

Rejoinder to the discussion of “The power of monitoring: how to make the most of a contaminated multivariate sample”

Andrea Cerioli¹  · Marco Riani¹ ·
Anthony C. Atkinson² · Aldo Corbellini¹

Accepted: 5 August 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

We thank the Editor, Tommaso Proietti, for the invitation to write a discussion paper and for encouraging such a wide ranging discussion. We also thank the Associate Editor in charge of the discussion, Alessio Farcomeni, for his very careful work.

We feel humbled by the quantity of insightful comments stimulated by our paper and that so many prominent researchers in the field of robust statistics were kind enough to contribute to the discussion. We are also highly surprised (and very glad) to see that the length of the discussion is twice the length of the original paper!

We thank all the discussants for their supportive comments and for their appreciation of our work. Therefore, we take the discussion as a good sign that the “philosophy” of monitoring will have more fans in the future. We believe that the discussions include contributions that are worth considering per se: improvements of existing methodologies; extensions to multi-parameter monitoring; a new ρ -function.

Writing this rejoinder is pleasant and difficult at the same time. It is pleasant because there has been general appreciation (and even enthusiasm) for the power of monitoring. It is difficult because there is very little about which to argue with the discussants: we

✉ Andrea Cerioli
andrea.cerioli@unipr.it

Marco Riani
mriani@unipr.it

Anthony C. Atkinson
a.c.atkinson@lse.ac.uk

Aldo Corbellini
aldo.corbellini@unipr.it

¹ Department of Economics and Management, University of Parma, Via Kennedy 6, 43125 Parma, Italy

² Department of Statistics, The London School of Economics, London WC2A 2AE, UK

agree with (and welcome) almost everything that has been suggested. Since there is so much support for our approach, we have kept our comments short; the discussions stand well as individual contributions. It is, at the time of writing, 65 years since Box (1953) brought the concept of robustness into the statistical literature. Although our modern approaches might not fall into his original robustness framework, it is exciting that there are so many new problems to be tackled and such enthusiasm for solving them.

We thank Agostinelli and Greco for the statement “Their monitoring approach may really give a new impulse to the use of robust methods in data analysis”. We do hope that they are good forecasters! We cannot help but recall the “false dawn” of robust statistics around the time of the Princeton Robustness Study (Andrews et al. 1972) that we discussed in the introduction to Cerioli et al. (2016). However, the results reported in this and other contributions to the discussion make us optimistic since monitoring releases data analysts from many tricky, and possibly arbitrary, decisions when contemplating a robust analysis.

We appreciate the extension of the idea of monitoring to the interesting domain of multivariate weighted likelihood estimation (Agostinelli and Greco 2018) for two main reasons. First, this extension reinforces the idea that monitoring is a general principle that can be applied to the specific methodology of interest as well as to the technique which is most suitable in a given application area (a general theme of all the discussions). Second, through monitoring, Agostinelli and Greco are able both to provide deeper insights into the properties of the weighted likelihood methodology and to improve their final estimates. The power of monitoring is demonstrated by the closeness of their results of our four examples to those we found using the FS and MCD.

We are grateful for Croux’s confirmation of our numerical results, as well as for his support of the general idea of monitoring. His deeper insights into the intricate relationship between efficiency and breakdown point should help to improve the interpretation of our monitoring plots.

Here are two specific answers:

- We welcome the suggestion of using our monitoring tools to assess alternative loss functions for multivariate location and scale estimation, as we did for regression (Riani et al. 2014), and to include further ρ -functions in future studies (see also the contributions of Maronna and Yohai and Raymaekers, Rousseeuw and Vranckx).
- We definitely agree with Croux that a very challenging task for future research is the derivation of theoretical results that can lead to correct statistical inference after adaptive selection of the tuning parameter, thus opening the door to the development of automatic monitoring procedures based on general robust estimators. We hope that our paper will stimulate talented researchers, such as those who have contributed to our discussion, to work on this topic.

We thank Farcomeni and Dotto for the statement that “monitoring might ... ultimately become the standard for applied sciences”. As with Agostinelli and Greco, we hope that they are good forecasters and, at the same time, hope to avoid hubris.

We enjoyed the plethora of ideas for the extension of monitoring to the important problem of robust clustering, a topic also considered by García-Escudero, Gordaliza,

Matrán and Mayo-Iscar. Farcomeni and Dotto apply the idea of monitoring to several problems:

- the choice of the reweighting probability in the reweighted version of TCLUS_T proposed by Dotto et al. (2018);
- the “snipping” level in robust clustering procedures that must be used in the case of cell-wise contamination, an important topic with modern high-dimensional data mentioned in several other discussions. See also Rousseeuw and Van Den Bossche (2018);
- robust clustering of both units and variables;
- robust fuzzy cluster-wise regression.

We are fascinated by their use of multi-parameter monitoring, which gives computational and graphical challenges that should become central in future research, a topic also raised by Raymaekers, Rousseeuw and Vranckx and, for robust clustering, by Riani et al. (2018).

García-Escudero, Gordaliza, Matrán and Mayo-Iscar strongly support the monitoring idea, that is that viewing a “full movie” is often better than viewing a “single frame”.

We appreciate their careful exposition of the extension of the idea of monitoring to the important problem of robust clustering, a topic also considered by Farcomeni and Dotto. García-Escudero, Gordaliza, Matrán and Mayo-Iscar nicely illustrate the benefits of monitoring mainly through the idea of reweighting the results of a robust cluster analysis (Dotto et al. 2018), for which the final trimming level is automatically determined from the data. For more direct extensions to cluster analysis of the monitoring approach of our paper we refer to Cerioli et al. (2018) and Riani et al. (2018).

We thank Heritier and Victoria-Feser for their really impressive review of robust methods and complex problems in which monitoring could potentially be useful. Moving from standard, one-population, multivariate models to more complex, but more realistic, conceptual frameworks is another crucial task in making robust methods appealing for scientists in other domains, as pointed out by several discussants. Our paper was not intended to be a review of the emerging streams of robust statistics. However, given this review, we would like to add to Heritier and Victoria-Feser’s §3 the reference Riani and Atkinson (2010) in which the Forward Search is combined with Mallows’ C_p .

This discussion nicely fills the gap created by our missing literature review. We hope that it may encourage new researchers to be involved and face the many challenges involved: there will assuredly be plenty of work for a new generation of statisticians!

We definitely agree with Maronna and Yohai that there should not necessarily be an opposition between high breakdown point and high efficiency. This is a major theme of our work and of the contributions to the discussion. What we have tried to do is a systematic exploration of this idea through the monitoring of breakdown point and we certainly welcome the existence of other (past or future) contributions in the same direction.

We also agree that our monitoring plots may become less interesting when the outliers are not distant (and indeed, there are examples of this in the paper). The

problem of bias for nearby contamination is common to many robust estimators, as pointed out also by Raymaekers, Rousseeuw and Vranckx, as may be the failure of theoretical results relating breakdown point and efficiency. Therefore we welcome the suggestion to extend our monitoring tools to other procedures, such as τ -estimators, and other loss functions that may be of interest in modern applications of robust statistical methods, such as high-dimensional data, skew distributions, etc. (see also the discussions by Croux and by Raymaekers, Rousseeuw and Vranckx). In this respect, we find it promising to see that in all the examples reported in the paper there is at least one method for which monitoring provides more illuminating results than a “static” analysis with fixed breakdown point.

It is difficult to provide a reply to Perrotta and Torti, given our many years of collaborative research. However, we warmly thank Perrotta and Torti, and the whole research team at the European Commission Joint Research Centre, for having exposed us to a wide range of applied problems that take on crucial importance in the working principles of our complex society. This exposure has greatly contributed to sharpening our view of applied robust statistics in general, and also the specific ideas presented in our paper.

Perrotta and Torti describe interesting applications of monitoring in regression, which include thinning out data points near the origin. In their thinned dataset in Figure 3 the outliers form 52.54% of the data. Despite this, the lower right-hand panel of Figure 5 shows a dramatic change in structure at a nominal bdp of 0.43. The monitoring plots of residuals for thinned trade data in Figures 7 and 8 show that indeed the thinned data have no interesting structure. We are intrigued by the smooth structure of the transition in the books data set. If this is indeed caused by data from several populations, we may perhaps have another weapon in our armoury of monitoring procedures.

We agree with Raymaekers, Rousseeuw and Vranckx, as with other discussants (e.g. Maronna and Yohai), that monitoring plots inherit the properties of the robust estimator under consideration. Our choice of estimators and ρ -functions was mainly based on popularity and availability of methods, and we greatly welcome additions to the existing battery of tools (as provided by many discussants). We enjoyed the *jeu d'esprit* leading to the new ρ -function proposed by Raymaekers, Rousseeuw and Vranckx, that makes the S-estimator behave similarly to the MCD in the case of nearby contamination. In this respect, monitoring can be used to assess the properties of this ρ -function and of different ones as is suggested by other discussants and as we did in Riani et al. (2014). There the conclusion was that the ρ -function was of secondary importance; it is salutary to be reminded otherwise.

We agree with Raymaekers, Rousseeuw and Vranckx that routine use of monitoring tools poses compelling challenges for future research:

- intensive computation, which may be solved either by software improvements (see also the discussion by Todorov) or by the adoption of computationally “cheaper” estimators (as suggested by Raymaekers, Rousseeuw and Vranckx)
- multi-parameter tuning in the case of more complex problems, especially in high dimensions: some preliminary attempts in this direction are provided by Farcomeni and Dotto, and by Riani et al. (2018).

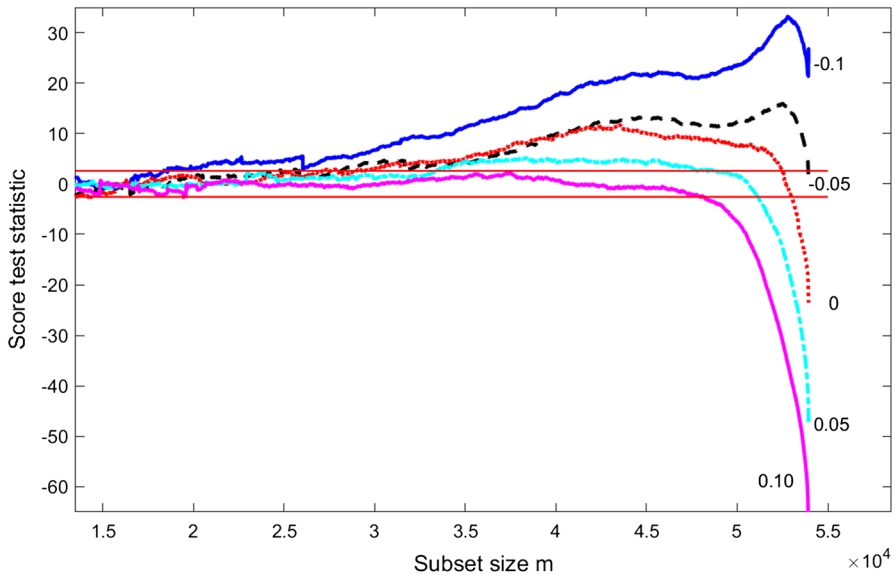


Fig. 1 Diamonds data. Fan plot; forward plot of the score statistic for the Box-Cox family of transformations ($\lambda = 0$ corresponds to the logarithmic transformation)

Sheather and McKean stress the importance of comparing ordinary (e.g. least-squares) and robust fits. We agree with this idea, which may be considered a “prototype” of monitoring: indeed our monitoring plots are intended to fill the gap between the two extremal pictures (given by the highly-robust and the classical diagnostics, respectively).

In their analysis of data on 53,900 diamonds Sheather and McKean regressed $\log(\text{price})$ on several linear predictors. We started our monitoring analysis with a predictor including $\log(\text{carat})$, depth, x and y , focusing on the correct transformation of price. We used the monitoring version of the approximate score statistic of Atkinson (1973) for the Box-Cox transformation (Box and Cox 1964) that was introduced by Riani and Atkinson (2000). The plot of statistic value against subset size in the Forward Search is called a “fan plot”.

The score statistic has, approximately, a standard normal distribution. Figure 1 shows the forward plots of the statistics for five values of the transformation parameter λ : -0.1 , -0.05 , 0 (the log transformation) 0.05 and 0.1 . The vertical scale of the figure goes from -65 to 35 . The value of the statistic for $\lambda = -0.05$ is inside the 99% limits for the normal distribution at the end of the search, but lies outside from around $m = 30,000$ to almost the end. The value for the log transformation is outside both at the end and for much of the search. Only for $\lambda = 0.1$ is the value inside for most of the search, until $m = 48,000$. It is clear from this analysis that either the Box-Cox transformation is not appropriate or that there is a more complicated structure in the data. The break that is clear at around $m = 28,000$ in Sheather and McKean’s Figure 2 shows slightly in our plot at values around this range, with some variation as a function of λ .

This analysis shows the information that can be gained by monitoring a different aspect of robust data analysis from those considered elsewhere in our paper and its discussion. Another interesting area that we feel will benefit from monitoring is the robust analysis of spatial data (Cerioli and Riani 2003; Filzmoser et al. 2014).

Todorov addresses some tricky computational issues, from the point of view of a potential user of monitoring tools on complex real-world problems (see also the discussion by Raymaekers, Rousseeuw and Vranckx): a recent extension to programs available in *fsdaR* is to robust clustering in the spirit of García-Escudero, Gordaliza, Matrán and Mayo-Iscar.

We appreciate the extension of monitoring to the analysis of compositional data, a topic which is far from our current research interests, but which is central to many applications. Again this reinforces the idea that monitoring is a general principle that could be applied to very diverse methodologies according to specific research and application interests.

Acknowledgements The work has been partially supported by the European Commission's Hercule III programme 2014–2020 through the Automated Monitoring Tool project. This research benefits from the HPC (High Performance Computing) facility of the University of Parma, Italy. M.R. gratefully acknowledges support from the CRoNoS project, reference CRoNoS COST Action IC1408.

References

- Agostinelli C, Greco L (2018) Weighted likelihood estimation of multivariate location and scatter. *Test*. <https://doi.org/10.1007/s11749-018-0596-0>
- Andrews DF, Bickel PJ, Hampel FR, Tukey WJ, Huber PJ (1972) Robust estimates of location: survey and advances. Princeton University Press, Princeton
- Atkinson AC (1973) Testing transformations to normality. *J R Stat Soc Ser B* 35:473–479
- Box GEP (1953) Non-normality and tests on variances. *Biometrika* 40:318–335
- Box GEP, Cox DR (1964) An analysis of transformations (with discussion). *J R Stat Soc Ser B* 26:211–246
- Cerioli A, Riani M (2003) Robust methods for the analysis of spatially autocorrelated data. *Stat Methods Appl* 11:335–358
- Cerioli A, Atkinson AC, Riani M (2016) How to marry robustness and applied statistics. In: Di Battista T, Moreno E, Racugno W (eds) *Topics on methodological and applied statistical inference*. Springer, Heidelberg, pp 51–64
- Cerioli A, Farcomeni A, Riani M (2018) Wild adaptive trimming for robust estimation and cluster analysis. *Scand J Stat*. <https://doi.org/10.1111/sjos.12349>
- Dotto F, Farcomeni A, García-Escudero LA, Mayo-Iscar A (2018) A reweighting approach to robust clustering. *Stat Comput* 28:477–493
- Filzmoser P, Ruiz-Gazen A, Thomas-Agnan C (2014) Identification of local multivariate outliers. *Stat Pap* 55:29–47
- Riani M, Atkinson AC (2000) Robust diagnostic data analysis: transformations in regression (with discussion). *Technometrics* 42:384–398
- Riani M, Atkinson AC (2010) Robust model selection with flexible trimming. *Comput Stat Data Anal* 54:3300–3312
- Riani M, Cerioli A, Atkinson AC, Perrotta D (2014) Monitoring robust regression. *Electron J Stat* 8:646–677
- Riani M, Atkinson AC, Cerioli A, Corbellini A (2018) Robust methods via monitoring for clustering and multivariate data analysis. Submitted
- Rousseeuw PJ, Van Den Bossche W (2018) Detecting deviating data cells. *Technometrics* 60:135–145