Marco Riani, Aldo Corbellini, Anthony C. Atkinson

The use of prior information in very robust regression for fraud detection

Article (Accepted version)
(Refereed)

Original citation: Riani, Marco and Corbellini, Aldo and Atkinson, Anthony C. (2018) *The use of prior information in very robust regression for fraud detection.* International Statistical Review.
ISSN 0306-7734
DOI: 10.1111/insr.12247

The Use of Prior Information in Very Robust Regression for Fraud Detection

Marco Riani[*] and Aldo Corbellini[†]
Dipartimento di Economia, Università di Parma, Italy
Anthony C. Atkinson[‡] Department of Statistics
London School of Economics, London WC2A 2AE, UK

November 9, 2017

Abstract

Misinvoicing is a major tool in fraud including money laundering. We develop a method of detecting the patterns of outliers that indicate systematic mis-pricing. Since the data only become available year by year, we develop a combination of very robust regression and the use of 'cleaned' prior information from earlier years which leads to early and sharp indication of potentially fraudulent activity that can be passed to legal agencies to institute prosecution. As an example we use yearly imports of a specific seafood into the European Union. This is only one of over one million annual data sets, each of which can currently potentially contain 336 observations. We provide a solution to the resulting big data problem which requires analysis with the minimum of human intervention. *Keywords:* big data, data cleaning, forward search, MM-estimation, misinvoicing, money laundering, seafood, timeliness.

[*] Email: mriani@unipr.it

[†] Email: aldo.corbellini@unipr.it

[‡] Email: a.c.atkinson@lse.ac.uk

1 Introduction

It is estimated (Economist, 2014) that, in 2011, \$950 billion flowed illegally out of poor countries into rich ones, mostly due to money laundering associated with the traffic in illegal drugs and arms trading. A basic technique is misinvoicing. In our paper we develop a form of very robust regression that uses "cleaned" data from previous years to give improved analyses of the data for the current year. In addition to the importance of improved methods for fraud detection, our paper extends a form of Bayesian regression to incorporate different amounts of prior information about the parameters of the linear model and the error variance.

As an example we look at data on three years importation of a specific seafood into the European Union from one country on the American continent. There are 165 monthly observations in the first year. However the problem is vast, with around 220 potential source countries, monthly data and over 1,000 categories of goods (although not all countries are sources of all goods). To cope with this example of Big Data we need robust methods that function semi-automatically on our relatively small problem, without the need for close personal intervention. In this way the big data problem of analysing a very large number of such data sets becomes feasible.

The observations are regression data of quantity against value with a few missing observations. In our particular example there is a linear relationship followed by the majority of the data, a few outliers and a second, lower, line with fewer observations. This line is an indication of potential fraud - by incorrectly recording import prices, import duties and taxes such as VAT can partially be avoided. Conversely, in other sets of data we see suspiciously high invoice prices, which allow illicit money to be laundered into legal bank accounts. In economic theory, the efficient-market hypothesis asserts that, in a well organized, reasonably transparent market, the market price is generally equal to or close to a fair value. Marked departures from this value are an indication of inefficiency in the market, in this case fraud.

To prosecute such behaviour it is necessary to demonstrate, as far as possible, the incontrovertible existence of outliers. This is very different from the standard intent of robust data analyses, where the purpose is to establish a single relationship between much of the data and a model; the remaining data are then either downweighted or trimmed.

As a method of very robust regression, we use the Forward Search (FS) (Atkinson and Riani, 2000). Atkinson *et al.* (2010) describe more recent developments in the theory of the FS. Comparisons of the most recent version with other forms of robust regression are in Riani *et al.* (2014). Johansen and Nielsen (2016) focus on outlier detection. An advantage of the FS is that it is fully automatic, avoiding the specification of such parameters as breakdown points or efficiencies. Also, since the FS uses least squares to fit a regression model to carefully chosen subsets of data, it is relatively straightforward to adapt the method for the incorporation of prior information.

Our paper starts in §2 with a description of the frequentist FS, which is used to analyse the data from 2002. This analysis indicates that 15 of the 165 observations are outliers. However, a scatterplot of the data suggests that not all of these are fraudulent, let alone being sufficiently outlying to provide judicially convincing evidence of fraud. We use the sufficient statistics from a cleaned version of the data to provide prior information for the analysis of the data from 2003, continuing the process from year to year until overwhelming evidence of fraud has accumulated.

We have two ways of cleaning the data, one for the model for the mean and the other for the variance. In the next year we use the non-outlying observations from the FS to determine the parameters β of the linear model which provide an estimate of the fair value. Trimming so many observations however indicates too many outliers to be helpful in fraud detection. Experience from those preparing legal evidence suggests that courts are most comfortable with evidence presented in the form of raw residuals, that is differences between observed and fitted values without scaling for leverage and the estimate of error variance. Accordingly, we use a relatively generous fixed threshold around the fitted regression line to indicate which of the outliers should be excluded from the central part of the data. We use all observations within this threshold to provide the prior estimate of the error variance σ^2. Use of a fixed threshold is justified since there is little interest in detecting fraudulently declared small transactions. This choice of the estimated variance is motivated by the analysis in the on-line supplement (Riani *et al.*, 2018) which shows that the error distribution gives rise to a large number of very small residuals, which can cause robust procedures to identify a large number of outliers.

For least squares without trimming, incorporation of prior information from previous years comes by inclusion of the sufficient statistics of the previous regression. This is also well-established in Bayesian regression (Chaloner and Brant, 1988). However, there are two difficulties in the application of this method in the present case. One is that the estimate of σ^2 from the FS is based on a central subset of observations, and so has to be adjusted before combination with the prior estimate, an adjustment which leads to a weighted form of least squares. The other difficulty is that we have one prior sample for the estimate of β and a larger one for the estimation of σ^2. In §3 we describe the incorporation of prior information into the FS.

The Bayesian analysis of data for 2003 and 2004 is in §4. At the end of the analysis of three years' data, a set of potentially fraudulent observations is clearly established in a sufficiently unambiguous form to be passed to the agency responsible for legal proceedings. Convincingly, they all relate to imports into one member state of the European Union. We show in §4.3 that choice of the threshold is not crucial to the identification of these observations, a wide range of values providing evidence of the outliers, provided the threshold is not too small. The value is to be decided in consultation with subject-matter experts. In the on-line supplement (Riani *et al.*, 2018) we summarise some other analyses of the data from 2002. These comparisons illustrate the dependency of S and MM estimates on the parameters, such as breakdown point, used in the algorithms. The forward search does not require such specifications. It is important that throughout we are developing a method for general departures from the regression model, rather than being interested in modelling the linear structure of departures we find in our data.

We conclude in §6 with some comments on other methods of robust regression that allow for the incorporation of prior information. We further comment on fair value, which may not be constant over time, on quantity and value, and on other forms of trade data, including some in which heteroskedasticity is present. We also mention recent developments of the forward search which render it highly efficient for the analysis of single large sets of regression data.

An important aspect of our solution is timeliness. We are now able to analyse the data in real time. But, when the data we analyse were collected, member states of the European Union only made data available (through Eurostat) on a monthly basis, with about three months delay. It is intended that, from the end of 2018, the data will be provided with greatly enhanced speed and regularity. Our methods will allow efficient exploitation of this improved flow of data.

2 The Frequentist Forward Search

2.1 Parameter Estimation

For analysis of data from the first year, we use a forward search without prior information.

In the regression model

$$y = X\beta + \epsilon, \tag{1}$$

y is the $n \times 1$ vector of responses, X is an $n \times p$ full-rank matrix of known constants, with ith row x_i^{T}, and β is a vector of p unknown parameters. The normal theory assumption is that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$. As we show in the on-line supplement, this assumption needs some modification for the trade data.

The least squares estimator of β is $\hat{\beta}$. Then the vector of n least squares residuals is $e = y - \hat{y} = y - X\hat{\beta} = (I - H)y$, where $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the 'hat' matrix, with diagonal elements h_i and off-diagonal elements h_{ij}. The residual mean square estimator of σ^2 is $s^2 = e^{\mathrm{T}}e/(n-p) = \sum_{i=1}^{n} e_i^2/(n-p)$.

In order to detect outliers and departures from the fitted regression model, FS uses least squares to fit the model to subsets of m observations. The initial subset of m_0 observations is chosen robustly, for example by least trimmed squares. The subset is increased from size m to $m+1$ by forming the new subset from the observations with the $m+1$ smallest squared residuals. For each m ($m_0 \leq m \leq n-1$), we calculate deletion residuals $r_i(m)$, defined in (5) of Appendix 2, for observations not in the subset. The presence of outliers is detected using the smallest absolute residual $r_{\mathrm{imin}}(m)$ from these $n-m$ values (6).

To provide a test for outliers with known properties, we need a reference distribution for the $r_i(m)$. Under normal theory assumptions, if we estimated σ^2 from all n observations, the statistics would have a t distribution on $n-p$ degrees of freedom. However, in the search we select the central m out of n observations to provide the estimate $s^2(m)$, so that the variability is underestimated. To allow for estimation from this truncated distribution, the estimated variance has to be scaled up to give the approximately unbiased estimate of variance $\hat{\sigma}^2(m) = s^2(m)/c(m,n)$. In the robustness literature, the important quantity $c(m,n)$ is called a consistency factor. See Riani *et al.* (2009) for a derivation from the general method of Tallis (1963) and §3 for the extension to a Bayesian analysis.

The central feature of the forward search is the conceptual use of forward plots of minimum deletion residuals in the detection of outliers and other anomalous behaviour. These plots are calibrated against pointwise distributions of the order statistics of the deletion residuals. For normal theory errors the distribution of the order statistics is found applying standard results to the absolute values of t-distributed variables. To avoid the large effect of repeated testing for outliers, simple rules on the number of point-

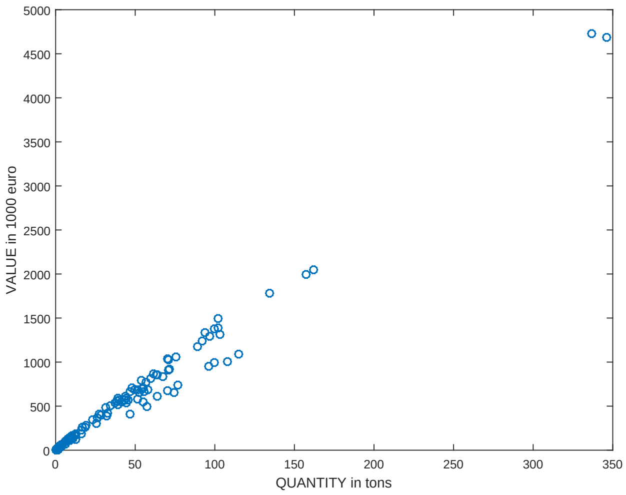

Figure 1: 2002 data. Scatter plot of value, in thousands of Euros, against quantity, in tons, for monthly imports of seafood Q into EU member states from a single exporting country

wise exceedances of percentage points are used to give samplewise rules for the detection of outliers with known size, in our case 1%. We follow the rule used for multivariate data by Riani *et al.* (2009). However, with the non-normal error structure of our data in which over half of the observations often have very small residuals, this procedure gives rise to the identification of a large number of small outliers that are not of interest. We instead use a Bonferroni correction.

If σ^2 is estimated on ν degrees of freedom, when the errors are normally distributed the deletion residuals $r_i(m)$ follow a t distribution on ν degrees of freedom. Since the test is for an outlier in a sample of size $m+1$, we use the Bonferroni bound $t_{\{\nu,\alpha/(m+1)\}}$ with, in our calculations, $\alpha = 1\%$. However, we test using $r_{\text{imin}}(m)$, the absolute value of the residual, so that the appropriate envelope is the folded t distribution. The difference from using a folded normal is negligible.

2.2 Data Analysis

Figure 1 shows a scatter plot of the data from the first year (2002). There are 165 observations on the value and quantity of monthly imports of seafood Q into the European Union from the single country on the American continent. Since not all member states report data for all months, there are some missing values. As would be expected, the value increases with quantity. However, there appear to be at least two lines in the plot, the lower one, including around twelve observations, may be an indication of fraudulent under-recording of the true value of the shipments.

The upper panel of Figure 2 shows the results of the FS with Bonferroni bounds. There is a marked increase in the value of the minimum deletion residual at $m = 150$, indicating that there are 15 outliers in the data. However, it is not clear from this plot that all these observations are indeed important as outliers.

In the lower panel of the plot, the non-outlying observations that are accepted by the FS are marked with crosses. However, there are three indicated outliers that are close to the main upper line, including one of the two observations with a quantity around 340. We need a way to augment the statistical indication of outlyingness with a practical measure, the fixed threshold around the regression line described in §1.

In this analysis we take this threshold as 300. In §4.3 we investigate the sensitivity of the method to the value of this threshold which, of course, will depend on the goods generating the data being analysed.

In Figure 2 we have marked with the symbol X (magenta in the pdf version) intermediate observations which were identified as outliers by the FS but have raw residuals less than 300. Circles (red) are used to mark indicated outliers that have larger raw residuals. As the plot shows, the three observations close to the majority relationship are no longer suspected of being fraudulent. We also lose some outliers for small quantities, but still note the largest five observations on the lower line as outliers, indicated by circles.

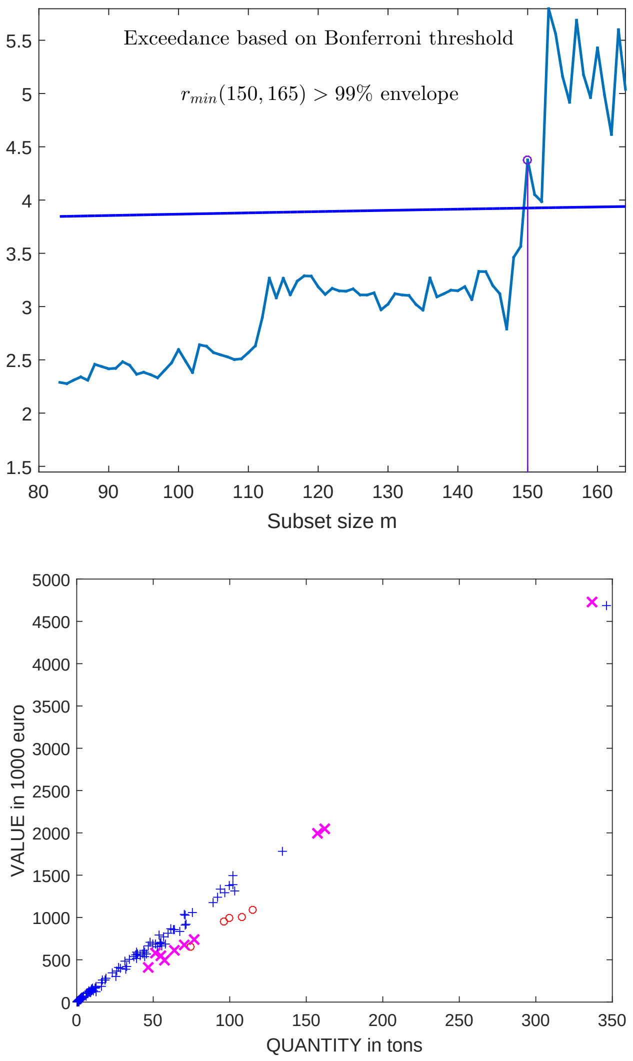

Figure 2: 2002 data. Upper panel: forward plot of minimum deletion residuals. The Bonferroni bound indicates 15 outliers. Lower panel: scatter plot. + non-outlying observations, X intermediate observations within a threshold of ±300, O outliers

The non-outlying observations in year 2002 (blue and magenta) provide prior information for the analysis of data from the next year. As a result of the data analysis of this section, we have two forms of prior information. That for β only uses the 150 'good' observations determined by the FS to be non-outlying. However, we have argued that the variance of this set of observations is too small. We therefore augment this set by the intermediate observations lying within the threshold, to give a set of 160 observations marked in the figure by crosses and the symbol X. These serve as prior observations for σ^2 in our analysis of the data from 2003.

3 Prior Information from Previous Observations

3.1 Fictitious Observations and the Posterior Distribution of β

The conjugate prior distribution for the parameter β in the regression model (1) is multivariate normal and that for the variance σ^2 is inverse gamma. It is standard, for example, Koop (2003, p. 18), to treat this prior information as coming from n_0 fictitious observations analysed by least squares. However, as a result of the analyses of data from previous years, we have two different sets of non-fictitious prior observations. There are $n_{0,1}$ prior observations for the value of β and $n_{0,2}$ prior observations for σ^2, with $n_{0,2} \geq n_{0,1}$. The $n_{0,k}$ observations $y_{0,k}$ $(k = 1, 2)$ arise from a matrix of explanatory variables $X_{0,k}$. Then the two sets of data consist of the $n_{0,k}$ prior observations plus n actual observations. The search in this case now proceeds from $m = 0$, when the prior observations provide the parameter values for all n residuals from the data. The search then continues as outlined above but with the prior observations always included amongst those used for parameter estimation; their residuals are ignored in the selection of successive subsets.

In addition to the two sets of prior data, there is one further complication in this procedure. The $n_{0,k}$ prior observations are treated as a sample with variance σ_0^2. However, the m observations in the FS subset of the actual data are, as in §2, from a truncated distribution of m out of n observations and so asymptotically have a variance $\sigma^2/c(m,n)$. An adjustment must be made before the two samples are combined. This becomes a problem in weighted least squares (for example, Rao 1973, p. 230). Let y_k^+ be the $(n_{0,k}+m) \times 1$ vector of responses from the prior data and the subset, with X_k^+ the corresponding matrix of explanatory variables. The covariance matrix of the independent observations is $\sigma^2 G$, with G a diagonal matrix; the first $n_{0,k}$ elements of the diagonal of G equal one and the last m elements have the value $c(m,n)$. The information matrix for the $n_{0,k}+m$ observations is

$$(X_k^{+\mathrm{T}} W X_k^+)/\sigma^2 = \{X_{0,k}^{\mathrm{T}} X_{0,k} + X(m)^{\mathrm{T}} X(m)/c(m,n)\}/\sigma^2, \tag{2}$$

where $W = G^{-1}$. In the least squares calculations we need only to multiply the elements of the sample values of y and X by $c(m,n)^{-1/2}$.

Let the prior estimate of β be $\hat{\beta}_0$, that is the least squares estimate of β from $n_{0,1}$ prior observations. The estimate including m sample observations can, from (2), be written

$$\begin{aligned}\hat{\beta}(m) &= (X_1^{+\mathrm{T}} W X_1^+)^{-1} X_1^{+\mathrm{T}} W y_1^+ \\ &= \{X_{0,1}^{\mathrm{T}} X_{0,1} + X(m)^{\mathrm{T}} X(m)/c(m,n)\}^{-1} \times \\ &\quad \{X_{0,1}^{\mathrm{T}} y_{0,1} + X(m)^{\mathrm{T}} y(m)/c(m,n)\}. \end{aligned} \tag{3}$$

3.2 Estimation of Variance in the Forward Search

The estimate of σ^2 requires $S_{0,2}$, the residual sum of squares of the $n_{0,2}$ good and intermediate observations around the model with linear parameter value $\hat{\beta}_0$. This prior estimate is adjusted for regression of $y(m)$ on $X(m)$ to give the posterior estimate $\hat{\sigma}^2(m)$.

Let $\tau = 1/\sigma^2$. The prior distribution of τ is gamma with parameters a and b, that is $p(\tau) \propto \tau^{a-1} e^{-b\tau}$. The mean of this gamma distribution is a/b.

The estimate of σ^2 from the $n_{0,2}$ observations is $\sigma_0^2 = S_{0,2}/\nu_0$ on $\nu_0 = n_{0,2} - p$ degrees of freedom. Then, in the gamma distribution for τ the prior values of the parameters are

$$a_0 = (n_{0,2} - p)/2 \quad \text{and} \quad b_0 = S_{0,2}/2,$$

whence $\sigma_0^2 = b_0/a_0$. The prior distribution of β conditional on τ is $N\{\beta, (1/\tau)R^{-1}\}$, where $R = X_{0,1}^T X_{0,1}$.

Then, in an extension of the results of Chaloner and Brant (1988),

$$\begin{aligned} a(m) &= (n_{0,2} + m - p)/2 \quad \text{and} \\ b(m) &= b_0 + \frac{1}{2}\left[\{y(m) - X(m)\hat\beta(m)\}^T y(m)/c(m,n) + \{\hat\beta^0 - \hat\beta(m)\}^T R\,\hat\beta^0\right], \end{aligned}$$

so that $\hat\sigma^2(m) = b(m)/a(m)$.

3.3 Algebra for the Bayesian Forward Search

The algebra for the FS with prior information is similar to that of the frequentist search, except that information from the $n_{0,1}$ and $n_{0,2}$ prior observations is always included in the search. Subsets are selected from the n observations for the current year.

Let $S^*(m)$ be the subset of size m found by FS, for which the matrix of regressors is $X(m)$. Weighted least squares on this subset of observations (3) yields parameter estimates $\hat\beta(m)$ and $\hat\sigma^2(m)$, the mean square estimate of σ^2 on $n_0 + m - p$ degrees of freedom. The residuals for all n observations, including those not in $S^*(m)$, are

$$e_i(m) = y_i - x_i^{\mathrm{T}}\hat\beta(m) \qquad (i = 1, \ldots, n). \tag{4}$$

The search moves forward with the augmented subset $S^*(m+1)$ consisting of the observations with the $m+1$ smallest absolute values of $e_i(m)$. To start, except for the first year, we take $m_0 = 0$, since the prior information specifies the values of β and σ^2.

To test for outliers the deletion residuals are calculated for the $n - m$ observations not in $S^*(m)$. These residuals are

$$r_i(m) = \frac{\{y_i - x_i^{\mathrm{T}}\hat{\beta}(m)\}\{c(m,n)^{-0.5}\}}{\sqrt{\hat{\sigma}^2(m)\{1 + h_i(m)\}}} = \frac{e_i(m)\{c(m,n)^{-0.5}\}}{\sqrt{\hat{\sigma}^2(m)\{1 + h_i(m)\}}}, \tag{5}$$

where, from (3), the leverage $h_i(m) = x_i^{\mathrm{T}}\{X_{0,1}^{\mathrm{T}}X_{0,1} + X(m)^{\mathrm{T}}X(m)/c(m,n)\}^{-1}x_i$, except for the first year when $h_i(m) = x_i^{\mathrm{T}}\{X(m)^{\mathrm{T}}X(m)\}^{-1}x_i$. Let the observation nearest to those forming $S^*(m)$ be i_{min} where

$$i_{\text{min}} = \arg\min_{i \notin S^*(m)} |r_i(m)|.$$

To test whether observation i_{min} is an outlier we use the absolute value of the minimum deletion residual

$$r_{\text{imin}}(m) = \frac{e_{\text{imin}}(m)\{c(m,n)^{-0.5}\}}{\sqrt{\hat{\sigma}^2(m)\{1 + h_{\text{imin}}(m)\}}}, \tag{6}$$

as a test statistic. If the absolute value of (6) is too large, the observation i_{min} is considered to be an outlier, as well as all other observations not in $S^*(m)$.

This FS provides the value of $\hat{\beta}$, based on $n_{0,1} + m_1^*$ observations, and so the fitted regression line to which the threshold is applied. For the next year we set $n_{0,1} \leftarrow n_{0,1} + m_1^*$. The variance σ^2 is estimated from the $n_{0,2}$ prior observations plus the m_2^* observations lying within the threshold, without any further search, and $n_{0,2} \leftarrow n_{0,2} + m_2^*$.

4 Data Analysis with the Bayesian Forward Search

4.1 2003

We incorporate the posterior information from the analysis of the data from 2002 into the analysis of the 167 observations for 2003, which have a similar structure to those for 2002. From the Bayesian analysis of data for that year we thus take $n_{0,1} = 150$ and $n_{0,2} = 160$. The resulting forward plot of deletion residuals is in the upper panel of Figure 3. The first outlier is at $m = 153$. The bottom-hand panel of the figure shows the 153 observations not indicated as outliers during the search, again plotted with a cross. The figure also shows that there are 6 observations that lie within the threshold of ±300 from the robust line. These again are marked X. The interesting observations, from the perspective of fraud detection, are the 8 observations marked O which lie on a line with a lower slope than the others.

4.2 2004

Finally,we consider in greater detail the Bayesian analysis of data from 2004, for which there are 168 observations. We now have prior information from the non-outlying observations from both 2002 and 2003; thus $n_{0,1} = 303$ and $n_{0,2} = 319$.

The upper plot of Figure 4 shows the central observations and outliers from the Bayesian FS for the 168 observations for 2004. This search suggests 17 outliers. However, four of them lie within the threshold of ±300, leaving 13 outliers, one of which lies close to the upper line. The remaining 12 observations lie on an extremely clear line which warrants further investigation. This investigation showed that all 12 observations came from a single member state. This pattern was so striking that a successful prosecution was eventually instituted. See our earlier comments on timeliness.

Comparison of this analysis with the frequentist analysis using just the data from 2004 amply illustrates the clarity obtained by the incorporation of prior information. The scatterplot of outliers from the analysis without prior information in the lower panel of Figure 4 shows 19 outliers. These include the observations falling on the lower line also revealed in the upper panel. However, the remaining outlying observations lie both above and below the upper line, thus obscuring the structure of the main relationship. Although thresholding reveals the structure, the procedure is arbitrary and may have reduced impact as legal evidence. An important advantage of

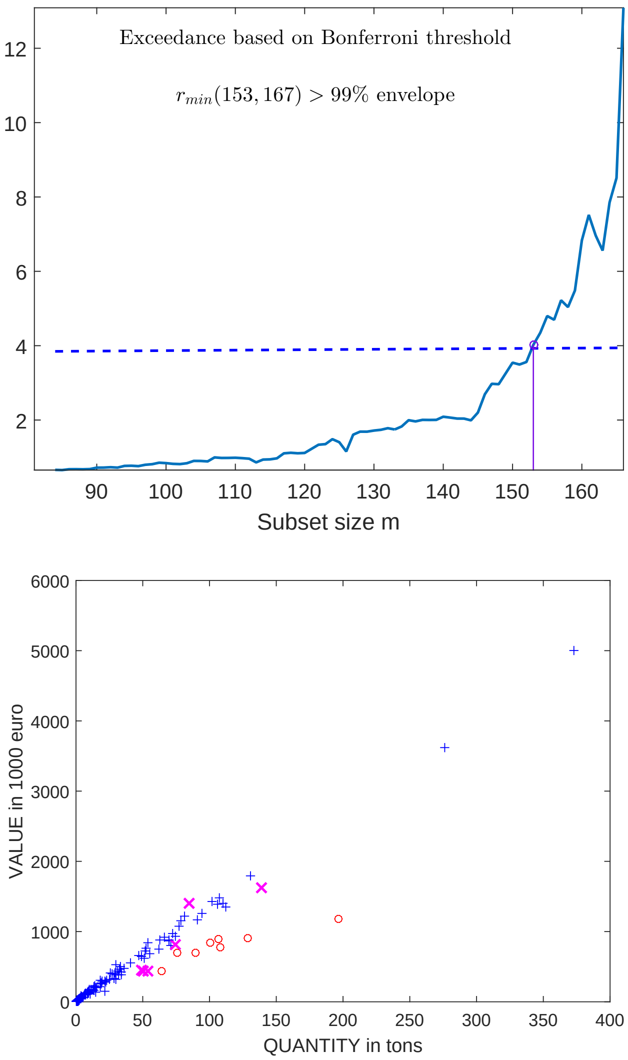

Figure 3: 2003 data, Bayesian search. Upper panel: forward plot of minimum deletion residuals. The Bonferroni bound indicates 14 outliers. Lower panel: scatter plot. + non-outlying observations, X intermediate observations within a threshold of ±300, O outliers

incorporation of prior information is the greater weight of evidence, compared to the analysis of each year separately, leading to more powerful tests of hypotheses about the evidence of fraud provided by the analysis of outliers. This effect is especially important if there are years when a particular good is only lightly traded.

The forward plots of deletion residuals for the two searches are given in Figure 5. The left-hand panel shows the frequentist search, in which the values of the residuals rise to a peak and then decline. This is clear evidence of masking caused by a cluster of outliers; as the observations from the lower line enter the FS subset, the parameter estimates are distorted and the remaining outlying observations seem less remote. There is no such feature in the Bayesian analysis in the right-hand panel where the prior information is sufficiently strong that the points on the lower line remain outlying.

4.3 Determining the Threshold.

An important part of our data analysis has been the threshold for physically significant outliers, to be determined by subject matter specialists. The determination is only required once for each good. The value has been taken equal to 300. We show the importance of the threshold by analysing the data with two other values, 100 and 500.

The left-hand panel of Figure 6 shows the scatterplot of data and residuals when the threshold is set at 100, when 26 outliers are suggested by the FS. The threshold indicates that 7 of these are to be taken as representative data. The remaining outliers include not only the lower line but five observations that the larger threshold of 300 indicates belong to the main population around the upper line. In this case the threshold is too small.

In the right-hand panel of the figure, the threshold is 500. Now all the points near the upper line are accepted and the lower line is clear. However, compared with the threshold of 300 that we used, the four smallest observations from the lower line are also accepted as genuine, as opposed to one in our analysis. Although visual inspection of such plots enables adjustments to

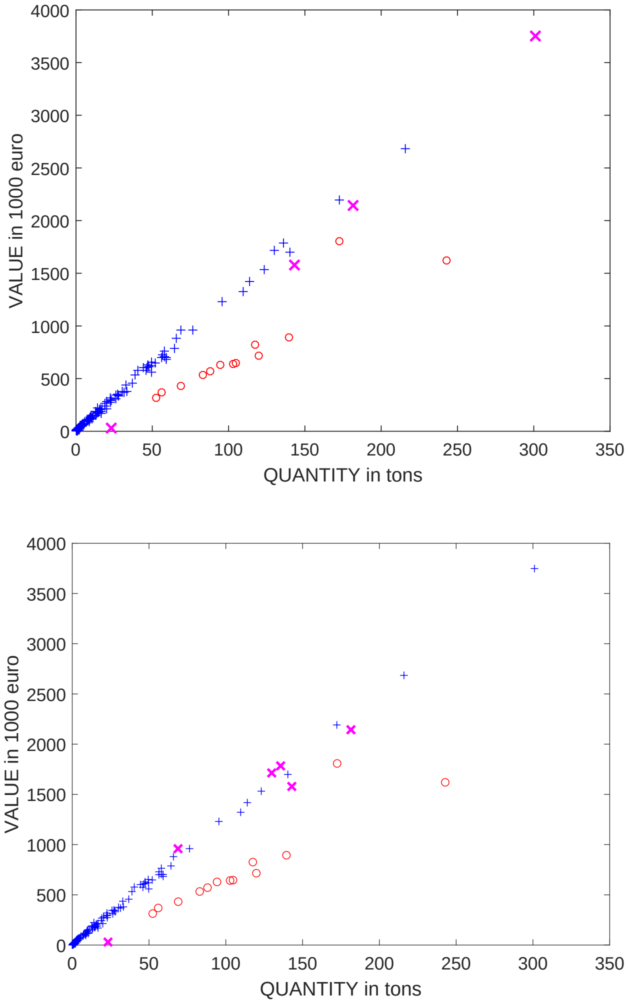

Figure 4: 2004 data, scatter plots of outliers. Upper panel, Bayesian FS using prior information from 2002 and 2003. Lower panel: frequentist FS. + non-outlying observations, X intermediate observations within a threshold of ±300, O outliers

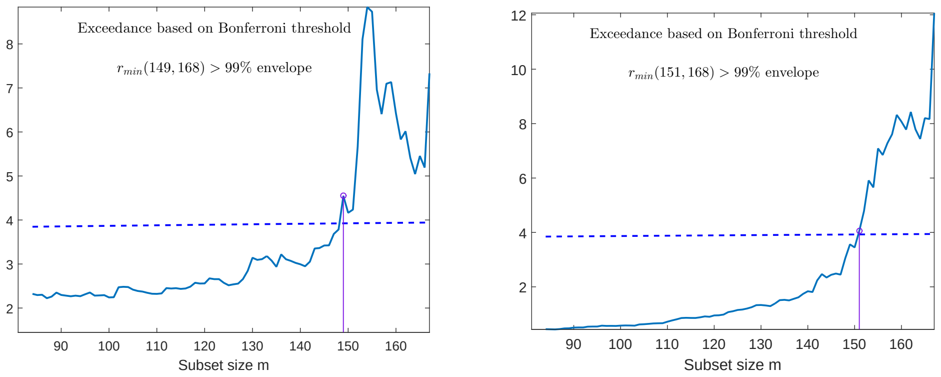

Figure 5: 2004 data, forward plots of minimum deletion residuals. Left-hand panel, frequentist search, showing the effect of masking. Right-hand panel, Bayesian search. Note the difference in vertical scales

be made to the behaviour of the procedure, we require a method that works automatically to indicate anomalous structures.

However, the analysis is not sensitive to the precise value of the threshold. Repeating our analysis with values of 240 and 340 leads to identical results to those when the threshold equalled 300. Too small a threshold indicates too many outliers and may obscure the structure of the data whereas too large a threshold may lead to procedures with reduced statistical power. Analysis with several values for the threshold may be informative.

5 2002: Other Analyses

A major argument both for the use of a Bonferroni correction to identify outliers and the inclusion of a threshold in the analysis was that much of the data lay virtually on a straight line with almost no error. In the online supplement (Riani *et al.*, 2018) we use least squares regression to illustrate this property of the error distribution. Then we consider two robust estimation procedures recommended by Maronna *et al.* (2006) which are potential alternatives to our FS-based analysis. Our results for S estimation show the strong dependence of the number of outliers detected on the breakdown point

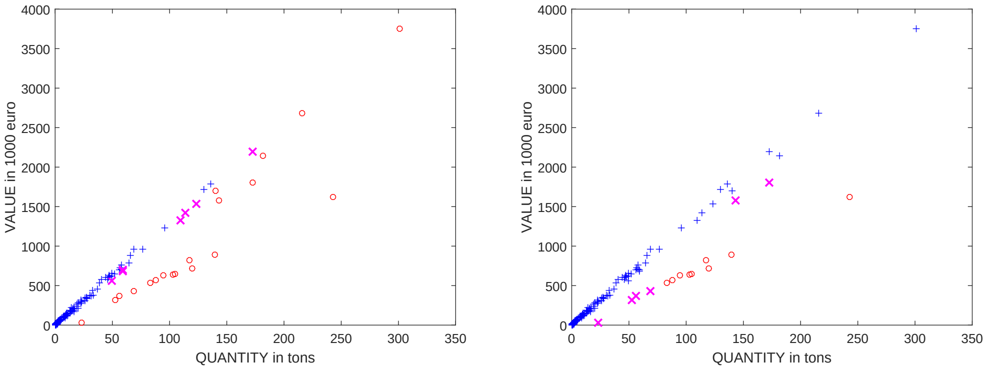

Figure 6: 2004 data, effect of threshold, Bayesian analysis. Scatter plots of outliers. Left-hand panel, threshold = 100, right-hand panel, threshold = 500. + non-outlying observations, X intermediate observations within the threshold, O outliers

specified in advance of the analysis. These results are in line with those from the monitoring of different forms of robust regression in Riani *et al.* (2014) which show, for three examples, just how sensitive the S-estimate can be to the choice of breakdown point. For MM estimation it is necessary to specify the efficiency of estimation of the parameter β. As this efficiency varies from 0.85 to 0.95 the number of declared outliers decreases from 39 to 32. In this context, the diagnostic advantage of least squares compared to straightforward robust procedures is that it does not produce large residuals from well-behaved data. See Cook and Hawkins (1990) for an example in multivariate analysis where a robust method leads to “outliers everywhere”.

6 Discussion

The critical dependence of MM and S estimates on constants chosen by the data analyst renders them problematic for semi-automatic routine use in monitoring large data sets. However, the use of prior information from year to year should serve to stabilize these methods. Some Bayesian methods for robust regression, for example Liu (1996), replace normally distributed errors with longer tailed ones, such as the t-distribution. These methods have lower breakdown points than the maximum value of 50% for the methods compared here. In contrast, we suggest a two-stage procedure in which the weights of the observations from frequentist very robust regression are combined with prior information. In our context of using prior information from previous observations, it is straightforward to modify the expressions for information matrices and parameter estimates in §3 to include weights from estimation methods other than FS. If W_R is the $n \times n$ diagonal matrix of the weights from a robust regression we replace the information matrix for the subset m, that is $X(m)^{\mathrm{T}}X(m)$ by $X^{\mathrm{T}}W_RX$ and the sufficient statistic $X(m)^{\mathrm{T}}y(m)$ by $X^{\mathrm{T}}W_Ry$. A comparison of their procedure with the FS is in section 6 of Atkinson *et al.* (2018). For the particular set of data analysed, the two analyses find virtually identical sets of outliers.

Robust methods, including the FS, can be computationally intensive. Here we have used the FS for relatively small individual sets of data. However, Riani *et al.* (2015) describe a version suitable for the analysis of large data sets. The principal improvements in speed come from a recursive implementation of the procedure which exploits the information of the previous step. The output is a set of efficient routines for fast updating of the model parameter estimates, which do not require any data sorting, and fast computation of likelihood contributions, which do not require matrix inversion or QR decomposition. It is shown that the new algorithms enable a reduction of the computation time by more than 80%. Furthermore, the running time now increases almost linearly with the sample size

Part of our argument for the thresholding procedure was that of the idea of a fair value for the goods being imported. Figure 7 illustrates this idea through scatter plots of the data for the three years. All have a similar structure and the upper line, calculated by the FS, has virtually constant slope: for the three years the values are 13.50, 13.29 and 12.57. Although the slope of the line for the majority of the data varies little over the years, the structure of the outliers is different. All are linear, but that for 2003

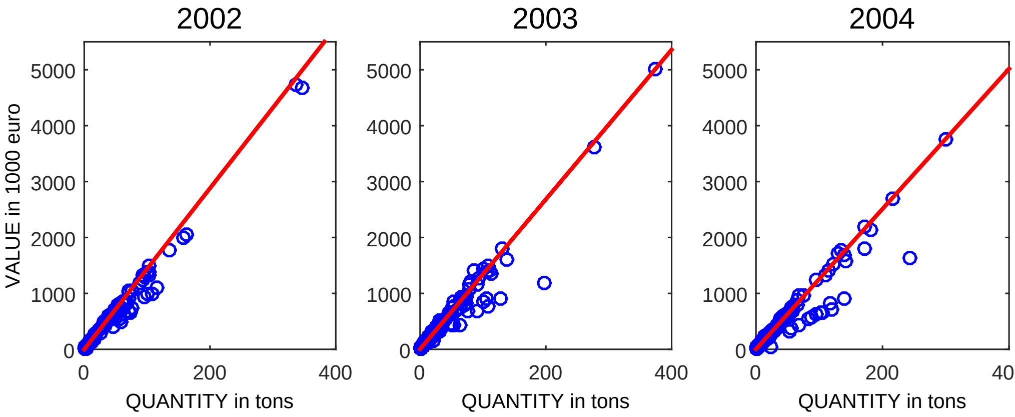

Figure 7: 2002-2004 data. Fair value. Annual regression lines from the FS. There are respectively 165, 167 and 168 observations

has the lowest slope, 5.880, with the others being 9.5875 and 6.7491. These are blatant departures from the fair value - a saving of 56% for 2003. The values of R^2 for these regressions, 92, 97 and 99% indicate that no attempt was made to disguise the fraud. However, our procedure is unaffected by the form of the outlying observations, depending on the their distance from the line of fair value, provided we successfully identify this. A surprising feature of the period over which the data were collected is the stability of the fair value. For goods for which this value is not so constant over years, a sector inflation (or deflation) factor can be used to adjust the value of the good before analysis. In other applications it has been found helpful also to have a moving window for the fair value, typically calculating it from data from no more than three consecutive years.

A strange feature of the trade data is the number of different forms encountered. The seafood data analysed in this paper have a relatively simple structure of two lines, a very few other outliers and an error distribution giving a large number of small observational errors. Data for other goods may have something of the same structure, perhaps with more outliers, but show appreciable heteroscedasticity, the variance increasing with the mean. The FS can also be used to provide heteroskedastic very robust regression, but currently without the incorporation of prior information.

Acknowledgements.

We are grateful to Domenico Perrotta, Francesca Torti and Andrea Cesara of the European Union Joint Research Centre in Ispra, Italy, for involving us in this problem, for providing the data and for information and discussions of this and related problems in fraud detection. Methods based on our approach are in routine use at the Centre; the estimates are disseminated to the Anti-Fraud Office of the European Union (OLAF) and the customs services of the member states. In particular we thank OLAF for support through the Hercule III programme and the AMT project series.

MR and ACA gratefully acknowledge support from the CRoNoS project, reference CRoNoS COST Action IC1408.

Routines for the robust Bayesian regression described here, and the data on seafood, are included in the FSDA toolbox downloadable from http://fsda.jrc.ec.europa.eu/ or http://www.riani.it/MATLAB.

References

Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer–Verlag, New York.

Atkinson, A. C., Riani, M., and Cerioli, A. (2010). The forward search: theory and data analysis (with discussion). *Journal of the Korean Statistical Society*, **39**, 117–134. doi:10.1016/j.jkss.2010.02.007.

Atkinson, A. C., Corbellini, A., and Riani, M. (2018). Robust Bayesian regression with the forward search: Theory and data analysis. *TEST*. (In Press).

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651–659.

Cook, R. D. and Hawkins, D. M. (1990). Comment on Rousseeuw and van Zomeren (1990). *Journal of the American Statistical Association*, **85**, 640–4.

Economist (2014). Uncontained. *Economist, UK Edn*, **411**(8885), 59–60.

Johansen, S. and Nielsen, B. (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, **43**, 321—348.

Koop, G. (2003). *Bayesian Econometrics.* Wiley, Chichester.

Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association*, **91**, 1219–1227.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods.* Wiley, Chichester.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications, 2nd edition.* Wiley, New York.

Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, **71**, 447–466.

Riani, M., Cerioli, A., Atkinson, A. C., and Perrotta, D. (2014). Monitoring robust regression. *Electronic Journal of Statistics*, **8**, 642–673.

Riani, M., Perrotta, D., and Cerioli, A. (2015). The forward search for very large datasets. *Journal of Statistical Software*, **67**(1), 1–20.

Riani, M., Corbellini, A., and Atkinson, A. C. (2018). On-line supplement for "The use of prior information in very robust regression for fraud detection". (Submitted).

Tallis, G. M. (1963). Elliptical and radial truncation in normal samples. *Annals of Mathematical Statistics*, **34**, 940–944.

On-line Supplement for "The Use of Prior Information in Very Robust Regression for Fraud Detection"

Marco Riani[*] and Aldo Corbellini[†]
Dipartimento di Economia, Università di Parma, Italy
Anthony C. Atkinson[‡] Department of Statistics
London School of Economics, London WC2A 2AE, UK

November 9, 2017

1 Other Analyses of the Data from 2002

1.1 Least Squares

A major argument both for the use of a Bonferroni correction to identify outliers and the inclusion of a threshold in the analysis was that much of the data lay virtually on a straight line with almost no error. We use a least squares regression to illustrate this property of the error distribution. Then we consider two robust estimation procedures recommended by Maronna *et al.* (2006) which are potential alternatives to our FS-based analysis. In this context, the diagnostic advantage of least squares compared to robust procedures is that it does not produce large residuals from well-behaved data.

[*] Email: mriani@unipr.it
[†] Email: aldo.corbellini@unipr.it
[‡] Email: a.c.atkinson@lse.ac.uk

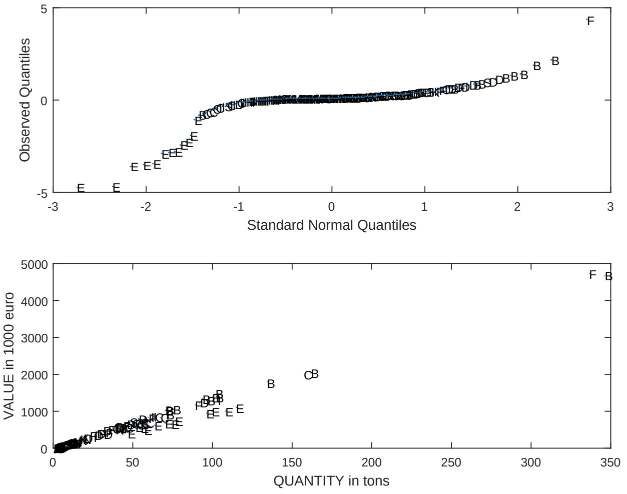

Figure 1: 2002 data, least squares analysis with arbitrary letters indicating member state. Upper panel, QQ-plot of standardized residuals, lower panel scatter plot of data

See Cook and Hawkins (1990) for an example in multivariate analysis where a robust method leads to "outliers everywhere".

The QQ-plot of standardized residuals for year 2002 in the upper panel of Figure 1, clearly shows this structure; the central part of the data (around 130 observations) is virtually horizontal. At the lower-end of the distribution there is what appears to be a clear set of outliers. The larger observations in the upper tail of the distribution are less obviously outlying, although they clearly come from a distribution with a larger standard deviation than the central observations.

Although the successful prosecution mentioned by Riani *et al.* (2018) led to the identification of European Union member state E, the data from the other member states remain unattributed; the symbols in the plot are marked with an arbitrary letter for each member state. The 12 most negative residuals all come from member E. The lower panel of Figure 1 shows a scatterplot of the data. The separation of the E group is clear, forming the lower line we have been continuously identifying. For the highest values, a single observation from member state F lies above a closely related observation from member state B. However, because the E group slightly reduces the slope of the overall least squares regression line, there is a large positive residual from F in the upper panel of the figure.

Although the least squares analysis has allowed us to interrogate two plots and discover part of the structure of the data, the significance of the results is not adequate for legal purposes. The upper panel of Figure 2 shows the plot of deletion residuals in observation order. The symbols reveal that the 12 observations from each member state are given together, in fact in time order. The E group give a set of negative residuals. There is also the positive residual we have noted for F. However, the Bonferroni bounds in the figure (to give an overall 1% test size for the sample) only reveal three outliers. The comparison with the results of the frequentist FS in the right-hand panel of Figure 2 of Riani *et al.* (2018) is revealing. There 15 outliers were revealed and the concern was that too many were being found. Here, the masking to which LS regression is subject shows how the residuals for the outliers have been rendered less extreme by use of a test in which the estimate of σ^2 is too large. The plot also exhibits the prevalence of small residuals.

Interesting insight into the structure of the data comes from the lower panel of the figure, which shows the quantities for each of the 165 transactions. Apart from the two large transactions, half a dozen member states account for nearly all the trade. The remaining observations are small (although not identically zero). It is these small transactions that give rise to the structure of the random variability, with many observations very close to the line fitted in the first part of the FS.

1.2 S Estimation

We now briefly describe the results of other robust analyses of the data from 2002 and compare them with that from the FS. Section 2 of Riani *et al.* (2014) summarizes the common structure and differences of the methods of very robust regression described in detail by Maronna *et al.* (2006). In the

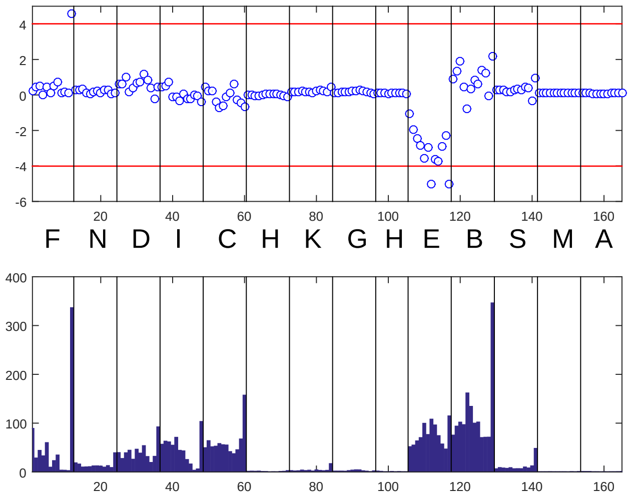

Figure 2: 2002 data, least squares analysis with member state indicators. Upper panel, index plot of deletion residuals with 99% Bonferroni bound; lower panel monthly value of imports

final section we sketch how prior information can be incorporated into these methods.

In least squares estimation, the value of $\hat{\beta}$ does not depend on the estimate of σ^2. The same is not true in M estimation and derived procedures in which observations with large residuals are downweighted by a function ρ, the extent of downweighting depending on the value of σ. In our calculations in this paper we take ρ as the Tukey biweight. Other choices could have been the hyperbolic or Hampel functions (Hampel *et al.*, 1981; Hoaglin *et al.*, 1983). S-estimates are a special case of M estimates introduced by Rousseeuw and Yohai (1984) in which the scale estimate is optimized for a specified breakdown point which cannot be less than 0.5.

First we analyse the data from 2002 with the breakdown point of the S estimator set to 0.25. When we use the simultaneous 99% confidence interval for outlier detection we obtain a pattern of outliers that is exactly identical to that from the frequentist FS. There are the same three outliers from the upper line and all the observations from country E. Of course, as we argued

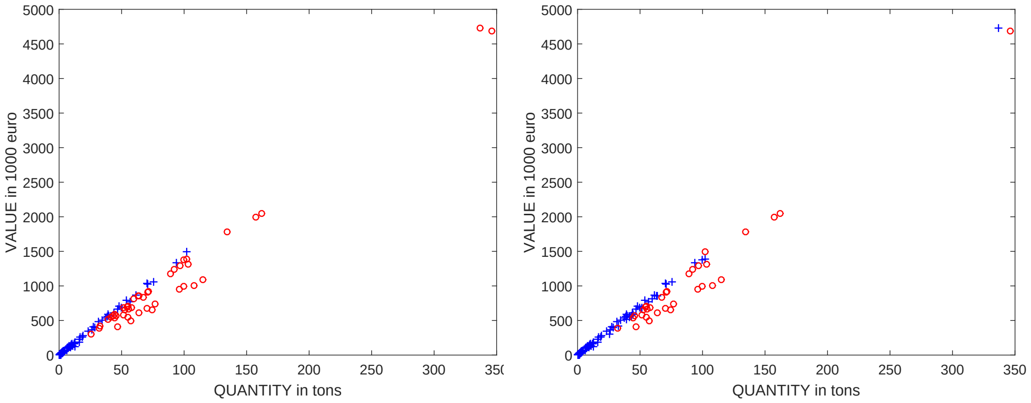

Figure 3: 2002 data, scatterplots with outliers (O). Left-hand panel, S estimator with breakdown point 0.5; right-hand panel, MM estimator with efficiency 0.95.

when introducing the threshold, this set of outliers is too large.

The breakdown point is a parameter to be chosen by the data analyst. We now repeat the analysis with a value of 0.5, the maximum value with physical meaning. There are now an amazing 44 outliers. The scatterplot in the left-hand panel of Figure 3 shows that the largest 5 observations from the main line are determined to be outlying as well as many more on that line and all those from country E and others lying just below the fitted line.

This behaviour is somewhat disturbing. When the breakdown point is 0.25 the analysis is in line with that of the FS, which adaptively chooses the amount of trimming in the light of the data and can, when appropriate, have a breakdown point of 0.5. However, 0.5 is not appropriate here; when the S-estimator is forced to have such a high breakdown point, outliers are, indeed, found everywhere. Figure 4 plots the number of outliers found as a function of breakdown point over the range 0.1 to 0.5. The number of outliers increases steadily with the breakdown point, with an abrupt change around a value of 0.38. These results are in line with those from the monitoring of different forms of robust regression in Riani *et al.* (2014) which show, for three examples, just how sensitive the S-estimate can be to the choice of breakdown point.

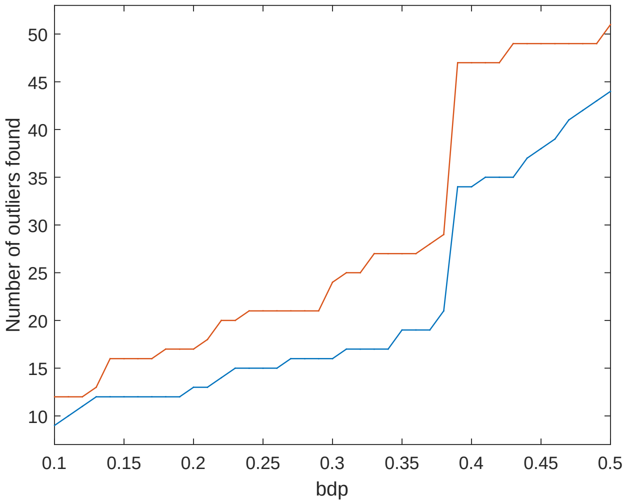

Figure 4: 2002 data, S estimation. Number of units declared as outliers as the breakdown point (bdp) varies. Upper curve, individual confidence level of 0.99, lower line simultaneous confidence level of 0.99

1.3 MM Estimation

Explicit asymptotic relationships between the breakdown point and efficiency of S estimators are presented by Riani *et al.* (2014, §2.2); as one increases, the other decreases. In an attempt to break out of this relationship, Yohai (1987) introduced MM estimation, which extends S estimation. In the first stage the breakdown point of the scale estimate is set at 0.5, thus providing a high breakdown point. This fixed estimate is then used in the estimation of β, which is chosen to have high efficiency.

Maronna *et al.* (2006, p. 126) suggest a value of 0.85 for the efficiency parameter in MM-estimation. When we combine this with the simultaneous interval for outlier testing we obtain 39 outliers. Increasing the efficiency to 0.9 gives 37 outliers, with a final increase to 0.95 indicating 32 outliers. The scatterplot of this last set of outliers is in the right-hand panel of Figure 3. The sets of outliers found by the two procedures plotted in the figure are, not surprisingly, very similar, given the similarity in the number found. A strange, although unimportant, difference is that MM estimation with an efficiency of 0.95 does not identify the large observation from F as an outlier.

References

Cook, R. D. and Hawkins, D. M. (1990). Comment on Rousseeuw and van Zomeren (1990). *Journal of the American Statistical Association*, **85**, 640–4.

Hampel, F. R., Rousseeuw, P. J., and Ronchetti, E. (1981). The change-of-variance curve and optimal redescending M-estimators. *Journal of the American Statistical Association*, **76**, 643–648.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester.

Riani, M., Cerioli, A., Atkinson, A. C., and Perrotta, D. (2014). Monitoring robust regression. *Electronic Journal of Statistics*, **8**, 642–673.

Riani, M., Corbellini, A., and Atkinson, A. C. (2018). The use of prior information in very robust regression for fraud detection. (Submitted).

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis: Lecture Notes in Statistics 26*, pages 256–272. Springer Verlag, New York.

Yohai, V. J. (1987). High breakdown-point and high efficiency estimates for regression. *The Annals of Statistics*, **15**, 642–656.