Stat Methods Appl (2018) 27:605–608
https://doi.org/10.1007/s10260-017-0415-x

CrossMark

ORIGINAL PAPER

Comments on "The power of monitoring: how to make the most of a contaminated multivariate sample"

L.A. García-Escudero[1] · A. Gordaliza[1] · C. Matrán[1] · A. Mayo-Iscar[1]

Accepted: 26 November 2017 / Published online: 30 November 2017

Abstract These are comments on the invited paper "The power of monitoring: How to make the most of a contaminated multivariate sample" by Andrea Cerioli, Marco Riani, Anthony Atkinson and Aldo Corbellini.

Keywords Clustering · Robustness · Monitoring

1 Introduction

First of all, we would like to thank and congratulate A. Cerioli, M. Riani, A. Atkinson and A. Corbellini for their interesting contribution. Authors know that we strongly agree on their conviction that, also in Data Analysis, viewing a "full movie" is often better than viewing a "single frame". In Cerioli et al. (2018), authors present convincing examples where the dynamical view of data can improve classical (even robust) methods in revealing the underlaying data structure. The development and dissemination of useful monitoring tools have been a continuous motivation in the authors research and we sincerely thank them for providing those valuable tools for the practitioner.

We are going to briefly review some monitoring tools, proposed in the Robust Cluster Analysis framework, which show our agreement with the authors point of view.

Through a sequence of works, our research group has developed robust clustering techniques (see Cuesta-Albertos et al. 1997; García-Escudero et al. 2008; Cuesta-Albertos and Matrán 2008) with the aim of addressing the well-known lack of

✉ L.A. García-Escudero
lagarcia@eio.uva.es

[1] IMUVA and Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, Universidad de Valladolid, 47011 Valladolid, Spain

robustness of traditional clustering approaches. The consideration of trimming (self-determined by the data) is the key ingredient in these approaches. We allow to discard a fixed fraction α of the "most outlying" observations. Choosing the correct trimming level α may be seen as a kind of tradeoff between robustness and efficiency, as also discussed in Cerioli et al. (2018). In addition, choosing the number of clusters k is one of the more relevant, complex and widely addressed problems in Cluster Analysis.

Nowadays, it is widely recognized that Cluster Analysis cannot be viewed as a fully automated procedure. Therefore, we do not think that any monitoring process would be able to return a unique undoubted optimal (α, k) pair of parameters. However, we believe that monitoring the effects of moving the (α, k) on the clustering results/performance is a sensible way to obtain a reduced list of sensible (α, k) couples.

With the previous ideas in mind, in García-Escudero et al. (2003), we proposed the careful monitoring of the so-called trimmed k-variation curves. The trimmed k-variation curves are obtained by plotting $(\alpha, k) \mapsto V_k(\alpha)$, where $V_k(\alpha)$ is the smallest value taken by the α-trimmed k-means (see Cuesta-Albertos et al. (1997) for details) target function given (α, k). We see that $V_k(\alpha)$, viewed as a function of α, decreases smoothly if k is a sensible choice for the number of clusters. On the other hand, changes in the rate of decrease can be noticed when k is not adequate or just when we start trimming outliers. Changes in the rates of decrease are better visualized by using their numerical second derivatives.

The use of the trimmed k-variation curves ideally assumes, as (trimmed) k-means implicitly do, that clusters to be detected are spherical with similar scatters. However, we may be interested in more heterogeneous clustering procedures, so allowing non-spherical and/or heteroscedastic clusters. In that case, we could use trimmed k-means with a high α just to detect few observations in the most central "core" regions of clusters and start adding the closest ones in a controlled way (see García-Escudero and Gordaliza 2007). Proceeding in this way, the variance-covariance matrices in each cluster can be estimated under normality assumptions for the cluster components. This process needs monitoring in order to avoid the inclusion of outlying observations. The correct monitoring of the cluster scales (i.e., the determinants of the variance–covariance matrices) plays a key role in this approach. An iterative procedure is also introduced in Cuesta-Albertos and Matrán (2008), which starts from central "core" regions, by applying maximum likelihood mixture principles.

Another proposal for handling heterogeneous clusters was introduced in García-Escudero et al. (2008) throughout the TCLUST methodology. TCLUST combines trimming and a maximal ratio constraint for the cluster variance–covariance matrices' eigenvalues. It was proposed to restrict this maximal ratio to be smaller than a fixed constant $c \geq 1$. The constant c serves to control the allowed differences in clusters' scatters, within and across clusters, and to avoid the detection of non-interesting "spurious" solutions. Although (initially) TCLUST was not a mixture modeling approach, some π_g weights were included in the associated classification likelihood maximization. In that statement, certain π_g weights can be set close to 0 if k is larger than the "true" number of clusters in our data set. Building on this, the "ctlcurves" in García-Escudero et al. (2011) are based on monitoring the TCLUST's target function when moving (α, k), for a fixed maximal eigenvalue ratio c. The approach also takes into account how the choice of α and k is clearly dependent on c. For instance, a set of very scattered outliers can be considered as an additional cluster (so increasing k and decreasing α) for high c values.

The results of applying TCLUST can be also improved throughout the iterative reweighting approach recently introduced in Dotto et al. (2017), when k is known. The reweighting process allows to recover back some observations that had been wrongly trimmed (for instance, after applying a high preventive trimming level). The procedure is closely related to García-Escudero and Gordaliza (2007) but α and c are automatically determined by the dataset itself. Dotto et al. (2017) shows that the resulting final choices for α and c are not very dependent on the initialization whenever that initialization contains a small proportion of observations from each cluster and outliers are not included in it.

A modified BIC criterium has been also introduced in Cerioli et al. (2017) which can be applied in both classification and mixture likelihood problems. The main idea follows from noticing that higher c values result in more unconstrained and "complex" models. In Cerioli et al. (2017), this extra model complexity is taken into account within the penalty term added to the log-likelihood. An appropriate monitoring of the modified BIC and the associated cluster partitions produces a reduced and ranked list of sensible partitions. Then, the researcher has to find the one that better fits his/her clustering purposes within that list. Although the methodology is presented in the $\alpha = 0$ case, we believe that it can be surely modified to cover more general robust clustering problems. Trimmed BIC modifications (and their monitoring) have been already considered in Neykov et al. (2007) and Gallegos and Ritter (2010).

Summarizing, we certainly agree on the authors's claim about the "power of monitoring" in contaminated multivariate samples, as authors have nicely shown. In our comment, we have also tried to illustrate how (robust) Cluster Analysis can also benefit from those monitoring ideas.

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 and FEDER, Grant VA005P17.

References

Cerioli A, García-Escudero LA, Mayo-Iscar A (2017) Riani M (2017) Finding the number of normal groups in model-based clustering via constrained likelihoods. J Comput Graph Stat. https://doi.org/10.1080/10618600.2017.1390469

Cerioli A, Riani M, Atkinson AC, Corbellini A (2018) The power of monitoring: how to make the most of a contaminated multivariate sample. Stat Methods Appl. https://doi.org/10.1007/s10260-017-0409-8

Cuesta-Albertos JA, Gordaliza A, Matrán C (1997) Trimmed k-means: an attempt to robustify quantizers. Ann Stat 25:553–576

Cuesta-Albertos JA, Matrán C (2008) Robust estimation in the normal mixture model based on robust clustering. J R Stat Soc Ser B Stat Methodol 70:779–802

Dotto F, Farcomeni A, García-Escudero LA, Mayo-Iscar A (2017) A reweighting approach to robust clustering. Stat Comput. https://doi.org/10.1007/s11222-017-9742-x

Gallegos MT, Ritter G (2010) Using combinatorial optimization in model-based clustering under spurious outliers and cardinality constraints. Comput Stat Data Anal 54:637–654

García-Escudero LA, Gordaliza A (2007) The importance of the scales in heterogeneous robust clustering. Comput Stat Data Anal 51:4403–4412

García-Escudero LA, Gordaliza A, Matrán C (2003) Trimming tools in exploratory data analysis. J Comput Graph Stat 12:434–449

García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2008) A general trimming approach to robust cluster analysis. Ann Stat 36:1324–1345

García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2011) Exploring the number of groups in robust model-based clustering. Stat Comput 21:585–599

Neykov N, Filzmoser P, Dimova R, Neytchev P (2007) Robust fitting of mixtures using the trimmed likelihood estimator. Comput Stat Data Anal 52:299–308