(n)}$ are the sorted r_i values. The choice made for α is considered as being appropriate if it is close to a value α_0, such that the values $r_{(i)}$ increase quickly when $i/n < \alpha_0$ and the increase becomes slower when $i/n > \alpha_0$.

For instance, in Figure 7,(b), (d) and (f), we have plotted these $r_{(i)}$ values for $\alpha = 0.02, 0.2$ and 0.1. The value $\alpha = 0.02$ is clearly not a good choice because the $r_{(i)}$ values are still increasing quickly at this value of α, as can be seen in Figure 7,(b). A value $\alpha = 0.2$ is not a good choice either, because the shaded region in Figure 7,(d) includes a values where the increase is quite slow. However, we can see in Figure 7,(f) that $\alpha = 0.1$ is close to be a sensible choice for α. In fact, recall that 10% was the true contamination level for

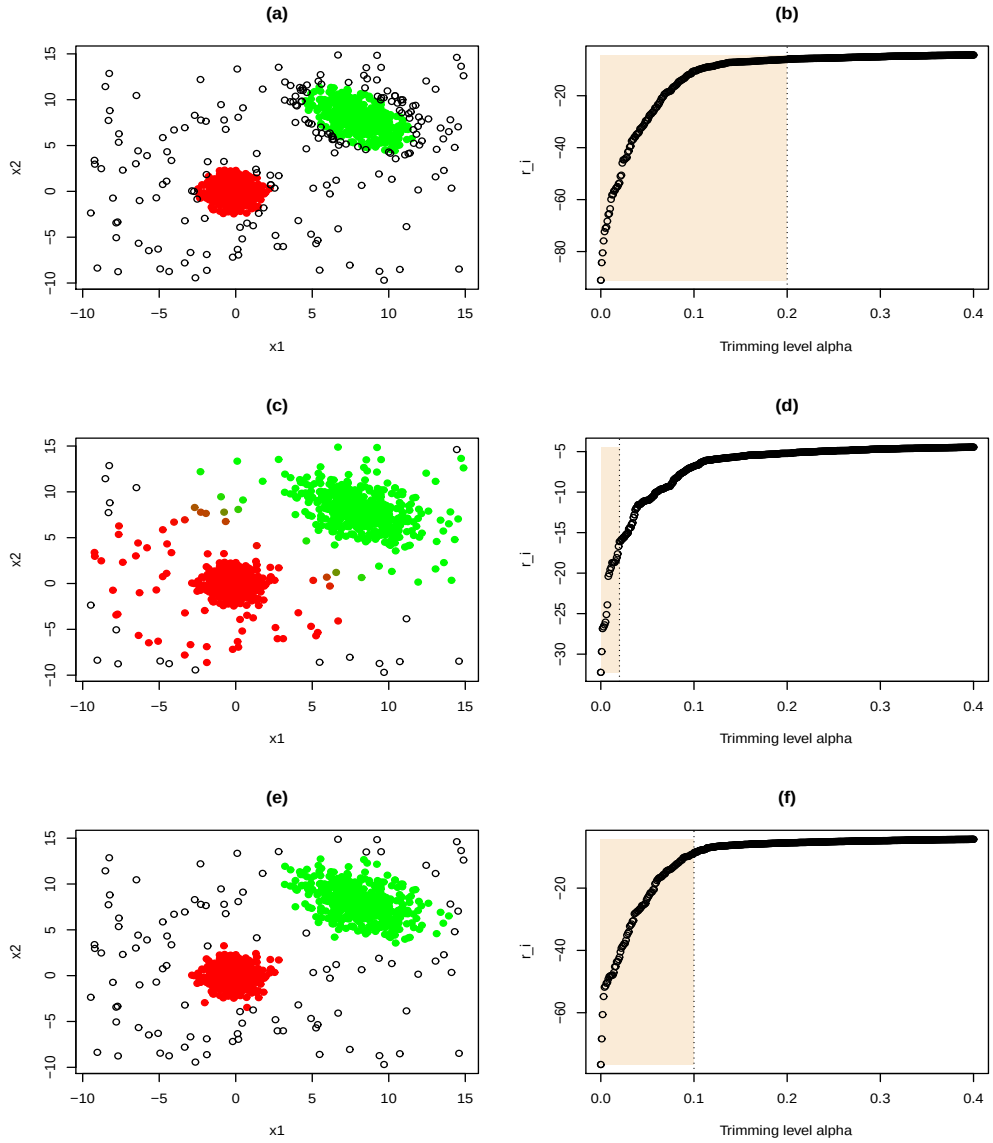

Figure 7: Different clustering solutions depending on the chosen trimming level ($\alpha = 0.02, 0, 2$ and 0.1) and the plots of the sorted r_i values for each different choice of α.

this simulated data set.

It is not compulsory to be extremely precise with an "exact" choice of α because the parameters p_j, m_j and S_j do not change notably, for instance, when α is slightly overestimated. Starting with a slightly overestimated α,

it is not difficult to make a better choice for the trimming level by carefully analyzing the trimmed observations which are close to the non-trimmed ones. From our point of view, in general, it is recommended to be initially conservative when choosing α.

We are currently investigating the possibility of unsupervised methods for choosing the trimming level α. However, it is important to note that the problem of choosing α is closely related to the choice of parameters k and c [see 16]. Addressing all these determinations in a unified manner still requires an active role to be played by the researcher, as a final decision may be very subjective, and, thus, it is not clear that a fully unsupervised strategy can be found.

To clarify previous claims, let us consider again the data set in Figure 1. If we allow for a large value of c (i.e., huge differences among cluster scatters), then $k = 3$ and $\alpha = 0$ is a sensible choice. But, on the other hand, it is surely better to choose $k = 2$ and $\alpha = 0.1$ when c is small. The fuzzy adaptation of the "classification trimmed-likelihood curves" introduced in [16] might be considered as an exploratory tool to help the researcher make sensible simultaneous choices for k, α and c (see Section 7).

Another possibility that may be explored follows from monitoring some cluster validity index [e.g., the density criterion in 17] against the trimming level α, as proposed in [21].

6. A real data example

The "Swiss Bank Notes" data set in [11] includes $p = 6$ variables measuring certain features in the printed image of 100 genuine and 100 counterfeit old Swiss 1000-franc bank notes.

Figure 8,(a) shows a scatterplot of the fourth ("Distance of the inner frame to lower border") against the sixth variable ("Length of the diagonal") and Figure 8,(b) shows a scatterplot of the first ("Length") against the fourth. In these two plots, the classification of bills in [11] is shown by using symbols "`G`" for the genuine bills and "`F`" for the forged ones. They also commented that the group of forged bills was not a homogeneous group and they pointed out [11, pg. 265] a list with 15 "anomalous" forged bills that surely follow from a different forgery pattern. These 15 bills are shown in Figure 8,(a) surrounded by circle symbols. Other authors have reported this inhomogeneity in the group of forged bills [see, e.g., 5]. Figures 8,(a) and (b)

also show an observation surrounded by a "□" symbol that corresponds to a "genuine" bill that would fit better in the group of "forged" bills [see 11].

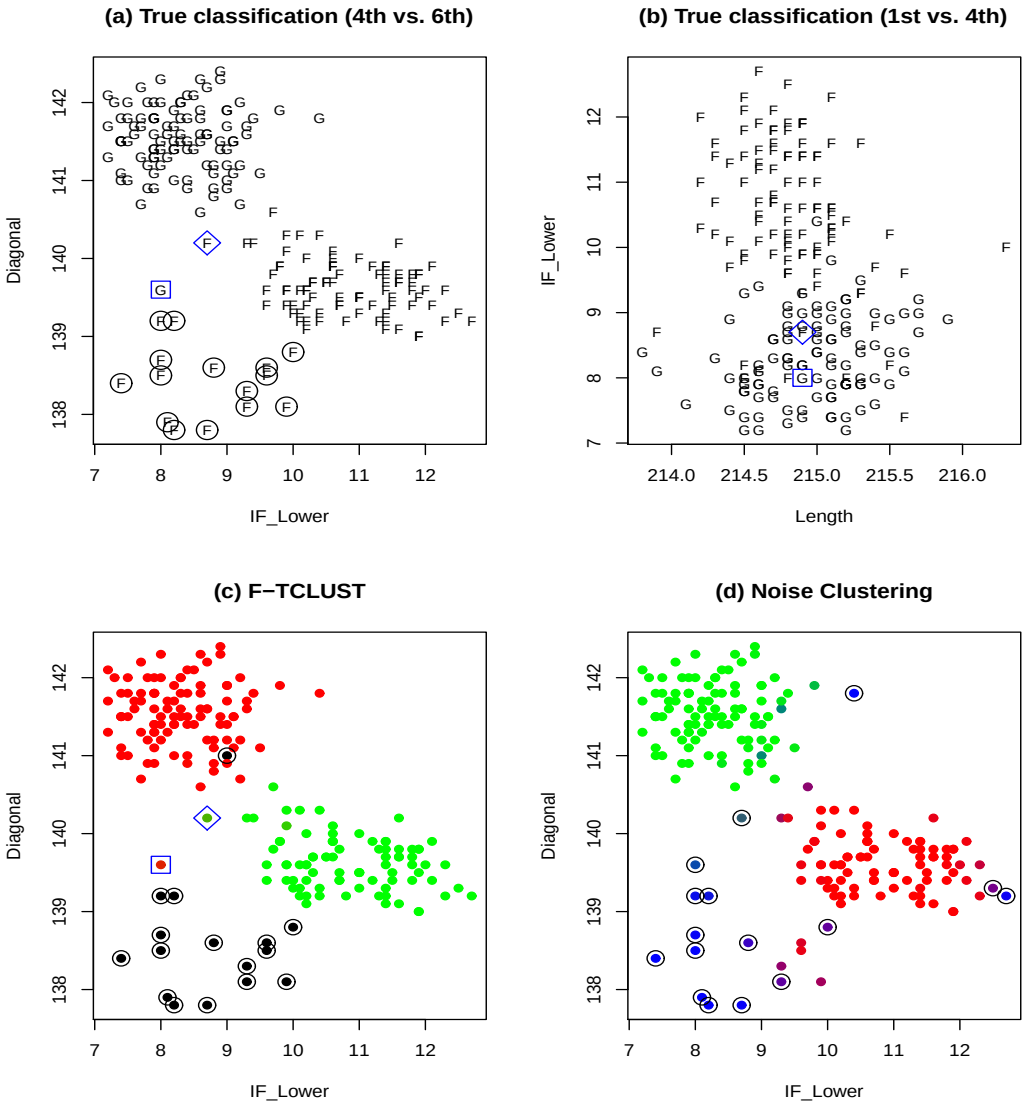

Figure 8: A scatterplot of the fourth and sixth variables of the "Swiss Bank Notes" data set (a) and the first against the fourth (b) with "G" standing for genuine bills and "F" for the forged ones. 15 "anomalous" forged bills are surrounded by circles in (a). The F-TCLUST clustering results are summarized in (c) where trimmed observations are surrounded by circles. The result of the "noise clustering" appears in (d) where observations assigned to the "noise cluster" are surrounded by circles.

Figure 8,(c) shows the F-TCLUST clustering results with parameters $k = 2$, $\alpha = 0.08$, $m = 1.3$ and $c = 10$. Membership values are summarized by using a mixture of "red" and "green" colors and the trimmed observations are shown as "black points" surrounded by circles. Taking into account the prior knowledge of the existence of $15 + 1$ "anomalous" bills, we have chosen $\alpha = 0.08$ (which yields $200 \cdot 0.08 = 16$ trimmed observations). We can see that the genuine bills are clustered into the "red" cluster and the forged ones into the "green" cluster. Apart from one wrongly trimmed genuine bill, the F-TCLUST trims those 15 forged bills listed as "anomalous" in [11].

Although we have used the six variables when applying the F-TCLUST, only two variables are represented in Figure 8,(c).

The two observations with the most "fuzzy" assignments are surrounded by "$\diamond$" and "$\square$" symbols in Figure 8,(a), (b) and (c). The observation corresponding to a forged bill surrounded by the "$\diamond$" symbol has a membership value of 0.703 to the cluster including the forged bills, and 0.297 to the cluster with the genuine bills. We can see in Figure 8,(b) that its assignment decision is not straightforward because, although it is a forged bill, it has values in the variable "Distance of inner frame to the lower border" more compatible with those corresponding to the genuine bills. The observation surrounded by a "$\square$" symbol is the previously commented non-typical genuine bill that had already been reported in [11]. This bill has a membership value of 0.871 to the cluster made up of genuine bills, while the rest of the genuine bills have membership values close to 1.

Figure 8,(d) shows the fuzzy clustering results obtained when applying the "noise clustering" approach [6] with $k = 2$ groups and $m = 1.3$. A mixture of 3 colors (red, green and blue) is used to represent the membership values of the 2 clusters and the "noise cluster". The "blue" color depends on the membership values corresponding to the "noise cluster". The value of the noise distance parameter δ has been chosen in such a way that exactly 16 observations are considered noisy ones. If u_{i3} is the membership value of observation x_i with respect to the "noise cluster", we consider that x_i is a noisy observation whenever $u_{i3} > u_{i1}$ and $u_{i3} > u_{i2}$. The 16 observation declared as noisy ones are shown surrounded by circles in Figure 8,(d). Although the "noise clustering" approach provides very sensible clustering results (discovering the two main groups of forged and genuine bills and most of these 15 "anomalous" forged bills), it does not exactly recover the list of 15 "anomalous" forged bills that were listed in [11].

Although we have considered a pre-fixed trimming level $\alpha = 0.08$, we

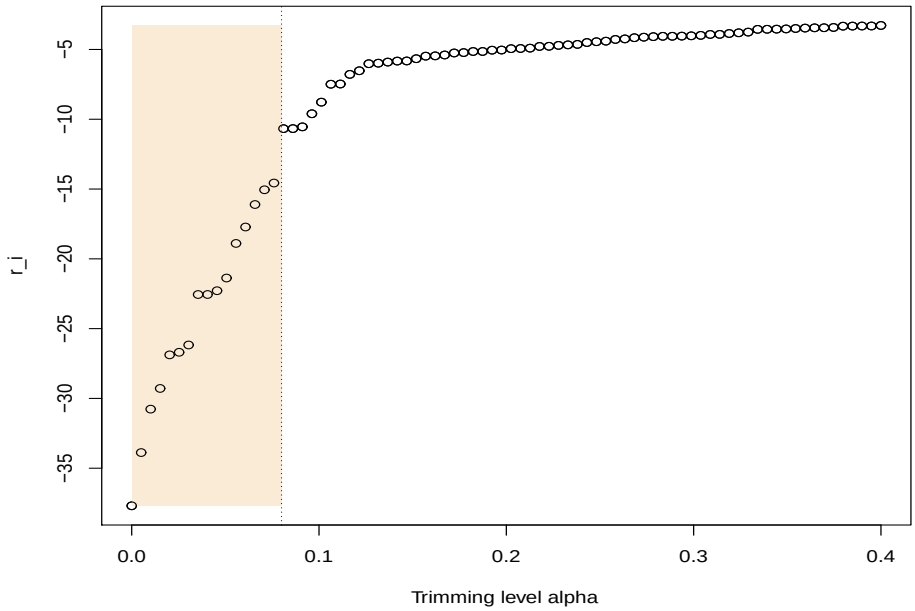

Figure 9: Plot of the r_i values when choosing $\alpha = 0.08$ for the "Swiss Bank Notes" data set.

can also see that a choice of α close to this value could be justified through the examination of Figure 9. Since we are quite close to a "high contrast" clustering output, no particular change of scale seems to be needed for this data set. A large value of $c = 10$ is chosen for the eigenvalue ratio constraint because there are no reasons for being exclusively interested in only detecting spherical clusters.

7. Conclusions and future research lines

In this paper, we have presented the so-called F-TCLUST robust fuzzy clustering approach which is based on a "maximum-likelihood" principle. The possibility of trimming a fixed proportion α of observations (self-determined by data) is also considered. An eigenvalue ratio constraint on the eigenvalues of the scatter matrices controls the relative shape and size of the clusters and serves to avoid the detection of spurious clusters. A computationally feasible algorithm is proposed which does not notably increase the computing time with respect to other similar algorithms in the literature.

The proposed methodology has a high flexibility by allowing very different choices of the tuning parameters involved in its statement. Although some

future research is clearly needed to make these selections easier for the user, the role played by each of these parameters has been explained and some general guidelines for their choices have been given.

Possibilistic fuzzy clustering methods [23] are also well-suited for addressing the problem of noisy data in fuzzy clustering. These methods are based on relaxing the constraint $\sum_{j=1}^{k} u_{ij} = 1$ in such a way that outlying observations could have arbitrarily low membership values for all the clusters. It is well-known that these approaches tend to produce coincident clusters when $k > 1$ and, therefore, it is better to see them more as "mode-seeking" procedures than as "partitioning" ones [1, 24]. The F-TCLUST method may undoubtedly be viewed as a "partitioning" procedure when equal weights are assumed (by using (1)) but it also has a certain "mode-seeking" behavior when allowing different weights (by using (3)). Note that, instead of finding coincident groups as possibilistic clustering methods do, the F-TCLUST can find clusters with weights p_j close to 0 when the chosen value of k is larger than needed (Figure 6,(b)).

As was also commented, a preventive (higher than needed) trimming level has no disastrous effect in the clustering results (Figure 8,(b)). [22] also noticed this fact for another fuzzy clustering method with a fixed trimming proportion, and he also showed the dangerous effect that a slightly decreased noise distance δ could have in "noise clustering" methods.

As a future promising research line, it could be interesting to evaluate the performance of the "classification trimmed-likelihood curves" [16] in the fuzzy clustering set up. This approach is based on the graphic representation of the maximum values attained by the target function (3) when moving parameters α and k. For instance, Figure 10 shows the classification trimmed-likelihood curves obtained when $k = 1, 2, 3$ and 4 and $\alpha \in [0, 0.3]$ when $m = 1.3$ and $c = 50$. Although a detailed explanation of how these curves may be interpreted is not given here, by following the interpretation of these curves as in [16], we could see that $k = 3$ is a good choice for the number of groups when $\alpha = 0$. We can also see that $k = 2$ is a good choice when the trimming level $\alpha = 0.1$ allows 10% of noise to be discarded. In any case, by examining these curves, there is no point in increasing k from 3 to 4.

As commented in Section 5, all parameters α, k and c can be seen as related (e.g., a fixed k would imply some specific α and the other way around). The use of the classification trimmed-likelihood curves may be a useful tool for helping users to make sensible simultaneous choices for all these parameters.

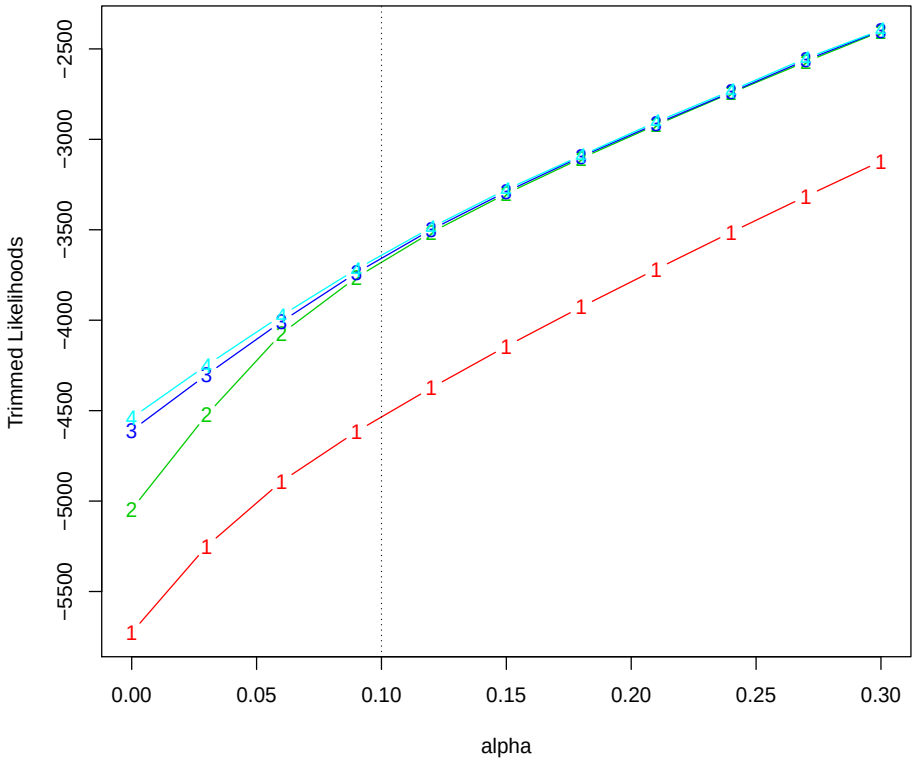

Figure 10: Classification trimmed likelihood curves for the data set in Figure 1 when $k = 1, 2, 3$ and 4 and $\alpha \in [0, 0.3]$.

References

[1] Barni, M., Cappellini, V. and Mecocci, A. (1996), "Comments on A Possibilistic Approach to Clustering," *IEEE Transactions on Fuzzy Systems*, **4**, 393-396.

[2] Bezdek, J.C. (1981), *Pattern Recognition with Fuzzy Objective Function Algoritms*, Plenum Press, New York.

[3] Borgelt, C. and Kruse, R. (2005). "Fuzzy and probabilistic clustering with shape and size constraints." In *Proceedings of the 11th International Fuzzy Systems Association World Congress, IFSA05, Beijing, China*, 945950.

[4] Choi, M.-S. and Krishnapuram, R. (1996), "Fuzzy and robust formula-

tion of maximum-likelihood-based gaussian mixture decomposition," , *In IEEE Conference on Fuzzy Systems*, 1899-1905.

[5] Cook R.D. (1999), "Graphical detection of regression outliers and mixtures." In *Proceedings of the International Statistical Institute 1999. ISI*, Finland.

[6] Davé, R.N. (1991). "Characterization and detection of noise in clustering", *Pattern Recognition Letters*, **12**, 657664.

[7] Davé, R.N. and Krishnapuram, R. (1997). "Robust clustering methods: a unified view". *IEEE Transactions on Fuzzy Systems*, **5**, 270-293

[8] Davé, R.N. and Sen, S. (1997). "Noise Clustering Algorithm Revisited, *In Proceedings of the Biennial Workshop NAFIPS 1997, Syracuse*, 199-204.

[9] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, **39**, 138.

[10] Dunn, J.C. (1974). "A fuzzy relative of the ISODATA Process and its use in detecting compact well-separated clusters", *Journal of Cybernetics*, **3**, 32-57.

[11] Flury, B. and Riedwyl, H. (1988), *Multivariate Statistics. A Practical Approach*, Chapman and Hall, London-New York.

[12] Fritz, H., García-Escudero, L.A. and Mayo-Iscar, A. (2012), "A fast algorithm for robust constrained clustering". Preprint available at `http://www.eio.uva.es/infor/personas/algorithm_web.pdf`.

[13] Gallegos, M. and Ritter, G. (2009), "Trimming algorithms for clustering contaminated grouped data and their robustness.", *Advances in Data Analysis and Classification*, **10**, 135167.

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

[15] García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2010). "A review of robust clustering methods", *Advances in Data Analysis and Classification*, **4**, 89-109.

[16] García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2011), "Exploring the number of groups in robust model-based clustering", *Statistics and Computing*, **21**, 585-599.

[17] Gath, I. and Geva, A.B. (1989), "Unsupervised optimal fuzzy clustering.", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 773 781.

[18] Gustafson, E.E. and Kessel, W.C. (1979). "Fuzzy Clustering with a Fuzzy Covariance Matrix". *Proceedings of the IEEE lnternational Conference on Fuzzy Systems, San Diego, 1979*, 761-766.

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

[20] Hathaway, R.J. and Bezdek, J.C. (1993). "Switching regression models and fuzzy clustering". *IEEE Transactions on Fuzzy Systems*, **1**, 195-204.

[21] Kim, J., Krishnapuram, R. and Davé, R. (1996) "Application of the least trimmed squares technique to prototype-based clustering". *Pattern Recognition Letters*, **17**, 633-641.

[22] Klawonn, F. (2004). "Noise clustering with a fixed fraction of noise". *Applications and Science in Soft Computing*. Springer, Berlin-Heidelberg-New York, 133138.

[23] Krishnapuram, R. and Keller, J.M. (1993), "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, **1**, 98-110.

[24] Krishnapuram, R. and Keller, J.M. (1996), "The possibilistic C-means algorithm: Insights and recommandations," *IEEE Transactions on Fuzzy Systems*, **4**, 385393

[25] Łeski, J. (2003). "Towards a robust fuzzy clustering", *Fuzzy Sets and Systems*, **137**, 215-233.

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

[27] Miyamoto, S. and Mukaidono, M. (1997). "Fuzzy c-means as a regularization and maximum entropy approach" *Proceedings of the 7th International Fuzzy Systems Association World Congress (IFSA'97)*, **2**, 86-92.

[28] Neykov, N., Filzmoser, P., Dimova, R. and Neytchev, P. (2007), "Robust fitting of mixtures using the trimmed likelihood estimator", *Computational Statistics and Data Analysis*, **52**, 299-308.

[29] Rehm, F., Klawonn, F. and Kruse, R. (2007). "A novel approach to noise clustering for outlier detection". *Soft Computing*, **11**, 489-494.

[30] Rousseeuw, P.J., Trauwaert, E. and Kaufman, L. (1995). "Fuzzy clustering with high contrast". *Journal of Computational and Applied Mathematics,* **64**, 81-90.

[31] Rousseeuw, P.J., Kaufman, L. and Trauwaert, E. (1996). "Fuzzy clustering using scatter matrices". *Computational Statistics and Data Analysis*, **23**, 135-151.

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

[33] Ruspini, E. (1969). "A new approach to clustering". *Information and Control*, **15**, 22-32.

[34] Trauwaert, E., Kaufman, L. and Rousseeuw, P.J. (1991). "Fuzzy clustering algorithms based on the maximum likelihood principle", *Fuzzy Sets and Systems*, **42**, 213-227.

[35] Wu, K.-L. and Yang, M.-S. (2002). "Alternative c-means clustering algorithms", *Pattern Recognition*, **35**, 2267-2278.

[36] Yang, M.-S. (1993). "On a class of fuzzy classification maximum likelihood procedures" *Fuzzy Sets and Systems* **57**, 365337.

[37] Yang, M.-S. and Wu, K.-L. (2006). "Unsupervised possibilistic clustering", *Pattern Recognition*, **39**, 5-21.