Robust approaches for fuzzy clusterwise regression based on trimming and constraints

Luis Angel García-Escudero[1], Alfonso Gordaliza[1], Francesca Greselin[2], and Agustín Mayo-Iscar[1]

Abstract Three different approaches for robust fuzzy clusterwise regression are reviewed. They are all based on the simultaneous application of trimming and constraints. The first one follows from the joint modeling of the response and explanatory variables through a normal component fitted in each cluster. The second one assumes normally distributed error terms conditional on the explanatory variables while the third approach is an extension of the Cluster Weighted Model. A fixed proportion of "most outlying" observations are trimmed. The use of appropriate constraints turns these problem into mathematically well-defined ones and, additionally, serves to avoid the detection of non-interesting or "spurious" linear clusters. The third proposal is specially appealing because it is able to protect us against outliers in the explanatory variables which may act as "bad leverage" points. Feasible and practical algorithms are outlined. Their performances, in terms of robustness, are illustrated in some simple simulated examples.

1 Introduction

The detection of clusters around linear subspaces, instead of just around points or centroids, is often needed in Cluster Analysis. This problem is meaningful not only because clusters are frequently arranged this way but also because sometimes it is interesting to discover different relations between a response variable and some other explanatory variables within each cluster. These problems are commonly known as clusterwise linear regression or switching regression models. Some others seminal references are [19],

Deptartamento de Estadística e Investigación Operativa and IMUVA, Universidad de Valladolid, Valladolid, Spain. e-mail: lagarcia,alfonsog,agustinm@eio.uva.es · Department of Statistics and Quantitative Methods, Milano-Bicocca University, Milano, Italy. e-mail: francesca.greselin@unimib.it

[21], [26] and [4]. All those "hard" or 0-1 clustering procedures partition the data into G completely disjoint clusters. Alternatively, fuzzy clustering methods provide nonnegative membership values of observations to clusters and overlapping clusters are so generated [25, 2, 17]. This fuzzy approach can be certainly useful in clusterwise regression applications. There already exist many proposals addressing clustering around linear subspaces problem from a fuzzy clustering point of view. For instance, [17] provides an adaptation of the fuzzy c-means in [2] by minimizing a weighted sum of distances of each observation from the estimated regression line and where these weights depend on the fuzzy membership values. See also [18] and [29] and references therein.

Robustness is also a desirable property for (fuzzy) clustering techniques due to the well-know harmful effect that (even a small fraction) outlying observations may have in them. Several methods have been recently proposed to improve clustering techniques robustness performance. For instances, many proposals can be found in [10, 6, 22] (hard) and in [3, 1] (fuzzy).

In this work, we are going to review three recent approaches for robust fuzzy clusterwise regression derived from considering a maximum likelihood approach with trimming and constraints. These methods can see as extensions of that introduced in [7]. Trimming is probably the simpler and easier to understand way to achieve robustness. Particularly, we consider an impartial trimming approach where "impartial" means that the data set itself tell us which are observations that should be trimmed as in [9]. When an observation with index i is detected as an outlier, we set membership values $u_{ig} = 0$ for every $g = 1, ..., G$. This is in contrast with [29] which sets $u_{ig} = 1/G$ for outlying observations. A fuzzy Classification Maximum Likelihood approach is applied in the three considered approaches. The maximization of fuzzified likelihoods is not a new idea in fuzzy clustering [15, 28, 23, 27]. It is important to fix some type of constraint on the scatters parameters because that maximization is a mathematically ill-defined problem otherwise. Therefore, appropriate constraints on the scatter parameters must be added. These constraints are also useful to avoid the detection of non-interesting ("spurious") local maxima.

In the three reviewed methods, the third one is particularly appealing because it simultaneously protects us against "vertical outliers" and even "bad leverage" points. This approach, recently introduced in [13], is a trimmed and fuzzified version of the Cluster Weighted Model (CWM) in [14].

2 Three different approaches

Let $\widetilde{\mathbf{X}} = (\mathbf{X}', Y)'$ be a random vector in $\mathbb{R}^d \times \mathbb{R}$, where the first d components $\mathbf{X}$ are the values taken by the explanatory variables or covariates, and Y is the value taken by a response variable. Let us assume that $\{\widetilde{\mathbf{x}}_i\}_{i=1}^n =$

$\{(\mathbf{x}_i', y_i)'\}_{i=1}^n$ is a random sample of size n, drawn from $\widetilde{\mathbf{X}}$. We use the notation $\phi_d(\cdot; \mathbf{m}, \mathbf{S})$ for the density of the d-variate Gaussian distribution with mean vector $\mathbf{m}$ and covariance matrix $\mathbf{S}$ and $\{\lambda_l(\mathbf{S})\}_{l=1}^d$ are the set of eigenvalues of the $d \times d$ matrix $\mathbf{S}$.

2.1 FTCLUST-based approach

The simplest approach follows from the application of the FTCLUST methodology introduced in [7] in dimension $d+1$. We propose maximizing

$$\sum_{i=1}^{n}\sum_{g=1}^{G} u_{ig}^m \log\left(\pi_g \phi_{d+1}(\widetilde{\mathbf{x}}_i; \widetilde{\boldsymbol{\mu}}_g, \widetilde{\boldsymbol{\Sigma}}_g)\right), \tag{1}$$

where the $u_{ig} \in [0,1]$ membership values are required to satisfy

$$\sum_{g=1}^{G} u_{ig} = 1 \text{ if } i \in \mathcal{I} \text{ and } \sum_{g=1}^{G} u_{ig} = 0 \text{ if } i \notin \mathcal{I}, \tag{2}$$

for a subset $\mathcal{I} \subset \{1, 2, ..., n\}$ with $\#\mathcal{I} = [n(1-\alpha)]$. The parameter $\alpha \in [0, 1)$ is the fixed trimming level and $m \geq 1$ is the fuzzifier parameter. Note that observations with indexes outside $\mathcal{I}$ do not contribute to the summation in (1). This target function (1) is unbounded as we can easily seen just by taking $|\widetilde{\boldsymbol{\Sigma}}_g| \to 0$. Thus, as done in [7], we introduce an additional constraint when maximizing (1) that forces the set eigenvalues of the scatter matrices to satisfy

$$\lambda_{l_1}(\widetilde{\boldsymbol{\Sigma}}_{g_1}) \leq c\lambda_{l_2}(\widetilde{\boldsymbol{\Sigma}}_{g_2}) \quad \text{for } 1 \leq l_1 \neq l_2 \leq d+1 \text{ and } 1 \leq g_1 \neq g_2 \leq G. \tag{3}$$

This type of constraints are an extension of those in [20, 9]. The use of constraints on the scatter parameters goes back to the seminal work by [16].

Let $\widetilde{\boldsymbol{\mu}}_{\mathbf{g}}$ and $\widetilde{\boldsymbol{\Sigma}}_g$ be the vectors and matrices obtained from the previous constrained optimization problems with

$$\widetilde{\boldsymbol{\mu}}_{\mathbf{g}} = \begin{pmatrix} \boldsymbol{\mu}_1^g \\ \mu_2^g \end{pmatrix} \text{ and } \widetilde{\boldsymbol{\Sigma}}_g = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11}^g & \boldsymbol{\Sigma}_{12}^g \\ \hline \boldsymbol{\Sigma}_{21}^g & \Sigma_{22}^g \end{array}\right). \tag{4}$$

From these (optimal) vectors and matrices, we obtain G linear structures as

$$y = \mu_2^g + \boldsymbol{\Sigma}_{21}^g(\boldsymbol{\Sigma}_{11}^g)^{-1}(\mathbf{x} - \boldsymbol{\mu}_1^g) \text{ for } g = 1, ..., G. \tag{5}$$

The constant $1 \leq c < \infty$ guarantees that the constrained maximization of (1) is a mathematically well-defined problem and serves to avoid the detection of "spurious" local maximizers. This type of constraints are an extension

of those in [20, 9]. Some weights π_g are also included in (8). We are thus considering a fuzzy classification EM-type approach as in [28]. These weights are interesting when the number of clusters is misspecified because weights can be set close to 0 whenever G is larger than needed [9, 7].

The problem with that approach is that linear clusters are generally, by definition, elongated clusters. Therefore, eigenvalues close to 0 on the $\widetilde{\mathbf{\Sigma}}_g$ matrices often appear in most of clusterwise regression problems. This fact implies that large c values for the eigenvalues ratio constraint are required. Unfortunately, those large c values do not protect us correctly against "spurious" local maximizers. Moreover, the FTCLUST's good robustness properties are lost with such large c values.

2.2 Robust fuzzy clusterwise regression

A different approach, which directly takes into account the underlying linear relations within each group is reviewed in this section. In clusterwise regression, it is frequently assumed that the conditional relation between Y given $\mathbf{X} = \mathbf{x}$ in the g-th group can be written as $Y = \mathbf{b}_g'\mathbf{x} + b_g^0 + \varepsilon_g$ with $\varepsilon_g \sim N_1(0, \sigma_g^2)$. In that case, a robust fuzzy clusterwise regression approach can be derived through the maximization of

$$\sum_{i=1}^{n}\sum_{g=1}^{G} u_{ig}^m \log\left(\pi_g \phi_1(y_i; \mathbf{b}_g'\mathbf{x}_i + b_g^0, \sigma_g^2)\right), \tag{6}$$

where the u_{ig} membership values and the π_g weights satisfy the same requirements as those in Section 2.1. Vector $\mathbf{b}_g$ and constant b_g^0 are, respectively, the regression slopes vector and the intercept for the g-th cluster. Again, constraints on the residual variances can be set as

$$\sigma_{g_1}^2 \leq c_\varepsilon \sigma_{g_2}^2 \qquad \text{for every } 1 \leq g_1 \neq g_2 \leq G, \tag{7}$$

for a fixed $1 \leq c_\varepsilon < \infty$ constant. These constraints again convert the maximization of (6) into a mathematically well-defined problem (see what happens when $\sigma_g^2 \to 0$). This approach has been introduced in [11].

We have applied this methodology, for a simulated data set, with $\alpha = 0$ and $c_\varepsilon = 10$ in Figure 1,(a) and with $\alpha = 0.05$ and $c_\varepsilon = 10$ in Figure 1,(b). The simulated data set includes a small 5% fraction of background scattered noise. As seen in Figure 1,(a), the detected linear structures when $\alpha = 0$ are not the correct ones and many misclassified observations are found.

This approach provides improved robustness performance by applying trimming that certainly protect us against "vertical outliers" (outliers only in y). However, as will be seen in Section 2.3, it does not provide great protection against "leverage points" (outliers in $\mathbf{x}$). It is well known that leverage

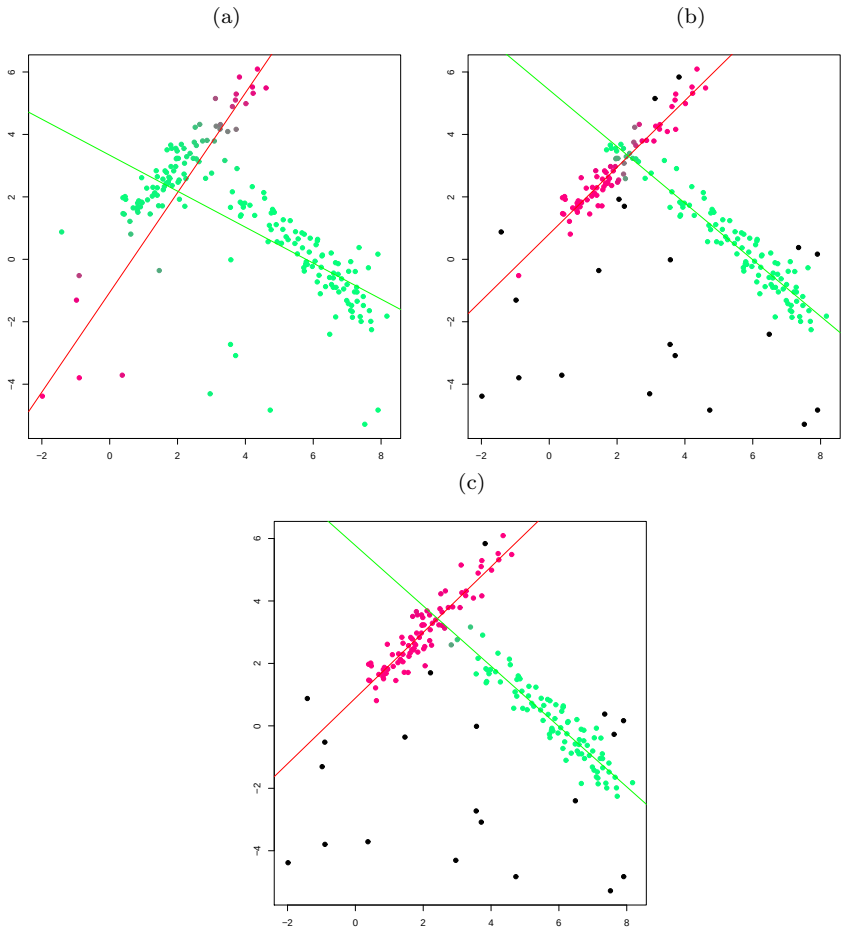

Fig. 1 (a) Fuzzy clusterwise regression with $\alpha = 0$ (b) Fuzzy clusterwise regression with $\alpha = 0.05$ (c) Fuzzy robust CWM with $\alpha = 0.05$ Fuzzy membership values are represented by using a mixture of red and green colors.

points can be extremely harmful in Regression Analysis. Additional protection, in that case, can be obtained by applying a "second trimming" stage as described in [10], which can be straightforwardly adapted to the fuzzy clustering framework.

2.3 Robust fuzzy cluster-weighted model (CWM)

Finally, a third approach is obtained throughout the "fuzzification" and "robustification" of the Cluster Weighted Model (CWM in the sequel) introduced

in [14]. This approach has been recently proposed in [13] as a fuzzification of the "hard" robust CWM in [12]. We just focus on the the linear CWM with Gaussian components where the conditional relationship between Y given $\mathbf{X} = \mathbf{x}$ in the g-th group is $Y = \mathbf{b}_g'\mathbf{x} + b_g^0 + \varepsilon_g$ with $\varepsilon_g \sim N_1(0, \sigma_g^2)$ but we also assume that $\mathbf{X} \sim N_d(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$.

Under these assumptions, we now consider the maximization of

$$\sum_{i=1}^{n}\sum_{g=1}^{G} u_{ig}^m \log\left(\pi_g \phi_1(y; \mathbf{b}_g'\mathbf{x}_i + b_g^0, \sigma_g^2)\phi_d(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g))\right), \tag{8}$$

with the same notation as in the statements of the two previous problems. We have that (8) is unbounded, and consequently, we introduce two further constraints as done in [12]. The first one has to do with the eigenvalues of the $\boldsymbol{\Sigma}_g$ matrices throughout

$$\lambda_{l_1}(\boldsymbol{\Sigma}_{g_1}) \leq c_X \lambda_{l_2}(\boldsymbol{\Sigma}_{g_2}) \quad \text{for every } 1 \leq l_1 \neq l_2 \leq d \text{ and } 1 \leq g_1 \neq g_2 \leq G. \tag{9}$$

A second constraint is added on the regression error terms as

$$\sigma_{g_1}^2 \leq c_\varepsilon \sigma_{g_2}^2 \qquad \text{for every } 1 \leq g_1 \neq g_2 \leq G.. \tag{10}$$

Notice that the two (not necessarily equal) constants $1 \leq c_X < \infty$ and $1 \leq c_\varepsilon < \infty$ serve to avoid "spurious" solutions whenever they are not excessively high ones. Moreover, a very flexible methodology is obtained because of the asymmetric treatment given to the marginal and conditional distributions.

Figure 1,(c) shows the results of applying the fuzzy robust CWM with $\alpha = 0.05$ and $c_X = c_\varepsilon = 10$ for the same simulated data set as above. The methodology in Section 2.2 was perfectly able to recover the two underlying linear structures (recall Figure 1,(b)). However, the cluster assignments are not so satisfactory because some observations which clearly belong to the cluster in the left have higher membership values to the cluster in the right. This issue is due to the fact that they are very close to the regression line fitted by using mainly the observations in the cluster in the right when this line is being elongated. On the other hand, the fuzzy robust CWM take advantage of the information conveyed in the marginal $\mathbf{X}$ distribution and so it is able to obtain more sensible membership values.

A second interesting feature of this fuzzy robust CWM is that it addresses the previously commented problems with "bad leverage" points in a very natural way because these observations take anomalous values on the explanatory variables. Therefore, their contribution to (8) is not very large and they are trimmed. For instance, we see in Figure 2,(a) how a 5% fraction of concentrated observations ($y \simeq 4$) are acting as bad leverage points when using the fuzzy clusterwise regression even though we had chosen a trimming level $\alpha = 0.05$ (equal to true contamination level) for it. The robust fuzzy

CWM, with the same trimming level, succesfully trim bad leverage observations.

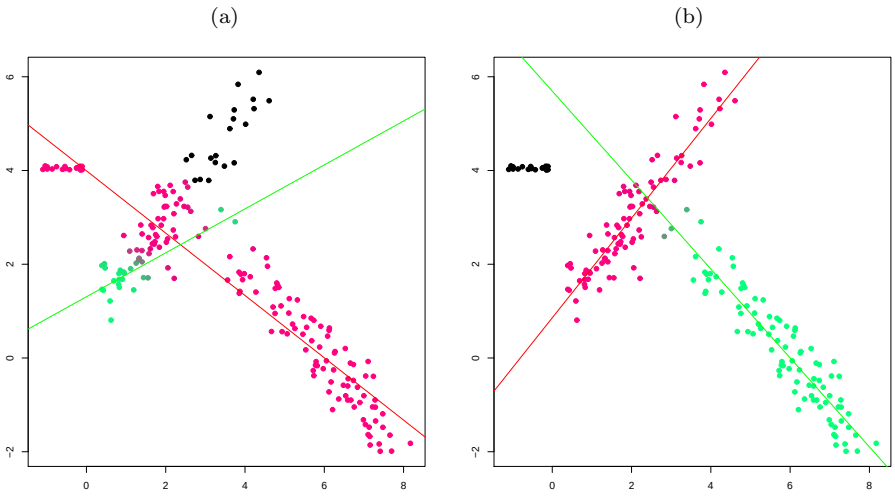

Fig. 2 (a) Robust fuzzy clusterwise regression with $\alpha = 0.05$ and $c_\varepsilon = 10$ for a data set with a 5% fraction of concentrated noise ($y \simeq 4$) (b) Robust fuzzy CWM with $\alpha = 0.05$ and $c_X = c_\varepsilon = 10$ for the same data set.

3 Algorithms and tuning parameters

In this section, we briefly outline the proposed algorithms to implement the previously reviewed approaches. Note that the target function in all of them can be written as

$$\sum_{i=1}^{n}\sum_{g=1}^{G} u_{ig}^{m} \log(\pi_g \varphi(\widetilde{\mathbf{x}}_i;\theta_g)), \tag{11}$$

where function $\varphi(\cdot)$ and the set of θ_g parameters change depending on the method applied.

1. *Initialization:* Initial θ_g and π_g parameters are randomly generated. Small random subsamples from our original data set are used to obtain these initial parameters.
2. *Iterative steps:* Repeat the following steps until convergence or reaching a maximal number of iterations:

2.1. *Membership values:* If $\max_{g=1,\ldots,G} \pi_g \varphi(\widetilde{\mathbf{x}}_i;\theta) \geq 1$, then

$$u_{ig} = I\big\{\pi_g \varphi(\widetilde{\mathbf{x}}_i;\theta_g) = \max_{q=1,\ldots,k} \pi_q \varphi(\widetilde{\mathbf{x}}_i;\theta_q)\big\}, \tag{12}$$

where $I\{\cdot\}$ is the 0-1 indicator function. If $\max_{g=1,...,G} \pi_g \varphi(\widetilde{\mathbf{x}}_i; \theta_g) < 1$, we set

$$u_{ig} = \left(\sum_{q=1}^{G} \left(\frac{\log(\pi_g \varphi(\widetilde{\mathbf{x}}_i; \theta_g))}{\log(\pi_q \varphi(\widetilde{\mathbf{x}}_i; \theta_q))} \right)^{\frac{1}{m-1}} \right)^{-1}. \tag{13}$$

2.2. *Trimmed observations:* Compute

$$r_i = \sum_{g=1}^{G} u_{ig}^m \log(p_g \varphi_g(\mathbf{x}_i, y_i; \theta)) \tag{14}$$

and sort them as $r_{(1)} \leq r_{(2)} \leq ... \leq r_{(n)}$. Set membership values $u_{ig} = 0$, $g = 1, ..., G$, for all the indexes i such that $r_i < r_{([n\alpha])}$.

2.3. *Update parameters:* Use previous u_{ig} to update weights as

$$\pi_g = \sum_{i=1}^{n} u_{ig}^m \Big/ \sum_{i=1}^{n} \sum_{g=1}^{G} u_{ig}^m. \tag{15}$$

and $\boldsymbol{\mu}_g$ (analogously, $\widetilde{\boldsymbol{\mu}}_g$) as

$$\boldsymbol{\mu}_g = \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{x}_i}{\sum_{i=1}^{n} u_{ig}^m}. \tag{16}$$

Update intercepts and slope vectors by computing

$$\mathbf{b}_g = \left(\frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{x}_i \mathbf{x}_i'}{\sum_{i=1}^{n} u_{ig}^m} - \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{x}_i}{\sum_{i=1}^{n} u_{ig}^m} \cdot \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{x}_i'}{\sum_{i=1}^{n} u_{ig}^m} \right)^{-1} \cdot \left(\frac{\sum_{i=1}^{n} u_{ig}^m y_i \mathbf{x}_i}{\sum_{i=1}^{n} u_{ig}^m} - \frac{\sum_{i=1}^{n} u_{ig}^m y_i}{\sum_{i=1}^{n} u_{ig}^m} \cdot \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{x}_i}{\sum_{i=1}^{n} u_{ig}^m} \right),$$

and

$$b_g^0 = \frac{\sum_{i=1}^{n} u_{ig}^m y_i}{\sum_{i=1}^{n} u_{ig}^m} - \mathbf{b}_g' \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{x}_i}{\sum_{i=1}^{n} u_{ig}^m}. \tag{17}$$

All previous formulae are typical in fuzzy clustering. The most difficult part is how to update the constrained scatter parameters. To update σ_g^2 and Σ_g^2, we start from the weighted sample covariance matrices

$$T_g = \frac{\sum_{i=1}^{n} u_{ig}^m (\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)'}{\sum_{i=1}^{n} u_{ig}^m}, \tag{18}$$

and the weighted residual variances

$$d_g^2 = \frac{\sum_{i=1}^{n} u_{ig}^m (y_i - b_g^0 - \mathbf{x}_i' \mathbf{b}_g)^2}{\sum_{i=1}^{n} u_{ig}^m}. \tag{19}$$

Then, to update $\mathbf{\Sigma}_g$ (analogously, $\widetilde{\mathbf{\Sigma}}_g$), the singular-value decomposition $T_g = U_g' E_g U_g$ is considered, with U_g being an orthogonal matrix and $E_g = \text{diag}(e_{g1}, e_{g2}, ..., e_{gd})$ a diagonal matrix. As done in [8, 7], these eigenvalues must be optimally truncated. The optimal truncation value is obtained by minimizing a real valued function. Analogously, in case that the d_j^2 error residual variances do not satisfy the required constraint, the d_j^2 must be optimally truncated too [11].

3. *Return* the set of θ_g of parameters yielding the highest value of (11) obtained after all the random initializations and iterative steps.

Note that trimming is done through "concentration steps" [24] and imposing the required constraint on the scatter parameters is an important ingredient of this algorithm.

As can be seen, several parameters have to be chosen when applying the proposed methods in real data problems. The estimated θ_g parameters do not necessarily dependent critically on all the tuning parameters. For instance, a trimming level slightly greater than the one needed to remove contamination is not necessarily problematic. However, monitoring the sizes of the sorted r_i values in (14) is useful to set sensible α values. Regarding the constraints on the scatter parameters, our suggestion is not choosing excessively high values for both c_X and c_ε (at least in the approaches described in Section 2.2 and 2.3). The choice of the fuzzifier parameter m depends on the desired degree of fuzziness in the clustering solution. Unfortunately, as happens with other likelihood-based fuzzy clustering approaches, the effect of m is affected by the scale of the measured variables (see [7, 11]). The joint monitoring of the proportions of "hard assignments" and "relative entropies" ($\sum_{g=1}^{G}\sum_{i=1}^{n} u_{ig}\log u_{ig}/[n(1-\alpha)]\log(G)$) provide useful heuristical tools aimed at addressing this issue.

References

1. Banerjee, A. and Davé, R.N. (2012), "Robust clustering," *WIREs Data Mining and Knowledge Discovery*, **2**, 2959.
2. Bezdek, J.C. (1981), *Pattern Recognition with Fuzzy Objective Function Algoritms*, Plenum Press, New York.
3. Davé, R.N. and Krishnapuram, R. (1997). "Robust clustering methods: a unified view". *IEEE Transactions on Fuzzy Systems*, **5**, 270-293.
4. DeSarbo, W.S. and Cron, W.L. (1988). "A Maximum Likelihood Methodology for Clusterwise Linear Regression", *Journal of Classification*, **5**, 249-282.
5. Dotto, F., Farcomeni, A., García-Escudero, L.A. and Mayo-Iscar, A. (2016), "A Fuzzy Approach to Robust Clusterwise Regression," Accepted for publication in *Advances in Data Analysis and Classification* DOI 10.1007/s11634-016-0271-9.
6. Farcomeni, A., Greco, L., (2015) *Robust Methods for Data Reduction* Chapman and Hall/CRC
7. Fritz, H., García-Escudero, L.A. and Mayo-Iscar, A. (2013) "Robust constrained fuzzy clustering," *Information Sciences*, **245**, 38–52.

8. Fritz H, García-Escudero LA, Mayo-Iscar A (2013) "A fast algorithm for robust constrained clustering" *Computational Statistics & Data Analysis,* **61** 124–136.
9. 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.
10. 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.
11. García-Escudero, L.A., Gordaliza, A., San Martín, R. and Mayo-Iscar, A. (2010) "Robust Clusterwise linear regresssion through trimming." *Computational Statistics and Data Analysis*, **54**, 3057-3069.
12. García-Escudero, L.A., Gordaliza, A., Greselin, F., Ingrassia, S. and 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.
13. García-Escudero, L.A., Greselin, F. and Mayo-Iscar, A. (2017) "Robust fuzzy cluster weighted modeling", *Submitted*.
14. Gershenfeld, N., Schoner, B., Metois, E., (1999), "Cluster-weighted modelling for time-series analysis", *Nature*, **397**, 329–332.
15. 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.
16. Hathaway, R. (1985) "A constrained formulation of maximum-likelihood estimation for normal mixture distributions" *The Annals of Statistics*, **13**, 795–800.
17. Hathaway, R.J. and Bezdek, J.C. (1993). "Switching regression models and fuzzy clustering". *IEEE Transactions on Fuzzy Systems*, **1**, 195-204.
18. Honda, K., Ohyama, T., Ichihashi, H. and Hotsu, A., FCM-type switching regression with alternating least square method, in *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ 2008)*, 122-127 (2008).
19. Hosmer, D.W. Jr. (1974), "Maximun Likelihood estimates of the parameters of a mixture of two regression lines." *Communications in Statstics* **3**, 995-1006.
20. Ingrassia S, Rocci R (2007) Constrained monotone EM algorithms for finite mixture of multivariate Gaussians, Computational Statistics & Data Analysis, 51:5339–5351.
21. Lenstra, A.K., Lenstra J.K., Rinnoy Kan, A.H.G., Wansbeek, T.J. (1982) "Two lines least squares" *Annals of Discrete Mathematics* **16**, 201-211
22. Ritter, G. *Robust Cluster Analysis and Variable Selection*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 2015.
23. Rousseeuw, P.J., Kaufman, L. and Trauwaert, E. (1996). "Fuzzy clustering using scatter matrices". *Computational Statistics and Data Analysis*, **23**, 135-151.
24. Rousseeuw, P.J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, **41**, 212-223.
25. Ruspini, E. (1969). "A new approach to clustering". *Information and Control*, **15**, 22-32.
26. Späth, H. (1982), "A Fast Algorithm for Clusterwise Regression" *Computing* **29**, 175-181.
27. 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.
28. Yang, M.-S. (1993). "On a class of fuzzy classification maximum likelihood procedures" *Fuzzy Sets and Systems* **57**, 365337.
29. Wu, W.L., Yang, M.S., Hsieh, N.J. (2009). "Alternative fuzzy switching regression", in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, **1**.