Stat Methods Appl (2018) 27:641–649
https://doi.org/10.1007/s10260-017-0420-0

COMMENT

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

Domenico Perrotta[1] · **Francesca Torti**[1]

Accepted: 24 December 2017 / Published online: 12 January 2018

Abstract We contribute to the discussion of an article where Andrea Cerioli, Marco Riani, Anthony Atkinson and Aldo Corbellini review the advantages of analyzing multivariate data by monitoring how the estimated model parameters change as the estimation parameters vary. The focus is on robust methods and their sensitivity to the nominal efficiency and breakdown point. In congratulating with the authors for the clear and stimulating exposition, we contribute to its discussion with an overview of what we experienced in applying the monitoring in our application domain.

Keywords Forward Search · MM-estimation · S-estimation · Density estimation · Thinning · International trade data

1 Prologue

Andrea Cerioli, Marco Riani, Anthony Atkinson and Aldo Corbellini (hereafter CRAC) are passionate supporters of the data analysis approach proposed for this discussion (Cerioli et al. 2018), which consists in monitoring the model parameters estimated for a reasonable range of values of the key parameters of the estimation method, and selecting those producing the best results. Robust estimation algorithms depend on several tuning constants, producing effects that should be monitored. CRAC focus on the key parameters used to specify the maximum possible breakdown or efficiency to be achieved. We would like to complement their exposition with recent

✉ Francesca Torti
francesca.torti@ec.europa.eu

Domenico Perrotta
domenico.perrotta@ec.europa.eu

[1] European Commission, Joint Research Centre, Ispra, Italy

applications of the monitoring approach to datasets relevant for international trade analysis and anti-fraud, which bring new statistical challenges not yet fully addressed.

2 Monitoring trade data

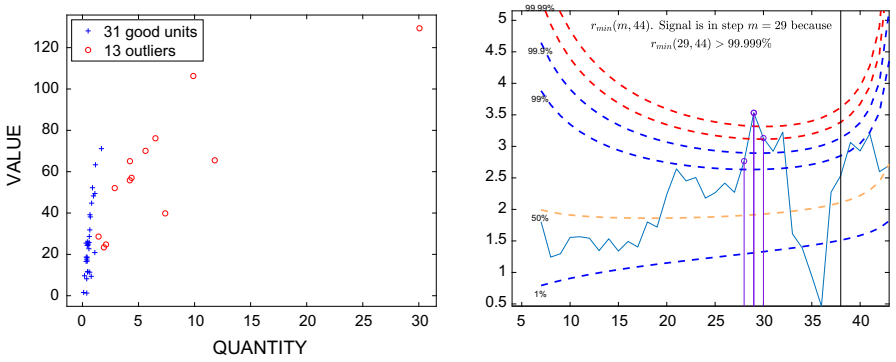
CRAC have introduced us to a particular monitoring instance, the Forward Search (FS, Atkinson and Riani 2000), more than ten years ago. We studied together the application of monitoring to other established robust regression estimators (Riani et al. 2014). Currently, we use different forms of monitoring in the routine analysis of large amounts of regression datasets relevant to European Union policies, such as international trade and anti-fraud. We have many more reasons for supporting enthusiastically the approach than drawbacks to signal.

We compute on a monthly basis robust estimates of "fair prices" for goods imported in the European Union from third countries. The estimates are used by customs and anti-fraud services to combat illegal practices. The financial impact for the budget of the EU is very big and the fair prices must be somehow "certified", in view of their use in Court cases. We are therefore studying appropriate statistics or indicators to summarize the sensitivity of the robust fair price estimate to the choice of the estimation method and the related parameters and tuning constants. To this end the monitoring is a precious instrument, although we are facing with two main disadvantages: one is the substantial computation time (which increases with the sample size and number of parameters monitored) and the other is the lack of clear instruments to summarize automatically in a unique statistic or indicator the rich collection of monitored results.

We illustrate the need of monitoring the stability of the fair price estimates, even for small datasets, in Fig. 1, which has to do with imports of "sports footwear with insoles of a length of less than 24 cm". We see the results obtained with the SAS `PROC ROBUSTREG` and the MATLAB FSDA Toolbox (Riani et al. 2012, 2015) with four methods: FS, LTS, S and MM. We could verify that the reason for the discordance between the two sets of results originates from the different default parameter values adopted by the two environments. The figure also reports in the caption the FS estimate, which is in line with the results obtained by the other estimators with the FSDA defaults. The typical forward plot associated to the FS (in the top-right panel) shows a sharp decrease of the statistic monitored (the minimum deletion residual), which is sign of structure in the data that the other methods have not addressed properly because of the unlikely choice of the key parameters that determine their robustness and efficiency.

3 The effect of concentrated non-contaminated observations

We introduce in the discussion another complication that occurs rather often in trade data, consisting in large proportions of non-contaminated observations falling in a small data region. To our knowledge, this problem was addressed in robust statistics only recently, with Heikkonen et al. (2013) and Cerioli and Perrotta (2014) showing that the effect of a high-density region can be so strong to override the benefits of robust devices such as trimming methods for robust clustering. We show that the monitoring plots do not make exception and become completely uninformative in



	use of SAS defaults		use of FSDA defaults	
	$\hat{\beta}$		$\hat{\beta}$	
LTS	12.73	$h = (3n + p)/4)$	51.97	$h = 0.5 \cdot (n + p + 1)$
S	13.25	$bdp = 0.25$	47.99	$bdp = 0.5$
MM	13.02	$eff = 0.85$, $s_{0,LTS}$	40.96	$eff = 0.95$, $s_{0,S}$

Fig. 1 Top panels: a trade dataset analysed with the Forward Search, giving rise to an estimated import price of $\hat{\beta} = 42.62$ euro per Kg. Bottom panel: estimates obtained with LTS, S and MM with different values for the breakdown point (bdp), efficiency (eff) and the fraction of observations over which the objective function is minimized (h). In the table, $s_{0,LTS/S}$ indicates that the initial scale value is computed respectively with LTS or S, with their default parameters

presence of highly concentrated data. The proposal of Cerioli and Perrotta (2014) in these cases, is to sample a much smaller subset of observations which preserves the cluster structure and also retains the main outliers of the original data set. This goal is achieved by defining the retention probability of each point as an inverse function of the estimated density function for the whole data set.

Consider for example the datasets of Fig. 2, which for the sake of clarity will be called respectively "Books" and "Jewellery" datasets. They are both characterized by a densely populated area in a "small trade" region of no practical interest in the anti-fraud context. In the case of the Books dataset, units are so concentrated that only 0.02% of the data is retained, while the general data pattern is preserved. Note that the initial size of these datasets can be so large to make analyses computationally very demanding (the application of the FS to the 33304 books import flows went out of memory after running several hours on a 2.1 GHz Xeon processor with 16 Gb of memory).

To illustrate the issues that we encounter in monitoring this type of datasets, we use first a trade-like dataset simulated for a separate assessment exercise, still in progress. It is represented in Fig. 3 before and after thinning (left and right panel rspectively). Along the lines of CRAC, we monitor the S and MM estimators fit on the original data. To be more general than in the case illustrated in Section 2, from now on we will use a model with intercept. The forward plot trajectories appear as uninformative flat lines, shown in Fig. 4 for the MM estimator.

After thinning, when the same monitoring is applied on the retained units, the forward plots of the S estimator (right panel of Fig. 5) show that, when the breakdown point is chosen between 0.5 and 0.45 the outliers are very well identifiable. On the

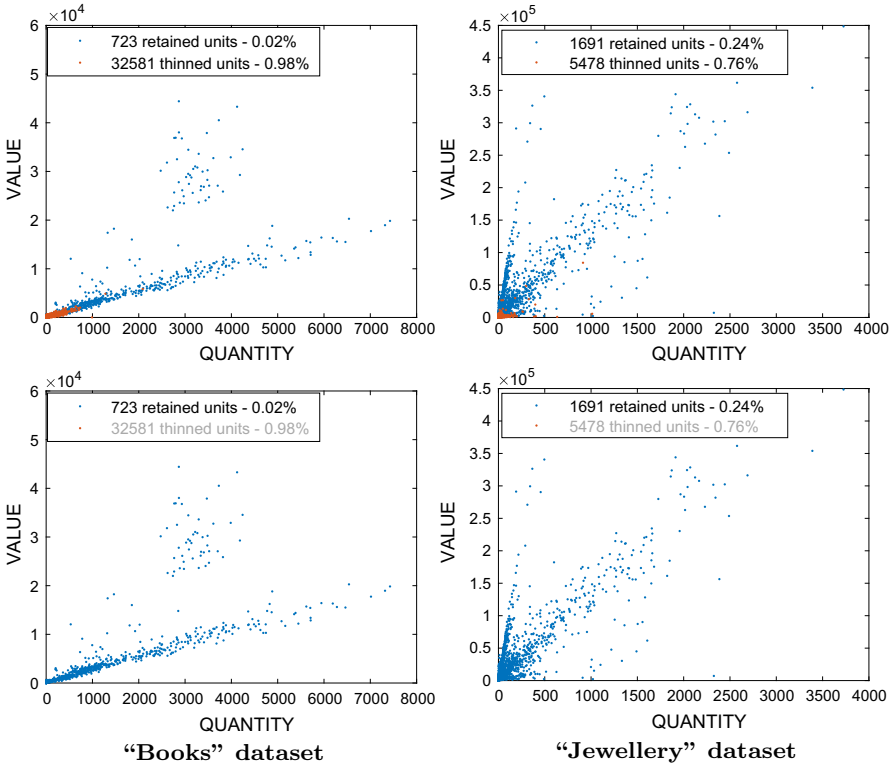


Fig. 2 Two trade datasets with one or more populations, outliers and dense area near the origin of the coordinated axes (the "small trade" area). Left panel: "Printed books, brochures, leaflets and similar printed matter". Right panel: "Imitation jewellery of base metal, whether or not plated with precious metal"

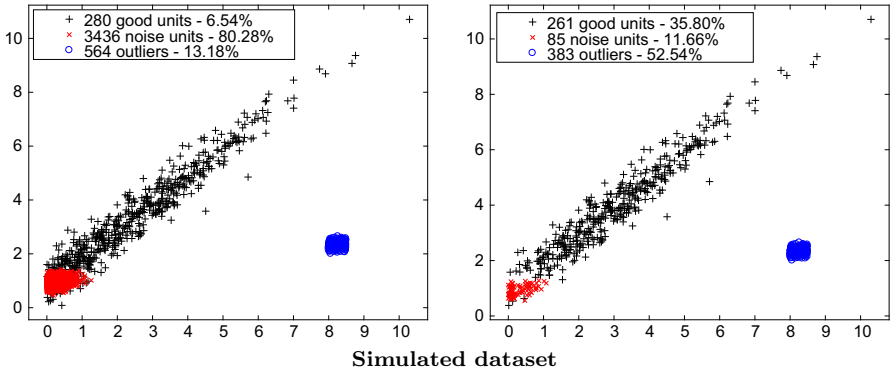


Fig. 3 Left panel: a 4280-unit simulated dataset. Main component formed by two superimposed linear groups, one with 280 good units (black crosses) and another with 3436 noise units (red crosses). The x values of both components are half-normal distributed, with parameters $\sigma = 3$ and $\sigma = 0.3$ respectively. The intercepts of both components are set to zero while the slopes are 1 and $(\arctan(10) \approx 0.18)$. Errors variances are respectively 0.02 and 0.2. The gaussian contamination on the right is formed by 564 outliers. Right panel: 729 retained observations after thinning, with indication of the group proportions in the legend

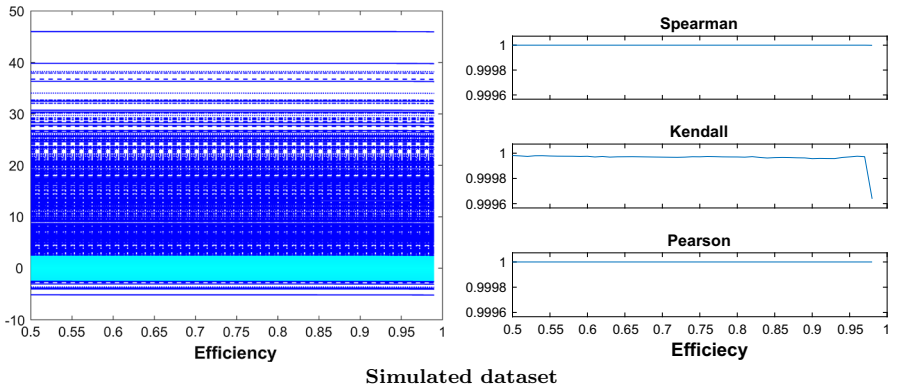


Fig. 4 Left panel: MM Forward plot of scaled residuals obtained on the original simulated dataset (left panel of Fig. 3), for different values of efficiency. Right panel: Spearman, Kendall and Pearson correlation coefficients among the MM residuals in the left panel

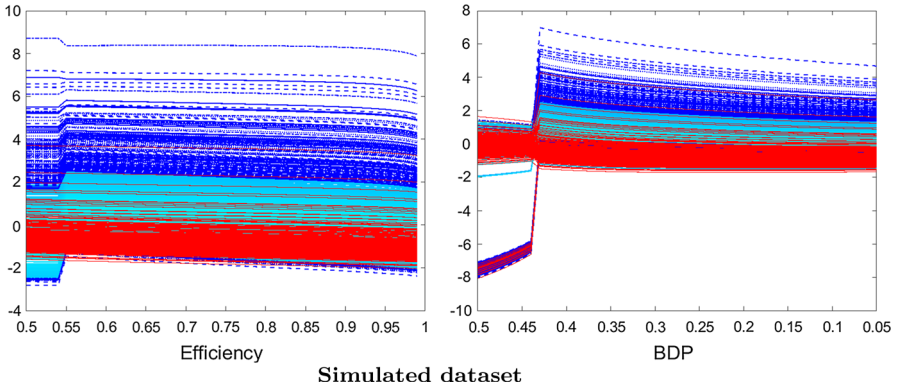


Fig. 5 MM and S forward plots of scaled residuals obtained on the 729 units retained by the thinning step in the simulated dataset (right panel of Fig. 3). The monitoring for MM (left panel) is in a range of efficiency values, while for the S (right panel) it is over the breakdown point values. The trajectories of the residuals of the 729 retained units are represented in blue. The residuals of the 3551 thinned units have been computed and added in red

contrary, for smaller breakdown point values, masking occurs. By checking the id number of the units in the lower group of trajectories between 0.5 and 0.45 breakdown point (we used for this an interactive data tooltip of our FSDA toolbox), we could verify that they correspond to the group of outliers. Same information in the forward plot of the MM residuals (left panel of Fig. 5) is more difficult to grasp, but the presence of structure in the data is now very clear in comparison with the flat plot of Fig. 4 obtained with the original dataset. Along the lines of CRAC, we provide the corresponding correlation forward plots in Fig. 6, where the structural change points are well identified.

Results obtained on the two real trade datasets represented in Fig. 2 are less definite. In both cases, the monitoring of the S-residuals on the whole datasets (left panels of Figs. 7 and 8) shows the effect of masking when inappropriate breakdown point

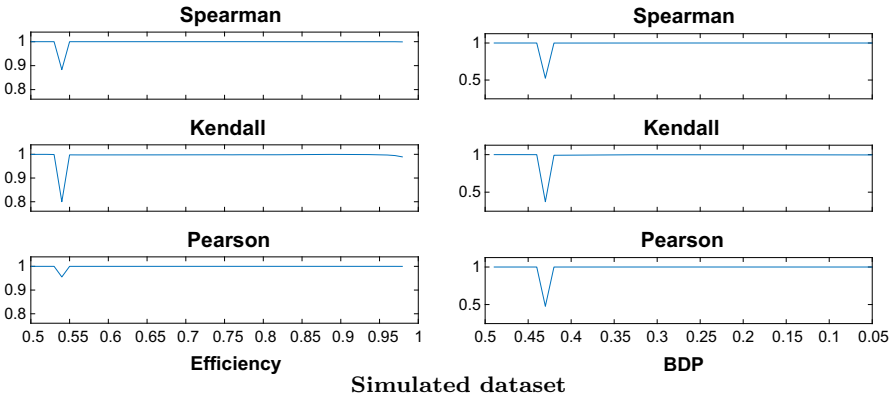


Fig. 6 Forward plot of Spearman, Kendall and Pearson correlation coefficients among the MM and S residuals (left and right panel respectively) of Fig. 5

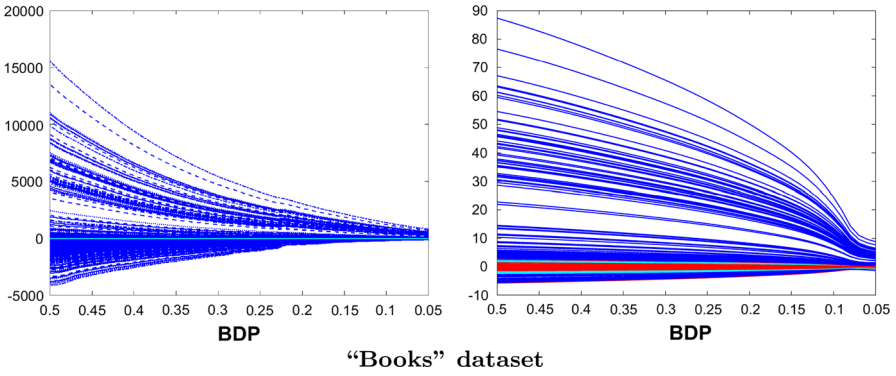


Fig. 7 Left panel: S forward plot of scaled residuals obtained on the original 33304 units. Right panel: S forward plot obtained on the 723 units retained by the thinning step. The trajectories of the residuals of the 723 retained units are represented in blue. The residuals of the 32581 thinned units have been computed and added in red

values are chosen, with residual trajectories which become closer and closer. After the thinning step, the monitoring on the retained data clearly shows the presence of structure in the data. In the right panel of Fig. 7 a drastic decrease of the residuals occurs below a certain breakdown value, around 10%, suggesting that masking is occurring when the outliers present, which indeed are roughly 10% (about 70 high price outliers among a total of 723 units), start distorting the estimates. In the right panel of Fig. 8 the sudden decrease corresponding to a breakdown point of 0.45 indicates the presence of two major groups in the data.

Note, in both figures, the different scales of the monitored residuals in the original and thinned datasets. To understand the nature of this effect we have monitored the intercept and slope values estimated in the two cases. Figure 9, which refers to the Books dataset, shows that the intercept is close to 0 if all data are fit, while with the retained units it is between 100 and 350, depending on the breakdown point, with

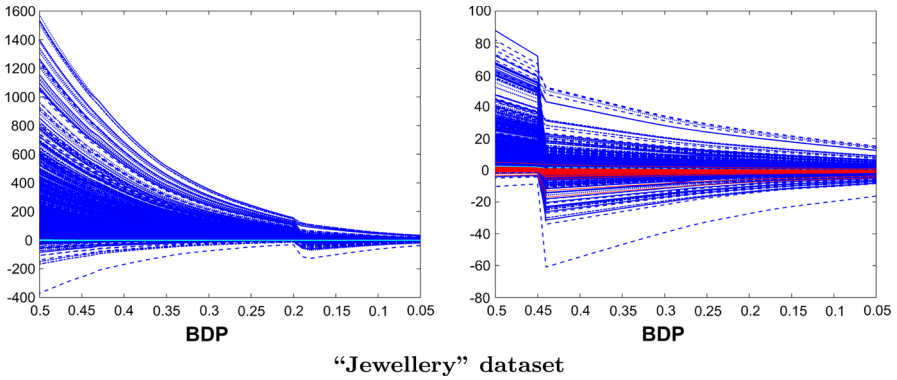


Fig. 8 Left panel: S forward plot of scaled residuals obtained on the original 7169 units. Right panel: S forward plot obtained on the 1691 units retained by the thinning step. The trajectories of the residuals of the 1691 retained units are represented in blue. The residuals of the 5478 thinned units have been computed and added in red

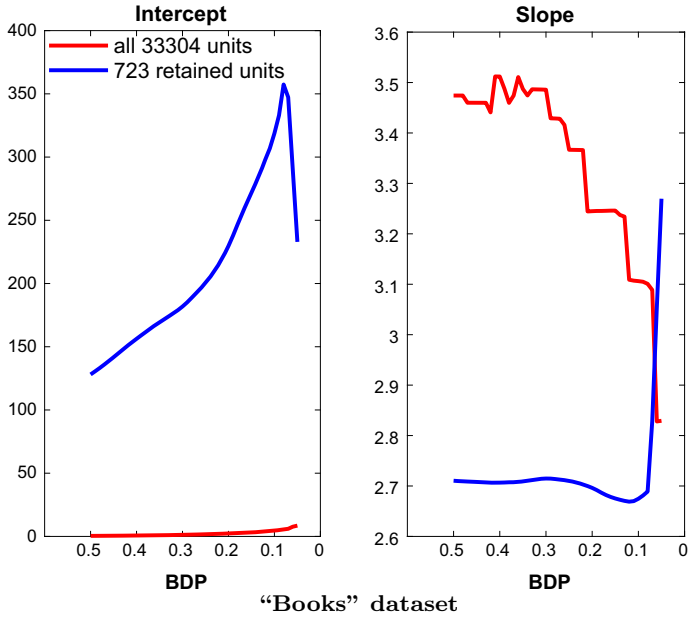


Fig. 9 Monitoring of the S-regression parameters fit on the Books data, before and after thinning (red and blue line respectively)

obvious inflation effect on the residuals. The corresponding slopes for a standard 0.5 breakdown point are respectively around 3.5 and 2.7. We could verify that the most reasonable slope (obtained with a robust fit using a model without intercept, to estimate the import price of the books) is 2.8, which is very close to the S fit on the retained units. Finally note that also the monitoring of the estimated regression parameters shows that something is occurring for a breakdown point approximately equal to 0.1.

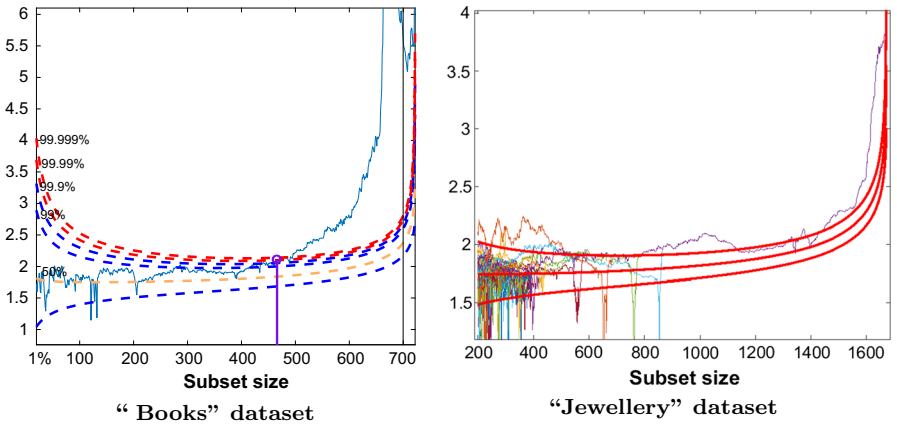


Fig. 10 Left panel: Forward Search plot of minimum deletion residual among observations not in the subset obtained on the 723 retained units of the Books dataset. Right panel: Forward Search random start plot of minimum deletion residual among observations not in the subset on the 1691 retained units of the Jewellery dataset

The monitoring of the minimum deletion residuals with the FS provides similar information about the two trade datasets. The left panel of Fig. 10 clearly shows that the Books dataset is formed by one main population and a set of outliers. The right panel of the same figure, which report the same monitoring (from step 200) for many random starts on the Jewellery dataset, shows different sets of trajectories indicating the presence of multiple groups.

4 Closure

For CRAC, *the monitoring* is more than a particular way of dealing with data: they often like to state that it is a truly data analysis philosophy, which comes from the belief that data can be completely understood only by appraising the effect on a fitted model of each statistical unit, or sub-groups of units. In this discussion we have provided other evidence that the monitoring is, at least, a very powerful instrument to summarize lot of information in one single plot.

References

Atkinson CA, Riani M (2000) Robust diagnostic regression analysis. Springer, Berlin. https://doi.org/10.1007/978-1-4612-1160-0

Cerioli A, Perrotta D (2014) Robust clustering around regression lines with high density regions. Adv Data Anal Classif 8(1):5–26. ISSN 1862-5355

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 (1). In press

Heikkonen J, Perrotta D, Riani M, Torti F (2013) Issues on clustering and data gridding. Springer, Berlin, pp 37–44

Riani M, Perrotta D, Torti F (2012) FSDA: a MATLAB toolbox for robust analysis and interactive data exploration. Chemom Intell Lab Syst 116(Supplement C):17–32

Riani M, Cerioli A, Atkinson AC, Perrotta D (2014) Monitoring robust regression. Electron J Stat 8(1):646–677. https://doi.org/10.1214/14-EJS897

Riani M, Perrotta D, Cerioli A (2015) The forward search for very large datasets. J Stat Softw Code Snippets 67(1):1–20