Anthony C. Atkinson, Andrea Cerioli and Marco Riani

Discussion of “asymptotic theory of outlier detection algorithms for linear time series regression models” by Johansen and Nielsen

Article (Accepted version) (Refereed)

Original citation:
Atkinson, Anthony C., Cerioli, Andrea and Riani, Marco (2016) Discussion of “asymptotic theory of outlier detection algorithms for linear time series regression models” by Johansen and Nielsen. Scandinavian Journal of Statistics, 43 (2). pp. 349-352. ISSN 0303-6898
DOI: 10.1111/sjos.12210

Discussion of “Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models” by Johansen and Nielsen

Anthony C. Atkinson
Department of Statistics, London School of Economics
Andrea Cerioli and Marco Riani
Department of Economics, University of Parma

Key words: Fan plot; forward search; Mahalanobis distance; monitoring; robustness

We enjoyed reading Johansen and Nielsen (2016) which describes some recent asymptotic theory with potential for applications in data analysis. We first comment on the Forward Search in regression and finally consider the similar mathematical problems that arise in robust estimation for multivariate data when residuals are replaced by Mahalanobis distances. In between, our major empirical contribution is a forward analysis of the Fulton fish market data which leads to straightforward conclusions.

Johansen and Nielsen stress the strong relationship between outlier detection and robust estimation. This depends on the use of the binary weights or indicator functions v_i in Johansen and Nielsen's §5.1. Smooth downweighting, as in M-estimation, is excluded. The aim is to partition the data into two sets: “good” data for which the $v_i = 1$ and “bad” data, or outliers, for which $v_i = 0$. Given the partition, least squares can be used to estimate the parameters and, using these estimates, the outliers are revealed. The subject is mathematically and computationally interesting because, if there are several outliers, they may not be revealed by initially fitting a model to all the data (that is starting with all $v_i = 1$); the parameter estimates may be so perturbed by the outliers that the outliers appear to be part of the good data, a phenomenon known as masking.

The title of the paper makes clear the asymptotic nature of the results obtained. It doesn't make clear that all calculations are under the null hypothesis; outliers perhaps only making an appearance in the Fulton fish market data. When Johansen and Nielsen come to their list of problems in their §11 they will need to consider both the size and power of the outlier test. Riani *et al.* (2014a) used the average power, that is the average number of observations correctly detected as being contaminated. Of course, high power can be obtained by a test with a higher than nominal size, so both quantities need to be considered.

The purpose of the study of Riani *et al.* (2014a) was to compare the performance of various robust estimators in regression when viewed as outlier detection tests. The novelty was to consider comparative performance as the outlier pattern, typically a cluster, moved in a parameterised way from being remote from the majority of the data to being close to it and then moved away again. Johansen and Nielsen might like to include something like this parametric trajectory of outliers in their list of further problems in their §11. Theoretically, this emphasized that there are often two asymptotic assumptions in the study of robust estimation: one is that the sample size $n \to \infty$. The other is consideration of what happens to estimators as the outliers move to the edge of the parameter space, that is the breakdown of an estimator.

The Forward Search algorithm is straightforward. Its use to detect outliers less so. The rule in Johansen and Nielsen's (18) can be thought of as the "crude" rule, responding to the first crossing of a boundary. Their Figure 4 shows how the number of outliers detected in outlier free samples can be large and depend on the part of the search monitored. Simulations in Atkinson and Riani (2006), adapting and extending a sophisticated simulation method of Buja and Rolke (2003), also considered rules responding to one or several successive values above the pointwise boundary, with similarly disappointing results.

A second problem with the crude rule is that, with many outliers, masking may cause the first crossing to occur in the centre of the search, while, at the same time, there is little or no indication of departures in the later stages. To overcome these problems, Riani *et al.* (2009) developed a complicated empirical rule which covers the whole search from $m > 3p$, using a variety of pointwise intervals derived from a scaled-F distribution and, sometimes, envelopes from a series of sample sizes. The regression version of this procedure, which uses minimum deletion residuals that contain an expression for leverage, has excellent size (1%) and good power over the ranges explored. We would greatly welcome an insightful comparison between our empirical rule and the asymptotic approach developed in this paper.

Another specific, but we feel important, remark on calibration of cutoff values is given by Johansen and Nielsen below their (17). A similar argument was exploited in Riani *et al.* (2014b) for S and MM-estimators, two instances of smooth downweighting, to obtain relationships between efficiency and breakdown point. This leads to the monitoring of the performance of methods of robust regression as the breakdown point or efficiency vary (Riani *et al.*, 2014a), made possible by the efficient routines of the FSDA Matlab Toolbox available at `http://www.riani.it`. Of course, an advantage of the Forward Search is that we do not have to specify breakdown point or efficiency in advance.

Now for some data analysis. Two data transformations are plotted in Johansen and Nielsen's Figure 1 and the results of a number of aggregate tests reported. In the spirit of the Forward Search we like to see several transformations considered and the values of tests monitored throughout the search. If there are outliers, which, if any, inferences do they affect?

Figure 1 about here

Figure 1 is a "fan plot" (Atkinson and Riani, 2000), that is a forward plot of an approximate score test for the parameter in the Box-Cox transformation of the response in the regression. The central band contains 99% of a standard normal distribution. This shows that $\lambda = 1$ (no transformation) is thoroughly rejected, as are values of -1 and -0.5. The log transformation ($\lambda = 0$) is rejected at the end of the search when the last five observations enter the subset; 0.25 is an acceptable value, but 0.5 scarcely so with a value of -2.41 for the statistic for all the data.

Figure 2 about here

Figure 2 shows results for the analysis on the log scale. The left-hand panel is the plot of scaled residuals. This stable plot reveals that the three most extreme residuals are negative. Brushing the plot shows where these points lie in the scatterplots of y against x_1 and x_2. The right-hand plot shows the forward plot of absolute minimum deletion residuals with 99% pointwise envelopes based on a sample size of 110. Although the three most extreme residuals cause the trajectory to increase, it does not go outside the upper limit.

Figure 3 about here (No new paragraph)

Figure 3 shows similar results for the square root transformation. Now the two largest residuals in the left-hand panel are positive. The plot is again stable. Introduction of these two observations causes a smaller increase in the trajectory plotted in the right-hand panel than occurs when $\lambda = 0$. The final figure

Figure 4 about here (No new paragraph)

is for the analysis with $\lambda = 0.25$. As might be expected from this intermediate parameter value, the largest residuals are both positive and negative. The scatterplots in the centre panel show that the observations are not far from the rest of the data, and the plot of absolute minimum deletion residuals at the end of the search lies close to the median value.

There are three comments about our analysis. One is that we see none of the instability hinted at in §10 of Johansen and Nielsen's paper. The second is on the width of our pointwise intervals, which are sufficiently wide not to detect any outliers for the three transformations considered in detail. This points again to the importance of an extensive comparison between asymptotic and finite sample bands. The third is that we have ignored the autoregressive structure when using the forward search.

We conclude our discussion by stressing that we agree with Johansen and Nielsen about the importance of controlling the size of any outlier detection test. Our experience has been that too liberal cutoff values can produce deleterious practical consequences in many areas, with a plethora of false signals. We also provided some corrections in the case of multivariate data, where Mahalanobis distances replace scaled residuals. Our improved methods involve both distributional results for finite samples and powerful ways of dealing with multiple outliers tests (Cerioli *et al.*, 2009; Cerioli, 2010; Cerioli and Farcomeni, 2011). However, we acknowledge that a unified asymptotic theory of multivariate outlier detection has yet to come. We thus look forward to Johansen and Nielsen's contribution also in this field.

References

Atkinson, A. C. & Riani, M. (2000). *Robust diagnostic regression analysis*. Springer–Verlag, New York.

Atkinson, A. C. & Riani, M. (2006). Distribution theory and simulations for tests of outliers in regression. *J. Comput. Graph. Statist.*, **15**, 460–476.

Buja, A. & Rolke, W. (2003). Calibration for simultaneity: (re)sampling methods for simultaneous inference with applications to function estimation and functional data. Technical report, The Wharton School, University of Pennsylvania.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *J. Amer. Statist. Assoc.*, **105**, 147–156.

Cerioli, A. & Farcomeni, A. (2011). Error rates for multivariate outlier detection. *Comput. Statist. Data Anal.*, **55**, 544–553.

Cerioli, A., Riani, M., & Atkinson, A. C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Stat. Comput*, **19**, 341–353.

Johansen, S. and Nielsen, B. (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, **43**, ???—???

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **71**, 447–466.

Riani, M., Cerioli, A., Atkinson, A. C., & Perrotta, D. (2014a). Monitoring robust regression. *Electron. J. Stat.*, **8**, 642–673.

Riani, M., Cerioli, A., & Torti, F. (2014b). On consistency factors and efficiency of robust S-estimators. *TEST*, **23**, 356–387.

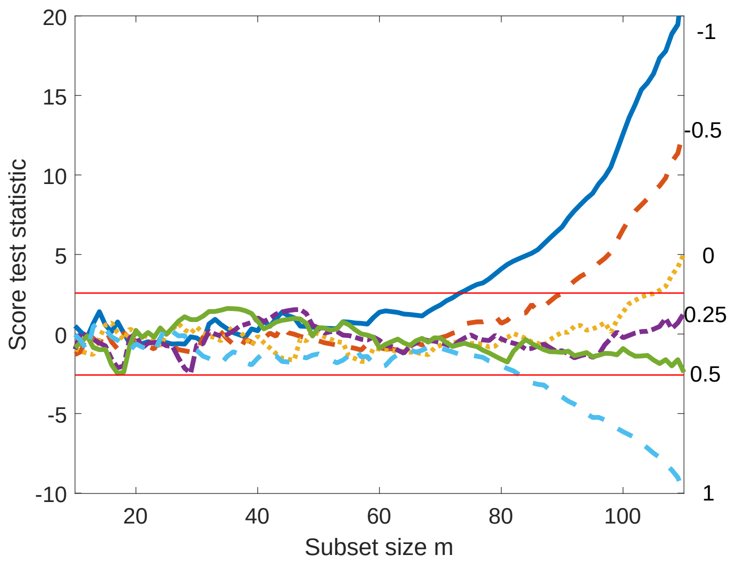

Figure 1: Fan plot for Fulton fish market data.

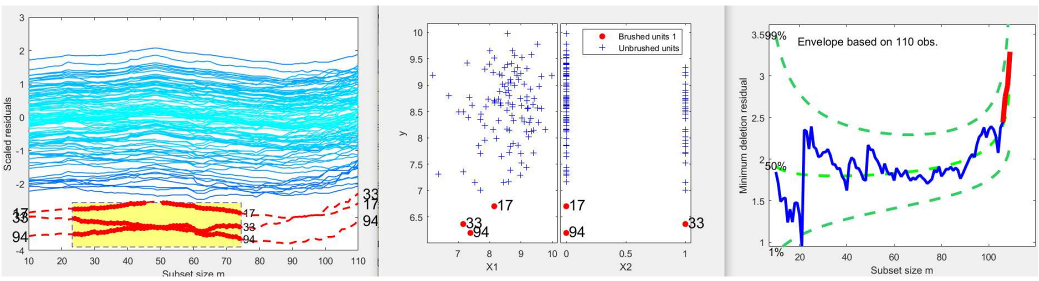

Figure 2: Fish data: analysis on the log scale, $\lambda = 0$. Left: plot of scaled residuals. Centre: brushing the scatterplots. Right: forward plot of absolute minimum deletion residuals.

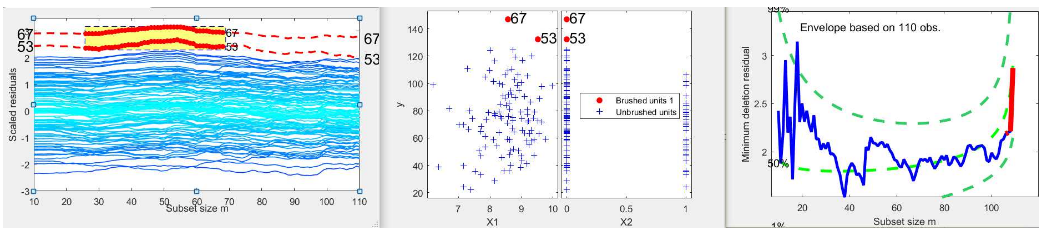

Figure 3: As Figure 2, but now the analysis is with the square root transformation, $\lambda = 0.5$.

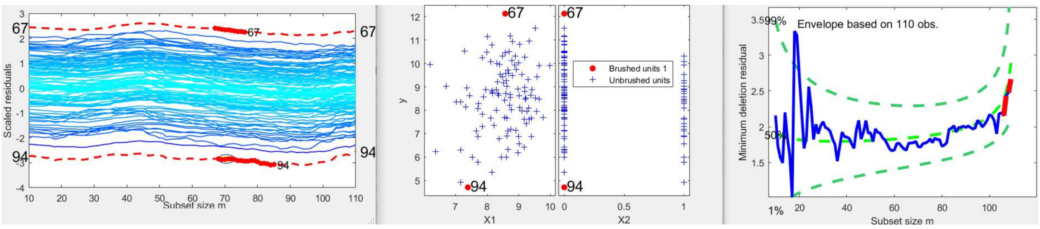

Figure 4: As Figure 2, but now the analysis is with $\lambda = 0.25$.