Fuzzy clustering through robust Factor Analyzers

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

Abstract In fuzzy clustering, data elements can belong to more than one cluster, and membership levels are associated with each element, to indicate the strength of the association between that data element and a particular cluster. Unfortunately, fuzzy clustering is not robust, while in real applications the data is contaminated by outliers and noise, and the assumed underlying Gaussian distributions could be unrealistic. Here we propose a robust fuzzy estimator for clustering through Factor Analyzers, by introducing the joint usage of trimming and of constrained estimation of noise matrices in the classic Maximum Likelihood approach.

1 Introduction

Clustering can be considered the most important unsupervised learning problem. It is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called clusters. A cluster is therefore a collection of objects which are similar to one another and thus can be treated collectively as one group. Clustering algorithms may be classified into Exclusive (or Crisp, Hard), Overlapping, Hierarchical and Probabilistic. To recall some well known examples, K-means [12] is an exclusive clustering algorithm, Fuzzy C-means [2] is an overlapping clustering algorithm, Single-linkage [1] is an agglomerative Hierarchical clustering and lastly Mixture of Gaussian is a probabilistic clustering algorithm. In the present work, we will introduce a Fuzzy version of Mixtures of Gaussian Factor Analyzers.

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

Starting from Wee and Fu's seminal work [16], fuzzy clustering has received an increasing attention by researchers from several fields in the last fifty years. The aim is to discover a limited number of homogeneous clusters in such a way that the objects are assigned to the clusters according to the so-called membership degrees ranging in the interval [0, 1]. In real applications, the data is bound to have noise and outliers, and the assumed models such as Gaussian distributions are only approximations to reality. Unfortunately, one of the main limitations of all clustering algorithms is that they are not robust to noise: a small fraction of outlying data may drastically deteriorate the clustering ability. Hence we will provide robustness properties to our estimator for Gaussian Factor Analyzers, by trimming those observations that are less plausible under the estimated model. According to [10], a robust procedure can be characterized by the following: 1) it should have a reasonably good efciency (accuracy) at the assumed model; 2) small deviations from the model assumptions should impair the performance only by a small amount; and 3) larger deviations from the model assumptions should not cause a catastrophe. We will see that our proposal satisfies the three properties.

2 Fuzzy clustering through Gaussian Factors

Suppose that we have n observations $\{\mathbf{x}_1 \dots \mathbf{x}_n\}$ in $\mathbb{R}^p$ and we want to fuzzy-classify them into k clusters. Therefore, our aim is to obtain a collection of non-negative membership values $u_{ij} \in [0, 1]$ for all $i = 1 \dots n$ and $j = 1 \dots k$. Increasing degrees of membership are allowed when $u_{ij} \in (0, 1)$, while $u_{ij} = 1$ indicates that object i fully belongs to cluster j and, conversely, $u_{ij} = 0$ means that it does not belong to this cluster. We will denote an observation as fully trimmed if $u_{ij} = 0$ for all $j = 1 \dots k$ and, thus, this observation has no membership contribution to any cluster.

Further, we want to employ Factor Analysis and suppose that, as in many phenomena, the p observed variables could be explained by a few unobserved ones. Factor Analysis is an effective method of summarizing 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 to account for that part of the variability which is unique to it (not in common with other variables). We will assume that the distribution of $\mathbf{x}_i$ can be given as

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U}_i + \mathbf{e}_i \quad \text{for } i = 1, \dots, n, \tag{1}$$

where $\boldsymbol{\Lambda}$ is a $p \times d$ matrix of *factor loadings*, the *factors* $\mathbf{U}_1, \dots, \mathbf{U}_n$ are $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ distributed independently of the *errors* $\mathbf{e}_i$. The latter are indepen-

dently $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ distributed, and $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix. The diagonality of $\boldsymbol{\Psi}$ is one of the key assumptions of factor analysis: the observed variables are independent given the factors. Note that the factor variable $\mathbf{U}_i$ models correlations between the elements of $\mathbf{x}_i$, while the errors $\mathbf{e}_i$ account for independent noise for $\mathbf{x}_i$. We suppose that $d < p$. Under these assumptions, $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma}$ has the form

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{2}$$

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

$$\sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m \log \phi(\mathbf{x}_i; \mathbf{m}_j, \boldsymbol{S}_j), \tag{3}$$

where $\phi(\cdot; \mathbf{m}, \boldsymbol{S})$ is the density of the multivariate Gaussian with mean $\mathbf{m}$ and covariance $\boldsymbol{S}$, and the membership values $u_{ij} \geq 0$ are assumed to satisfy

$$\sum_{j=1}^{k} u_{ij} = 1 \quad \text{if} \quad i \in \mathcal{I} \quad \text{and} \quad \sum_{j=1}^{k} u_{ij} = 1 \quad \text{otherwise,} \tag{4}$$

for a subset

$$\mathcal{I} \subset 1, 2, \ldots, n \quad \text{with} \quad \#\mathcal{I} = [n(1-\alpha)], \tag{5}$$

where $\mathbf{m}_1, \ldots, \mathbf{m}_k$ are vectors in $\mathbb{R}^p$, and $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_k$ are positive semidefinite $p \times p$ matrices satisfying the decomposition in (2), i.e. $\boldsymbol{S}_j = \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j' + \boldsymbol{\Psi}_j$. With reference to the diagonal elements $\{\psi_k\}_{k=1,\ldots,p}$ of the noise matrices $\boldsymbol{\Psi}_j$, it is required that

$$\psi_{j_1 h} \leq c_{noise} \ \psi_{j_2 l} \qquad \text{for every } 1 \leq h \neq l \leq p \text{ and } 1 \leq j_1 \neq j_2 \leq k \tag{6}$$

The constant c_{noise} is finite and such that $c_{noise} \geq 1$, to avoid the $|\Sigma_g| \to 0$ case. This constraint can be seen as an adaptation to MFA of those introduced in [11], [5], and is similar to the mild restrictions implemented for MFA in [7]. They all go back to the seminal paper of [9].

Notice that $u_{i1} = \ldots = u_{ik} = 0$ for all $i \notin \mathcal{I}$, so these observations do not contribute to the summation in the target function (3).

Our fuzzy method is based on a maximum likelihood criterium defined on a specific underlying statistical model, as in many other proposal in the literature.

After the introduction of trimmed observation, the second specific features of the proposed methodology is the application of the eigenvalue ratio constraint in (6). This is needed to avoid the unboundedness of the the objective function (3), whenever one of the $\mathbf{m}_j$ is equal to one of the observations $\mathbf{x}_i$, setting $u_{ij} = 1$, and for a sequence of scatter matrices $\boldsymbol{S}_j$ such that $|\boldsymbol{S}_j| \to 0$.

This problem is recurrent in Cluster Analysis whenever general scatter matrices are allowed, and has been already noticed in fuzzy clustering, among other authors, by [8]. In our approach, the unboundedness problem is addressed by constraining the ratio between the largest and smallest eigenvalues of the so-called noise matrices $\boldsymbol{\Psi}_j$. Larger values of c lead to an almost unconstrained fuzzy clustering approach.

It is well known that the use of an objective function like that in (3) tends to provide clusters with similar sizes, or more precisely, 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 (3) by

$$\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m \log p_j \phi(\mathbf{x}_i; \mathbf{m}_j, \boldsymbol{S}_j), \tag{7}$$

where $p_j \in [0,1]$ and $\sum_{j=1}^{k} p_j = 1$ are some weights to be maximized in the objective function, as in the entropy regularizations in [14]. Once the membership values are known, the weights are optimally determined as $p_j = \sum_{i=1}^{n} u_{ij}^m / \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m$ (see [**?**], for a detailed explanation). Finally, considering (7) as our target function, and performing trimming and constrained estimation along the EM algorithm we obtain a robust approach to fuzzy clustering through factor analyzers.

More precisely, we consider an AECM algorithm, where we incorporates a concentration step, as in many high-breakdown point robust algorithms like [15], before each E-step. After selecting the set of observations that contributed the most to the target function (concentration step), at each iteration, given the values of the parameters, the best possible membership values are obtained (E-step). Afterwards, the parameters are updated by maximizing expression (7) on the parameters (M-step). The name of AECM (that appeared in the literature for the case of mixtures of Gaussian factor analyzers, see [13]) comes from the fact that the M-step is performed alternatively on a partition of the parameter space. When updating the $\boldsymbol{S}_j$ matrices the constraint on the eigenvalue ratios are imposed accordingly, along the lines of [3].

Finally, it is worth to remark that the general approach presented herein encompasses the soft robust clustering method introduced in [6], and leads to hard clustering for $m = 1$. For $m > 1$ it provides fuzzy clustering.

3 Numerical results

We present here a first experiment on synthetic data, to show the performance of the proposal. We choose a two component population in $\mathbb{R}^{10}$, from which we draw two samples. Aiming at providing a plot of the obtained results, we work with unidimensional factors (otherwise we could not find a unique space, for the two components, to represent the data). The first population

X_1 is defined as follows:

$$X_{11} \sim \mathcal{N}(0,1) + 4 \qquad X_{12} \sim 5 * X_{11} + 3 * \mathcal{N}(0,1) - 6;$$

and the second population X_2 is given as:

$$X_{21} \sim \mathcal{N}(0,1) + 4 \qquad X_{22} \sim X_{21} + 2 * \mathcal{N}(0,1) + 19.$$

After drawing 100 points for each component, to check the robustness of our approach, we add some pointwise contamination X_3 to the data, by drawing 10 points as follows

$$X_{31} \sim \mathcal{N}(0,1) + 4 \qquad X_{32} \sim 50 + 0.01 * \mathcal{N}(0,1);$$

and 10 more points, denoted by X_4, where

$$X_{41} \sim \mathcal{N}(0,1) + 6 \qquad X_{42} \sim -20 + 0.01 * \mathcal{N}(0,1).$$

Finally, we complement the data matrix with $X_{ij} \sim \mathcal{N}(0,1)$ for $i = 1, \ldots, 4$ and $j = 3, \ldots, 10$. In this way we have built a dataset where one factor is explaining the correlation among the 10 variables, in each component.

Figure 1 shows that the estimation is robust to the most dangerous outliers, in the form of pointwise contamination. In all our results we observed that small deviations from the model assumptions impair the performance only by a small amount, and that good efficiency is obtained on data without contamination.

References

1. Cattell R (1944) A note on correlation clusters and cluster search methods, Psychometrika, 9.3: 169-184.
2. Dunn JC (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics, 3: 32–57.
3. Fritz H, García-Escudero LA, Mayo-Iscar A (2013) A fast algorithm for robust constrained clustering, Computational Statistics & Data Analysis, 61: 124–136.
4. Fritz H, García-Escudero LA, Mayo-Iscar A (2013). Robust constrained fuzzy clustering, Information Sciences, 245: 38–52.
5. García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2008) A General Trimming Approach to Robust Cluster Analysis. The Annals of Statistics, 36: 3,1324–1345.
6. García-Escudero LA, Gordaliza A, Greselin F, Ingrassia S, Mayo-Iscar A (2016) The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers, Computational Statistics & Data Analysis, 99:131–147.
7. Greselin F, Ingrassia S (2015) Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers, Statistics and Computing, 25: 215–226.
8. Gustafson EE, Kessel WC (1979) Fuzzy clustering with a fuzzy covariance matrix, in: Proceedings of the IEEE lnternational Conference on Fuzzy Systems, San Diego, pp. 761–766.

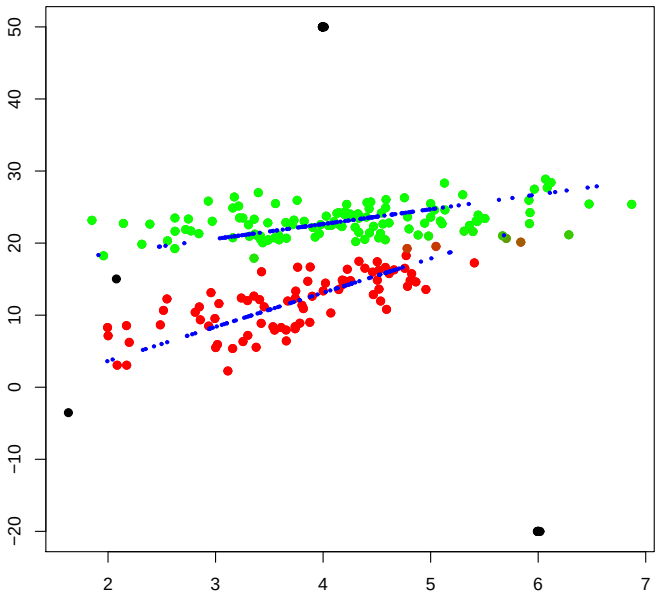

Fig. 1 Fuzzy classification of the synthetic data. Blue points are the projections of the 10-dimensional data in the latent factor space of the first component. Black points are trimmed units. The strenght of the membership values is represented by the color saturation level.

9. Hathaway R (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions, The Annals of Statistics,13: 2, 795–800.
10. Huber PJ (1981) Robust Statistics. New York, Wiley.
11. Ingrassia S, Rocci R (2007) Constrained monotone EM algorithms for finite mixture of multivariate Gaussians, Computational Statistics & Data Analysis, 51:5339–5351.
12. MacQueen JB (1967) Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1: 281–297 .
13. McLachlan GJ, Peel D (2000) Finite Mixture Models, John Wiley & Sons, New York.
14. Miyamoto S, Mukaidono M (1997) Fuzzy c-means as a regularization and maximum entropy approach, Proceedings of the 7th International Fuzzy Systems Association World Congress (IFSA97), 2: 86–92.
15. Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator, Technometrics, 41: 212–223.
16. Wee WG, Fu KS (1969) A Formulation of Fuzzy Automata and Its Application as a Model of Learning Systems, IEEE Transactions on Systems Science and Cybernetics, 5, 3, pp. 215–223. doi: 10.1109/TSSC.1969.300263