Statistical Methods & Applications manuscript No.
(will be inserted by the editor)

The power of (extended) monitoring in robust clustering

Alessio Farcomeni and Francesco Dotto

Received: date / Accepted: date

Abstract We complement the work of Cerioli, Riani, Atkinson and Corbellini by discussing monitoring in the context of robust clustering. This implies extending the approach to clustering, and possibly monitoring more than one parameter simultaneously. The cases of trimming and snipping are discussed separately, and special attention is given to recently proposed methods like double clustering, reweighting in robust clustering, and fuzzy regression clustering.
Key Words: Double Clustering, Fuzzy clustering, Multidimensional monitoring, Reweighting, Snipping, Tuning

1 Introduction

First of all, let us commend the authors of Cerioli et al. (2018) for a very clear and thought-provoking paper.

For the past several years the authors have given substantial contributions in the area of robust statistics. Their extensive work on the forward search has lead them to realize that monitoring robust procedures has an intrinsic informative power. In this paper they give some ideas on this, and illustrate with several interesting examples. We believe that monitoring might indeed be very informative, and might also contribute to lead several innovative and sound methods in robust statistics to be more often used in practice, and ultimately become the standard for applied sciences.

It is apparent in our opinion that monitoring is closely related to sensitivity analysis. The idea of tilting tuning parameters is certainly not new, and is often the simplest route to follow when no optimal choice is available (e.g., Farcomeni and Greco (2015), Farcomeni (2009)). The welcome and very innovative concept in Cerioli et al. (2018) is that the informative content obtained by monitoring goes beyond the mere aid to choice of tuning parameters.

Sapienza - University of Rome and University of Roma Tre

Monitoring has also an intuitive appeal which might make robust estimators more widely accepted, as it immediately allows the user to understand to what extent data are contaminated and classical estimators are (not) reliable (Farcomeni and Ventura 2012).

We believe that one should not necessarily monitor focusing on a single parameter. Several aspects can be varied simultaneously (generating a 3-D, or in general a p-D movie). The biggest challenge is computational, as in most cases robust estimators shall be computed from scratch for each new set of tuning parameters. In this sense, monitoring opens the issue of obtaining simple and quick updating rules when tuning parameters are tilted (e.g., increased) only slightly. Another issue opened by the idea of monitoring is that, as the authors are well aware, it becomes much more difficult to study the theoretical properties of the monitored procedures (Cerioli et al. 2014).

In the remaining part of this contribution to the discussion we will focus on the power of monitoring in robust clustering, through examples.

We extend the monitoring to clustering in a completely natural way: we compute minimal Mahalanobis distances as the minimal Mahalanobis distance of trimmed observations *from their closest centroid*, and maximal Mahalanobis distances as the maximal distances of non-trimmed values from their assigned centroid. Other indicators, like the average silhouette width, can be used directly. In the following we will mostly focus on monitoring with respect to the trimming level.

2 Monitoring and reweighting

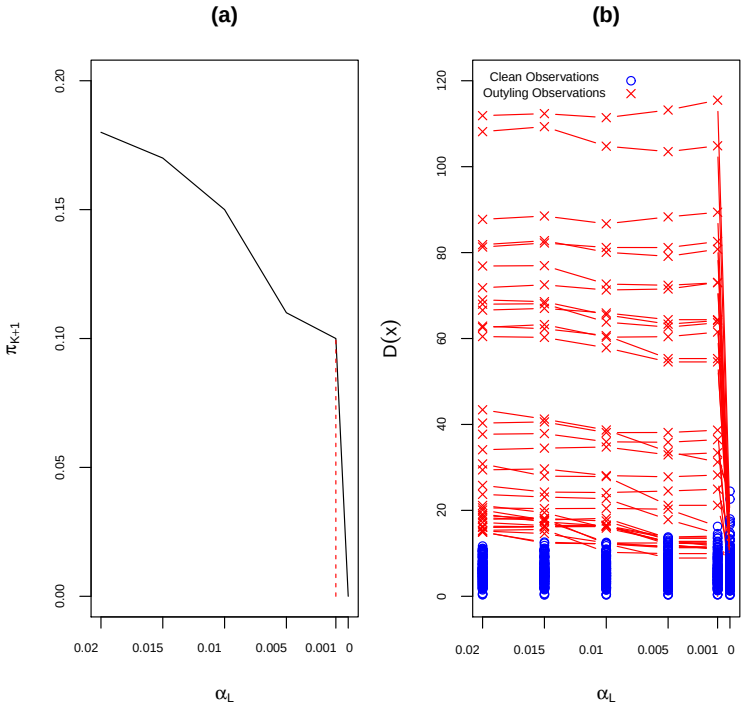
Reweighting is a general principle in robust statistics according to which, after an initial very robust but possibly inefficient estimator has been obtained, discarded or downweighted observations are used to improve efficiency. In most cases reweighting is a two-step procedure, but it can also be iterative (Cerioli 2010; Dotto et al. 2017).

Dotto et al. (2017) introduced `rtclust`, a robust model based clustering method based on reweighting. The procedure is initialized with a robust clustering partition based on a very high trimming level and is then iterated throughout a sequence of decreasing trimming levels. At each step parameters are updated and the clean observations, initially flagged as outlying, are possibly reinserted in the active set. Each observation is moved to the active set if its Mahalanobis distance is lower than a threshold. The latter is obtained by fixing an opportune quantile $(1 - \alpha_L)$ of the target distribution of the Mahalanobis distance. The parameter α_L establishes how far outliers are supposed to be with respect to the bulk of data. Thus such choice is pretty subjective and strongly depends on the context of application. Code is available from the authors upon request.

In this section we apply the philosophy of Cerioli et al. (2018) to explore features of data by means of monitoring the outcomes of `rtclust` with respect to α_L. We use the well known Swiss banknotes dataset. The involved data

describe six features (length, left and right height, lower and upper distance of inner frame to closest border, diagonal length) of 200 Swiss 1000-franc bank notes. It is *a priori* known that there are two balanced groups, made by genuine and counterfeit notes, respectively. Additionally Flury (1988) pointed out that the group of forged bills is not homogeneous, as 15 observations arise from a different pattern and are, for that reason, outliers. Dotto et al. (2017) applied `rtclust` to the Swiss bank notes dataset, fixing $\alpha_L = 0.001$.

Fig. 1: Swiss banknote data: monitoring estimated proportion of outliers π_{K+1} and individual Mahalanobis distances as obtained through `rtclust`.



In Figure 1, panel (a), we plot on the x-axis values for the parameter α_L, while, on the y-axis, the final estimated proportion of outliers (namely π_{K+1}). It is straightforward to see that the estimated contamination level slowly decreases (roughly from 0.17 to 0.10) as α_L varies within the range $[0.02 - 0.005)$. On the contrary, as we move within the range $[0.005 - 0.001)$ an even slower decline is seen. Finally, for $\alpha_L \leq 0.001$ the estimated contamination level suddenly drops towards 0. This clearly suggests that (i) there is contamination, (ii) there might be two or even three separated groups of outliers, with one group of approximately 0.12 (y value with $\alpha_L = 0.005$) minus 0.1 (y value

with $\alpha_L = 0.001$) percent being soft outliers, and the remaining 20 being more clearly placed far from the bulk of the data. Note that, as `rtclust` (and reweighting procedures in general) are not direct outlier detection devices, we shall not make any claim as to outlyingness of the discarded observations; but only claim that the resulting estimates are robust.

In Figure 1, panel (b), we monitored the Mahalanobis distance of each observation to the assigned/closest centroid as a function of α_L. In blue bullets there are observations flagged as clean, while, with red crosses we show observations discarded from the estimation set. This plot is a more detailed, but equally clear, account of what we have seen discussing panel (a) of the same figure. There are two gross outliers, a group of clearly separated outliers, and a third group whose Mahalanobis distances are large but not clearly separated from the bulk of the data. Note that as soon as α_L is decreased too much, the estimates become unstable (which is a clear sign of contamination).

3 Monitoring and high-dimensional clustering: snipping

Farcomeni (2014a,b) developed *snipping*, a technique for parsimoniously trimming sparse entries of a data matrix, rather than entire rows, while clustering. This is done in the spirit of dealing with entry-wise outliers, based on the seminal paper of Alqallaf et al. (2009), and is particularly useful for moderately to high dimensional data. Snipped k-means and `sclust` (which is the model-based version) require the user to specify in advance the number of groups k and a trimming level α. A fraction $np\alpha$ of entries, where n is the sample size and p the number of variables, will be discarded; and each of the n rows (without exceptions, unless all p measurements for one or more subjects are discarded) will be assigned to one of the k groups.

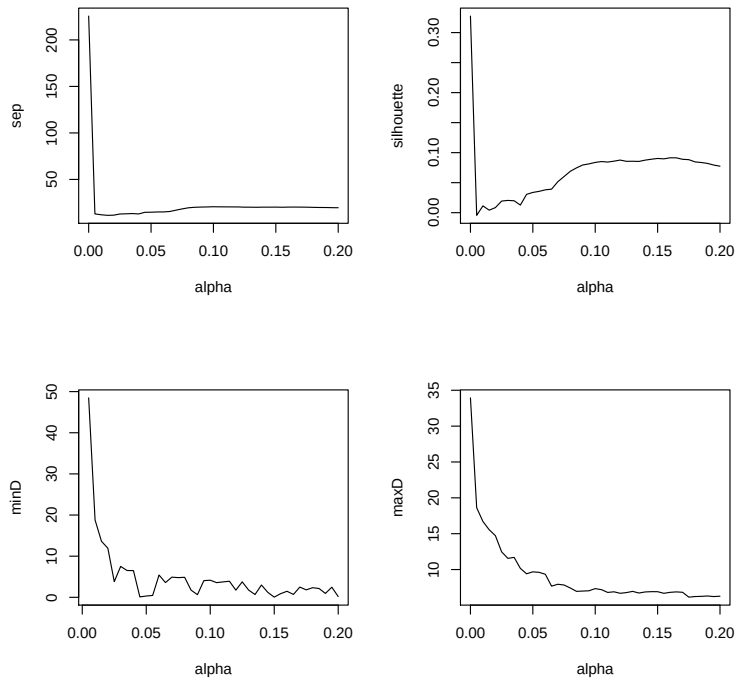
`R` functions to implement the methods, also for the case of robust location and scatter estimation ($k = 1$), are available in the package `snipEM`.

We here revisit an example based on a data set of $n = 380$ daily measurements of $p = 38$ indicators, taken at a urban waste water treatment plant. These include levels of Zinc, pH, chemical and biological demand of oxygen, suspended and volatile solids, and sediments; taken at each of four different locations (input, first settler, second settler, output). As in previous analyses we fix $k = 3$ and monitor the outcome of `sclust` as a function of the removal of $np\alpha$ entries of the data matrix, for several values of α.

In order to work with snipping procedures we might simply define monitored measures at entry level. Hence, k-th Mahalanobis distance of entry X_{ij} will correspond to the univariate distance $|X_{ij} - m_{kj}|/s_{kj}$, where m_{kj} is the j-th entry of the k-th centroid and s_{kj} the j-th standard deviation for the k-th centroid. In Figure 2 we report, as a function of α, the unweighted standard deviation of the estimated centroids (`sep`aration, left upper panel), the average silhouette width (`silhouette`, right upper panel), the minimal Mahalanobis distance of trimmed entries from their closest centroid (`minD`, left lower

panel) and the maximal Mahalanobis distance of untrimmed entries from their assigned centroid (`maxD`, right lower panel).

Fig. 2: Water treatment data: monitoring of separation, average silhouette width, minimal and maximal Mahalanobis distance as obtained through `sclust`.



It can be seen from Figure 2 that there is a very small fraction of gross outliers (say, less than 1%) which form a cluster if not trimmed and make monitored statistics explode for $\alpha < 1\%$. Separation and silhouette after the drop are slightly increasing (up to about $\alpha = 10\%$), indicating that about another 8-9% of the entries can be considered as mild outliers and/or bridge points. Note that the silhouette decreases again for $\alpha > 16\%$, a clearly too large snipping level. Interestingly enough, as reported by Farcomeni (2014b), for $5\% \leq \alpha \leq 15\%$ estimates are fairly stable. Mahalanobis distances in the lower panels are a little more unstable. The maximal Mahalanobis distances are in substantial agreement with our previous discussion, showing a very steep drop for $\alpha < 1\%$, a quick but less steep decrease for $1\% \leq \alpha < 10\%$, and a slower decrease for larger values. The sudden drop for $\alpha \cong 17.5\%$ might indicate that for larger values of α groups are re-arranged, that is, some observations after removal of appropriate entries are moved from one group to another.

4 Monitoring two tuning parameters: the case of robust double clustering

Farcomeni (2009) developed trimmed double k-means. Double clustering aims at simultaneously grouping rows and columns of a data matrix. The robust approach of Farcomeni (2009) is based on separate impartial trimming of rows and columns. This was then extended to the case of snipping in Farcomeni and Greco (2015), which describes snipped double k-means.

`R` functions to implement the methods are freely available as web based supplementary material of Farcomeni and Greco (2015), on the publisher's website.

Robust double clustering gives a clear motivation for simultaneous monitoring of more than one tuning parameter, as both the number of trimmed rows and columns can change. Monitoring shall proceed through 3-D plots, 2-D contour plots (as in our example), or by appropriately tabulating results.

We here monitor the analysis of a microarray data matrix for $n = 200$ genes measured under $p = 20$ conditions. Data are freely available in `R` package `biclust`. We fix $k_1 = 3$ row groups and $k_2 = 2$ column groups, and different row trimming levels $\alpha_1 = 0, 1/200, 2/200, \ldots, 40/200$ and column trimming levels $\alpha_2 = 0, 1/20, 2/20, 3/20, 4/20$. For each combination of α_1 and α_2 we run the trimmed double k-means algorithm, and compute the `sep`aration as above (corresponding to the standard deviation of the estimated centroid matrix), the within sum of square (`withinSS`) which is the sum of the squared distances of each untrimmed entry with respect to its assigned centroid, and as usual the minimal and maximal Euclidean distances (`minD` and `maxD`). In double clustering the minimal and maximal Euclidean distances shall be defined entry-wise, as for instance different columns of the same row can belong to different clusters. Let m_{rc} denote the estimated centroid for an entry in row cluster r and column cluster c. Suppose the i-th row is assigned to cluster r_i and the j-th row to cluster c_j, with $r_i = 0, \ldots, k_1$ and $c_j = 0, \ldots, k_2$; where $r_i = 0$ ($c_j = 0$) indicates a trimmed row (column). Then, the minimal Euclidean distance shall be defined as

$$\min_{ij:\ c_j=0\ |r_i=0} \min_{r=1,\ldots,k_1} \min_{c=1,\ldots,k_2} |X_{ij} - m_{rc}|$$

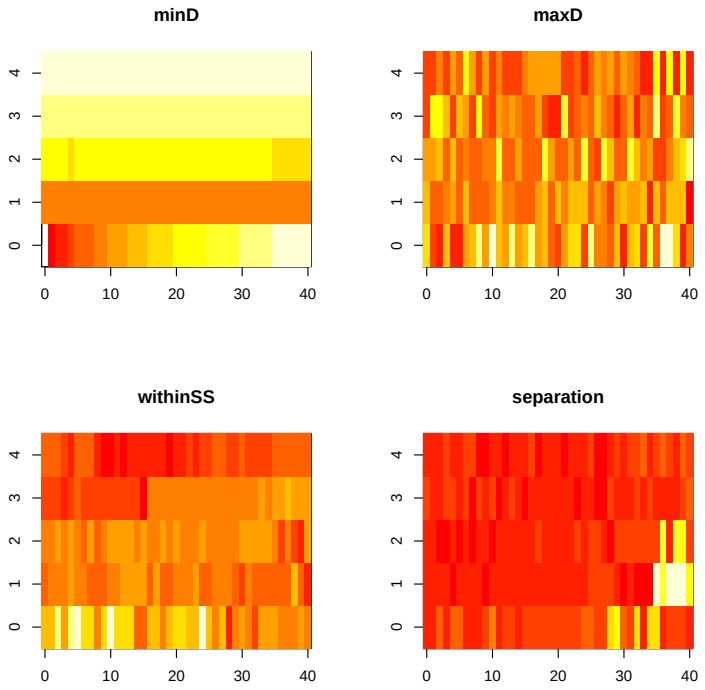
and the maximal Euclidean distance as

$$\max_{ij:\ c_j\neq 0 \& r_i \neq 0} |X_{ij} - m_{r_i c_j}|.$$

In Figure 3 we report, as a function of $n\alpha_1$ and $p\alpha_2$, the separation (right lower panel), the within sum of squares (left lower panel), the minimal Euclidean distance of trimmed entries from their closest centroid (left upper panel) and the maximal Euclidean distance of untrimmed entries from their assigned centroid (right upper panel).

It can be seen from the figure that the minimal Euclidean distance (which is not defined for $\alpha_1 = \alpha_2 = 0$) rapidly increases, with minimal values for

Fig. 3: *E. Coli* data: monitoring of minimal and maximal Euclidean distance (`minD`, `maxD`)), within sum of squares (`withinSS`) and standard deviation of the centroids (`separation`)



$\alpha_2 = 1/20$. About 8 row outliers are clearly identified when $\alpha_2 = 0$, while removal of one or more entire columns makes row trimming less important. This is a clear indication that the 8 gross outliers are probably generated by a single condition (that is, during one of the $p = 20$ experiments). The maximal Mahalanobis distance is much more difficult to interpret, and this is probably an obvious limitation of monitoring which does not necessarily give interpretable information. The within sum of squares and separation are well in agreement with the minimal Mahalanobis distances, even if the information is less apparent.

5 Monitoring three tuning parameters: the case of robust fuzzy regression models.

Dotto et al. (2016) introduce a robust fuzzy regression clustering model based on trimming. Linear clustering models are based on identifying k groups of units, each forming a separate linear structure. Each unit is assigned to the group minimizing the regression error (i.e. its squared residuals from the es-

timated regression line). The methodology proposed in the aforementioned paper aims at maximizing the objective function given by

$$\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^{m} \log\left(f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2)\right) \tag{1}$$

where $f(\cdot; \mu, \sigma^2)$ is the p.d.f of a normal distribution with mean μ and standard deviation σ, and m is a fuzzification parameter. In fuzzy clustering each observation can potentially be a member, with varying degrees of memberships, of each cluster. This is controlled by the parameter m, and measured by $u_{ij} \in [0, 1]$. The latter is a membership value of the observation i to the cluster j. The parameter m values in the range $[1, +\infty)$. Letting $m \to \infty$ implies equal membership values $u_{ij} = 1/k$ regardless of the data; while when $m = 1$ crispy weights $\{0, 1\}$ are always obtained and all (untrimmed) observations are hard-assigned to one and only one cluster. The weights u_{ij} satisfy the following equalities:

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

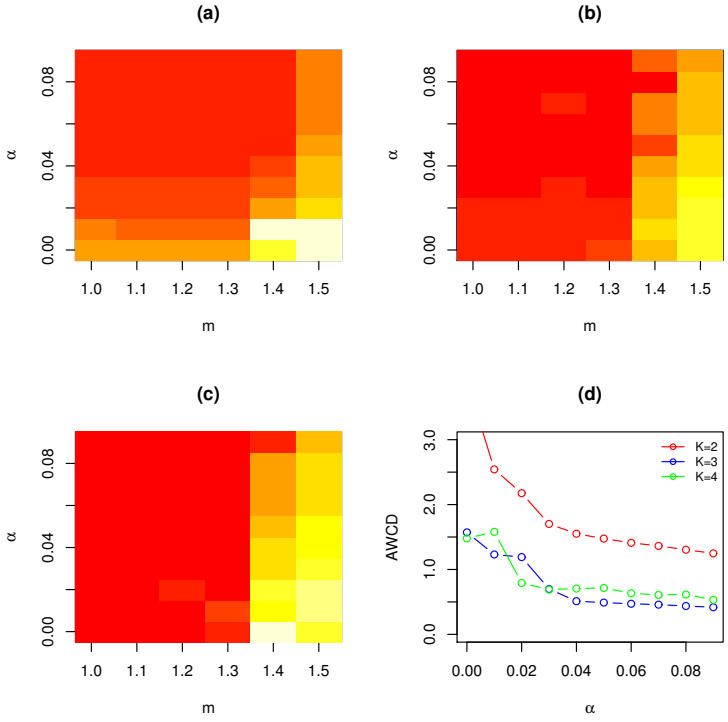
for a subset $\mathcal{I} \subset \{1, 2, ..., n\}$ whose cardinality is $n(1-\alpha)$. The $n\alpha$ observations which are not included in the subset $\mathcal{I}$ are the ones flagged as outlying and, as a consequence, receive $u_{ij} = 0$ membership for each $j = 1, 2, \ldots, k$. Equation (1) contains two tuning parameters: m and α. Additionally, the number of clusters k can be seen as a tuning parameter and in our experience in fuzzy regression clustering it is strongly dependent on m, and more dependent on α than what usually happens in other contexts. These parameters are intertwined and a long discussion is given in Dotto et al. (2016) on their tuning. For these reasons, we will compare simultaneous monitoring of α and m for different values of k.

We illustrate monitoring with the real data analyzed in García-Escudero et al. (2010) and Dotto et al. (2016). The dataset is made of 362 measurements of height and diameter of *Pinus Nigra* trees located in the north of Palencia (Spain), and the aim is to explore the linear relationship between these two quantities.

In our context we need to use a monitoring quantity which is tailored for the special case of fuzzy clustering. We have chosen to use the Average Within-Cluster Distance (AWCD), which is basically a weighted average of the distance of each observation to each cluster (Campello and Hruschka 2006). We here generalize AWCD to fuzzy linear clustering by using squared residuals as distances.

We fixed a grid of values for the parameters m and α and evaluated the results by using a heatmap reporting the value of the AWCD for each combination. Additionally we fixed three different candidate values for k (i.e. $\{2, 3, 4\}$) and plotted a heat map for each candidate value.

Fig. 4: *Pinus Nigra* data: monitoring the AWCD with respect to m, α and k. In panel (a), AWCD for different combinations of m and α when $k = 2$; (b): $k = 3$, (c): $k = 4$. In panel (d), AWCD as a function α for $k = 2, 3, 4$ when $m = 1.3$.



In Figure 4, panels (a), (b), and (c), we show the AWCD when $k = 2, 3, 4$, respectively. It is quite clear that reasonable values for m are within the range $[1 - 1.3]$, as when $m > 1.3$ very large AWCD are generally obtained (especially for small α). When $m \in [1 - 1.3]$, there are clear differences between panel (a) and the other two. In panel (a), AWCD are generally larger than the two other panels for any fixed m and α, indicating that the right number of clusters is $k = 3$ (as there is no advantage in increasing k further). The very high AWCD values obtained for $m > 1.3$ make it hard to distinguish further. We could here restrict to $m \in [1 - 1.3]$ and repeat monitoring, but we prefer simply monitoring with respecto to α for fixed $m = 1.3$ and $k = 2, 3, 4$. This is done in panel (d) of Figure 4. Panel (d) clearly shows how very low values of the AWCD index are reached as the proportion of trimmed observations is ≥ 0.04, and that basically there are no differences in performance when comparing $k = 3$ with $k = 4$. It shall be noted that panel (d) is very similar to classification trimmed likelihood curves (García-Escudero et al. 2011), even if a different quantity than the likelihood at convergence is monitored. As a final consideration point out that there is substantial agreement between the

monitoring used for tuning parameter choice in Dotto et al. (2016) and the (hierarchical) monitoring adopted here, even if this different quantities are monitored.

6 Conclusions

In this paper we applied the philosophy of monitoring, presented in Cerioli et al. (2018), to four recent methodological contributions related to robust cluster analysis. We mostly focused on monitoring the trimming level. Even though different trimming principles were adopted, in all cases the trimming level is directly related to both breakdown point and efficiency.

References

Alqallaf, F., Van Aelst, S., Yohai, V. J., and Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics* **37,** 311–331.

Campello, R. J. and Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems* **157,** 2858–2875.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association* **105,** 147–156.

Cerioli, A., Farcomeni, A., and Riani, M. (2014). Strong consistency and robustness of the forward search estimator of multivariate location and scatter. *Journal of Multivariate Analysis* **126,** 167–183.

Cerioli, A., Riani, M., Atkinson, A. C., and Corbellini, A. (2018). The power of monitoring: How to make the most of a contaminated multivariate sample. *Statistical Methods & Applications* to appear.

Dotto, F., Farcomeni, A., García-Escudero, L. A., and Mayo-Iscar, A. (2016). A fuzzy approach to robust regression clustering. *Advances in Data Analysis and Classification* pages 1–20.

Dotto, F., Farcomeni, A., García-Escudero, L. A., and Mayo-Iscar, A. (2017). A reweighting approach to robust clustering. *Statistics and Computing* pages 1–17.

Farcomeni, A. (2009). Robust double clustering. *Journal of Classification* **26,** 77–101.

Farcomeni, A. (2014a). Robust constrained clustering in presence of entry-wise outliers. *Technometrics* **56,** 102–111.

Farcomeni, A. (2014b). Snipping for robust k-means clustering under component-wise contamination. *Statistics and Computing* **24,** 909–917.

Farcomeni, A. and Greco, L. (2015). *Robust Methods for Data Reduction*. Chapman and Hall/CRC Press, Boca Raton.

Farcomeni, A. and Ventura, L. (2012). An overview of robust methods in medical research. *Statistical Methods in Medical Research* **21,** 111–133.

Flury, B. (1988). *Multivariate statistics: a practical approach*. Chapman & Hall, Ltd.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2011). Exploring the number of groups in model-based clustering. *Statistics and Computing* **21,** 585–599.

García-Escudero, L. A., Gordaliza, A., Mayo-Iscar, A., and San Martín, R. (2010). Robust clusterwise linear regression through trimming. *Computational Statistics & Data Analysis* **54,** 3057–3069.