Robust, fuzzy, and parsimonious clustering based on mixtures of Factor Analyzers

Luis Angel García-Escudero[1], Francesca Greselin[2], and Agustin Mayo Iscar[3]

Abstract A clustering algorithm that combines the advantages of fuzzy clustering and robust statistical estimators is presented. It is based on mixtures of Factor Analyzers, endowed by the joint usage of trimming and the constrained estimation of scatter matrices, in a modified maximum likelihood approach. The algorithm generates a set of membership values, that are used to fuzzy partition the data set and to contribute to the robust estimates of the mixture parameters. The adoption of clusters modeled by Gaussian Factor Analysis allows for dimension reduction and for discovering local linear structures in the data. The new methodology has been shown to be resistant to different types of contamination, by applying it on artificial data. A brief discussion on the tuning parameters, such as the trimming level, the fuzzifier parameter, the number of clusters and the value of the scatter matrices constraint, has been developed, also with the help of some heuristic tools for their choice. Finally, a real data set has been analyzed, to show how intermediate membership values are estimated for observations lying at cluster overlap, while cluster cores are composed by observations that are assigned to a cluster in a crisp way.

Key words: Fuzzy clustering; Robust clustering; Unsupervised learning; Factor analysis

Department of Statistics and Operational Research and IMUVA, University of Valladolid (Spain) lagarcia@eio.uva.es · Department of Statistics and Quantitative Methods, Milano-Bicocca University (Italy) francesca.greselin@unimib.it · Department of Statistics and Operational Research and IMUVA, University of Valladolid (Spain) agustin@med.uva.es

1 Introduction

Fuzzy clustering is a method of data analysis and pattern recognition which allocates a set of observations to clusters in a "fuzzy" way, more formally, constructs a "membership" matrix whose (i, g)-th element represents "the degree of belonging" of the i-th observation to the g-th cluster.

In this sense, the clusters are "fuzzy sets" as defined in [50]. In our case, the fuzzy sets will be based on mixtures of Gaussian factor analyzers, hence they have a strong probabilistic meaning and our rules of operation are not those proposed by Zadeh but those that come naturally from probability theory.

Factor analysis is a widely employed method to be used when, as it happens in many phenomena, several observed variables could be explained by a few unobserved ones, exploiting their correlations. It is a powerful method, whose scope of application is unfortunately limited by the assumption of global linearity for representing data. To widen its applicability, [23] and [28] introduced the idea of combining one of the basic form of dimensionality reduction - factor analysis - with a basic method for clustering - the Gaussian mixture model, thus arriving at the definition of mixtures of factor analyzers. At the same time, [44, 45] and [3] considered the related model of mixtures of principal component analyzers, for the same purpose. Further references may be found in [36] (chapter 8). Factor analysis is related to principal component analysis (PCA) [45]; however, these two methods have many conceptual and algorithmic differences, with the most significant being that the former aims at modelling *correlation* between variables and searches for underlying linear structures, the second focuses on their *variances* and identifies a reduced set of linear transformations of variables, maximizing their variance.

With reference to other techniques for data reduction, we want in particular to mention [6], where it has been shown that the principal components corresponding to the larger eigenvalues do not necessarily contain information about group structure. Therefore, data reduction and clustering separately may not be a good idea. Data reduction can ameliorate clustering and classification results, but combining variable selection *and* clustering can give improved results. Mixtures of factor analysers are designed exactly for simultaneously performing clustering and dimension reduction. Using this approach, we aim at finding local linear models in clusters, that could be particularly useful for datasets with a high number of observed variables. Beyond its parsimony, this method often provides a better insight into the structure of the original data.

In real datasets, noise and outliers contaminate the data collection, and any assumed model is only an approximation to reality. Unfortunately, one single outlier - if located sufficiently far away - can completely ruin the results of many (hard and fuzzy) clustering methods, in the sense that at least one of the component parameters estimate can be arbitrarily large. Remarkably, the fuzzy community has been traditionally very interested in robustness issues in Cluster Analysis (see, e.g., the large amount of references in the two

review papers [11, 1]). In fuzzy clustering, the aim is to obtain a collection of membership values $u_{ig} \in [0,1]$ for all $i = 1 \dots n$ and $g = 1 \dots G$, where increasing degrees of membership are meant by higher values of the u_{ig}. We may understand why robustness in so critical in fuzzy clustering when we observe that for an outlying observation $\mathbf{x}_i$, generally lying far from the G clusters, we could obtain $u_{ig} \sim 1/G$, while we would expect $u_{ig} \sim 0$ for $g = 1, \dots, G$, to convey a very low plausibility to belong to any cluster. In addition, if one outlying observation $\mathbf{x}_i$ is placed in a very distant position but closer to cluster g than the others, then typically $u_{ig} \sim 1$ and $\mathbf{x}_i$ would influence heavily the parameters' estimation of cluster g.

Hence we put forward a robust methodology based on trimming, to identify outliers and to assign them zero membership values. In our approach, we will indicate observation $\mathbf{x}_i$ as fully trimmed if $u_{ig} = 0$ for all $g = 1, \dots, G$ and, thus, this observation has no membership contribution to any cluster. The proposed idea trace back to the seminal paper [9], and is called *impartial trimming*, or trimming self-determined by the data. We look for a method having a reasonably good efficiency (accuracy) at the assumed model; for which small deviations from the model assumptions should impair the performance only by a small amount; and for which larger deviations from the model assumptions should not cause a catastrophe [29].

Among the several methodologies for robust fuzzy clustering, some well-known approaches are the "noise clustering" in [10], the use of more robust discrepancy measures [48, 34] and the "possibilistic" clustering in [33]. References concerning robustness in hard (crisp) clustering can be found in [19, 14, 38].

An interesting robust proposal have been introduced in [7], where the Student's t-mixture (instead of the Gaussian), is exploited for modeling factors and errors, allowing for outlier downweighting during the model fitting procedure. Alternatively, outliers can be accomodated in the model by considering additional mixture components. It turns out that the two alternatives, while adding stability in the presence of outliers of moderate size, do not possess a substantially better breakdown behavior than estimation based on normal mixtures [27]. In other words, one only observation could completely breakdown the estimation. Hence a model-based alternative with good breakdown properties is still missing in the literature.

On the other hand, a further issue inherent to the choice of Gaussian Mixtures (GM) is originated by the unboundedness of the likelihood, turning the maximization of the objective function into an ill-posed problem. Therefore, following a wide literature stream, based on [26, 30, 18, 24], we will adopt a constrained estimation of the component covariances, to avoid singularities and to reduce the appearance of spurious solutions. In particular, the joint use of trimming and constrained estimation for mixtures of factor analyzers (MFA) have previously been studied for non fuzzy clustering in [21]. The advantage of MFA with respect to GM is due to the reduction of the number of free parameters to be estimated. In each component, a reduced number

of latent variables, called *factors* and lying in a lower-dimensional subspace, explain the observed variables, through a linear submodel. Therefore, there is scope for a robust fuzzy clustering approach, along the lines of [16], and [12] for fuzzy robust clusterwise regression.

The original contribution of the present paper is to combine: *i*) robust estimation techniques; *ii*) fuzzy clustering; *iii*) dimension reduction through factor analysis; *iv*) the flexibility of having hard and soft assignments for units (the "hard contrast" property introduced in [39]). The interplay between these four features provides a novel powerful model with a parsimonious parametrization, specially useful in higher dimensional fuzzy clustering problems.

The outline of the paper is as follows. In Section 2 we introduce a fuzzy version of robust mixtures of Gaussian MFA, considering how to implement fuzziness, trimming and constrained estimation. Section 3 presents an efficient algorithm for its estimation. Afterwards, in Section 4, we present results on synthetic data to show the effectiveness of the proposal. A brief discussion on the role of the parameters involved in the fuzzy robust procedure, such as the trimming level, the fuzzifier parameter, the number of clusters and the value of the scatter matrices constraint, has been developed, also with the help of some heuristic tools for their tuning. An application to real data is provided in Section 5, also in comparison with competing methods. Conclusions and further work are delineated in Section 6.

2 Mixtures of Gaussian factors analyzers for fuzzy clustering

Let $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ denote a random sample of size n, composed by random vectors taking values in $\mathbb{R}^p$. The random sample could arise from medical records, aerial photos, market trends, library catalogs, galactic positions, fingerprints, cash flows, chemical constituents, demographic therefore several applications could be benefited by this approach. Our goal is the identification of a probabilistic model which can provide a reasonable explanation of the process generating the data. Supposing that, as it happens in many phenomena, the p observed variables could be explained by a few unobserved ones, we adopt Factor Analysis. Factor analysis allows to summarize the variability between a number of correlated features, through a much smaller number of unobservable, hence named *latent*, factors. Under this approach, each single variable (among the p observed ones) is assumed to be a linear combination of d underlying common factors, with an accompanying error term accounting for that part of the variability which is unique to it (not in common with other variables). Considering a mixture of Gaussian factors, the distribution of $\mathbf{X}_i$ is given as

$$\mathbf{X}_i = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{U}_{ig} + \mathbf{e}_{ig} \quad \text{with probability } \pi_g \; i = 1, \ldots, n, \; g = 1, \ldots, G, \; (1)$$

where $\boldsymbol{\mu}_g$ denotes the mean, $\boldsymbol{\Lambda}_g$ is a $p \times d$ matrix of *factor loadings*, and the *factors* $\mathbf{U}_{1g}, \ldots, \mathbf{U}_{ng}$ are $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ distributed independently of the *errors* $\mathbf{e}_{ig}$. The errors are independent identically distributed $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$, with a $p \times p$ diagonal matrix $\boldsymbol{\Psi}_g$. In factor analysis the observed variables are independent given the factors: hence the diagonality of $\boldsymbol{\Psi}_g$. Note that the factor variable $\mathbf{U}_{ig}$ models correlations between the elements of $\mathbf{X}_i$, while the errors $\mathbf{e}_{ig}$ account for independent noise for $\mathbf{X}_i$, in group g. The π_g, called mixing proportions, are non-negative and sum up to 1. We suppose that $d < p$.

Unconditionally, the density of each observation $\mathbf{X}_i$ is a mixture of G normal densities, in proportions $\pi_1, \ldots, \pi_G$, as

$$f_{MFA}(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \quad (2)$$

where $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g, \text{ for } g = 1, \ldots, G. \quad (3)$$

Therefore, mixtures of Gaussian factors concurrently perform clustering and, within each cluster, local dimensionality reduction. These models should be especially useful in high-dimensional problems. For fixed q, the number of parameters grows linearly in dimension p, as their covariances have $pq - q(q-1)/2 + p$ free parameters [36]. A more detailed discussion of this topic is provided in a final remark of this Section.

Our robust fuzzy method is based on a maximum likelihood criterion defined on a specific underlying statistical model, as in many other proposal in the literature. The next three steps are to modify the classical model in (2), to incorporate fuzzy belonging of the units, impartial trimming for outlier handling, and constrained estimation of the component scatters. Hence, given a trimming proportion $\alpha \in [0, 1)$, a constant $c \geq 1$ and a value of the fuzzifier parameter $m > 1$, a robust constrained fuzzy clustering is obtained via the maximization of the following objective function

$$\sum_{i=1}^{n} \sum_{g=1}^{G} u_{ig}^m \log \phi(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g). \quad (4)$$

Fuzziness: Soft membership values $u_{ig} \geq 0$ in (4) allow for overlapping clusters, and are assumed to satisfy

$$\sum_{g=1}^{G} u_{ig} = 1. \tag{5}$$

Trimming: To incorporate impartial trimming, we modify (5) into

$$\sum_{g=1}^{G} u_{ig} = 1 \quad \text{if} \quad i \in \mathcal{I} \quad \text{and} \quad \sum_{g=1}^{G} u_{ig} = 0 \quad \text{otherwise,} \tag{6}$$

for a subset $\mathcal{I}$ that identifies the "regular" units, that are the more plausible ones, under the currently estimated model. Specifying a fixed proportion of observations to be trimmed has been also considered for robust fuzzy clustering purposes in [31, 32].

Constrained estimation: In (4), whenever $\mathbf{x}_1 = \boldsymbol{\mu}_1$, setting $u_{11} = 1$, and taking a sequence of scatter matrices $\boldsymbol{\Psi}_1$ such that $|\boldsymbol{\Psi}_1| \to 0$, the likelihood become unbounded. This is a recurrent issue in Cluster Analysis, when general scatter matrices are allowed for the clusters. This fact has been already noticed in fuzzy clustering, among other authors, by [25] where the authors propose to fix the $|\boldsymbol{\Sigma}_g|$ values in advance. In our approach, the unboundedness problem is addressed by constraining the ratio between the largest and smallest eigenvalues, i.e. the diagonal elements $\{\psi_{gl}\}_{l=1,\ldots,p}$ of the noise matrices $\boldsymbol{\Psi}_g$, as follows

$$\psi_{g_1 l_1} \leq c \ \psi_{g_2 l_2} \qquad \text{for every } 1 \leq g_1 \neq g_2 \leq G \text{ and } 1 \leq l_1 \neq l_2 \leq p. \tag{7}$$

The constant $c \geq 1$ is finite, and a larger value of c leads to a less constrained fuzzy clustering approach. Geometrically, c controls the maximal ratio among the lengths of the axes of the equidensity ellipsoids of the errors $\mathbf{e}_{ig}$ to be smaller than $\sqrt{c}$. Note that we are simultaneously controlling departures from sphericity and differences among the scatters of the component errors, by means of (7). This constraint can be seen as an adaptation to mixtures of Gaussian factors of those introduced in [30, 18], and is similar to the mild restrictions implemented for the model considered in [24]. They all go back to the seminal paper [26].

It is well known that an objective function like the one in (4) is inclined to provide clusters with similar "sizes", say having similar values of $\sum_{i=1}^{n} u_{ig}^m$. If this effect is not desired, it is better to replace the previous objective function by

$$\sum_{i=1}^{n} \sum_{g=1}^{G} u_{ig}^m \log\left[p_g \phi(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right], \tag{8}$$

where $p_g \in [0,1]$ and $\sum_{g=1}^{G} p_g = 1$ are weights that have to be taken into account in the maximization of the target function. Differentiating the target function with respect to p_g and equating it to 0, it is trivial to see that, once the u_{ig} membership values are known, the weights p_g are optimally determined as

$$p_g = \sum_{i=1}^{n} u_{ig}^m / \sum_{i=1}^{n} \sum_{g=1}^{G} u_{ig}^m. \tag{9}$$

The introduction of the p_g weights in the target function for fuzzy clustering is done in [49] (see display (2) in the cited work), but it goes back to [43] in hard 0-1 clustering. The idea appeared under the name of "penalized classification likelihoods" in [4], where Bryant motivated the usage of the p_g weights to determine which components are really present, or not, in the mixture: when the number of groups G is larger than needed, some p_g may be set close to 0 along the estimation. The interested reader can find a more detailed explanation of this last claim in [16] (see subsection "Weights and number of clusters" within Section 5) and in [20], where this approach is further developed and illustrated in the fuzzy clustering framework and in the hard clustering case, respectively.

We conclude this section by commenting on the parsimony gained when adopting MFA, in comparison to GM, measured in terms of the reduction of the number of free parameters to be estimated. For each of the G component of the mixture of Factor Analysers, we have to estimate the p-dimensional component means $\boldsymbol{\mu}_g$, the $p \times d$ matrices $\boldsymbol{\Lambda}_g$ of factor loadings, and the diagonal p-dimensional noise matrices $\boldsymbol{\Psi}_g$, along with the mixing proportions π_g $(g = 1, \ldots, G-1)$, on putting $\pi_G = 1 - \sum_{i=1}^{G-1} \pi_g$. Note that in the case of $d > 1$, there is an infinity of choices for $\boldsymbol{\Lambda}_g$, since model (1) is still satisfied if we replace $\boldsymbol{\Lambda}_g$ by $\boldsymbol{\Lambda}_g \mathbf{H}'$, where $\mathbf{H}$ is any orthogonal matrix of order q. As $d(d-1)/2$ constraints are needed for $\boldsymbol{\Lambda}_g$ to be uniquely defined, the number of free parameters, for the g-th covariance matrix of the mixture, is $p + pd - \frac{1}{2}d(d-1)$. In the case of Gaussian Mixture (GM) model-based approach, each data point $\mathbf{x}$ is taken to be a realization of the mixture probability density function, as defined in (2), but with full covariance matrices $\boldsymbol{\Sigma}_g$ having $\frac{1}{2}p(p+1)$ parameters. To measure the gained parsimony, we observe that the reduction in the number of free parameters to be estimated in MFA with respect to GM is equal to

$$G\left[\frac{1}{2}p(p+1) - (p + pd - \frac{1}{2}d(d-1))\right] = \frac{G}{2}\left[(p-d)^2 - (p+d)\right]. \tag{10}$$

Note that (10) can be large when $p >> d$.

3 An efficient algorithm for fuzzy robust clustering based on mixtures Gaussian factor analyzers

The algorithm for implementing the robust estimator is presented in detail in this section. We deal, more precisely, with a fuzzy Classification version of the Alternating Expectation - Conditional Maximization algorithm (AECM), which is an extension of the EM algorithm, suitable for the factor structure of the model. It is based on a partition of the parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$, because the target function is easy to be maximized for $\boldsymbol{\theta}_1 = \{u_{ig}, p_g, \boldsymbol{\mu}_g, g = 1, \ldots, G\}$ given $\boldsymbol{\theta}_2 = \{u_{ig}, p_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g,\ g = 1, \ldots, G\}$ and viceversa. At each stage of the iterations, the optimal membership values are obtained first, then parameters are updated by maximizing the target function on them, along the lines of [25] or [41], where fuzzy clustering is performed with general covariance matrices. The algorithm should be initialized several times and run till convergence (or till reaching a maximum number of iterations), retaining the best solution in terms of the objective function.

1. *Initialization:* Each iteration begins by selecting initial values for $\boldsymbol{\theta}^{(0)}$ where $\boldsymbol{\theta}^{(0)} = \{u_{ig}^{(0)}, p_g^{(0)}, \boldsymbol{\mu}_g^{(0)}, \boldsymbol{\Lambda}_g^{(0)}, \boldsymbol{\Psi}_g^{(0)}; g = 1, \ldots, G\}$. To this aim, following a common practice in robust estimation, we draw G small random subsamples from the data and fit a Gaussian factor analysis model on each of them. In more detail, $p + 1$ units are randomly selected (without replacement) for group g from the observed data set. The obtained subsample, denoted by $\mathbf{x}^g$, is arranged in a $(p + 1) \times p$ matrix, and the vector with the columnwise means are the initial $\boldsymbol{\mu}_g^{(0)}$. We may consider model (1) in group g as a regression of $\mathbf{x}^g$ with intercept $\boldsymbol{\mu}_g$, regression coefficients $\boldsymbol{\Lambda}_g$, explanatory variables given by the latent factors $\mathbf{U}_{ig}$, and regression errors $\mathbf{e}_{ig}$. The missing information here are the factors which, under the assumptions for the model, are realizations from independently $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ distributed random variables. Hence, we draw $p + 1$ random independent observations from the d-variate standard Gaussian, to fill a $(p+1) \times d$ matrix $\mathbf{u}^g$. Then we set $\boldsymbol{\Lambda}_g^{(0)} = ((\mathbf{u}^g)'\mathbf{u}^g)^{-1}(\mathbf{u}^g)'\mathbf{x}_c^g$, where $\mathbf{x}_c^g$ is obtained by centering the columns of $\mathbf{x}^g$. To provide a restricted random generation of $\boldsymbol{\Psi}_g$, the $(p + 1) \times p$ matrix $\varepsilon_g = \mathbf{x}_c^g - \boldsymbol{\Lambda}_g^{(0)}\mathbf{u}^g$ is computed, and the diagonal elements of $\boldsymbol{\Psi}_g^{(0)}$ are set equal to the variances of the p columns of the ε_g matrix. After repeating these steps for $g = 1, \ldots, G$, if the obtained matrices $\boldsymbol{\Psi}_g^{(0)}$ do not satisfy the required inequalities (7), the constrained maximization described in step 2.2.3 must be applied. Afterwards, $p_1^{(0)}, ..., p_G^{(0)}$ in the interval $(0, 1)$ and summing up to 1 are randomly chosen. Finally, each initial $u_{ig}^{(0)}$ membership value is obtained by raising at power m the posterior probability that the unit i belongs to the component g when considering the fitted model for those initial parameters.

2. *Iterative steps:* The following steps 2.1–2.6. are alternatively executed until a maximum number of iterations, `MaxIter`, is reached.

2.1 *First cycle*: $\boldsymbol{\theta}_1 = \{u_{ig}, p_g, \boldsymbol{\mu}_g, g = 1, \ldots, G\}$. We have to maximize the objective function (8) over $\boldsymbol{\theta}_1$, with $\boldsymbol{\Lambda}_g, \boldsymbol{\Sigma}_g \in \boldsymbol{\theta}_2$ held fixed at their previous evaluation.

2.1.1 *Membership values:* Based on the current parameters, and setting $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$, if

$$\max_{1 \le g \le G} \; p_g \; \phi(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \ge 1,$$

then we make a *hard* assignment:

$$u_{ij} = I\left\{ p_j \; \phi(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \max_{1 \le g \le G} \; \left[p_g \; \phi(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right] \right\}$$

where $I\{\cdot\}$ is the 0-1 indicator function, which takes value 1 if the expression within the brackets holds; otherwise we make a *fuzzy* assignment:

$$u_{ij} = \left[\left(\frac{\log\left[p_j \; \phi(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right]}{\sum_{g=1}^{G} \log\left[p_g \; \phi(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right]} \right)^{\frac{1}{(m-1)}} \right]^{-1}.$$

2.1.2 *Trimming:* Evaluated the n quantities

$$r_i = \sum_{g=1}^{G} u_{ig}^m \log\left(p_g \phi(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right), \tag{11}$$

we sort them to obtain $r_{(1)} \le \ldots \le r_{(n)}$, and their α-quantile $r_{([n\alpha])}$. To incorporate impartial trimming, we modify the membership values as

$$u_{ig} = 0 \text{ for every } g = 1, ..., G \text{ whenever } i \notin \mathcal{I}, \tag{12}$$

for $\mathcal{I} = \left\{ i : r_{(i)} \ge r_{([n\alpha])} \right\}$.

2.1.3 *Update parameters:* Given the u_{ig} membership values obtained in the previous step, then the p_g weights are optimally determined as (9). The component location parameters $\boldsymbol{\mu}_g$ are obtained as weighted means, with weights u_{ig}^m, for $g = 1, \ldots, G$, as

$$\boldsymbol{\mu}_g = \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{x}_i}{\sum_{i=1}^{n} u_{ig}^m}.$$

2.2 *Second cycle*: $\boldsymbol{\theta}_2 = \{u_{ig}, p_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g,\ g = 1, \ldots, G\}$, and we have to maximize the objective function (8) over $\boldsymbol{\theta}_2$, with $\boldsymbol{\theta}_1$ held fixed at their previous evaluation.

2.2.1 *Membership values:* as in step 2.1.1

2.2.2 *Trimming:* as in step 2.1.2

2.2.3 *Update parameters:* After re-evaluating the optimal weights p_g from the updated membership values, we have to obtain optimal $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$. With some matrix algebra, and setting

$$\boldsymbol{\gamma}_g = \boldsymbol{\Lambda}_g'(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g)^{-1},$$

it is straightforward to obtain the updated ML-estimates

$$\begin{aligned}\boldsymbol{\Lambda}_g &= \boldsymbol{S}_g\boldsymbol{\gamma}_g{}'[\boldsymbol{\gamma}_g\boldsymbol{S}_g\boldsymbol{\gamma}_g{}' + \mathbf{I}_q - \boldsymbol{\gamma}_g\boldsymbol{\Lambda}_g]^{-1}\\ \boldsymbol{\Psi}_g &= \operatorname{diag}\left\{\boldsymbol{S}_g - \boldsymbol{\Lambda}_g\boldsymbol{\gamma}_g\boldsymbol{S}_g\right\}\end{aligned}$$

where $\boldsymbol{S}_g$ denotes the weighted sample scatter matrix in group g:

$$\boldsymbol{S}_g = \frac{\sum_{i=1}^n u_{ig}^m \left(\mathbf{x}_i - \boldsymbol{\mu}_g\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_g\right)'}{\sum_{i=1}^n u_{ig}^m}.$$

During the iterations, due to the updates, it may happen that the matrices

$$\boldsymbol{\Psi}_g = \operatorname{diag}\left\{\boldsymbol{S}_g - \boldsymbol{\Lambda}_g\boldsymbol{\gamma}_g\boldsymbol{S}_g\right\} = \operatorname{diag}\left(\psi_{g1}, ..., \psi_{gp}\right)$$

do not satisfy the required constraint (7). In this case, we set

$$[\psi_{gk}]_t = \min\left\{c \cdot t,\, \max(\psi_{gk}, t)\right\}, \quad \text{for } g = 1, \ldots, G;\ k = 1, \ldots, p,$$

and fix the optimal threshold value t_{opt} by minimizing the following real valued function:

$$f(t) = \sum_{g=1}^{G} p_g \sum_{k=1}^{p} \left(\log\left([\psi_{gk}]_t\right) + \frac{\psi_{gk}}{[\psi_{gk}]_t}\right). \tag{13}$$

As done in [15, 16], t_{opt} can be obtained by evaluating $2pG+1$ times $f(t)$ in (13). Thus, $\boldsymbol{\Psi}_g$ is finally updated as

$$\boldsymbol{\Psi}_g = \operatorname{diag}\left([\psi_{g1}]_{t_{\text{opt}}}, ..., [\psi_{gp}]_{t_{\text{opt}}}\right), \tag{14}$$

where p_g is defined as in (9). It is worth remarking that the given constrained estimation provides, at each step, the parameters $\boldsymbol{\Psi}_g$ that maximize the target function in the constrained parameter space.

3 *Evaluate target function:* After performing the iterations, the value of the target function (8) is evaluated. If convergence has not been achieved before reaching the maximum number of iterations, `MaxIter`, the results are discarded.

The set of parameters yielding the highest value of the target function (among the multiple runs) and the associated membership function u_{ig} are returned as the final output of the algorithm. Notice, in passing, that a high number of initializations and a high value for `MaxIter` are seldom required, as it will be seen in Section 4.

Some remarks featuring the algorithm follow. In our initialization strategy, the initial values for the parameters in each population are based on small subsamples, aiming at ideally covering, in an increasing number of trials, the full parameter space. Our proposal is based on the idea that a small subsample has to be ideally drawn for each group and then the information extracted from the subsample is completed with random data generated under the model assumptions. This "free" exploration of the parameter space may also lead to spurious solutions or even singularities along the iterations, and the constraint on the scatter matrices plays the role of protecting against those undesired solutions. By considering many random initializations, we are confident that the optimal parameters, in terms of the target function, can be approached inside the restricted parameter space. The number of random initializations should increase with the number of groups G, the dimension p and in the case of very different group sizes.

The proof of the optimality of the membership values adopted in step 2.1.1 can be found in [16]. It is also worth remarking that the usual monotone convergence of the target function in the proposed fuzzy robust AECM algorithm holds true when incorporating trimming and constrained estimation for $\boldsymbol{\Psi}_g$ matrices. To prove this, notice that, when performing the trimming step, the optimal observations have been retained, i.e. the ones with the highest contributions to the objective function. The set $\mathcal{I}$ identifies the "regular" units, that are the more plausible ones, under the currently estimated model. Secondly, the optimality of the estimation of the parameters $\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ has been shown in [21]. Being the target function monotonically non-decreasing along the iterations, the algorithm allows one to find a local maximum. By considering several random initializations, we aim at reaching the global maximum. Finally, we want to remark the crucial effects of trimming, for robustness purposes. By setting $u_{i1} = \ldots = u_{iG} = 0$ for all $i \notin \mathcal{I}$, and improving the model estimation along the iterations, the more outlying observations do not contribute to the parameter estimate in the target function (4). This type of trimming is the basis of the "concentration steps" applied in high-breakdown robust methods [40].

4 Numerical results on artificial data

We present here some experiments on synthetic data, to show the performance of the proposal. Our methodology has been developed to be as general as possible; its generality comes at the cost of having to set in advance four quantities: the number of clusters G, the fuzzifier parameter m, the trimming level α and the value of the constant c for the eigenvalue ratio of the noise matrices. Along with our numerical experiments, we provide a first discussion on the combined effects of these parameters, as well as on their tuning.

We choose a two component population in $\mathbb{R}^7$, from which we draw two samples. Aiming at providing a plot of the obtained results, we looked at a model reducing the original 7-dimensional variables into 2 unidimensional factors (otherwise we could not find a unique bivariate space, for the two components, to represent the data). The first two variables in the first group of observations $\mathcal{X}_1$ are defined as follows:

$$X_{11} \sim \mathcal{N}(0,1) + 4 + 0.5 \cdot \mathcal{N}(0,1) \text{ and } X_{12} \sim 5 * X_{11} - 19 + 3 \cdot \mathcal{N}(0,1);$$

and in the second group of observations $\mathcal{X}_2$ are given as:

$$X_{21} \sim \mathcal{N}(0,1) + 4 + 0.5 \cdot \mathcal{N}(0,1) \text{ and } X_{22} \sim X_{21} + 10 + 2 \cdot \mathcal{N}(0,1).$$

All these "$\mathcal{N}(0,1)$" distributions are taken as independent ones. With this choice, we have:

$$\boldsymbol{\Lambda}_1 = (1,5)', \quad \boldsymbol{\Lambda}_2 = (1,1)', \quad \boldsymbol{\Psi}_1 = \text{diag}(0.25, 9), \quad \boldsymbol{\Psi}_2 = \text{diag}(0.25, 4),$$

and

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.25 & 5 \\ 5 & 34 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.25 & 1 \\ 1 & 5 \end{bmatrix}.$$

After drawing 100 points for each component, to check the robustness of our approach, we add some pointwise contamination $\mathcal{X}_3$ to the data, in the form of 11 points close to the point (4,50), as follows

$$X_{31} \sim 4 + 0.03 \cdot \mathcal{N}(0,1) \qquad X_{32} \sim 50 + 0.03 \cdot \mathcal{N}(0,1);$$

and 11 more points, denoted by $\mathcal{X}_4$, close to the point (6,-20), with

$$X_{41} \sim 6 + 0.03 \cdot \mathcal{N}(0,1) \qquad X_{42} \sim -20 + 0.03 \cdot \mathcal{N}(0,1).$$

Finally, we complement the data matrix, and the contamination, with standard normally distributed random variables $X_{ij} \sim \mathcal{N}(0,1)$ for $j = 3, \ldots, 7$. In this way, we build a dataset where one factor is explaining the correlation among the 7 variables, in each component of the "regular" data, while no linear structure is embedded in the contamination.

Figure 1 shows, in the left panel, a specimen of randomly generated data from the given mixture, while in the right panel it reports results obtained

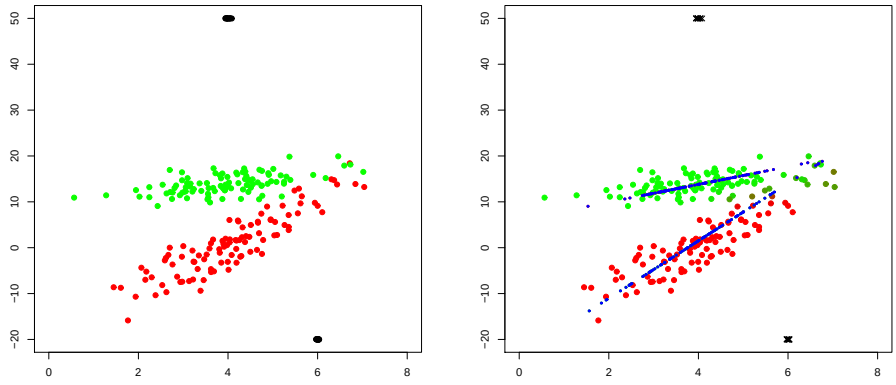


Fig. 1 Left panel: synthetic data (data drawn from $\mathcal{X}_1$ in green, from $\mathcal{X}_2$ in red, and contamination drawn from $\mathcal{X}_3$ and $\mathcal{X}_4$ in black). Only the two first coordinates are plotted. Right panel: Robust fuzzy classification of the synthetic data, with fuzzifier $m = 1.1$. Blue points are the factor scores of the 7-dimensional data in the estimated latent factor space. Black crosses (denoted by "×") are trimmed units. The strength of the membership values is represented by the color usage: a mixture of red and green colors that can turn into brownish colors.

from the proposed robust estimation with $\alpha = 0.1$. We see that the estimation is robust to the most dangerous outliers, like the pointwise contamination (although we have used the seven variables when applying the algorithm, only the first two variables are represented in the plots).

Noise matrices constraint: An important feature of our approach is represented by the constrained estimation of the cluster scatters matrices. To show its effectiveness to discard spurious maximizers, we firstly perform a fuzzy clustering with $c = 100$, allowing some controlled variability within and between clusters eigenvalues, and successively we perform an almost unconstrained estimation, with $c = 10^{10}$. Results for $c = 1$ have been shown in Figure 1. In the right panel of Figure 2, we see that using $c = 10^{10}$ and searching for three clusters, we got a poor solution, with a few observations following a random pattern, clustered apart (in blue).

The fuzzifier parameter m: To observe the effect of the fuzzifier parameter m on the clustering, we perform the robust estimation with $G = 2$ clusters, 30 random initializations, a maximal number of iterations equal to 50, $c = 100$, and $\alpha = 0.1$, and compare the obtained results considering different values for m. From the definition of the fuzzifier, we recall that $m = 1$ provides "hard" clustering through robust Gaussian MFA, as it has been done in a classification EM version of the procedure introduced in [21]. In the $m = 1$

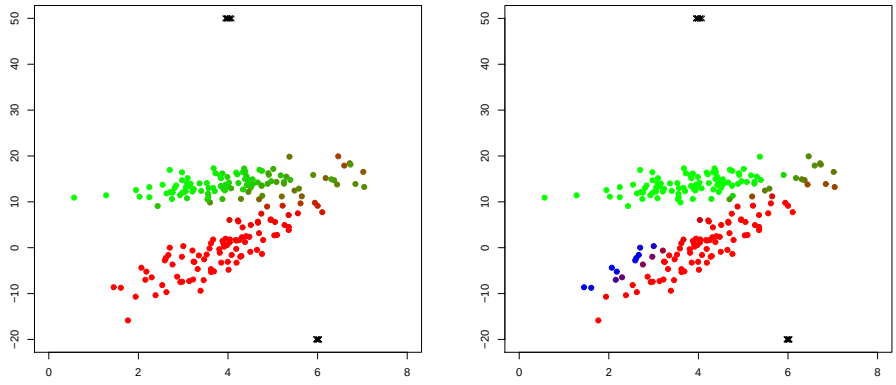


Fig. 2 Left panel: Fuzzy classification of the synthetic data, with $c = 10$ and $G = 2$. Right panel: Almost unconstrained fuzzy classification of the synthetic data, with $c = 10^{10}$ and $G = 3$, generating a spurious solution. Blue points are belonging to a third component.

case, each observation is fully assigned to a cluster or trimmed off, as shown in Figure 3 (a). On the opposite side, we have all the non-trimmed observations shared with approximately equal membership values for larger values of m, like $m = 1.2$, as can be seen in Figure 3 (f). Intermediate values of m yield perhaps more interesting membership values. By taking into account the updating step for the membership values in step 2.1.1 of the algorithm, it is important to see that "hard 0-1" clustering assignments can be done for observations clearly in the "cores" of the clusters. However, more fuzzy assignments (u_{ig} in the open $(0, 1)$ interval) may appear for observations outside these "cores". This type of property was named as "hard contrast" in [39], to overcome a known shortcoming of fuzzy clustering. It is desirable, indeed, that the cluster centers, which are weighted means of all objects, should not be influenced by observations that clearly do not belong to these clusters. An analogous effect would distort scatter matrices as well, computed inside the clustering algorithm. To avoid or at least highly reduce these unwanted effects, outlying objects should be discarded, clusters centers become crisp, whereas "doubtful" observations remain fuzzy assigned. To obtain such a solution, the choice of the value of m (also in relationship with an adequate scale) plays a key role.

Scale and fuzzifier parameter m: The proposed approach is equivariant with respect to location shifts and rotations, while it is not affine equivariant, due to the constraint on the noise term variance. Clearly, choosing a large value of c turns the procedure into an almost affine equivariant one, at the

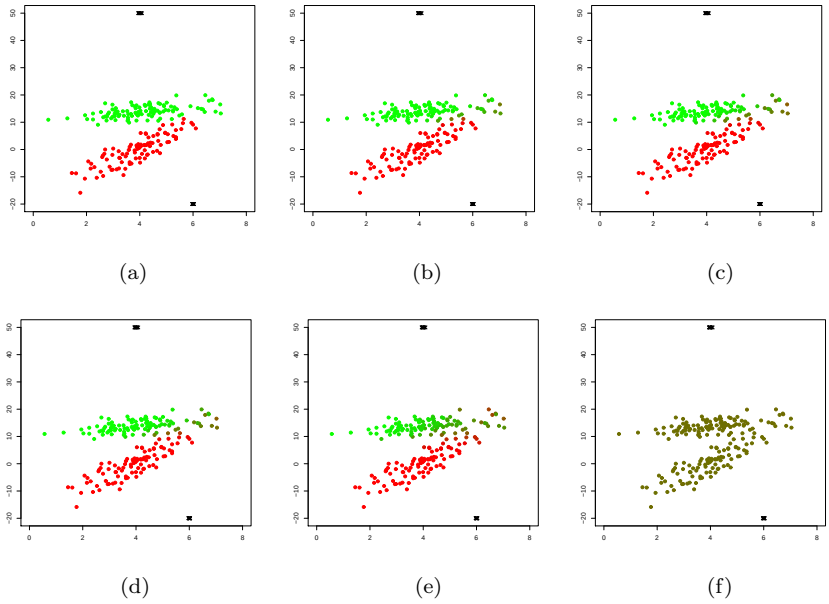


Fig. 3 Effects of the value of the fuzzifier parameter m on the cluster membership values (for panels (a), (b) and (c) we set $m = 1$, $m = 1.1$ and $m = 1.13$; while for panels (d), (e) and (f) we set $m = 1.16$, $m = 1.17$ and $m = 1.2$, respectively)

risk of incurring spurious solutions (as shown before). In this paragraph we want to discuss a second inherent issue, shared with other likelihood-based fuzzy clustering algorithms, as in [22], [46], [49] and [41]. Figure 4 shows in the first row panels the effect of the fuzzy clustering when all the variables in the artificial data are scaled by a constant factor, say divided by 2, in the second panel, and divided by 50 in the third one, considering $m = 1.17$. There is an interplay between the scale and the fuzzifier parameter, in such a way that the scale of the variable leads to changes in the results when $m > 1$, while when $m = 1$ (hard clustering) results are scale independent.

Hence, to suggest the user plausible values for the scale of the data and the fuzzifier parameter, in one only shot, we may base our analysis on some useful and well understandable quantities, like the percentage of hard assignment, and the relative entropy. The relative entropy is defined as

$$\frac{\sum_{g=1}^{G}\sum_{i=1}^{n} u_{ig}\log u_{ig}}{[n(1-\alpha)]\log G}.$$

Then, we apply a procedure by simultaneously changing the scale (horizontal axis with legend on the upper part of the plot in Figure 5), and the fuzzifier parameter m (legend in the lower horizontal axis), and draw the results in

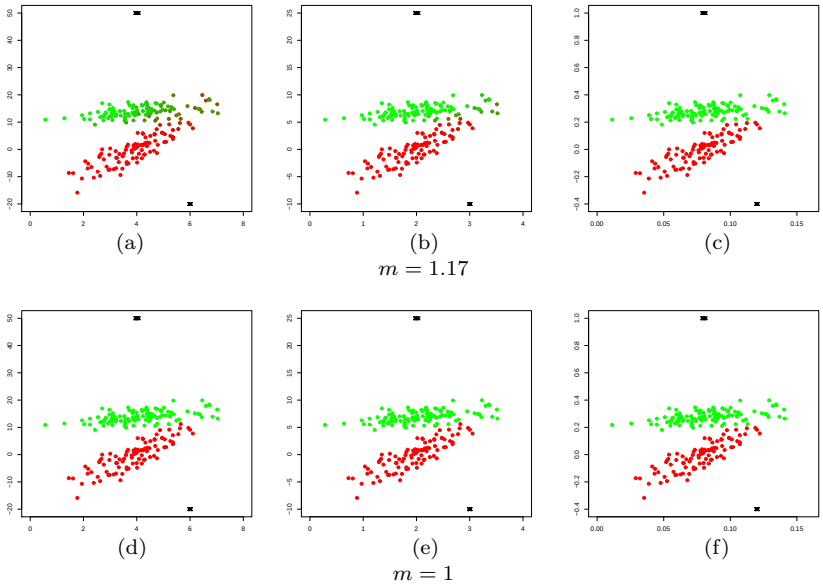


Fig. 4 In the upper panels, results of the robust fuzzy clustering with $m = 1.17$ obtained on the artificial data, when a change in scale is considered (from panel (a) to panel (b), all the variables have been scaled by a factor of 2, while in panel (c) by a factor of 50. In the lower panels, results of the hard clustering (with $m = 1$, same scaling as before, from left to right)

terms of hard assignment and relative entropy. Let us suppose that the user wants to obtain a clustering solution with 70% of observations crisply assigned and low relative entropy, say around 0.1. Inspecting the plot in Figure 5 we can derive guidelines for choosing a scale factor equal to 0.97 and a value $m = 1.7$ for the fuzzifier parameter. By a scale factor we mean a number which *scales* or multiply the values in a dataset.

Trimming level α: After discussing the robustness of the procedure to pointwise contamination, now we consider 22 noisy observations added to the artificial data as a result of a uniformly distributed random variable with support in a hyperrectangle including all the "good" observations. We want to show how the robust procedure performs, with different choices of the trimming level α. In our fuzzy clustering proposal, observations that are identified as non plausible ones, under the estimated model, are trimmed off. They do not contribute to parameter estimation, neither they belong to any cluster, in other words their membership level is equal to 0 for all the estimated clusters. We understand that the trimming level plays a key role in the robust estimation. To determine a correct trimming level, there are cases in which the researcher has some intuition or knowledge about

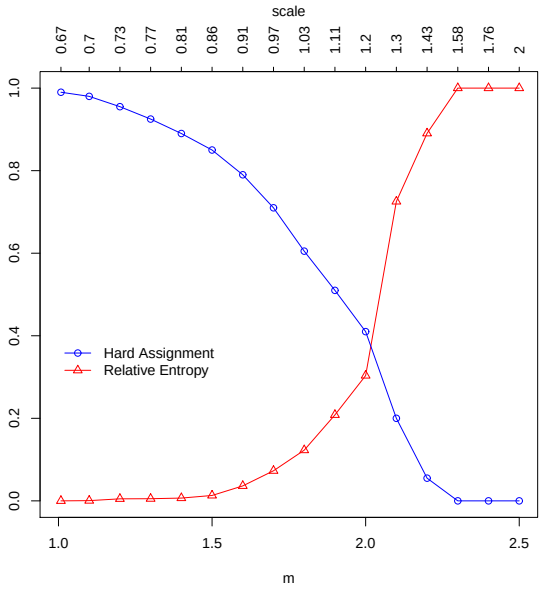


Fig. 5 Percentages of hard assignment (blue curve) and relative entropy (red curve) obtained when simultaneously changing the fuzzifier parameter and the scale.

the amount of noise in the data at hand, hence he can set the parameter α accordingly. If the contamination level is not known, we are interested in studying the effect of a value of α greater/lower than needed, on the robust estimation. Figure 6 shows, in the upper panel, the data contaminated by noise, where the contamination is represented by the 22 black observations. In the three lower panels of the same Figure 6, we show the results of the robust estimation, when a lower than needed ($\alpha = 0.05$, left panel), adequate ($\alpha = 0.1$, central panel) and greater than needed trimming level ($\alpha = 0.2$ right panel) have been adopted. We observe that the factor structure of the model is still recovered even for $\alpha = 0.2$ where the trimming level is twice its correct value; while for $\alpha = 0.05$ the estimation is poor (and for $\alpha = 0$ it is spoiled out; results are not shown for sake of brevity). Based on our experience, we would suggest to start with a (high) conservative choice of α and then decrease it, on the light of a careful analysis of the plausibility of a few trimmed observations close to non trimmed ones.

To help the researcher to set a reasonable value for the trimming level, it is useful to recall that the quantities $r_{(i)}$ defined in (11) measure the evidence of the belonging of observation $\mathbf{x}_i$ to the estimated model. Data with $r_{(i)}$ lower than $r_{([n\alpha)])}$ are identified as outlying observations and trimmed off. The information conveyed by the $r_{(i)}$ values can be employed to give guidance on the choice of the trimming level, as we can see in Figure 7, where the points

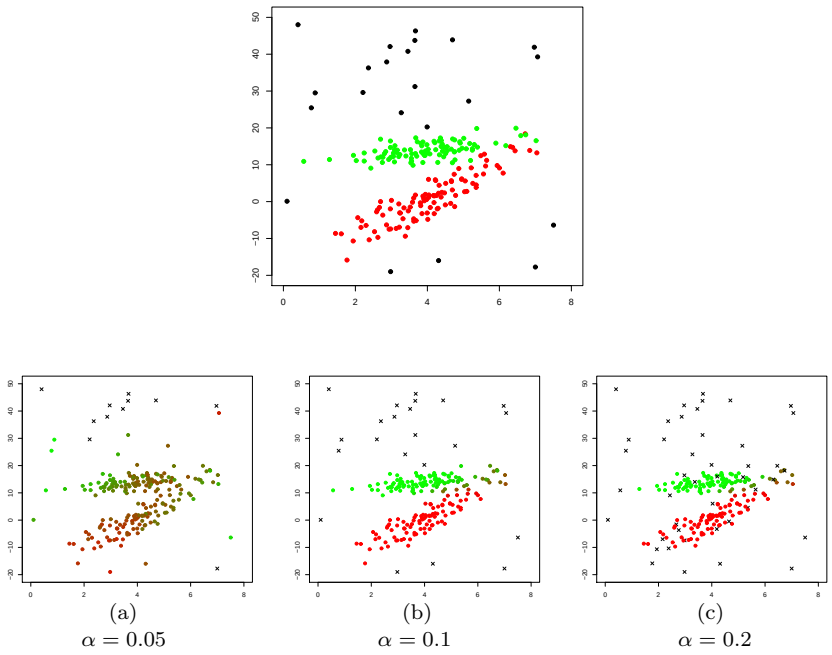
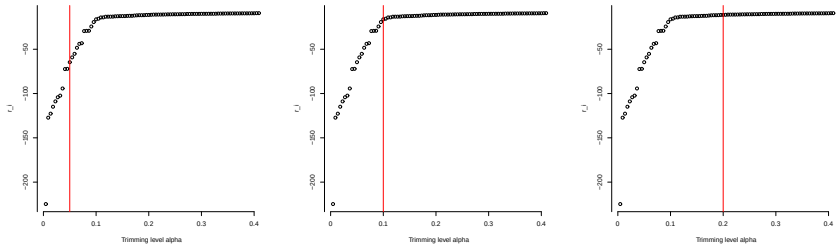


Fig. 6 Considering 22 noisy observations (in black) added to the original artificial dataset (upper panel), in the lower panels we show the results of the robust fuzzy clustering, with $m = 1.16$, when a different trimming level is employed. The colors here represent the fuzzy membership values, whereas trimmed observations are denoted by black crosses.

$\left\{(i/n, r_{(i)})\right\}_{i=1}^{n}$ have been plotted, highlighting also the α value employed for the estimation, with reference to the results shown in the three lower panels in Figure 6, respectively. The appropriate choice of α can be derived by finding an elbow in the plot, in other words a value α_0 such that $r_{(i)}$ is steep for $i/n < \alpha_0$ and the slope of the curve decreases for $\alpha > \alpha_0$. For instance, in the leftmost panel of Figure 7, we see that $\alpha = 0.05$ is not a good choice, because the curve $\left\{(i/n, r_{(i)})\right\}_{i=1}^{n}$ is still quickly increasing around $\alpha = 0.05$, and the red line "cut" it too early, much before its elbow. On the opposite rightmost panel, we see that $\alpha = 0.2$ would trim too much observations, and with some undesirable side effect, classifying points with almost equal contribution to the target function indifferently among the trimmed or among the non trimmed observations. The appropriate choice is suggested in the central panel, where $\alpha = 0.1$.

Number of groups G: The estimation of the number of groups is a highly discussed topic within the clustering community, due to the fact that it is intrinsically related to the definition of what a cluster is. The number of groups, say G, in the fuzzy mixture can be explored by means of the "classification



$\alpha = 0.05$ $\alpha = 0.1$ $\alpha = 0.2$

Fig. 7 Plots of the points $\left\{(i/n, r_{(i)})\right\}_{i=1}^{n}$, with reference to the robust fuzzy clustering shown in the lower panels of Figure 6. The employed trimming levels are highlighted by the vertical lines.

trimmed likelihood curves", once that a value for the parameter c has been set by the user (based on some previous knowledge that the user has on the data and/or on the type of clusters he is particularly interested in, depending on his/her final clustering purposes) and after incorporating the desired fuzziness by the choice of the fuzzifier parameter m. These curves are graphical tools, obtained by plotting the target function for a range of values of G and α that can be inspected to derive a reasonable choice for these parameters. They were introduced in hard clustering problems in [20]). Figure 8 shows the classification trimmed likelihood curves obtained for the artificial dataset with added uniform noise, previously shown in the lower row panels in Figure 6. We monitor the maximum value obtained by the objective function (8), depending on a number of groups from 1 to 4, and a trimming level from 0 to 0.2. We see that, when reaching the trimming level $\alpha = 0.1$, the results obtained for 2, 3 or 4 clusters almost virtually coincide. This means that with a lower trimming proportion, we would need a third or a fourth cluster to accommodate the noisy data. However, whenever the noisy observations have been identified and trimmed off, we are able to discover two main clusters in that artificial dataset.

5 An application to real data

It has been said that fuzzy clustering appears as a "hedge" against the uncertainty incorporated in any clustering model, as well as some a posteriori evaluation scheme providing a way to appraise algorithmically suggested clustering [2]. In this section, our aim is to apply the proposed approach to a real dataset concerning 202 athletes at the Australian Institute of Sport, and to highlight the advantages of fuzzy clustering. Data have been made publicly available in connection with the book [8]. Data refers to 102 male and 100 fe-

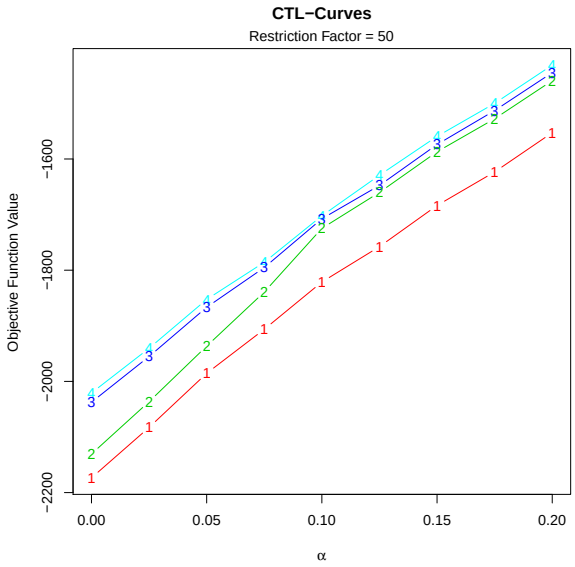


Fig. 8 The classification trimmed likelihood curves for the artificial data, with added 0.1% noise, obtained with $c = 50$ and $m = 1.1$.

male athletes collected at the Australian Institute of Sport. It consists of 202 observations on the following 11 biometrical or hematological variables, apart from gender and sport: red cell count (RCC), white cell count (WCC), Hematocrit (Hc), Hemoglobin (Hg), plasma ferritin concentration (Fe), body mass index weight/height (BMI), sum of skin folds (SSF), body fat percentage (Bfat), lean body mass (LBM), height (in cm, Ht) and weight (in kg, Wt). The dataset is available within the `sn` package in R. We aim at studying the joint linear dependence among these biometrical and blood composition variables, through a fuzzy clustering method, to highlight atypical patterns and athletes' features that would better require a fuzzy classification. We will exploit the conjecture that a strong correlation exists between the hematological and physical measurements. Therefore, a robust MFA may be estimated, assuming the existence of some underlying unknown factors (like nutritional status, hematological composition, body weight status indices, and so on) which jointly explain the observed measurements. We rely on previous work on this data in [21], where, based on a trimmed version of the Bayesian Information Criterion, the authors suggest to employ $G = 2$ clusters, and underlying $d = 6$ factor dimension.

Aiming at selecting the value of the trimming level α, we employ the information conveyed by the sorted $r_{(i)}$ values. In the associated plot, we look for a value of α that it is able to isolate a few observations having

the smallest contributions to the target function, when compared with the remaining data. This empirical approach, shown in the left panel of Figure 9, has suggested us the choice of $\alpha = 0.05$. Finally, to choose a scale factor and a value for the fuzzifier parameter m, we can base our decision on the desired percentage of hard assignments, jointly with the advisable relative entropy. The right panel of Figure 9 show us, once more, that the scale plays an important role in fuzzy clustering, due to its interplay with the fuzzifier m. As we are here dealing with multivariate data with disparate column scales (see Table 1), we need a different scale for each column. The value in the scale legend (upper horizontal legend) should be read as the multiple of the column standard deviation (or a robust variability measure) to be adopted for each variable in the dataset.

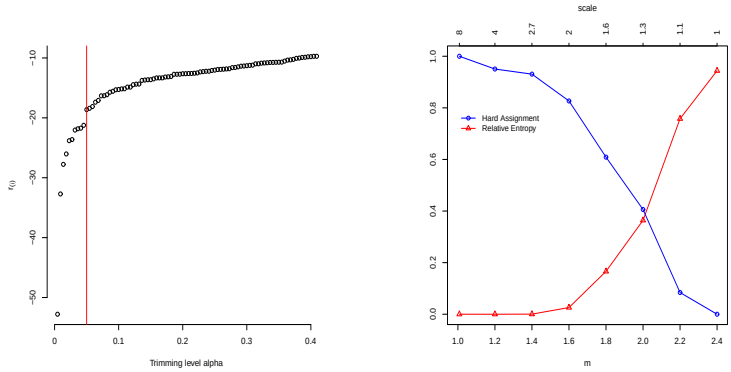


Fig. 9 Plots for guiding the choice of the parameters when analyzing AIS data with fuzzy clustering. In the left panel we see how to choose the trimming level $\alpha = 0.05$. The right panel shows how to pick the fuzzifier parameter m and the scale, based on the desired percentage of hard assignment and the relative entropy.

Finally, we have chosen three different values of the fuzzifier, say $m = 1.1$, $m = 1.4$ and $m = 1.8$ (on the scaled data) to show the effects of an increasing level of fuzziness (results in Figure 10).

Comparing the three plots, we may see that the observations in rows 11, 68, 75, 99, 113, 133, 160, 163, 166, and 178 of the AIS dataset are consistently trimmed in the three estimated models (labels have been added in the upper panel). In the central plot, the only difference is that observation in row 93 has been trimmed, instead of the one in row 113. This could suggest to analyze more in detail the data of these athletes. Just to give one example, we may inspect the information we have for the female athlete in row 11, who plays basketball and whose weight is 96.3 kg and height is 193.4 cm (just to mention two of the 11 variables). Median values of these two variables in the group of female athletes are, respectively, 68.1 kg and 178 cm. In [21], where clustering is based on robust MFA, two common factors have been

identified both in the group of male ad female athletes: a first *hematological factor*, with a very high loading on Hc, followed by RCC and Hg, and second factor, loading heavily on Ht, and in a lesser extent on Wt and LBM, that may be denoted as a *general nutritional status*. Hence observations that are consistently trimmed in the three plots are expected to have an unusual score on the factors.

The identification of outliers could have a very important role for the trainer or the physician that follow the team, to study a possible relationship among sport performances and exceptional values of some hematological and/or biometrical variables. Observing different membership levels, for this type of data, could also provide a particularly useful information. On the one side, athletes whose features are almost paradigmatic in their gender (or discipline) group are denoted by high levels of membership and can be easily identified. Like, for example, athlete in row 29, in the group of female, labeled in the central plot of Figure 10, colored with full red (with $u_{29,1} = 0.999$ and $u_{29,2} = 0.001$) has the features shown in first row of Table 1, that can be interpreted as "paradigmatic". Similar considerations hold for the athlete in row 188, in the group of male athletes, colored with full green (with $u_{188,1} = 0.008$ and $u_{188,2} = 0.993$), whose measurements can be found in the last row of Table 1, that can be considered as a prototype for male athletes.

Conversely, athletes having unusual scores on the underlying factors, are denoted by low membership levels, and this could be a valuable information to be maintained up-to-date for team coachers, to assign a proper training regime, as well as for sport nutritionists to indicate an adequate diet. Such an example is athlete in row 70 where $u_{70,1} = 0.511$ and $u_{70,2} = 0.489$ or athlete in row 130 with $u_{130,1} = 0.469$ and $u_{130,2} = 0.531$. Table 1 reports - in the second and third row - the observed variables for athletes that do not openly qualify for belonging to a group.

Table 1 Features for some athletes in the AIS dataset

row#	gender	sport	RCC	WCC	Hc	Hg	Fe	BMI	SSF	Bfat	LBM	Ht	Wt
29	F	Row	4.2	7.5	38.4	13.2	73	20.5	74.7	16.6	41.5	156.0	49.8
70	F	Field	4.6	5.8	42.1	14.7	164	28.6	109.6	21.3	68.9	175.0	87.5
130	M	B.Ball	4.5	5.9	44.4	15.6	97	20.7	41.5	7.2	73.0	195.3	78.9
188	M	Row	4.8	9.3	43.0	14.7	150	25.1	60.2	10.0	78.0	186.0	86.8

The reader accustomed to employ mixture models for clustering would perhaps comment that we may use the probability of belonging of a unit to a component of the mixture to draw such considerations. We want to stress here the advantage of using such a fuzzy approach with selection of the percentage of hard assignment, and where the fuzzifier directly controls how much clusters may overlap, to accomodate "doubtful" observations. We are here conjugating, in a fresh approach, the beneficial effects of crisp and fuzzy clustering to obtain a robust estimation.

At the end of this section, we would like to highlight the effectiveness of the proposed method, and compare it with other approaches in the literature. The natural benchmarks to our proposal are the mixtures of Factor Analyzers (MFA), and their robustification through trimming and constrained estimation in [21]. We may also want to compare with the fully parameterized Mixture of Gaussian (MG), in both robust and non-robust versions, as well as their fuzzy counterparts, to compare with previous work in [16]. Our first aim is to compare their clustering performances, reported in Table 2, when running the associated (trimmed and untrimmed) EM algorithm from 60 random initializations, with maximum number of iterations set to 30. To be more precise, the algorithm described in [21] is applied with $\alpha = 0$ and $c = 10^{10}$ in the non-robust case for MFA and with $\alpha = 0.05$ and $c = 50$ for robust MFA. We adopted the same choice of the parameters for the fully parameterized MG, estimated using the `tclust` package, for its fuzzy version described in [16], and for the proposed fuzzy MFA. To obtain Table 2, we have assigned even the trimmed observations, by resorting to the membership values that would have been obtained from the (robustly) estimated parameters and the proposed model. We can see that the proposed methodology has nice clustering performance, at least for the inspected values of the fuzzifier parameter m. Furthermore, we may appreciate the beneficial effect of using trimming and also, for the AIS data, the advantage of combining robust methods with models that are able to capture the underlying latent structure among variables. In other words, the model conveys how hematological factors and physical measurements play a slightly different role in male and in female athletes. Parsimony in model parameters has been pursued, too.

Table 2 Clustering performances of different models, in terms of number of misclassified units on the AIS dataset (different values of the fuzzifier m have been considered for fuzzy clustering)

robust fuzzy MFA			robust MFA	non-robust MFA	robust fuzzy MG			robust MG	non-robust MG
	m					m			
1.1	1.4	1.8			1.1	1.4	1.8		
2	4	10	3	7	9	8	9	8	17

6 Concluding remarks

In this paper, we propose a novel FCM-type fuzzy clustering scheme providing two significant benefits when compared with the existing approaches. First, it provides a well-established space dimensionality reduction framework for fuzzy clustering algorithms based on factor analysis, allowing concurrent performance of fuzzy clustering and, within each cluster, local dimensionality

reduction. Second, it exploits the outlier tolerance advantages of procedures base on impartial trimming to provide a novel, soundly founded, nonheuristic, robust fuzzy clustering framework. Third, our proposal allows for soft as well as hard robust clustering, hence it encompasses both methodologies. Fourth, it enable the property of "hard contrast" of the solution: outlying objects should be discarded, clusters centers become crisp, whereas "doubtful" observations remain fuzzy assigned.

This way, the proposed model yields a significant performance increase for the fuzzy clustering algorithm, as we experimentally demonstrate.

Comparing the fuzzy approach to classical model-based clustering via mixtures, we would like to stress that in both methodologies we obtain a "degree of belonging" of the i-th observation to the g-th cluster. To contrast the two approaches, we remark that in the fuzzy approach we can control the level of fuzziness and/or a percentage of units to be crisply assigned to groups, as well as a relative entropy to be fulfilled by the solution, while in the mixture approach they arise as a byproduct of the estimated model. Based on our findings, we observed that small deviations from the model assumptions impair the performance of the fuzzy classifier only by a small amount, and that good efficiency is obtained on data without contamination.

With respect to robustness, that is our second qualifying point, roughly speaking, it is worth to recall that two main approaches can be found in the literature. The first, based on mixture modeling, aims at fitting the outliers in the model, e.g. by considering additional mixture components or by allowing heavy-tailed components. The second, based on trimming, just try to trim them off, without assuming any distribution for them. Hence, in comparison with fuzzy mixtures of Student's t factor analyzers in [7], our proposal, being based on the constrained estimation of scatters and trimming off a proportion α of the most outlying observations, provides good breakdown properties of the estimators (see [42] and [17]), in the sense of [27]. Yet, the complexity of the EM algorithm is considerably reduced, as it does not require the complicated estimation of the degree of freedom parameter, for which a closed solution in the EM algorithm is not available. Another interesting feature of our approach is that trimming, as a by-product, leads to the identification of outlying observations.

The effectivity of our proposal is also due to the parameters involved along the fuzzy robust estimation, like the noise matrices constraint c, the fuzzifier parameter m, the trimming level α, the dimension of the underlying space d and the number of groups G. The entire Section 4 have been devoted to the discussion on the role of the tuning parameter, with the help of artificial data. Some tools for helping the user to make a reasonable choice for the trimming level, the number of clusters, as well as the fuzzifier parameter have been developed, also discussing their interplay. The usefulness of such tools has been exemplified, both on the artificial dataset and also through an application to a real data set. In the real data application, where athletes features are analyzed, we have also shown how the fuzzy approach provides

a way to identify athletes whose features are almost paradigmatic in their gender or discipline group, while athletes with low scores on the underlying factors, are a relevant information to be kept up-to-date for team coachers. Future research lines include developing more formal tools for helping the user to choose the several tuning parameters that this very flexible methodology requires.

References

1. A. Banerjee, R.N. Davé, Robust clustering, WIREs Data Mining Knowl Discov (2) (2012) 29–59.
2. J.C. Bezdek, Numerical taxonomy through fuzzy sets, J. Math. Biol. (1) (1974) 57–71.
3. C.M. Bishop, Latent variable models, in: Jordan, M.I. (Ed.), Learning in Graphical Models, Kluwer, Dordrecht, 1998, pp. 371–403.
4. P.G. Bryant, Large-sample results for optimization-based clustering methods, J. Classif. (8) (1991) 31–44.
5. R. Cattell, A note on correlation clusters and cluster search methods, Psychometrika 9(3) (1944) 169–184.
6. W. Chang, On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions, J. R. Stat. Soc. C-Appl. 32(3) (1983) 267–275.
7. S. Chatzis, T. Varvarigou, Factor analysis latent subspace modeling and robust fuzzy clustering using t-distributions, IEEE T. Fuzzy Syst. 17(3) (2009) 505–517.
8. R.D. Cook, S. Weisberg, An Introduction to Regression Graphics, John Wiley & Sons, New York, 1994.
9. J.A. Cuesta-Albertos, A. Gordaliza, C. Matrán, Trimmed k-means: an attempt to robustify quantizers, Ann. Stat. 25(2)(1997) 553–576.
10. R.N. Davé, Characterization and detection of noise in clustering, Pattern Recogn. Lett. 12 (1991) 657–664.
11. R.N. Davé, R. Krishnapuram, Robust clustering methods: a unified view, IEEE T. Fuzzy Syst., 5 (1997) 270–293.
12. F. Dotto, A. Farcomeni, L.A. García-Escudero, A. Mayo-Iscar, A fuzzy approach to robust regression clustering, Adv. Data Anal. Classif. (2016) 1–20.
13. J.C. Dunn, A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, J. Cybernetics 3 (1973) 32–57.
14. A. Farcomeni, L. Greco, Robust Methods for Data Reduction, Chapman and Hall/CRC, 2015.
15. H. Fritz, L.A. García-Escudero, A. Mayo-Iscar, A fast algorithm for robust constrained clustering, Comput. Stat. Data An. 61(2013) 124–136.
16. H. Fritz, L.A. García-Escudero, A. Mayo-Iscar, Robust constrained fuzzy clustering, Inform. Sciences 245 (2013) 38–52.
17. M.T. Gallegos, G. Ritter, Trimming algorithms for clustering contaminated grouped data and their robustness, Adv. Data Anal. Classif., 3 (2009) 135–167.
18. L.A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Iscar, A General Trimming Approach to Robust Cluster Analysis, Ann. Stat. 36 (3) (2008) 1324–1345.
19. L.A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Iscar, A review of robust clustering methods, Adv. Data Anal. Classif. 4 (2010) 89–109.
20. L.A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Iscar, Exploring the number of groups in robust model-based clustering, Stat. Comput. 21 (2011) 585–599.
21. L.A. García-Escudero, A. Gordaliza, F. Greselin, S. Ingrassia, A. Mayo-Iscar, The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers, Comput. Stat. Data An. 99 (2016) 131–147.

22. I. Gath, A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE T. Pattern Anal. 11(7) (1989) 773–780.
23. Z. Ghahramani, G.E. Hinton, The EM algorithm for factor analyzers, Technical Report No. CRG-TR-96-1, The University of Toronto, Toronto, 1996.
24. F. Greselin, S. Ingrassia, Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers, Stat. Comp. 25 (2015) 215–226.
25. E.E. Gustafson, W.C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in: Proceedings of the IEEE lnternational Conference on Fuzzy Systems, San Diego, 1979, pp. 761–766.
26. R. Hathaway, A constrained formulation of maximum-likelihood estimation for normal mixture distributions, Ann. Stat. 13(2) (1985) 795–800.
27. C. Hennig, Breakdown points for maximum likelihood estimators of location-scale mixtures, Ann. Stat. 32(4) (2004) 1313–1340.
28. G.E. Hinton, P. Dayan, M. Revow, Modeling the manifolds of images of handwritten digits, IEEE T. Neural Networ. 8 (1997) 65–73.
29. P.J. Huber, Robust Statistics, Wiley, New York, 1981.
30. S. Ingrassia, R. Rocci, Constrained monotone EM algorithms for finite mixture of multivariate Gaussians, Comput. Stat. Data An. 51 (2007) 5339–5351.
31. J. Kim, R. Krishnapuram, R. Davé, Application of the least trimmed squares technique to prototype-based clustering, Pattern Recogn. Lett. 17 (1996) 633–641.
32. F. Klawonn, Noise clustering with a fixed fraction of noise, in: Applications and Science in Soft Computing, Springer, Berlin-Heidelberg-New York, 2004, pp. 133–138.
33. R. Krishnapuram, J.M. Keller, A possibilistic approach to clustering, IEEE T. Fuzzy Syst. 1 (1993) 98–110.
34. J. Łeski, Towards a robust fuzzy clustering, Fuzzy Set. Syst. 37 (2003) 215–233.
35. J.B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, in: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1, 1967, pp. 281–297 .
36. G.J. McLachlan., D. Peel, Finite Mixture Models, John Wiley & Sons, New York, 2000.
37. S. Miyamoto, M. Mukaidono, Fuzzy c-means as a regularization and maximum entropy approach, in: Proceedings of the 7th International Fuzzy Systems Association World Congress, IFSA'97, 2, 1997, pp. 86–92.
38. G. Ritter, Robust Cluster Analysis and Variable Selection, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 2015.
39. P.J. Rousseeuw, E. Trauwaert, L. Kaufman, Fuzzy clustering with high contrast, J. Comput. Appl. Math. 64 (1995) 81–90.
40. P.J. Rousseeuw, K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, Technometrics 41 (1999) 212–223.
41. P.J. Rousseeuw, L. Kauffman, E. Trauwaert, Fuzzy clustering using scatter matrices, Comput. Stat. Data An. 23 (1996) 135–151.
42. C. Ruwet, L.A. García-Escudero, A. Gordaliza, A. Mayo-Iscar, On the breakdown behavior of the TCLUST clustering procedure, Test, 22 (3) (2013) 466–487.
43. M.J. Symons, Clustering criteria and multivariate normal mixtures, Biometrics 37 (1981) 35–43.
44. M.E. Tipping, C.M. Bishop, Mixtures of probabilistic principal component analysers, Technical Report No. NCRG=97=003, Neural Computing Research Group, Aston University, Birmingham, 1997.
45. M.E. Tipping, C.M. Bishop, Mixtures of probabilistic principal component analysers, Neural Comput. 11 (1999) 443–482.
46. E. Trauwaert, L. Kaufman, P. Rousseeuw, Fuzzy clustering algorithms based on the maximum likelihood principle, Fuzzy Set. Syst. 42(2) (1991) 213–227.
47. K.S. Wee, K.S. Fu, A Formulation of Fuzzy Automata and Its Application as a Model of Learning Systems, IEEE T. Syst. Sci. Cyb. 5(3) (1969) 215–223. doi: 10.1109/TSSC.1969.300263

48. K.L. Wu, M.S. Yang, Alternative c-means clustering algorithms, Pattern Recogn. 35 (2002) 2267–2278.
49. M.S. Yang, On a class of fuzzy classification maximum likelihood procedures, Fuzzy Set. Syst. 57 (1993) 365–375.
FUZZY SET SYST
50. L.A. Zadeh, Fuzzy sets, Inform. Control 8(3) (1965) 338–353.

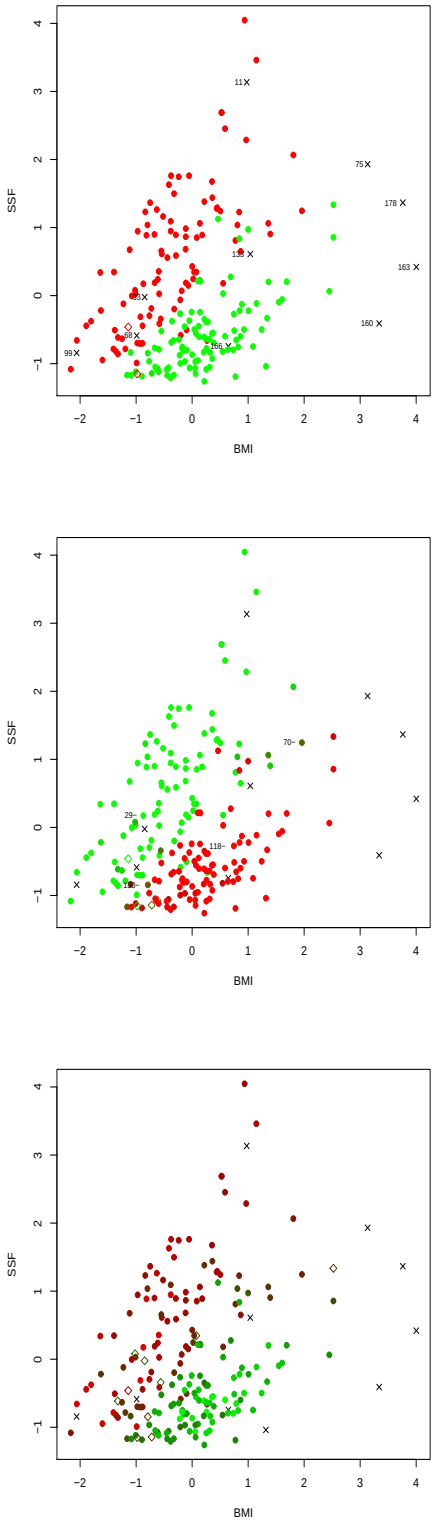


Fig. 10 Fuzzy clustering on the AIS dataset, with different choices of the fuzzifier parameter: in the upper panel $m = 1.1$, in the central panel $m = 1.4$, in the lower panel $m = 1.8$. By $\times$ we denote trimmed observations, by $\Diamond$ the misclassified ones