Stat Methods Appl (2018) 27:621–623
https://doi.org/10.1007/s10260-017-0418-7

COMMENT

Discussion of "The power of monitoring: how to make the most of a contaminated multivariate sample" by Andrea Cerioli, Marco Riani, Anthony C. Atkinson and Aldo Corbellini

Christophe Croux[1]

Accepted: 6 December 2017 / Published online: 8 January 2018

Abstract In this short note comments are given on the discussion paper of Cerioli, A., Riani, M., Atkinson, A. C., and Corbellini, A., entitled "The power of monitoring: How to make the most of a contaminated multivariate sample."

1 Introduction

I would like to congratulate the authors of Cerioli et al. (2018) on their paper. The power of monitoring for finding structure in multivariate data is clearly demonstrated with great visual displays. All results are fully reproducible from code on Marco Riani's website, where also the *cows* data set is shared with use. This is a real data set containing multiple outliers I've not seen being used before. The authors are well known from their work on the *forward search*, where an outlier diagnostic (or, more general, any type of diagnostic statistic) is monitored as a function of m, the sample size of the most 'central' m observations. An early reference is Atkinson and Riani (1997), and we recommend the book (Atkinson et al. 2013) for the use of the forward search in multivariate data analysis.

2 Monitoring plots

This paper deals with a different type of monitoring: here a plot is made of an outlier diagnostic versus a tuning parameter of a given robust estimator. The paper takes the setting of a normal multivariate location-scale model. The outlier diagnostic is the

✉ Christophe Croux
christophe.croux@edhec.edu

[1] EDHEC Business School, 24 Avenue Gustave Delory, CS 50411, 59057 Roubaix Cedex 1, France

Mahalanobis distance; since it can be computed for each observation, the monitoring plot contains n lines, with n the sample size. The tuning parameter can, for instance, be the trimming portion of the MCD estimator, the breakdown point of the S-estimator, or the efficiency of the MM-estimator. In the practice of robust statistics, such tuning constants are often fixed at some default choices: 25 or 50% trimming for MCD, 50% breakdown point for the S, and 95% efficiency for the MM-estimator. The monitoring plots allow to see how sensitive the results are with respect to these choices and, as discussed in detail in the paper, to find structure in the multivariate data. As such, one of the findings of the authors is that an 80% efficiency of an MM-estimator yields more stable results than the default choice, and I fully concur.

When computing the MCD estimator in statistical software, the standard implementation yields the reweighted version. The tuning parameter can then be taken as the trimming portion of the initial raw MCD estimator, or as the reweighting probability, as suggested by the authors. A third option is to take the efficiency of the reweighted MCD estimator, analogously as is done for monitoring the MM-estimator. This efficiency depends on the trimming portion δ in a highly nonlinear way (Croux and Haesbroeck 1999). For $p = 2$ and the maximal breakdown point we get an efficiency of only 25.3% for $\delta = 0.05$ (left plot Fig. 7), while the efficiency increases to 95.3% for $\delta = 0.001$ (right plot Fig. 7). Furthermore, while for the monitoring plot in Figure 4 the efficiency of the MM-estimator ranges from 50% to 1, the efficiency of the reweighted MCD estimator (left plot Figure 7) ranges from 25.3% to 1 and at the midpoint of 25% breakdown point the efficiency of the reweighted MCD is still only at 48.0%.

Both the MM and the S estimator require the choice of a loss function ρ, with Tukey's biweight as a default. I fully agree with the authors that monitoring the results for varying choices of the loss function is worthwhile, in particular for the multivariate location/scale model where other proposals for robust loss functions have been made (e.g. Rocke 1996). Note that Croux et al. (2011) showed that it is possible to choose ρ such that one combines 50% breakdown with an efficiency arbitrarily close to 1. Such a construction is not possible in the regression case.

3 Conclusion

The proposed monitoring tools rely on visual inspection of plots, and require intense data analysis for every separate data set by the statistician. Such an approach may be worthwhile in many different situations, but a fully automatic procedure including adaptive choice of the tuning constant could be envisaged as well. For the forward search such automatic procedures have been developed over the years, but my feeling is that this task is even more challenging here. For instance, correct statistical inference after adaptive selection of the tuning constant is an issue.

The selection of the tuning constants of robust multivariate methods is an important problem. The monitoring tools proposed in this paper partly circumvent this problem, and additionally allow to gain more insight in the structure and number of outliers. Let me thank the authors for sharing their ideas and methods with us. I enjoyed reading the paper.

References

Atkinson A, Riani M (1997) Bivariate boxplots, multiple outliers, multivariate transformations and discriminant analysis: the 1997 Hunter Lecture. Environmetrics 8(6):583–602

Atkinson AC, Riani M, Cerioli A (2013) Exploring multivariate data with the forward search. Springer, Berlin

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 13:179

Croux C, Dehon C, Yadine A (2011) On the optimality of multivariate S-estimators. Scand J Stat 38(2):332–341

Croux C, Haesbroeck G (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. J Multivar Anal 71(2):161–190

Rocke DM (1996) Robustness properties of s-estimators of multivariate location and shape in high dimension. Annals Stat 24:1327–1345