A robust clustering procedure with unknown number of clusters

Una procedura di cluster analysis robusta con un numero di cluster incognito

Francesco Dotto and Alessio Farcomeni

Abstract A new methodology for robust clustering without specifying in advance the underlying number of Gaussian clusters is proposed. The procedure is based on iteratively trimming, assessing the goodness of fit, and reweighting. The forward version of our procedure is initialized with a high trimming level and $K = 1$ populations. The procedure is then iterated throughout a fixed sequence of decreasing trimming levels. New observations are added at each step and, whenever a goodness of fit rule is not satisfied, the number of components K is increased. A stopping rule prevents our procedure from using outlying observations. Additional use of a backward criterion is discussed.

Abstract *In questo lavoro viene introdotta una metodologia per la cluster analysis robusta che non richiede la specificazione a priori del numero di cluster gaussiani. La procedura si basa, iterativamente, sul trimming, la valutazione della bontà di adattamento ed il reweighting. La sua versione forward viene inizializzata fissando un livello di trimming alto e $K = 1$ popolazioni sottostanti. In seguito la procedura viene iterata all'interno di una griglia fissata di livelli trimming decrescenti. Ad ogni passo vengono reinserite osservazioni e, laddove l'adattamento peggiori sostanzialmente, il numero di componenti K viene aumentato. Una regola di arresto garantisce che valori anomali non vengano usati per stimare i parametri. Si discute infine un criterio aggiuntivo di tipo backward.*

Key words: Trimming, Reweighting, Robustness

Francesco Dotto
Univeristy of Rome "Roma Tre", Via Silvio D'amico 77, 00145 Roma, e-mail: francesco.dotto@uniroma3.it

Alessio Farcomeni
Univerity of Rome "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma e-mail: alessio.farcomeni@uniroma1.it

1 Introduction

Model based clustering procedures for multivariate data can be inconsistent in presence of contamination. A trimming step is often used to guarantee robustness. Impartial trimming is based on discarding a fixed proportion α of observations lying far from their closest centroid. A detailed review of robust clustering may be found in part II of [7]. The procedure of [10] is based on pre-specifying α and the number of clusters K. Simultaneously fixing the tuning parameters is still an open problem. First, it shall be noticed that the two parameters are clearly intertwined. Indeed the optimal α depends on the chosen K and also the *vice versa* holds. We propose here to use iterative reweighting to obtain robust cluster analysis without having to specify in advance these two tuning parameters. Our approach is related to the forward search philosophy (e.g., [1]), but with substantial differences. The rest of the paper is as follows. In Section 2.1 we briefly review the `tclust` methodology and its reweighted version. Our new proposal for robust clustering is presented in Section 3 while its application to real data example is provided in Subsection 3.2. Finally Section 4 contains the concluding remarks and the further directions of research.

2 Trimming approach to cluster analysis and its reweighted version

2.1 The `tclust` methodology

Within this subsection we briefly present the `tclust` methodology, [10], whose `R` implementation is presented in [9] and its reweighted version introduced in [4]. Let $x_i \in \mathbb{R}^p$ be a sample point, $f(\cdot)$ the multivariate normal density, μ_j and Σ_j be location and scatter parameters, respectively, of the j-th group. Additionally let $g_{\psi_i}(\cdot)$ be the contaminating density and K the number of groups. Then the likelihood function associated to the spurious outliers model is given by:

$$\left[\prod_{j=1}^{K}\prod_{i\in R_j} f(x_i;\mu_j;\Sigma_j)\right]\left[\prod_{i\notin R_j} g_{\psi_i}(x_i)\right] \tag{1}$$

Additionally it must be pointed out that, in equation (1), $R = \bigcup_{j=1}^{K} R_j$ represents the set of the clean observations and is such that $\#R = \lceil n(1-\alpha)\rceil$ and only the clean data give a contribution to the likelihood function, while, noise component, whose likelihood is given by the right hand side of equation (1) give no contribution to the likelihood function. The maximum likelihood estimator of (1) exists if and only if the following condition on the contaminating density holds:

$$\arg\max_{\mathscr{R}} \max_{\mu_j,\Sigma_j} \prod_{j=1}^{k} \prod_{i\in R_j} f(x_i;\mu_j,\Sigma_j) \subseteq \arg\max_{\mathscr{R}} \prod_{i\notin \cup_{j=1}^{k} R_j} g_\psi(x_i) \tag{2}$$

As pointed out in [5], condition (2) states that identification of clean observations by maximization of the right hand term of (2) identifies the same observations as would identification of contaminated observations by maximizing the part of the likelihood corresponding to the noise. Thus, once clean observations are identified by maximizing the right hand term of (2), then the contaminated entries are optimally identified.
Additionally, if the condition (2) holds, the MLE of the likelihood function (1) has a simple representation and its maximization reduces to the maximization of:

$$\sum_{j=1}^{K} \sum_{i\in R_j} \log f(x_i;\mu_j,\Sigma_j) \tag{3}$$

2.2 The `tclust` without specifying α in advance

We now focus our attention to two tuning parameters, that are required to be fixed by the user in order to appy the `tclust` methodology: the trimming level α and the number of clusters K. In [4] is introduced a contribution, known as reweighted `tclust` or `rtclust` for the sake of brevity, designed to avoid the specification of the trimming level α. The idea behind such contribution is starting with a high trimming level α_0 the `tclust`, for which efficient computing is possible ([9]). Once the procedure is initialized, L decreasing trimming levels $\alpha_1 > \alpha_2 > \ldots > \alpha_L$ are fixed; then the `rtclust` algorithm proceeds, for each $l = 1,2\ldots,L$, as follows:

1. *Initialization:* Set the initial parameters' set $\pi_1^0,...,\pi_k^0,\pi_{k+1}^0$, $\mu_1^0,...,\mu_k^0$ and $\Sigma_1^0,...,\Sigma_k^0$ obtained by applying the `tclust` with a high trimming level α_0.
2. *Reweighting process:* Consider $\alpha_l = \alpha_0 - l\cdot\varepsilon$ with $\varepsilon = (\alpha_L - \alpha_0)/L$ for $l = 1,...,L$

   2.1 *Fill the clusters:* Given $\pi_1^{l-1},...,\pi_k^{l-1},\pi_{k+1}^{l-1}$, $\mu_1^{l-1},...,\mu_k^{l-1}$ and $\Sigma_1^{l-1},...,\Sigma_k^{l-1}$ from the previous step, let us consider

   $$D_i = \min_{1\le j\le k} d^2_{\Sigma_j^{l-1}}(x_i,\mu_j^{l-1}) \tag{4}$$

   and sort these values as $D_{(1)} \le ... \le D_{(n)}$. Take the sets

   $$A = \{x_i : D_i \le D_{([n(1-\alpha_l)])}\} \text{ and } B = \{x_i : D_i \le \chi^2_{p,\alpha_L}\}$$

   Now, use the distances in (4) to obtain a partition $A\cap B = \{H_1,...,H_k\}$ with

$$H_j = \left\{ x_i \in A \cap B \ : \ d_{\Sigma_j^{l-1}}(x_i, \mu_j^{l-1}) = \min_{q=1,\dots,k} d_{\Sigma_q^{l-1}}(x_i, \mu_q^{l-1}) \right\}.$$

2.2 *Update cluster weights* The proportion of contamination is estimated by computing

$$\pi_{k+1}^l = 1 - \frac{\#B}{n}.$$

Given $n_j = \#H_j$ and $n_0 = n_1 + ... + n_k$ the cluster weights are estimated by computing:

$$\pi_j^l = \frac{n_j}{n_0}\left(1 - \pi_{k+1}^l\right). \tag{5}$$

2.3 *Update locations and scatters:* Update the cluster centers by taking μ_j^l equal the sample mean of the observations in H_j and the scatter by computing the sample covariance matrix of the observations in H_j multiplied by its consistency factor.

3. *Output of the algorithm:* $\mu_1^L, ..., \mu_k^L$ and $\Sigma_1^L, ..., \Sigma_k^L$ are the final parameters estimates for the normal components. From them, final assignments are done by computing

$$D_i = \min_{1 \le j \le k} d^2_{\Sigma_j^L}(x_i, \mu_j^L),$$

for $i = 1, ..., n$. Observations assigned to cluster j are those in H_j with

$$H_j = \left\{ x_i : d_{\Sigma_j^L}(x_i, \mu_j^L) = \min_{q=1,\dots,k} d_{\Sigma_q^L}(x_i, \mu_q^L) \text{ and } D_i \le \chi^2_{p,\alpha_L} \right\}$$

and the trimmed observations are observations not assigned to any of these H_j sets (i.e., those observations with $D_i > \chi^2_{p,\alpha_L}$).

There are, in our opinion, two great advantages in using `rtclust`. First, as shown in th simulation study and the theoretical properties reported in [4], high robustness with high efficiency can be reached at the same time. Secondly no much tuning is required. Indeed the final estimated contamination level is independent to the initial trimming α_0 and the assumptions on constraint on the eigenvalues can be relaxed after the initialization. It shall be noticed that, besides the required number of groups K - to which are dedicated the further sections of the paper - the parameter α_L my need tuning too. Such parameter establishes how far the outliers are supposed to be placed with respect to the bulk of the data. Such choice is pretty subjective and strongly depends on the context of application only heuristics. We only recall the guidelines provided in [2] for generally tuning in robust statistics and the contribution provided in [6] where an example of the tuning of the parameter α_L is provided.

3 The `tclust` without specifying α and K in advance

3.1 Introduction

We now outline an automatic methodology based on reweighting that does not need the imposition of the desired number of groups by the user. We do so by applying the reweighting logic to both reinsert the wrongly discarded observations and increase the number of groups, if required. To do so, we resort the forward search philosophy outlined in [1]. In practice, we start by applying the `tclust` method with $K = 1$ and $\alpha = .9$ imposed. Then, we apply the reweighting approach outlined in [4] to estimate the true contamination level $\hat{\varepsilon}$ given $K = 1$ population. Once we can rely on a "precise" estimate of the contamination level we try to increase number of groups imposing $K_{try} = K + 1$ and a trimming level equal to $\hat{\varepsilon}$. The goodness of fit of this new proposed model is evaluated by computing the proportion of observations that are flagged as outlying in the new proposed model that were not flagged as outlying at the previous step. The underlying idea is the following. If a considerably high proportion of observations initially considered clean at the previous step, are recognized as outlying in the current step as a higher number of underlying groups is imposed, the this means that a high dense region of points (a potential cluster) has been trimmed off in the previous step. Algorithmically speaking we alternate the `tclust` and the `rtclust` up to convergence within the steps described in Algorithm 1. A graphical counterpart is provided in Figure 1.

Algorithm 1

`Initialization:`

1. *Fix:* $K_0 = 1, \alpha_0 = .9$ *and* $\rho \in [0.01, 0.05]$
2. *Let* mod_{rew} *be the output of* `rtclust` *with* K_0 *and* α_0 *imposed*

`Update:`

3. *Take* $\hat{\varepsilon}$ *estimated by the model* mod_{rew} *and set* $K_{try} = K + 1$.
4. *Launch the* `tclust` *with* $\alpha = \hat{\varepsilon}$ *and* $K = K_{try}$ *imposed.*
5. *Take* π_{new}*: the proportion of observations flagged as outlying by* mod_{try}.

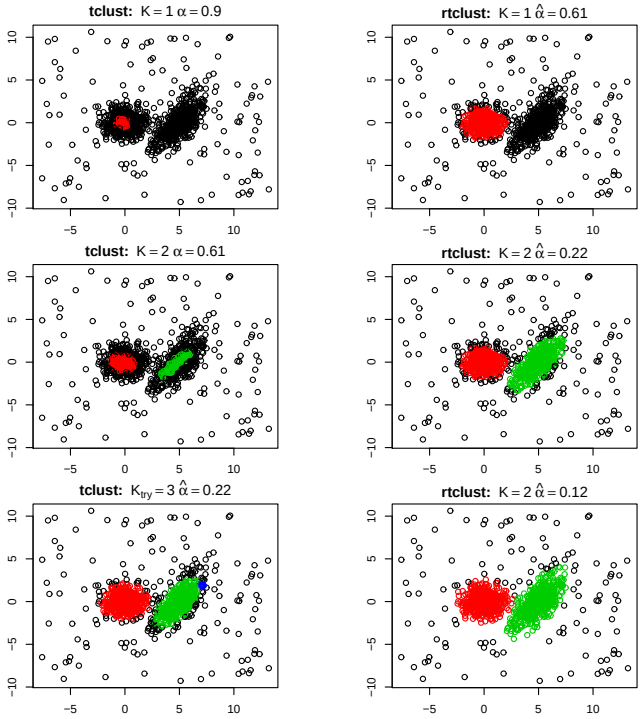
`Stopping rule:`

6. *If* $\pi_{new} \leq \rho$ *then* `stop`*. Else, if* $\pi_{new} > \rho$*:*
   - $K = K + 1$
   - *Calculate* mod_{rew} *by launching* `rtclust` *with K imposed.*
   - *Repeat steps 3-6.*

`Final output:`

7. *Return the output of* mod_{rew} *as the final output of the algorithm.*

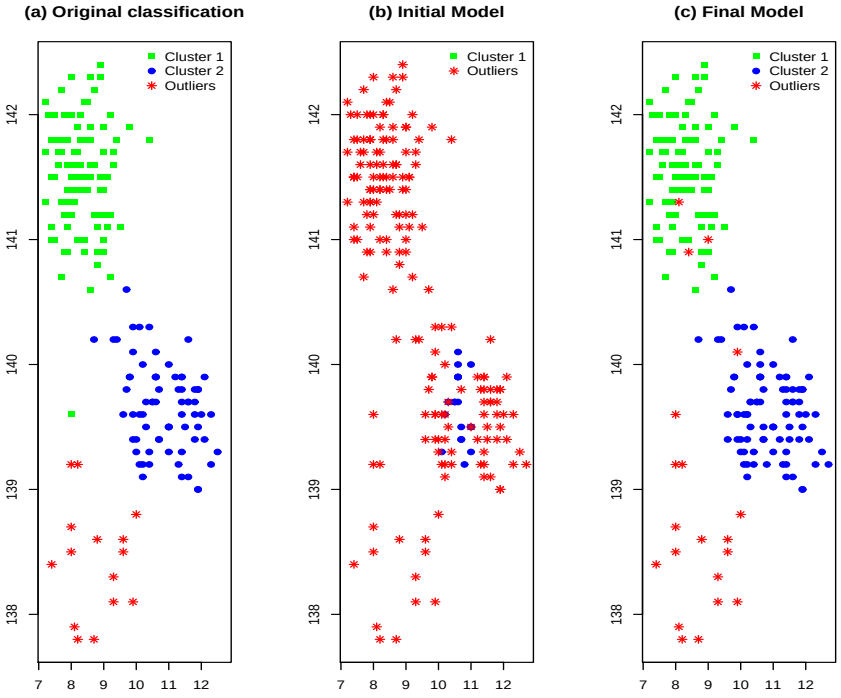
Fig. 1 The application of Algorithm 1 to a 2- dimensional simulated composed by $K = 2$ clusters and a proportion of contaminating points equal to 0.10.



3.2 A real data application

In this Section we apply the proposed iterative reweighting approach to the 6-dimensional "Swiss Bank Notes" data set presented in [8] which describes certain features of 200 Swiss 1000-franc bank notes divided in two groups: 100 genuine and 100 counterfeit notes. This is a well known benchmark data set. In [8], it is pointed out that the group of forged bills is not homogeneous since 15 observations arise from a different pattern and are, for that reason, outliers. As stated in Algorithm 1 we start by imposing a trimming level $\alpha_0 = .9$ and $K = 1$ clusters. The obtained results, that are briefly summarized in Figure 2, are in substantial agreement with characteristics described in [8]. Indeed $K = 2$ are automatically estimated by the algorithm while the estimated proportion of outliers is slightly overestimated: 10% of outliers are recognized by the algorithm while in [8] the declared percentage of outliers is equal to 7.5%.

Fig. 2 Fourth against the sixth variable of the Swiss Bank Notes data set. (a) The original classification. (b) The initial classification obtained by imposing $K = 1$ and $\alpha_0 = .9$. (c) The final output of Algorithm 1



4 Concluding Remarks

We outlined a robust procedure for clustering data that does not need the specification in advance of the required number of groups and of the proportion of outlying observations. We are aware of the fact that, as pointed out in [11], *There are no unique objective "true" or "best" clusters in a dataset. Clustering requires that the researchers define what kind of clusters they are looking for.* In conclusion, as pointed out in [3], we do not think that a fully automatized way to fix simultaneously all the parameters is to be expected. Indeed, the outlined methodology can be viewed as an additional tool to be combined with researchers' specification and a priori informations to provide a better understanding of the phenomenon of interest.

References

1. Atkinson, A.C., Riani, M., Cerioli, A.: Cluster detection and clustering with random start forward searches. Journal of Applied Statistics pp. 1–22 (2017)
2. Cerioli, A., Riani, M., Atkinson, A.C., Corbellini, A.: The power of monitoring: how to make the most of a contaminated multivariate sample. Statistical Methods & Applications pp. 1–29 (2018)
3. Dotto, F., Farcomeni, A., García-Escudero, L.A., Mayo-Iscar, A.: A fuzzy approach to robust regression clustering. Advances in Data Analysis and Classification **11**(4), 691–710 (2017)
4. Dotto, F., Farcomeni, A., García-Escudero, L.A., Mayo-Iscar, A.: A reweighting approach to robust clustering. Statistics and Computing **28**(2), 477–493 (2018)
5. Farcomeni, A.: Robust constrained clustering in presence of entry-wise outliers. Technometrics **56**, 102–111 (2014)
6. Farcomeni, A., Dotto, F.: The power of (extended) monitoring in robust clustering. Statistical Methods & Applications pp. 1–10
7. Farcomeni, A., Greco, L.: Robust methods for data reduction. CRC press (2016)
8. Flury, B., Riedwyl, H.: Multivariate Statistics. A Practical Approach. Chapman and Hall, London (1988)
9. Fritz, H., García-Escudero, L., Mayo-Iscar, A.: tclust: An R package for a trimming approach to cluster analysis. J Stat Softw **47** (2012). URL http://www.jstatsoft.org/v47/i12
10. García-Escudero, L., Gordaliza, A., Matrán, C., Mayo-Iscar, A.: A general trimming approach to robust cluster analysis. Ann Stat **36**, 1324–1345 (2008)
11. Hennig, C., Liao, T.F.: How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. Journal of the Royal Statistical Society: Series C (Applied Statistics) **62**(3), 309–369 (2013)