# Robust Fuzzy Clustering via Trimming and Constraints

Francesco Dotto[1], Alessio Farcomeni[1], Luis Angel García-Escudero[2], and Agustín Mayo-Iscar[2]

**Abstract** A methodology for robust fuzzy clustering is proposed. This methodology can be widely applied in very different statistical problems given that it is based on probability likelihoods. Robustness is achieved by trimming a fixed proportion of "most outlying" observations which are indeed self-determined by the data set at hand. Constraints on the clusters' scatters are also needed to get mathematically well-defined problems and to avoid the detection of non-interesting spurious clusters. The main lines for computationally feasible algorithms are provided and some simple guidelines about how to choose tuning parameters are briefly outlined. The proposed methodology is illustrated through two applications. The first one is aimed at heterogeneously clustering under multivariate normal assumptions and the second one migh be useful in fuzzy clusterwise linear regression problems.

## 1 Introduction

Hard clustering methods are aimed at searching meaningful partitions of a data set into $k$ disjoint clusters. Therefore, "0-1" membership values of observations to clusters are provided. On the other hand, fuzzy clustering methods provide nonnegative membership values which may generate overlapping clusters where every subject is shared among all clusters [28, 2].

It is known that the presence of an (even a small) amount of outlying observations can be problematic when applying traditional hard clustering methods. For instance, clearly differentiated clusters can be wrongly joined together and non-interesting clusters (made up of only few outlying observations) can be detected. This is also the case when applying many fuzzy

Sapienza University of Rome, Rome, Italy. e-mail: francesco.dotto@uniroma1.it, alessio.farcomeni@uniroma1.it · University of Valladolid, Valladolid, Spain. e-mail: lagarcia@eio.uva.es, agustim@eio.uva.es

clustering techniques. In fact, historically, the fuzzy clustering community was the first one to face this robustness issue. This is due to the fact that outliers may be approximately "equally remote" from all clusters and, thus, they may have similar (but not necessarily small) membership values.

References on robustness in hard clustering can be found in [10] and in two recent [7, 24] books. On the other hand, [4, 1] are good reviews on robust fuzzy clustering. These proposals in fuzzy clustering include "noise clustering" [3], the replacement of the Euclidean distance by other discrepancy measures [31, 22] or the use of "possibilistic" clustering [19].

Trimming has a long history as a simple way to provide robustness to statistical procedures. Its application in clustering needs to be done by taking into account the possibility of discarding "bridge points". A sensible way to perform trimming is to let the data decide which observations must be trimmed such that we find an optimal clustering for the non-trimmed ones. This is the "impartial" trimming approach adopted when using the TCLUST method [9]. This approach was extended in [8] to fuzzy clustering. This can be also seen as an extension of the "least trimmed squares" approach in fuzzy clustering [17]. Discarding a fixed fraction of data was also considered in [18].

One clear advantage of the methodology in [8] is that it allows the detection of non-necessarily spherically-shaped clusters. Additionally, the use of likelihoods in its statement allows its generalization to very different frameworks. The use of procedures based on likelihoods is not new in fuzzy clustering (see, e.g., [13, 12, 32, 25, 26, 30]). Note also that some type of constraint on the clusters' scatters is always needed. Otherwise, the defining problem would become a mathematically ill-posed one. By using these constraints, clusters with arbitrarily very different scatters are not allowed. The use of procedures based on likelihoods is also useful in clusterwise linear regression problems. Instead of detecting clusters just around centroids, it is often interesting to detect clusters around linear structures [15, 21, 29] (hard clustering) and [14, 16] (fuzzy clustering).

## 2 Methodology

Suppose that we have $n$ observations $\{x_1, ..., x_n\}$ in $\mathbb{R}^p$ and we want to group them into $k$ clusters in a fuzzy way. Therefore, our aim is to obtain a collection of nonnegative membership values $u_{ij} \in [0, 1]$ for all $i = 1, ..., n$ and $j = 1, ..., k$. A membership value 1 indicates that object $i$ fully belongs to cluster $j$ while a 0 membership value means that it does not belong at all to this cluster. However, intermediate degrees of membership are allowed when $u_{ij} \in (0, 1)$. We consider that an observation is fully trimmed if $u_{ij} = 0$ for all $j = 1, ..., k$.

Let us assume that $\varphi(\cdot; \theta_j)$ is a $p$-variate probability density function in $\mathbb{R}^p$ that depends on a set of parameters $\theta_j$. Given a fixed trimming level $\alpha \in [0, 1)$ and a fixed value of the fuzzifier parameter $m > 1$; a robust constrained fuzzy

clustering problem can be defined through the maximization of:

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

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

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

for a subset $\mathcal{I}\subset\{1,2,...,n\}$ with $\#\mathcal{I}=[n(1-\alpha)]$, when $\theta=(\theta_1,...,\theta_k)\in\Theta$, for a given parametric space $\Theta$, and the $p_j$'s are positive weights satisfying $\sum_{j=1}^{k}p_j=1$. Notice that $u_{i1}=...=u_{ik}=0$ for all $i\notin\mathcal{I}$, so these observations do not contribute to the summation in (1). The notation $[\cdot]$ is used for the floor function.

For instance, we may consider $\theta_j=(m_j,S_j)$ and

$$\varphi(x_i;\theta_j)=(2\pi)^{-p/2}|S_j|^{-1}\exp\big(-(x_i-m_j)'S_j^{-1}(x_i-m_j)/2\big). \tag{2}$$

In a clusterwise linear regression framework, if $x_i=(y_i,\mathbf{x}_i')$ with $y_i\in\mathbb{R}$ as the response variable value and $\mathbf{x}_i\in\mathbb{R}^{p-1}$ as the values taken by $p-1$ explanatory variables, then we can use $\theta_j=(\beta_j,s_j^2)$ and

$$\varphi(x_i;\theta_j)=(2\pi s_j^2)^{-1/2}\exp\big(-(y_i-\mathbf{x}_i'\beta_j)^2/(2s_j^2)\big). \tag{3}$$

In the target function (1), clusters' weights $p_j$'s are also included. This may be seen as an "entropy regularization" [23]. Including these weights is interesting when the number of clusters is misspecified, because some $p_j$ weights can be set close to 0 when $k$ is larger than the "true" number of clusters. Another possibility is to exclude these weights by directly assuming $p_1=...=p_k=1/k$. This would shrink assignments towards similar number of observations within each cluster.

It is important to note that the maximization of (1) when $k>1$ is commonly an ill-posed problem without any constraint on the scatter parameters. For instance, in the two previous problems, we can see that (1) becomes unbounded when $|S_j|\to 0$ or when $s_j^2\to 0$. Additionally, these constraints are useful to avoid the detection of non-interesting "spurious" solutions. Thus, in [8], it is proposed the use of an eigenvalue ratio constraint

$$\frac{\max_{j=1}^{k}\max_{l=1}^{p}\lambda_l(S_j)}{\min_{j=1}^{k}\min_{l=1}^{p}\lambda_l(S_j)}\leq c, \tag{4}$$

for a fixed constant $c\geq 1$, where $\{\lambda_l(S)\}_{l=1}^{p}$ denote the $p$ eigenvalues of the matrix $S$. In a similar way, the use of (3) with the constraint

$$\frac{\max_{j=1}^{k} s_j^2}{\min_{j=1}^{k} s_j^2} \leq c, \tag{5}$$

is proposed in [6] for fuzzy clusterwise linear clustering.

Therefore, if $\Theta_c \subseteq \Theta$ denotes the restricted parametric space, the maximization of (1) when $\theta \in \Theta_c$ yields the FTCLUST method ($\varphi(\cdot)$ as in (2) and (4)) and the FTCLUST-R method ($\varphi(\cdot)$ as in (3) and (5)).

## 3 Algorithm

The maximization of (1) under those constraints is not an easy problem. However, a feasible algorithm can be given:

1. *Initialization:* The procedure is initialized several times by randomly selecting initial $\theta_j$'s parameters. This can be done by selecting $k$ subsets of size $p+1$ in general position. Fitting $k$ simple models within each subsample allows to obtain these initial $\theta_j$'s. Weights $p_1, ..., p_k$ with $p_j \in (0, 1)$ and summing up to 1 are also randomly chosen.
2. *Iterative steps:* The following steps are executed until convergence or a maximum number of iterations is reached.

2.1. *Membership values:* If $\max_{j=1,...,k} p_j\varphi(x_i;\theta_j) \geq 1$, then

$$u_{ij} = I\big\{p_j\varphi(x_i;\theta_j) = \max_{q=1,...,k} p_q\varphi(x_i;\theta_q)\big\} \text{ (hard assignment)},$$

with $I\{\cdot\}$ as the 0-1 indicator function. If $\max_{q=1,...,k} p_q\varphi(x_i;\theta_q) < 1$, then

$$u_{ij} = \left(\sum_{q=1}^{k}\left(\frac{\log(p_j\varphi(x_i;\theta_j))}{\log(p_q\varphi(x_i;\theta_q))}\right)^{\frac{1}{m-1}}\right)^{-1} \text{ (fuzzy assignment)}.$$

2.2. *Trimmed observations:* Let

$$r_i = \sum_{j=1}^{k} u_{ij}^m \log(p_j\varphi(x_i;\theta_j)) \tag{6}$$

and $r_{(1)} \leq r_{(2)} \leq ... \leq r_{(n)}$ be these values sorted. The observations to be trimmed are those with indexes $\{i : r_i < r_{([n\alpha])}\}$. The membership values for those observations are redefined as $u_{ij} = 0$, for every $j$ if $r_i < r_{([n\alpha])}$.

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

$$p_j = \sum_{i=1}^{n} u_{ij}^m \Big/ \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m,$$

and the $\theta_j$'s are updated by maximizing (1) where the $u_{ij}$'s are those obtained in the previous step. For instance, this maximization implies the use of weighted means and weighted covariance matrices for the FTCLUST and the use of weighted least squares for the FTCLUST-R (weights $u_{ij}^m$ in both cases). In more general frameworks, a weighted likelihood should be maximized in a closed form or numerically.
It may happen that these so obtained $\theta_j$'s do not fall within $\Theta_c$. In this case, as done in [8] and [6], it is needed to modify them properly by using optimally truncated scatter parameters. I.e., if $\{d_l\}$ are these scatter parameters (eigenvalues in the case of the FTCLUST and error terms' variances in the case of FTCLUST-R), then we use

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

with $t$ being a threshold value. Note that these truncated $\{d_l\}$ do satisfy the required constraints and we only need to obtain the optimal threshold value $t_{opt}$ which maximizes (1). Sometimes, there are closed forms expressions for obtaining $t_{opt}$ (see [8] and [6]).

3. *Evaluate objective function* and return parameters yielding the highest (1).

This algorithm can be seen as a fuzzy extension of the classical EM algorithm [5] where "concentration steps", as those in [27], are also applied. Note also that it naturally leads to a fuzzy clustering method with "high contrast" [25] (a compromise between "hard" and "fuzzy" clustering methods).

## 4 Tuning parameters

The proposed methodology exhibits high flexibility but the price we pay is that of specifying several tuning parameters. In this section, we briefly discus about them and we give some practical guidelines for their choice.

*Fuzzifier parameter:* Parameter $m$ serves to control the degree of fuzziness in the obtained clustering. The $m = 1$ case provides "hard" or "crisp" clustering membership values. In fact, with $m = 1$, we recover the TCLUST method in [9] from the FTCLUST and the robust linear grouping in [11] (without second trimming) from the FTCLUST-R. However, there is an unexpected problem if $m > 1$ when applying fuzzy clustering approaches based on the maximum likelihood principle. This inherent problem has to do with the different effect of $m$ depending on the scale (i.e.,

when we replace $x_i$ by $S \cdot x_i$ for a given constant $S$). This problem can be addressed by choosing simultaneously $m$ and the scale of data ($S$) in such a way that we achieve some pre-specified "proportions of hard assignments" and "relative entropy". The relative entropy is defined as $\sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} \log u_{ij} / [n(1-\alpha)] \log(k)$.

*Trimming level:* The trimming level $\alpha$ is the proportion of observations discarded. Although an $\alpha$ value smaller than the true contamination level can be problematic, we can see that $\alpha$ (slightly) higher than needed most of times provides good $\theta_j$ estimates. Then, wrongly trimmed observations can be recovered back. Additionally, given a tentative $\alpha$ value and $r_{(1)} \leq ... \leq r_{(n)}$ being the sorted $r_i$ values in (6), we can check if this $\alpha$ was a sensible choice by seeing whether these $r_{(i)}$ increase quickly when $i/n < \alpha$ and increase slowly when $i/n > \alpha$.

*Constraint on the scatter parameters:* The constant $c$ serves to control the degree of "heteroscedasticity" in the obtained clusters. A large $c$ value allows for more different variances in the error terms when using FTCLUST-R. Large $c$ values also allows for more severe departures from sphericity in FTCLUST. The most constrained case $c = 1$ (with $\alpha = 0$ and "equal weights") yields the classical fuzzy $k$-means [2] when using FTCLUST and fuzzy $k$-regressions [14] when using FTCLUST-R.

## 5 An example

We conclude with an example of the application of FTCLUST to the "M5data" set in [9] (available at the `tclust` package in the CRAN repository). This data set is obtained from three normal bivariate distributions with different scales and proportions (see the "true" cluster labels in Fig.1(a)). One of the components strongly overlaps with another one and there is a 10% background noise. Fig.1(b) shows the very bad results obtained when applying FTCLUST with $\alpha = 0$ (all observations are wrongly shared with similar membership values). We can see in Fig.1(c) that the use $\alpha = .1$ and $c = 1$ gives better clustering results but it is unable to deal with the very different cluster scatters. Finally, Fig.1(d) shows the excellent results obtained $\alpha = .1$ and $c = 50$, i.e. a higher eigenvalues ratio constraint value.

**Acknowledgements** Research partially supported by the Spanish Ministerio de Economía y Competitividad, grant MTM2014-56235-C2-1-P, and by Consejería de Educación de la Junta de Castilla y León, grant VA212U13.

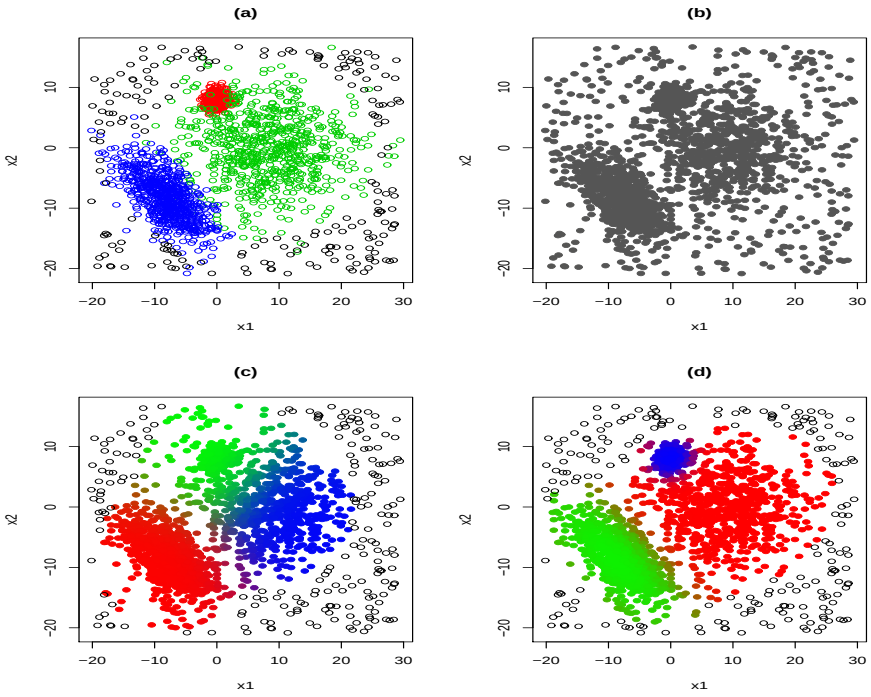


**Fig. 1** (a) "M5data" dataset with the true assignments. Results of applying FTCLUST with $\alpha = 0$ and $c = 1$ in (b), $\alpha = .1$ and $c = 1$ in (c) and $\alpha = .1$ and $c = 50$ in (d). A mixture of red, blue and green colors with intensities proportional to the membership values are used to summarize the clustering results and "o" are the trimmed observations.

## 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. (1991). "Characterization and detection of noise in clustering", *Pattern Recognition Letters*, **12**, 657664.
4. Davé, R.N. and Krishnapuram, R. (1997). "Robust clustering methods: a unified view". *IEEE Transactions on Fuzzy Systems*, **5**, 270-293
5. Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, **39**, 138.
6. Dotto, F., Farcomeni, A., García-Escudero, L.A. and Mayo-Iscar, A. (2016), "A Fuzzy Approach to Robust Clusterwise Regression," *submitted manuscript*.
7. Farcomeni, A., Greco, L., (2015) *Robust Methods for Data Reduction* Chapman and Hall/CRC
8. Fritz, H., García-Escudero, L.A. and Mayo-Iscar, A. (2013) "Robust constrained fuzzy clustering," *Information Sciences*, **245**, 38–52.

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 regresin through trimming." *Computational Statistics and Data Analysis*, **54**, 3057-3069.
12. Gath, I. and Geva, A.B. (1989), "Unsupervised optimal fuzzy clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 773 781.
13. 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.
14. Hathaway, R.J. and Bezdek, J.C. (1993). "Switching regression models and fuzzy clustering". *IEEE Transactions on Fuzzy Systems*, **1**, 195-204.
15. Hosmer, D.W. Jr. (1974), "Maximun Likelihood estimates of the parameters of a mixture of two regression lines." *Communications in Statstics* **3**, 995-1006.
16. Kuo-Lung, W., Miin-Shen ,Y. and June-Nan, H. (2009). "Alternative fuzzy switching regression" *Proceedings of the International MultiConference of Engineers and Computer Scientist* **1**
17. Kim, J., Krishnapuram, R. and Davé, R. (1996) "Application of the least trimmed squares technique to prototype-based clustering". *Pattern Recognition Letters*, **17**, 633-641.
18. Klawonn, F. (2004). "Noise clustering with a fixed fraction of noise". *Applications and Science in Soft Computing*. Springer, Berlin-Heidelberg-New York, 133138.
19. Krishnapuram, R. and Keller, J.M. (1993), "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, **1**, 98-110.
20. Krishnapuram, R. and Keller, J.M. (1996), "The possibilistic *C*-means algorithm: Insights and recommandations," *IEEE Transactions on Fuzzy Systems*, **4**, 385393
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. Łeski, J. (2003). "Towards a robust fuzzy clustering", *Fuzzy Sets and Systems*, **137**, 215-233.
23. Miyamoto, S. and Mukaidono, M. (1997). "Fuzzy *c*-means as a regularization and maximum entropy approach" *Proceedings of the 7th International Fuzzy Systems Association World Congress (IFSA'97)*, **2**, 86-92.
24. Ritter, G. *Robust Cluster Analysis and Variable Selection*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 2015.
25. Rousseeuw, P.J., Trauwaert, E. and Kaufman, L. (1995). "Fuzzy clustering with high contrast". *Journal of Computational and Applied Mathematics,* **64**, 81-90.
26. Rousseeuw, P.J., Kaufman, L. and Trauwaert, E. (1996). "Fuzzy clustering using scatter matrices". *Computational Statistics and Data Analysis*, **23**, 135-151.
27. Rousseeuw, P.J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, **41**, 212-223.
28. Ruspini, E. (1969). "A new approach to clustering". *Information and Control*, **15**, 22-32.
29. Späth, H. (1982), "A Fast Algorithm for Clusterwise Regression" *Computing* **29**, 175-181.
30. 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.
31. Wu, K.-L. and Yang, M.-S. (2002). "Alternative *c*-means clustering algorithms", *Pattern Recognition*, **35**, 2267-2278.
32. Yang, M.-S. (1993). "On a class of fuzzy classification maximum likelihood procedures" *Fuzzy Sets and Systems* **57**, 365337.