Electronic Journal of Statistics
Vol. 8 (2014) 646–677
ISSN: 1935-7524
DOI: 10.1214/14-EJS897

Monitoring robust regression

Marco Riani, Andrea Cerioli

Dipartimento di Economia, Università di Parma, Italy
e-mail: mriani@unipr.it; andrea.cerioli@unipr.it

Anthony C. Atkinson

The London School of Economics, London WC2A 2AE, UK
e-mail: a.c.atkinson@lse.ac.uk

and

Domenico Perrotta

European Commission, Joint Research Centre, Ispra, Italy
e-mail: domenico.perrotta@ec.europa.eu

Abstract: Robust methods are little applied (although much studied by statisticians). We monitor very robust regression by looking at the behaviour of residuals and test statistics as we smoothly change the robustness of parameter estimation from a breakdown point of 50% to non-robust least squares. The resulting procedure provides insight into the structure of the data including outliers and the presence of more than one population. Monitoring overcomes the hindrances to the routine adoption of robust methods, being informative about the choice between the various robust procedures. Methods tuned to give nominal high efficiency fail with our most complicated example. We find that the most informative analyses come from S estimates combined with Tukey's biweight or with the optimal ρ functions.

For our major example with 1,949 observations and 13 explanatory variables, we combine robust S estimation with regression using the forward search, so obtaining an understanding of the importance of individual observations, which is missing from standard robust procedures. We discover that the data come from two different populations. They also contain six outliers.

Our analyses are accompanied by numerous graphs. Algebraic results are contained in two appendices, the second of which provides useful new results on the absolute odd moments of elliptically truncated multivariate normal random variables.

MSC 2010 subject classifications: Primary 62J05, 62J20, 62G35; secondary 62P20.
Keywords and phrases: Forward search, graphical methods, least trimmed squares, outliers, regression diagnostics, rho function, S estimation, truncated normal distribution.

Received March 2014.

Contents

1. Introduction

Data rarely follow the simple models of mathematical statistics. Often, there will be distinct subsets of observations so that more than one model may be appropriate. Further, parameters may gradually change over time. In addition, there are often dispersed or grouped outliers. This distance between mathematical theory and data reality has led, over the last sixty years, to the development of a large body of work on robust statistics. By the time of Andrews *et al.* (1972) (the Princeton Robustness Study), according to Stigler (2010), it was expected that in the near future "any author of an applied article who did *not* use the robust alterative would be asked by the referee for an explanation". Now, a further forty years on, there does not seem to have been the foreseen breakthrough into the wider scientific universe. The purpose of our paper is to sketch what we see as some of the reasons for this failure and to suggest a system of interrogating robust analyses, which we call "monitoring", whereby we consider fits from very robust to highly efficient and follow what happens to aspects of the fitted model.

It has long been advocated that a very robust fit, asymptotically resistant to 50% of aberrant observations, be compared with a non-robust fit. For example, Rousseeuw and Leroy (1987), p. 111, present comparative index plots of least squares (LS) and least median of squares (LMS) residuals. Such comparisons are for two extreme forms of regression; the most and least robust. Hawkins and Olive (2002), Point 3, recommend using several estimators, both classical and robust. In our monitoring of robust regression we extend this suggestion by also including the intervening fits of intermediate robustness, monitoring

such quantities as residuals, parameter estimates and test statistics; we obtain information on the important changes in conclusions that come from differing assumptions about the degree of contamination in the data.

The procedure makes enthusiastic use of graphics. If the initial very robust fit and the final fit are thought of as providing two snapshots of the data, our monitoring can be thought of as providing a film of which the two snapshots are stills from the beginning and end of the film. The inclusion of outliers is usually signalled by a sudden change in residuals and a more gradual change in parameter estimates. Typically we require 50 robust regression fits per analysis; a computational burden only made possible by the efficiency of the FSDA robust library (Riani *et al.*, 2012) and by the recent technical advances of Riani *et al.* (2014b) described in §2.2, together with results in Appendix B.

Our monitoring of robust procedures has at least two consequences. One is that we produce insightful data analyses. The second is methodological. By considering a variety of procedures for robust regression, we are able to determine which are the critical parameters in determining the properties of the robust fit, distinguishing them from those that are only of secondary importance. We are thus able to provide comparatively simple prescriptions for robust regression.

Our paper is structured as follows. Section 2 introduces three classes of robust regression estimators and presents properties including the breakdown point and efficiency of estimation. The important family of soft trimming estimators, leading to downweighting of observations by a function ρ, derives from M estimation described in §2.2. The derived methods, S, MM and τ estimators, which differ in the way in which the error variance is estimated are the subject of the remainder of §2. In §3 we discuss choice of an appropriate form of robust regression, difficulties in numerical procedures and the interpretation of estimated parameters; we focus on standard errors of estimated regression coefficients. A second problem of interpretation is that of the effect of individual observations on inferences, which can be provided by the forward search (Riani *et al.*, 2014c).

Examples of the application of monitoring are in §4. For each set of data we explore monitoring for a total of 20 combinations of ρ function and estimation procedure. For each combination we look at the fitted regression for 50 values of breakdown point or efficiency, depending on which is more easily specified for the specific method. We also monitor the behaviour of hard trimming methods including least trimmed squares and the forward search. The first example, 'Stars data' has a simple structure with one explanatory variable and four outliers, well separated from the main body of the data. Even with this simple problem, we detect some differences in the performance of the methods. These become more pronounced as we move to more complicated examples. For the bank data, analysed in §4.4, there are 1,949 observations and 13 explanatory variables. Through the use of monitoring, this example very nicely illustrates the similarities and distinctions between the various forms of robust regression.

The conclusions from our exploration of monitoring are in §5. We recommend S estimation with either Tukey's bisquare or the optimal ρ function. Insight into the relationship between individual observations and inferences is best provided by the forward search. The first of three appendices describes the four ρ

functions that we use: Tukey's bisquare, optimal, hyperbolic and Hampel. The second provides algebra for application of the Hampel ρ function that avoids the necessity for the customary numerical integration. These new results render straightforward the routine application of this ρ function in data analysis.

2. Robust regression

We work with the customary regression model in which the n response variables y_i are related to the values of a set of p explanatory variables x by the relationship

$$y_i = \beta^T x_i + \epsilon_i \qquad i = 1, \ldots, n, \tag{2.1}$$

including an intercept term. The independent errors ϵ_i have constant variance σ^2.

2.1. Three classes of robust estimator and some properties

It is helpful to divide methods of robust regression into three classes.

1. Hard (0,1) Trimming. In Least Trimmed Squares (LTS: Hampel, 1975, Rousseeuw, 1984) the amount of trimming is determined by the choice of the trimming parameter h, $[n/2] + [(p+1)/2] \leq h \leq n$, which is specified in advance. The LTS estimate is intended to minimize the sum of squares of the residuals of h observations. For LS, $h = n$. In the generalization of Least Median of Squares (LMS, Rousseeuw, 1984) that we monitor, the estimate minimizes the median of h squared residuals. Comments on numerical matters associated with these optimizations are in §3.
2. Adaptive Hard Trimming. In the Forward Search (FS), the observations are again hard trimmed, but the value of h is determined by the data, being found adaptively by the search. Data analysis starts from a very robust fit to a few, carefully selected, observations found by LMS or LTS with the minimum value of h. The number of observations used in fitting then increases until all are included. (See Atkinson and Riani, 2000 and Riani *et al.*, 2014c for regression, Atkinson *et al.*, 2010 for a general survey of the FS, with discussion, and Cerioli *et al.*, 2014 for results on consistency).
3. Soft trimming (downweighting). M estimation and derived methods. The intention is that observations near the centre of the distribution retain their value, but the ρ function ensures that increasingly remote observations have a weight that decreases with distance from the centre.

We shall consider all three classes of estimator. The FS by its nature provides a series of decreasingly robust fits which we monitor for outliers in order to determine how to increment the subset of observations used in fitting. In this paper we place particular emphasis on monitoring the soft trimming estimators for four ρ functions, for which special numerical techniques are necessary (§2.2). Four properties of the estimators are of importance. Here we give the values for hard trimming. The extension to S estimation is in §2.2.

1. Breakdown point, bdp; the asymptotic proportion of observations that can go to ∞ without affecting the parameter estimates; bdp$= 1 - h/n$. We stress that this definition requires both that $n \to \infty$ and that the contaminating observations also become increasingly remote.
   As a result of monitoring we observe an *empirical breakdown point*, the point at which the fit switches from being robust to non-robust least squares. This important property depends both on the nominal properties of the estimator and on the particular data set being analysed.
2. The consistency factor K_σ required to rescale the estimate of σ^2. Let the estimator of σ^2 from the residual sum of squares of the central h observations be $\hat{\sigma}_h^2$. Since the sum of squares contains only the central h observations from a normal sample, the estimate needs scaling. Let K_σ be the ratio of the variance of the truncated normal distribution containing the central h/n portion of the full distribution to the variance of the untruncated distribution (see Croux and Rousseeuw (1992), equation (6.5), or the results of Tallis (1963) on elliptical truncation). To estimate σ^2 we accordingly take
   $$\tilde{\sigma}^2 = \hat{\sigma}_h^2 / K_\sigma. \tag{2.2}$$
   As $h \to n, K_\sigma \to 1$.
3. The efficiency of estimation. For normally distributed responses with explanatory variables x that follow some multivariate distribution, let the robust estimator of the parameter β_j of the linear model be $\tilde{\beta}_j$, with $\hat{\beta}_j$ the full-sample least squares estimator. If the observations for $\tilde{\beta}_j$ are selected at random (and so have the same distribution for x), the asymptotic efficiency of estimation of least squares relative to full-sample least squares is Eff $= \text{var}\hat{\beta}_j / \text{var}\tilde{\beta}_j = h/n$. For the trimmed observations used in robust estimation, the efficiency can be much less than this.
4. The asymptotic variance of any element of $\tilde{\beta}$ relative to least squares is the reciprocal of the efficiency Eff.

For hard trimming, once one of the values, for example the breakdown point bdp, has been selected, the other three follow. In the next section we present related results for S estimators.

2.2. M and S estimation

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.

Suppose σ is known and let the residuals for some estimate b of β be $r_i = y_i - b^T x_i$. Then the regression M-estimate of β is the value that minimizes the objective function

$$\sum_{i=1}^{n} \rho\{r_i(\beta)/\sigma\}, \tag{2.3}$$

where ρ is a function with properties given below that reduces the importance of observations with large residuals.

For robust M estimation, σ should also be estimated robustly. The M-estimator of scale $\tilde{\sigma}_M$ is found by solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i}{\sigma}\right)=\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i-\beta^T x_i}{\sigma}\right)=K, \tag{2.4}$$

in theory solved among all $(\beta,\sigma)\in\Re^p\times(0,\infty)$, where $0<K<\sup\rho$ (but see §3). If we take the minimum value of $\tilde{\sigma}_M$ which satisfies equation (2.4), we obtain the S-estimate of scale ($\tilde{\sigma}_S$) and the associated estimate of the vector of regression coefficients (Rousseeuw and Yohai, 1984). The estimator of β is called an S-estimator because it is derived from a scale statistic, although in an implicit way.

We now consider the properties of the class of functions ρ that we use. Rousseeuw and Leroy (1987), p. 139 show that if ρ satisfies the following conditions:

1. It is symmetric and continuously differentiable, and $\rho(0)=0$;
2. There exists a $c>0$ such that ρ is strictly increasing on $[0,c]$ and constant on $[c,\infty)$;
3. It is such that $K/\rho(c)=\text{bdp}, \text{with } 0<\text{bdp}\leq 0.5$,

(2.5)

the asymptotic breakdown point of the S-estimator tends to bdp when $n\to\infty$. As c increases, fewer observations are downweighted, so that the estimate of σ^2 approaches that for least squares and bdp $\to 0$. For consistency when the errors are normally distributed, we require

$$K=E_{\Phi_{0,1}}\left[\rho\left(\frac{r_i}{s}\right)\right] \tag{2.6}$$

where $\Phi_{0,1}$ is the cdf of the standard normal distribution.

It is however customary to rescale ρ (for example, Maronna *et al.*, 2006, p. 31). If $\rho(x)$ is normalized in such a way that $\rho(c)=1$, the constant K becomes the breakdown point of the S-estimator. If we fix bdp it follows from (2.5) and (2.6) that c and K are determined. The exact relationship will depend upon the function ρ. The four ρ functions that we use are described in Appendix A.

We monitor S estimators by looking over a grid of values of bdp. Riani *et al.* (2014b), §3.1, give computationally efficient calculations for finding the value of c for Tukey's bisquare once the value of bdp is specified. The calculations depend on the polynomial nature of the ρ function and require moments of truncated chi-squared random variables. For MM estimators we instead monitor efficiency. The calculations to find c again rely on expectations of truncated chi-squared variable and are given in their §3.2. The extension to the optimal loss function is given in their §7 – the calculations are similar to those for Tukey's bisquare since the ρ function is again of a polynomial form. We use numerical integration for the hyperbolic ρ function. New results on the absolute odd moments of the multivariate normal distribution under elliptical truncation, presented in Appendix B, allow us to avoid numerical integration for any ρ function that is a polynomial in absolute values of its argument. We apply the results to calculation of Hampel's ρ function.

An important final point is that the ρ functions for the mean in (2.3) and (2.5) may be different. However, in our numerical calculations for all estimators where such a choice exists, we use the same ρ for both the mean and the scale estimators.

2.3. MM and Tau estimation

The results of §2.2 establish an asymptotic relationship between the breakdown point and efficiency of S estimators; 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 for which K can be chosen to provide an estimator of β with a high efficiency. Maronna *et al.* (2006), p. 126, recommend a value of 0.85 for this efficiency, but, when we monitor MM estimates, we of course look over a range of values.

The final estimators we consider are an extension of S and MM introduced by Yohai and Zamar (1988). In τ estimation, unlike MM estimation, there is no global precalculated estimate of scale. Both β and σ are iteratively and alternatively estimated. In the general procedure the function ρ_0 used to estimate scale is chosen to give the maximum breakdown point for regression estimates. On the other hand the function ρ_1 used for estimation of β is chosen to give high efficiency. Sometimes a value as high as 0.95 is suggested. It is important to stress that these are asymptotic values for extremely well separated data; less fortunate forms of data can give rise, for example, to biased estimates. Although we use the same functional form for ρ_0 and ρ_1, the constants are chosen to give a range of breakdown points, over which we monitor the estimates for three values of efficiency.

3. Problems of formulation, of calculation and of interpretation

3.1. Formulation and calculation

A major disincentive to the routine use of standard robust methods is the number of decisions that have to be made before the analysis of the data begins. We now describe some of these.

The most difficult problem is often specification of the desired efficiency or, equivalently, breakdown point. Less formally, this is asking what proportion of outliers are expected in the particular set of data being analysed. The second is the nature of robust estimator that is required – in the regression case the choice between the four forms of estimator described in the previous section, together with the hard trimming methods. The third choice is that of the ρ function. We show how the monitoring proposed in our paper makes many of these decisions redundant and illuminates which remaining ones are important. However two further groups of problems remain.

The second group are those of calculation. The functions to be maximized when using any of these robust estimators are complicated, with many local maxima. In consequence, approximate methods are used. These are typically based on randomly sampled subsets of p observations (elemental sets) to which the model is fitted exactly. The fitted values are then used to evaluate the function to be maximized, perhaps after some refining iterations (concentration steps). In these the ρ function is used to evaluate weights for the n residuals at each iteration which are used to provide a new parameter estimate and so a new set of weights. There are several details which need to be decided. The conceptually simplest is the number of subsamples to extract and the number of concentration steps in each subsample. Our implementation for this paper follows the recommendations of the FSDA toolbox.

The final group of problems are statistical in nature. One is loss of the simplicity of distribution theory associated with least squares estimation and related tests. We illustrate this point in §3.2 for the t-tests for the parameters in the fitted linear model.

The other point is of extreme statistical importance, namely that the researcher loses information on the effect that each unit, outlier or not, has on the final proposed estimate. Although this is not a problem in the numerical application of robust methods, it can be seen as a scientific limitation. We endeavor to overcome this in our Example 3 by incorporating some information on the effect of individual units from the FS.

3.2. Robust standard errors

This section summarises results on the robust analogues of normal-theory t-statistics for the parameters in a linear model which we need for our numerical results in §§4.3 and 4.4. Under suitable regularity conditions (see, for example, Maronna *et al.*, 2006, §10.3), M estimates are asymptotically normal, thereby allowing for Wald-type tests and confidence intervals. The asymptotic covariance matrix of $\hat{\beta}_M$, the M estimator of the regression vector β, can be written as a product of three terms

$$cov(\hat{\beta}_M) = \sigma^2 \gamma V_X^{-1},$$

where σ is the scale parameter, V_X is a square symmetric matrix and γ is a correction factor which depends on the particular function $\psi(x) = \rho'(x)$.

The correction factor γ (Maronna *et al.*, 2006, p. 100) is given by $E(\psi^2)/(E(\psi'))^2$, estimated by

$$\hat{\gamma} = \frac{1}{n-p} \sum_{i=1}^{n} \psi\left(\frac{r_i}{\hat{\sigma}}\right)^2 \Bigg/ \left[\frac{1}{n} \sum_{i=1}^{n} \psi'\left(\frac{r_i}{\hat{\sigma}}\right)\right]^2 .$$

The factor $1/(n-p)$ is used instead of n in order to obtain the classical formula when $\psi(x) = x$ and $V_X = X^T X$, corresponding to LS.

Huber and Ronchetti (2009), §7.6, suggest three expressions to estimate V_X. The one which is most used in the literature is

$$\hat{V}_X = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \hat{w}_i} X^T \hat{W} X, \tag{3.1}$$

where $w_i = w(r_i/\hat{\sigma}) = \hat{\sigma}\psi(r_i/\hat{\sigma})/r_i$ and $W = diag(\hat{w}_1, \ldots, \hat{w}_n)$. Under the assumptions of a symmetric error distribution, a symmetric ρ function and a matrix X with all leverages equal to p/n, Huber (1973) showed that $\hat{\gamma}(X^TX)^{-1}$ contains a bias of order $O(p/n)$ and derived a correction $\hat{K}^2$ that makes $\hat{\gamma}(X^TX)^{-1}$ unbiased up to terms of order $O(p^2/n^2)$. This correction is

$$\hat{K}^2 = \left[1 + p \sum_{i=1}^{n} \left\{ \psi'\left(\frac{r_i}{\hat{\sigma}}\right) - \overline{\psi'}\left(\frac{r_i}{\hat{\sigma}}\right) \right\}^2 \Bigg/ \left\{ \sum_{i=1}^{n} \psi'\left(\frac{r_i}{\hat{\sigma}}\right) \right\}^2 \right]^2,$$

where

$$\overline{\psi'}\left(\frac{r_i}{\hat{\sigma}}\right) = \sum_{i=1}^{n} \psi'\left(\frac{r_i}{\hat{\sigma}}\right)/n.$$

Although this correction is only valid when $V_X = X^TX$, it is standard in the robust literature (see for example Koller and Stahel, 2011), to use it in combination with $\hat{V}_X$ defined as in (3.1).

Of course, all leverages will only be equal to p/n for some particular designed experiments satisfying D-optimality. Even if we have such data, we will be looking at subsets of observations and the condition will not hold. Given that, as Huber and Ronchetti (2009), p. 161 admit, there is no general agreement on the "desirability of safeguarding against leverage points in an automatic fashion". We use both

$$\hat{\sigma}^2 \hat{\gamma} (X^T X)^{-1} \tag{3.2}$$

and

$$\hat{\sigma}^2 \hat{\gamma} \hat{K}^2 \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \hat{w}_i} (X^T \hat{W} X)^{-1} \tag{3.3}$$

in the calculation of robust t-statistics.

We are here concerned with robustness against outlying observations. Croux *et al.* (2004) extend the discussion to t-statistics which are also robust against correlation and heteroskedasticity of the errors in (2.1).

4. Examples

We illustrate the use of monitoring with three examples of increasing complexity. For effective monitoring we require a method that moves from a very robust fit to least squares. Although all methods have such properties asymptotically, the comparisons of Riani *et al.* (2014c) show that the finite sample properties of the various methods vary depending on the distance between the main body of the data and the contamination. In particular, our results suggest we need to

avoid methods which are tuned to have a very high efficiency for the parameters of the linear model but which are liable to failure unless the contamination is extremely remote.

4.1. Correlation

In monitoring S estimators we vary the bdp from 0.5 to 0.01, as we do for τ estimates, but now for three values of efficiency. For MM estimates it is more convenient to monitor changes as the efficiency goes from 0.5 to 0.99. In all cases we look at plots of residuals. For simple structures, like our first example, there is a clear division of the solutions into a robust fit and a non-robust one, with a sharp break between them. For more complicated examples the point of transition is not so clearly visible. But in all cases we find that the structure of the plot is well summarized by the correlation of the ranks between the residuals at adjacent monitoring values. We consider three standard measures of correlation:

1. Spearman. The correlations between the ranks of the two sets of observations.
2. Kendall. Concordance of the pairs of ranks.
3. Pearson. Product-moment correlation coefficient (see Stigler (1989) for an account of Galton's contribution).

4.2. Example 1: Stars data

We start with an easy to understand data example that, nevertheless, presents some of the main points of our argument. The data are taken from Rousseeuw and Leroy (1987), p. 27 and form part of a Hertzsprung-Russell diagram of stars. This log-log plot has the effective surface temperature of the star as the explanatory variable and (logged) light intensity as the response. A typical plot has around 30,000 stars which fall into groups including "the main sequence", "white dwarves" and "giants" of several kinds. However, in our example, there are only 47 observations. Since there is only one explanatory variable, the structure is obvious on inspection.

These data have the canonical structure against which the methods of very robust regression were developed. Fig. 1 is a plot of the data. Note that we have plotted temperature values from low to high, which is the reverse of the standard diagram in astronomy. There are 41 observations which plausibly lie on the same regression line, two relatively close outliers, observations 7 and 9 and a cluster of four outliers, observations 11, 20, 30 and 34. This structure of the main sequence and giants is well established. Included in the figure are five linear fits: least squares, which is attracted towards the cluster of outliers, and four robust fits that fit predominantly to the main group. The steepest line is LTS, followed by S and then MM with efficiencies of 0.85 and 0.95. The higher the desired efficiency, the closer is the fit to least squares. There are thus

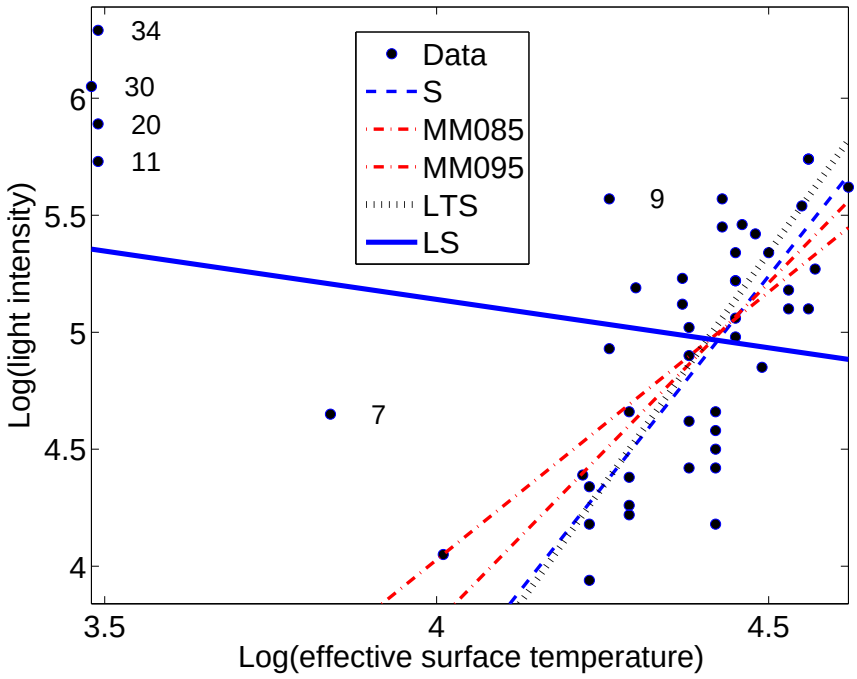

FIG 1. *Stars data: scatterplot and fitted regression lines: LS, least squares, which is attracted towards the cluster of outliers and four robust fits. The steepest line is LTS, followed by S and then MM with efficiencies of 0.85 and 0.95. The higher the desired efficiency, the closer is the fit to least squares.*

two very different groups of fitted lines, those from high breakdown estimators and that from a procedure with zero breakdown. Our monitoring plots each consider a single estimator as the coefficients are changed. Virtually all show a robust and a non-robust fit as the two extremes. The interesting comparison is at what empirical breakdown point this transition occurs. For hard trimming with six outliers out of 47 observations, we would hope to achieve the minimum breakdown point, for these data, of 12.8%, with correspondingly high efficiency. Methods that monitoring shows have a higher breakdown point require a more robust fit to reveal the data structure. We can expect that, for data with greater contamination, they may fail to reveal any contamination whatsoever.

Figure 2 shows a typical monitoring plot, in this case for S estimation. This is generated by a series of robust fits, starting from a breakdown point of 0.5 and decrementing the value by 0.01 up to 0.01. There are therefore 50 robust fits leading to the plot of scaled residuals in the figure, for which, in this case, we used Tukey's bisquare. The plot of scaled residuals for high breakdown fits clearly shows three sets of residuals: four very large (units 11, 20, 30 and 34), two intermediate (units 7 and 9) and the cluster of the remainder. There is some slight decrease in the values around a breakdown point of 0.2 and then an

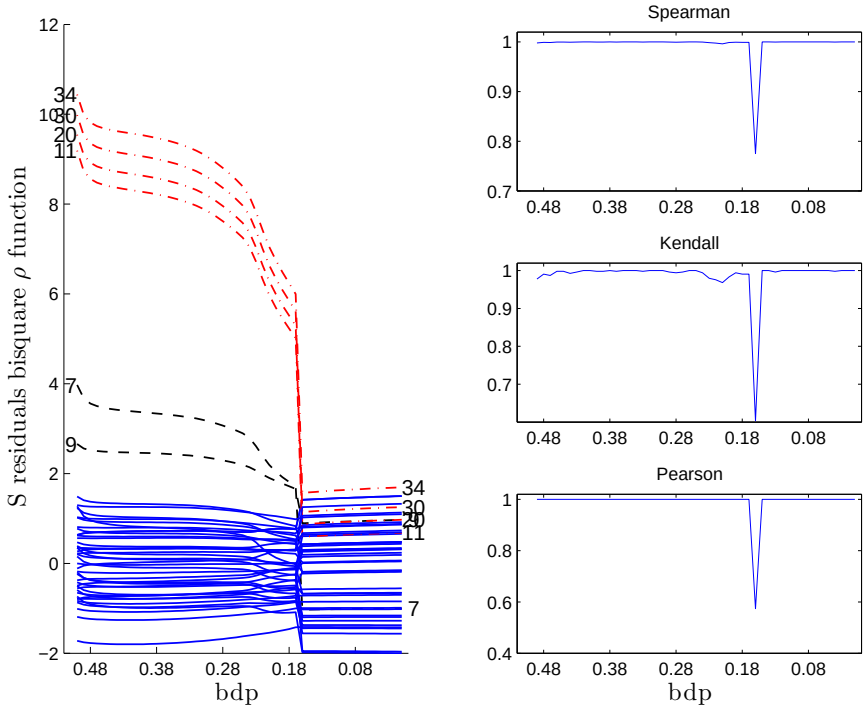

FIG 2. *Stars Data: S estimation, Tukey's bisquare ρ function. Left-hand panel, plot of scaled residuals. Three sets of residuals (the cluster of outliers, the two intermediate outliers and the main cluster of the data) are clearly seen for bdp values down to around 0.2. Right-hand panel, three measures of the correlations of adjacent residuals. The abrupt switch to LS at 0.16 is evident.*

abrupt change at 0.16 when the fit switches to least squares. The exact point of this change shows clearly in the right hand panel, the plot of correlation values.

For MM estimation we first find σ^2, with a breakdown point of 0.5, and then increase the efficiency of estimation of β from 0.5 to 0.99 with an increment of 0.01, the estimate of σ^2 remaining fixed. The results in Figure 3 show that the very robust fit predominates, the change to least squares occuring at an efficiency of 0.99. However, from Table 2 of Riani *et al.* (2014b), an efficiency of 0.99 corresponds to, for Tukey's bisquare, a breakdown point of 0.0570. The minimum value of the plot of correlations is in the last step when we compute the correlation between the residuals with eff = 0.98 and those for eff = 0.99.

We explored the τ residuals for three somewhat high values of efficiency of estimation of β: 0.85, 0.9 and 0.95. The plots are similar to those already given, except that the breakdown point associated to the step immediately before the change from very robust to least squares fits occurs increases from 0.14 to 0.17 and then 0.26. As the required efficiency of estimation of β is increased, the bdp of the procedure is reduced.

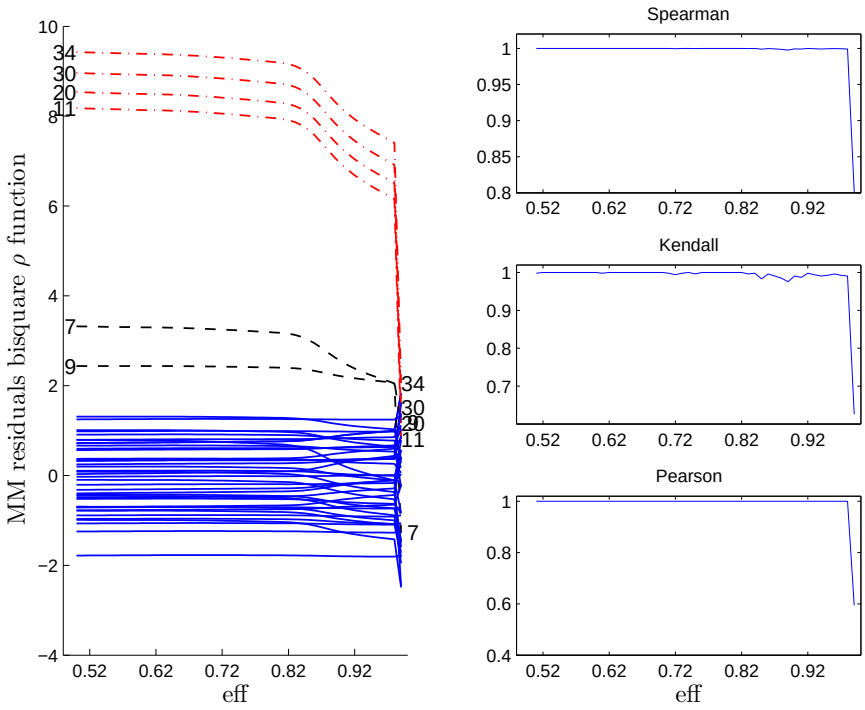

FIG 3. *Stars Data: MM estimation, Tukey's bisquare ρ function. Left-hand panel, plot of scaled residuals. The three sets of residuals are clearly seen for virtually all efficiency values. Right-hand panel, three measures of the correlations of adjacent residuals. The abrupt switch to LS is at eff* $= 0.99$*, corresponding to a bdp of 0.057.*

All of these robust procedures were also used with the other three ρ functions: hyperbolic, optimal and Hampel. For this example the choice is mostly not critical, apart from τ estimation. For S estimation with all three ρ functions the change in the plots is at a breakdown point of 0.17. For MM estimation with the Hampel function the first step associated to non-robust estimation is at an efficiency of 0.97 and at 0.98 for the hyperbolic. Although, for the optimal function there is a change at 0.93 causing a rescaling of the residuals, the three groups found by robust estimation remain over the whole range up to an efficiency of 0.99. Only τ estimation is somewhat sensitive to the form of the function.

For the hyperbolic and optimal ρ functions, the behaviour of the τ estimate is similar to that for the bisquare, as it is with the Hampel function and $\tau = 0.85$ or 0.9. However, with Hampel's ρ function, the method completely breaks down when the efficiency is set at 0.95, producing a uniform plot of least squares residuals. These results are summarised in Table 1.

It is also interesting to look at the plots of residuals that arise from LTS, LMS and the FS, but we leave this until we consider the more complicated structure of Example 2.

TABLE 1
Stars data. Empirical breakdown point (bdp) or efficiency (eff) for MM: five estimators and four ρ functions. The values are for the step before the switch to a non-robust fit

Estimator		Bisquare	Optimal	Hyperbolic	Hampel
S	bdp	0.17	0.17	0.17	0.17
$\tau = 0.85$	bdp	0.14	0.14	0.14	0.16
$\tau = 0.90$	bdp	0.17	0.16	0.16	0.20
$\tau = 0.95$	bdp	0.26	0.21	0.24	—[a]
MM	eff	0.98	0.99	0.97	0.96

[a] For this combination of τ and ρ function, only non-robust solutions were obtained during monitoring.

4.3. Example 2: AR 2000 data

In these data, introduced by Atkinson and Riani (2000), §1.2.2, there are three explanatory variables and 60 observations, with a structure of six masked outliers. There is no evidence of this structure in the scatterplot of the data in their Figure 1.5 and this is an example where LS shows little, apart from the slightly anomalous observation 43, which is not one of the six. The minimum breakdown we can expect from hard trimming is 6/60, i.e. 0.1. We again look at the structure of residuals during monitoring, but augment this with plots of the t-statistics for the three explanatory variables.

We start with Figure 4. This time we show the plot of S residuals for the optimal ρ function. Again there is a clear division of the plot into an initial, very robust, region which, at a breakdown point of 0.27 becomes close to the least squares fit. In the initial part of the plot the six remote outliers are clearly visible. However, there is also interesting fine detail in the plot.

For high values of the breakdown point, greater than 0.34, there are four groups of observations, with observations 7 and 39 having the most negative residuals. At a bdp of 0.34 the two central groups coalesce and observation 43 becomes as outlying as observations 7 and 39. At the next transition, the outliers with the largest positive and negative residuals form a single group, with observation 43 outlying. This structure remains stable for lower values of bdp.

So far we have not reported results for the hyperbolic and Hampel's ρ functions in any detail. Figure 5 shows monitoring plots of S residuals from these functions. The two plots are similar, if not identical. Compared with the plot for the optimal ρ function in Figure 4, they only show one abrupt transition, that is from four groups of observations to one plus a mild outlier, although observation 43 does become increasingly remote before this transition. However, either plot would serve to signal the difference between the very robust and non-robust fits.

Plots for the MM estimator and all ρ functions are also of this kind, with the optimal ρ function again showing two transitions. As with Figure 3, for MM estimation in the stars data, the transition to a non-robust fit occurs at a high efficiency, although not so high as in that figure. In the case of many τ estimators, the plots only show the least squares fit. We accordingly summarise these results in Table 2. The main conclusions are that the form of ρ function is unimportant for the S estimator. But the choice of ρ in this example, does seem

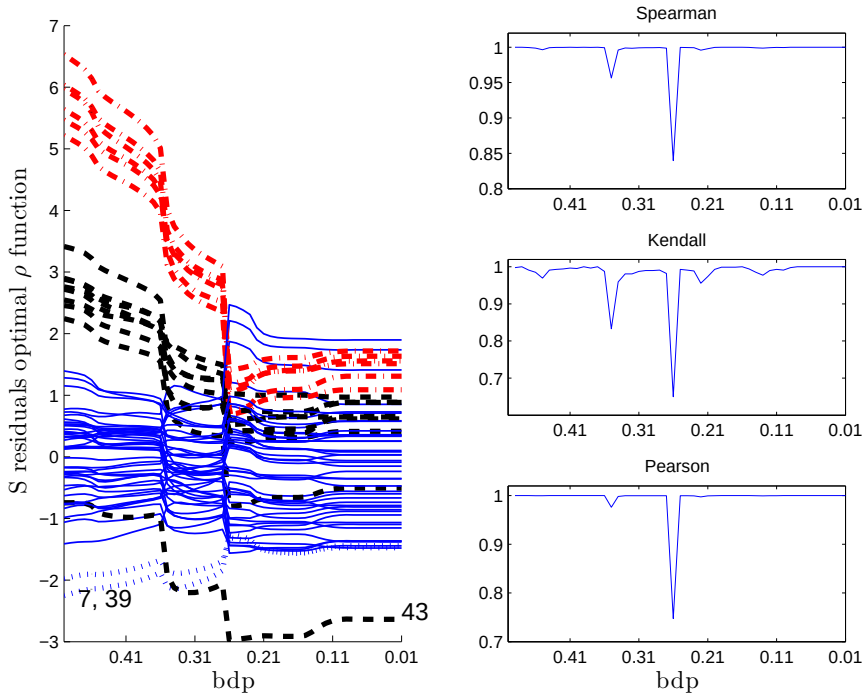

FIG 4. *AR Data: S estimation, optimal ρ function. Left-hand panel, plot of scaled residuals. Four groups of observations are evident for a bdp greater than 0.34 and three until the bdp is 0.27. Thereafter the plot shows the LS fit with one outlier. Right-hand panel, three measures of the correlation. Kendall's measure, in particular, shows the two transitions during monitoring.*

to have some effect on the performance of the MM estimator. However, if the purpose of the analysis is to establish different possible structures for the data, which are then to be further examined, the choice of ρ function is not crucial for these data when MM estimation is used. It is the results for the τ estimator that cause concern. For an efficiency of 0.85, the breakdown point is around 0.4. For higher values of efficiency it is either closer to 0.5, or so high that we only see a non-robust fit to the data. Use of the value of 0.95 for τ would lead to monitoring which gave no indication of any departure from the least squares model.

We now turn to hard trimming procedures. The residuals for LTS are in Figure 6. As h/n increases from just above 0.5 to 0.99 and the bdp correspondingly reduces, the residuals gradually decrease in magnitude as the number of observations used in fitting increases. This contrasts with the plots for the robust analysis of the stars data, when the plots are sensibly constant within sector. For the AR data, Figures 4 and 5, the plots are also constant in the last region of monitoring, although, particularly with the optimal ρ function, there is a decrease of the residuals for high bdp. The plot of residuals in Figure 6 shows all the detail of groups and their coalescence evident in Figure 4, together with

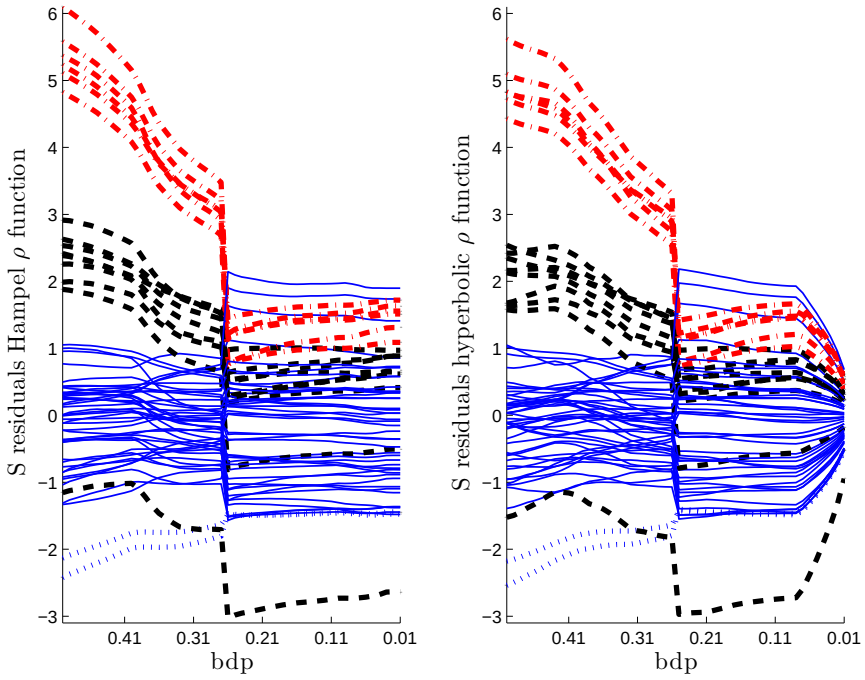

FIG 5. *AR Data: S estimation, plots of scaled residuals. Left-hand panel, Hampel's ρ function. Right-hand panel, hyperbolic ρ function. To be compared with Figure 4.*

TABLE 2

Atkinson and Riani (2000) Data. Empirical breakdown point (bdp), or efficiency (eff) for MM, during monitoring for the transition between very robust and least squares regression: five estimators and four ρ functions. The values are for the step before the switch to a non-robust fit

Estimator		Bisquare	Optimal	Hyperbolic	Hampel
S	bdp	0.27	0.27	0.26	0.27
$\tau = 0.85$	bdp	0.38	0.40	0.41	0.41
$\tau = 0.90$	bdp	0.45	0.48	—[a]	0.50
$\tau = 0.95$	bdp	—[a]	—[a]	—[a]	—[a]
MM	eff	0.91	0.97	0.90	0.87

[a] For these values of τ and ρ function, only non-robust solutions were obtained during monitoring.

a division of the main group for a bdp greater than 0.41. The six large residuals are clearly visible throughout, becoming included in the fit from a breakdown point of 0.09. This feature is clearly shown in the plots of correlation functions. (Recall that there are six major outliers in the 60 observations).

In LMS the median of asymptotically half the squared residuals is used as an estimate, rather than the sum of squares of the residuals as in LTS. To provide a generalization of LMS suitable for monitoring we vary the efficiency of estimation by minimizing the $100h/n$ percentage point of the distribution of

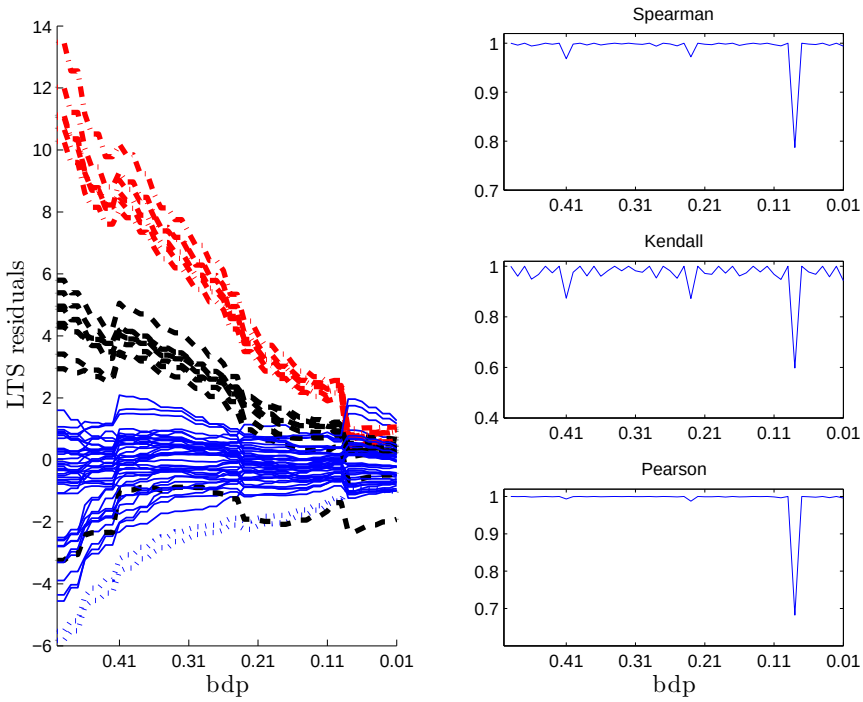

FIG 6. *AR Data: LTS estimation. Left-hand panel, plot of scaled residuals. The four groups of observations are evident for a bdp greater than 0.23, with the main group further split above 0.41. The groups then gradually move closer together, with the change to one group plus one outlier for a bdp of 0.09. Right-hand panel, three measures of the correlation. All show the final transition during monitoring.*

the squared residuals. The general structure of the plot in Figure 7 is similar to that of Figure 6 from a bdp of 0.41, showing the four groups but with much greater variation, reflecting an extension of the slow $n^{-1/3}$ rate of convergence of LMS (Rousseeuw and Leroy, 1987, p. 178). The transition to a non-robust fit is not clear in the plot of residuals, although it is indicated by the plots of correlation.

Finally we consider FS, a forward plot of the residuals for which is given in Figure 8. Now the abscissa is the number of observations m in the subset used for fitting. The scaled residuals in this plot tell a similar story to those from monitoring LTS for a bdp below 0.42 – chiefly that there are four groups, that observations 7, 39 and 43 behave differently and that there are six appreciable outliers from robust fits. In both LTS and FS the six outliers remain evident until near the end of monitoring, giving estimates with higher efficiency than those for the bdp of 0.27 with S estimation in Figure 4.

These three plots are different from those for the S estimates and variations shown in the earlier plots. One difference is that they are less stable, reflecting the effect of individual observations that are smoothed by mostly having zero

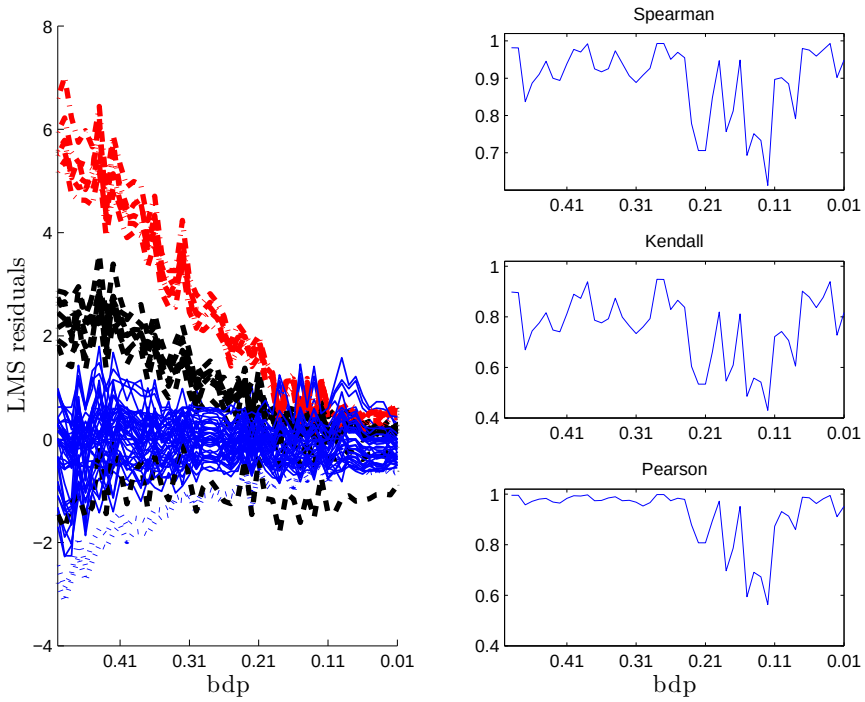

FIG 7. *AR Data: LMS estimation. Left-hand panel, plot of scaled residuals. The four groups of observations are evident for the highest values of bdp but gradually converge. The curves are much rougher than in the other plots. Right-hand panel, three measures of the correlation. All show some transition around a bdp of 0.12.*

or one weights in the other robust procedures. The second is that these three plots decline as efficiency increases. This effect is less marked in the case of FS. Since the residuals are scaled by the square root of the final estimate of σ^2 the differences in the plots reflect the changes in the estimates of β as monitoring evolves. These remain constant over long periods for the estimators with flexible trimming as the weights from the ρ function remain constant. However for LMS, LTS and the FS the estimates change for each new observation included in the fit.

Although the plots of the residuals for the groups of methods can be very different, the plots of t-statistics from monitoring and the FS are similar. Figure 9 shows that the very robust fit, in this case S estimation with the optimal ρ function, finds significant effects, in size order, for x_2, x_3 and x_1. However, the fits with low bdp show that x_1 is hardly, or not at all, significant, the values lying around -1.96, the lower limit of the 95% asymptotic confidence interval. Now x_3 is more significant that x_2. The two panels of the plot give the two versions of the robust t-statistics form the end of §3.2. In the 'traditional' form the errors are calculated from the covariance matrix in (3.2), whereas the modified form (3.3) includes a correction for the effect of robust estimation on $X^T X$.

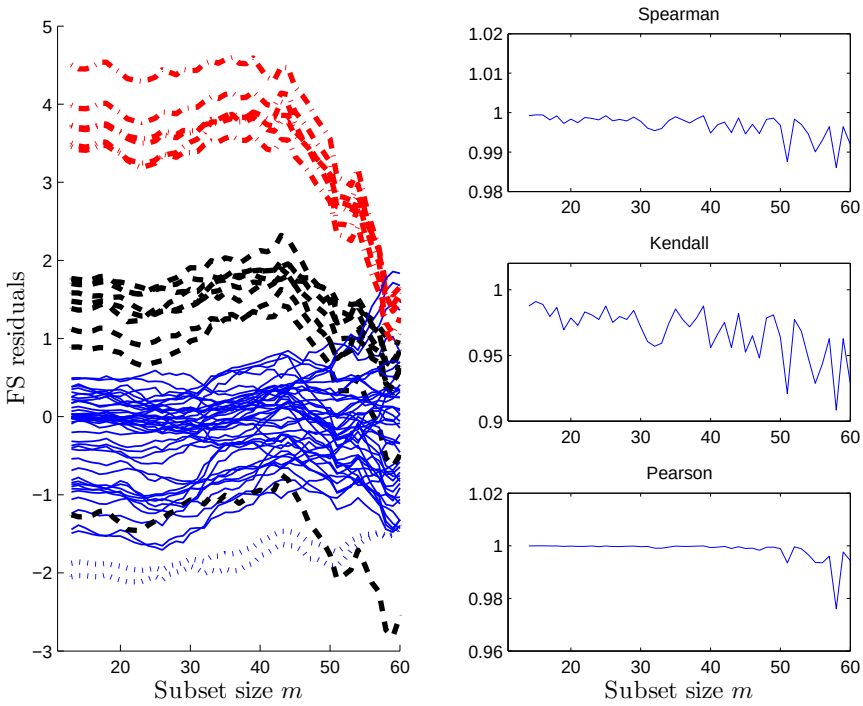

FIG 8. *AR Data: FS. Left-hand panel, plot of scaled residuals. The four groups of observations are initially evident, and particularly so around* $m = 40$. *The six outliers remain distinct until* $m = 55$. *Observation 43 becomes increasingly outlying from* $m = 43$ *and observations 7 and 39 are distinct until* $m = 58$. *Right-hand panel, three measures of the correlation. Kendall's measure, in particular, shows the gradual change in the outlyingness of the various units.*

In Figure 10 we present the related plot of t-statistics from the forward search (Atkinson and Riani, 2002). Although these statistics are less smoothed than those from S estimation and have a different horizontal scale, they lead to very similar conclusions about the significance of the variables in the robust and non-robust analyses. In fact, following the procedure of Huber and Ronchetti (2009) leads to values of the t-statistics in the lower panel of Figure 9 virtually identical to those in Figure 10. As with the plots of the residuals, the horizontal scales for the FS is relatively shrunk as m approaches n.

For these data, the results of this section show that the use of robust procedures with soft trimming leads to residuals which are sensibly constant over regions of the monitoring plot, although these regions depend on the exact details of the estimation method and ρ function that we employ. These residuals change sharply when outliers are included in the fit, changes which are readily detected by the correlation plots. However, this plot appears to be of lesser value for LTS, LMS and, especially, the FS. But, as we show in the final example, for the FS we monitor plots of outlier tests to indicate where interesting changes occur in the structure of the models fitted to the data.

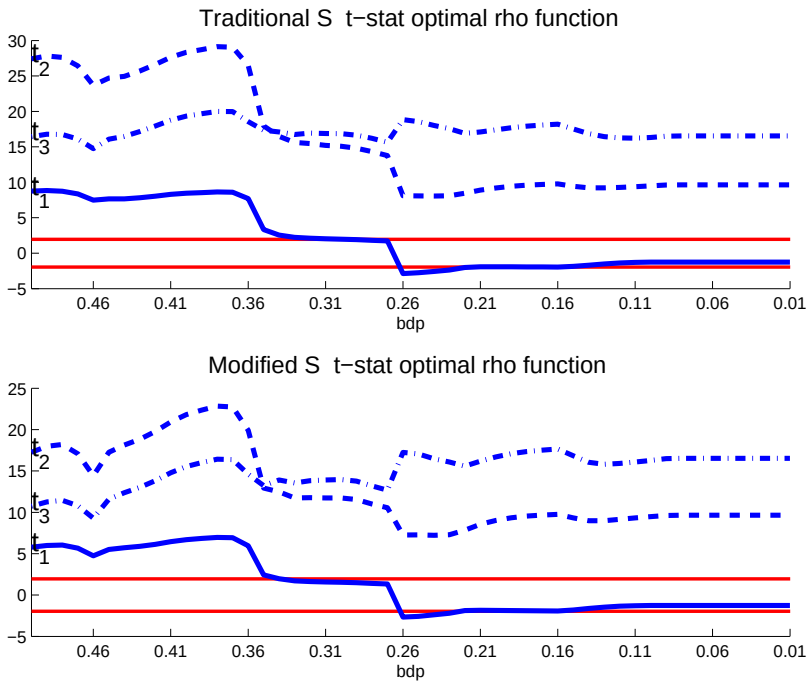

FIG 9. *AR data; S estimation with the optimal ρ function. Monitoring plots of two versions of t-statistics for the three parameters of the linear model. Upper panel, 'traditional' version from §3.2; lower panel, modified.*

4.4. Example 3: Bank data

We conclude with a more complicated example, in which there is no simple model for all the data. Furthermore, the aberrant observations do not form a simple cluster.

There are 1,949 observations, taken from a larger data set, on the amount of money made from personal banking customers over a year. There are 13 potential explanatory variables, listed in Appendix C, describing the services used by the customers, all of which are discrete, one being binary. The prime interest in the analysis is to discover which activities are particularly profitable. Since the response may be positive or negative, without a sharp lower bound, a power transformation (Box and Cox, 1964) is not an option for improving the agreement of the data with the regression model.

As a consequence of the results in §4.3, we only describe the results of S and MM estimation for these data, although we did also analyse them using τ estimation. We found that, for all four ρ functions, the MM estimates produced virtually constant plots of residuals, with change, if any at the last one or two monitoring points. Monitoring plots of S residuals, however, indicated a non-homogeneous structure in the data; Figure 11 plots the S residuals when the

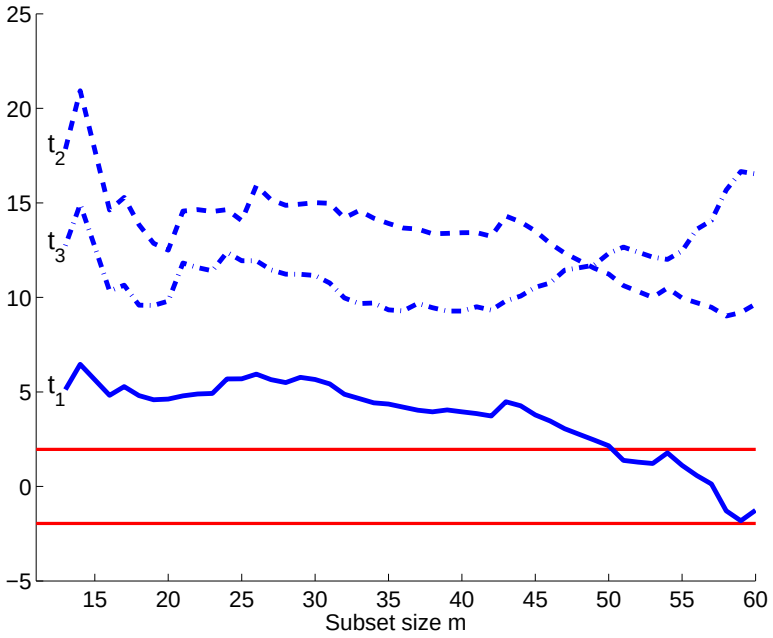

FIG 10. *AR data; FS estimation. Forward plots of t-statistics for the three parameters of the linear model.*

optimal ρ function is used. The residuals are clearly highly skewed, with a few large outliers, the pattern changing as the bdp decreases. The plot of correlation coefficients shows a clear dip at a bdp of 0.14.

These results are in agreement with those from the FS. Figure 12 shows a forward plot of the minimum deletion residuals that form a series of tests for the presence of outliers, together with the distributional bounds used in the rule for constructing a test of the required size over the whole sample. Here the simultaneous bound has a level of 1% (Riani *et al.*, 2014a). The resuperimposition of envelopes used to establish the number of outliers indicated that there are 255, agreeing with the bdp of 0.13.

We now consider the behaviour of the t-statistics from the two analyses and then use the FS to explore the fine structure of the data.

The t-statistics from S estimation based on the optimal ρ function are in Figure 13. If the data are not homogenous we would expect any changes in the plots to occur around a bdp of 0.13. The most dramatic changes seems to be for x_4, which becomes appreciably more significant, x_5 and x_{12} which lose significance, x_{10}, the significance of which decreases and x_9, which goes from having a slightly significant positive coefficient to one that is strongly negative. Such changes will be important when trying to decide on a model for the data with non-zero weights in the fits with higher bdp.

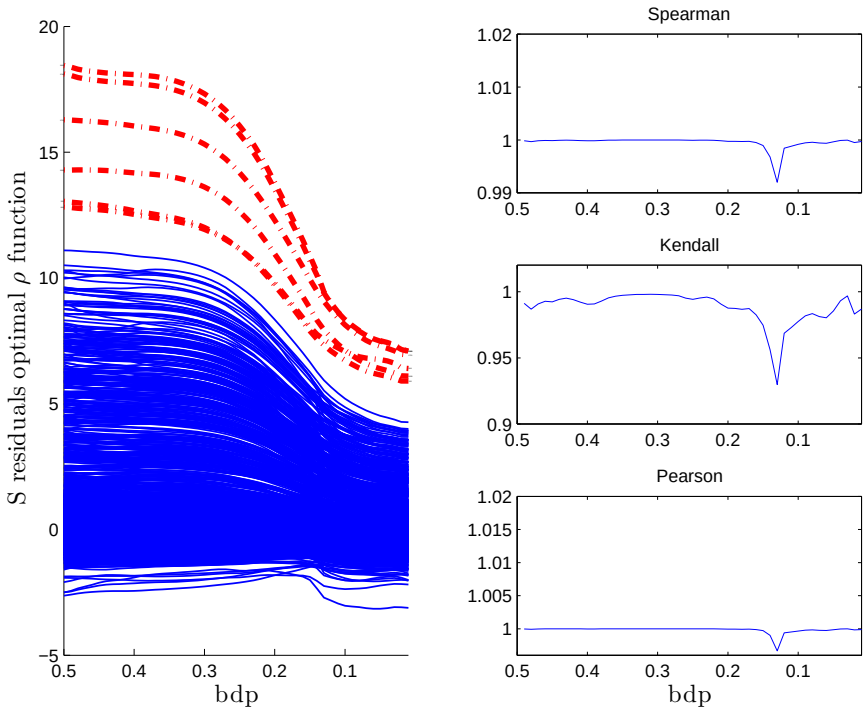

FIG 11. *Bank data: S estimation, optimal ρ function. Left-hand panel, plot of scaled residuals. These are highly skewed, with a few large outliers. Right-hand panel, three measures of the correlation. All show a clear transition at a bdp of 0.14.*

The plot of t-statistics from the FS in Figure 14, starting from $m = 1000$, is similar in general shape to that for S estimation. Although the vertical scales of the panels are not identical, the changes in importance of x_4, x_5, x_9, x_{10} and x_{12} are the same. As with the residual plots for analysis of the AR data, the plots for the S estimator are again smoother than those for the FS. Like the plots of t-statistics in Figures 9 and 10, the horizontal scale for FS is relatively shrunken as $m \to n$.

From a computational and numerical standpoint, these are not easy data to analyse. Particularly since all methods proceed by fitting subsets of the data, near collinearities and leverage points do occur. In more severe examples, a simple way to produce numerically stable solutions is to jitter the data, when the important inferential features will hopefully remain unchanged.

To close we briefly explore the two subsets into which we have so far broken the data; there is a larger set of 1,694 observations, which seem to be homogeneous and a remaining 255 which are merely "different" from the majority.

The left-hand panel of Figure 15 shows the scatterplots of y against each x for the larger portion into which the FS has divided the data. The right-hand panel shows the same plot for the remaining 255 observations. The two

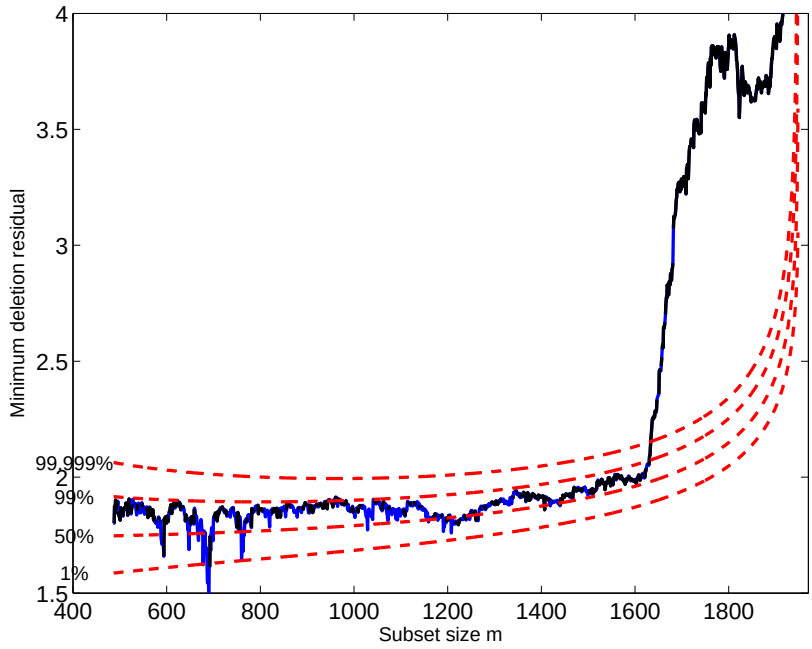

FIG 12. *Bank data. Forward plot of minimum deletion residuals, leading to identification of multiple outliers. The given levels of the envelopes are for pointwise tests of outlyingness.*

sets are clearly different. It is not just that the values of y in the right-hand panel are, in general, higher than those in the left-hand panel; they also have a different structure. For example, as the plots in Figure 14 of the t-statistics for x_1 (personal loans) and x_2 (financing and hire purchase) show, there is positive regression on these variables for the data in the left-hand panel. However, as the customers from the right-hand panel are included towards the end of the search, the regression decreases, as it does for x_{10} (credit cards). On the other hand, the plot suggests that the second group has a higher uptake of life insurance (x_4). A striking feature of the scatterplots in the right-hand panel of Figure 15 are the six large outliers, which are indeed visible in Figure 11.

Further analysis of these data might build a regression model for the set of 1,694 observations and do the same for the remaining 255 observations, having first checked whether they are homogeneous after exclusion of the six outliers. An important practical aspect would be to determine, if there are just two groups, whether they can be readily separated, either on the basis of the 13 explanatory variables used in the current analysis or by use of additional factors we have not included. The implication of our discussion of the values of the t-statistics is that the second group may be more affluent than the first. However, the bank is unlikely to have accurate information on the actual income of its clients.

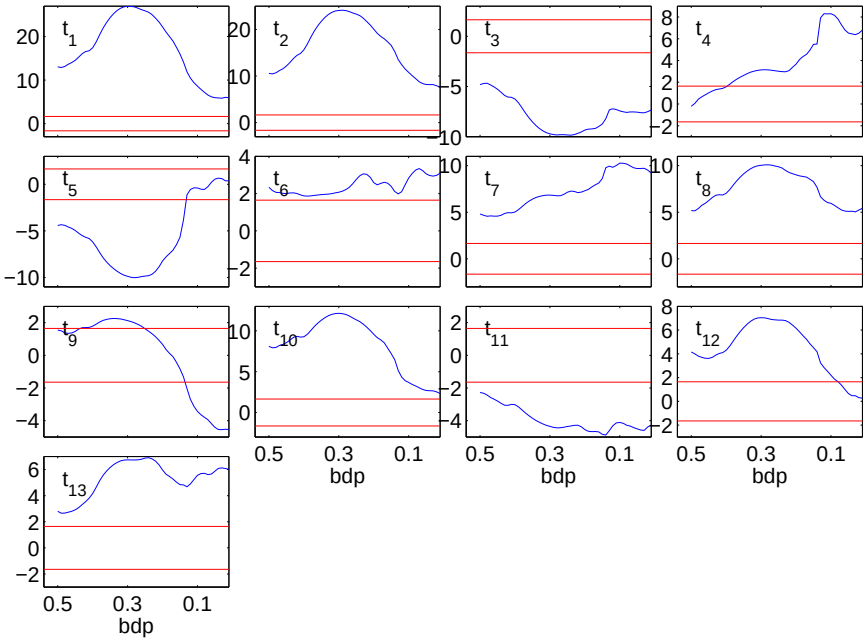

FIG 13. *Bank data, S estimation with the optimal ρ function. Monitoring plots of modified t-statistics* (3.3) *for the 13 parameters of the linear model.*

5. Comments and conclusions

As a result of monitoring robust regression for three data sets of increasingly complex nature, we have arrived at some simple conclusions. For monitoring we require robust methods that are robust to the choice of ρ function and estimation method, including the parameters that determine the efficiency, or breakdown point, of the method. A similar point on the necessity of robust estimators being themselves robust to differing circumstances is made by Croux *et al.* (2004).

Our results are helpfully informative about the choice between the various methods of robust regression. Methods, such as MM and τ estimation, that are tuned to give nominal high efficiency and a high breakdown point, fail with our most complicated example. We find that the most informative analyses come from S estimates combined with Tukey's biweight or the optimal ρ functions. The indication of the monitoring residual plot for the AR data with S estimation and the optimal ρ function in Figure 4 is that this ρ function can sometimes provide more information about the structure of the data than does Tukey's biweight

Finally, we commented in §3.1 that a major drawback to interpretation of the results of robust analyses was the loss of information on the effect of individual observations on inferences drawn from the data. Our analysis of the bank data shows how the forward search, in combination with the insights gained from

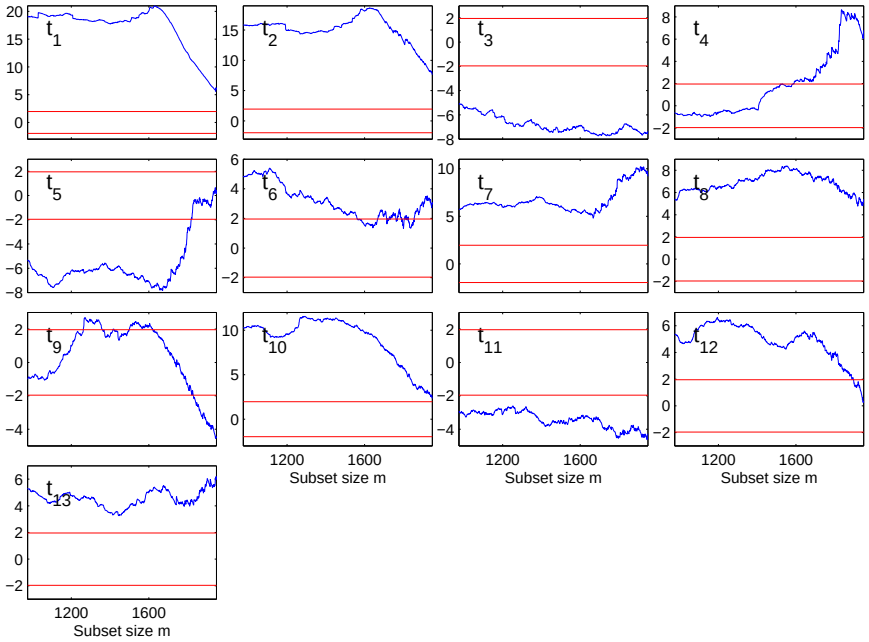

FIG 14. *Bank data; FS estimation. Forward plots of t-statistics for the 13 parameters of the linear model.*

robust regression, can be used to provide information on the data structures lying behind the need for robust procedures.

Acknowledgements

The authors thank the participants of the workshop on Advances in Robust Data Analysis and Clustering, held at the Joint Research Centre of the European Commission, Ispra (Italy), for stimulating discussions. The constructive comments of an anonymous referee and the Associate Editor are also gratefully acknowledged.

This work was partly supported by the project MIUR PRIN "MISURA-Multivariate models for risk assessment".

Appendix A: Rho functions

Perhaps the most popular ρ function for redescending M and S-estimates is **Tukey's Bisquare (or Biweight) function**

$$\rho(u) = \begin{cases} \frac{u^2}{2} - \frac{u^4}{2c^2} + \frac{u^6}{6c^4} & \text{if } |u| \leq c \\ \frac{c^2}{6} & \text{if } |u| > c, \end{cases} \tag{A.1}$$

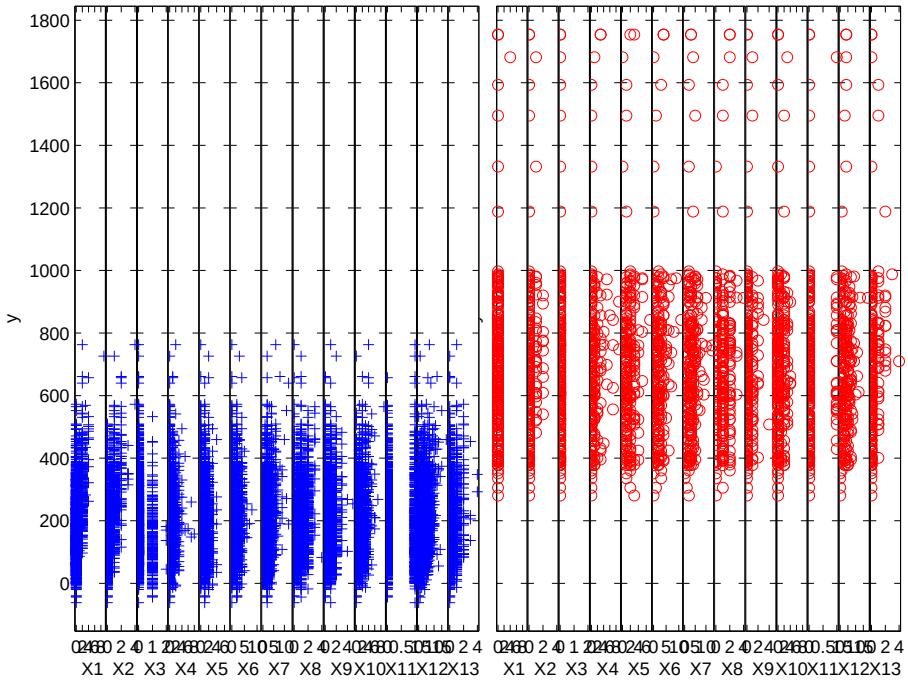

FIG 15. *Bank data. Scatterplots of y against individual explanatory variables for the two parts of the data. Left-hand panel, the main body of the data. Right-hand panel, the remaining, somewhat different, 255 observations.*

the first derivative of which vanishes outside the interval $[-c, +c]$. Therefore, for this function c is the crucial tuning constant, determining the efficiency or, equivalently, the breakdown point.

A similar, but less smooth, shape is shared by **Hampel's ρ function**

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \le c_1 \\ c_1|u| - \frac{1}{2}c_1^2 & \text{if } c_1 < |u| \le c_2 \\ c_1 \frac{c_3|u| - \frac{1}{2}u^2}{c_3 - c_2} & \text{if } c_2 < |u| \le c_3 \\ c_1(c_2 + c_3 - c_1) & \text{if } |u| > c_3. \end{cases} \tag{A.2}$$

The first derivative is piecewise linear and vanishes outside the interval $[-c_3, +c_3]$. Again c_3 is the crucial tuning constant, although Huber and Ronchetti (2009), p. 101 suggest that the slope between c_2 and c_3 should not be too steep.

Yohai and Zamar (1997) introduced a ρ function which minimizes the asymptotic variance of the regression M-estimate, subject to a bound on a robustness measure called contamination sensitivity. Therefore, this function is called the

optimal ρ function. It is given by

$$\rho(u) = \begin{cases} 1.3846\left(\frac{u}{c}\right)^2 & \text{if } |u| \le \frac{2}{3}c \\ 0.5514 - 2.6917\left(\frac{u}{c}\right)^2 + 10.7668\left(\frac{u}{c}\right)^4 - & \\ \quad -11.6640\left(\frac{u}{c}\right)^6 + 4.0375\left(\frac{u}{c}\right)^8 & \text{if } \frac{2}{3}c < |u| \le c \\ 1 & \text{if } |u| > c. \end{cases} \tag{A.3}$$

Now the first derivative vanishes outside the interval $[-c, +c]$. The resulting M-estimate minimizes the maximum bias under contamination distributions (locally for a small fraction of contamination), subject to achieving a desired nominal asymptotic efficiency when the data are normally distributed.

Hampel *et al.* (1981) considered a different optimization problem, by minimizing the asymptotic variance of the regression M-estimate, subject to a bound on the supremum of the Change of Variance Curve of the estimate. This led to the **Hyperbolic Tangent ρ function**, which, for suitable constants c, c_1, c_2, c_3 and c_4, is defined as

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \le c_1 \\ -2\frac{c_2}{c_3}\ln\cosh[\frac{1}{2}\sqrt{\frac{(c_4-1)c_3^2}{c_2}}(c-|u|)] + & \\ \quad +\frac{\xi^2}{2} + 2\frac{c_2}{c_3}\ln\cosh[\frac{1}{2}\sqrt{\frac{(c_4-1)c_3^2}{c_2}}(c-\xi)] & \text{if } \xi \le |u| \le c \\ \frac{\xi^2}{2} + 2\frac{c_2}{c_3}\ln\cosh[\frac{1}{2}\sqrt{\frac{(c_4-1)c_3^2}{c_2}}(c-\xi)] & \text{if } |u| > c, \end{cases} \tag{A.4}$$

where $0 < \xi < c$ is such that

$$\xi = \sqrt{[c_2(k-1)]}\tanh\left[\frac{1}{2}\sqrt{\frac{(k-1)c_3^2}{c_2}}(c-\xi)\right],$$

c_2 and c_3 satisfy suitable conditions, and c_4 is related to the bound in the Change of Variance Curve. These constants must be computed iteratively, by applying Newton-Raphson steps and numerical integration.

Appendix B: Absolute odd moments of the multivariate normal distribution under elliptical truncation

Unlike Tukey's bisquare and the optimal ρ functions, Hampel's function (A.2) includes absolute moments of random variables. This appendix concentrates on the relationship between untruncated and elliptically truncated absolute central odd moments in the multivariate normal distribution. The results extend those of Tallis (1963) who derived the moment generating function of the multivariate normal distribution under elliptical truncation.

As is well known, the central absolute moments coincide with plain moments for all even orders, but are nonzero for odd orders.

Definition 1. *The central moment of order k of the v dimensional random vector U is defined as (see, e.g., Kotz* et al.*, 2000, pp. 107–111)*

$$\mu_{k_1,k_2,\ldots,k_v}(U) = \mu_{1,2,\ldots,v}(U) = E\left[\prod_{j=1}^{v}(U_j - \mu_j)^{k_j}\right],$$

where $k_1 + k_2 + \cdots + k_v = k$, $k_j \geq 0$ and $\mu_j = E(U_j)$.

Definition 2. *If $U \sim N_v(\mu, \Sigma)$, we define as elliptical truncation in the interval $[b; c]$ the set of points $u \in R^v$ belonging to*

$$E = \left\{u | b \leq (u-\mu)'\Sigma^{-1}(u-\mu) \leq c\right\}, \qquad 0 \leq b < c.$$

We are interested in finding and discussing the expression of the central absolute moment of order k, when k is odd, under elliptical truncation

$$\mu_{k_1,k_2,\ldots,k_v}(|U|) = E\left[|(U_1-\mu_1)^{k_1}(U_2-\mu_2)^{k_2}\ldots(U_v-\mu_v)^{k_v}|\right].$$

Starting from the Definition 1, when $U \sim N(0, I_v)$ with the constraint that the region of integration is $b^2 < u'u < c^2$, we can write:

$$\mu_{k_1,k_2,\ldots,k_v}(|U|) = \int\int\ldots\int_{b^2<u'u<c^2} |u_1^{k_1}\ldots u_v^{k_v}| d\Phi(U_1)\ldots d\Phi(U_v).$$

This v-dimensional integral can be rewritten in the transformed space as a univariate integral as follows:

$$\begin{aligned}
\int_{b^2/2}^{c^2/2} y^{k/2} f_{\chi^2_v}(y)dy &= \int_{b^2}^{c^2} \frac{y^{\frac{v+k}{2}-1}e^{-y/2}}{2^{v/2}\Gamma(v/2)}dy \\
&= \int_{b^2/2}^{c^2/2} \frac{(2t)^{\frac{v+k}{2}-1}e^{-t}dt}{2^{v/2}\Gamma(v/2)}dy \\
&= 2^{k/2}\frac{\Gamma(\frac{v+k}{2})}{\Gamma(v/2)}\int_0^{c^2/2} \frac{t^{\frac{v+k}{2}-1}e^{-t}}{\Gamma(\frac{v+k}{2})}dt \\
&= 2^{k/2}\frac{\Gamma(\frac{v+k}{2})}{\Gamma(v/2)}\left[P\left(\frac{c^2}{2}, \frac{v+k}{2}\right) - P\left(\frac{b^2}{2}, \frac{v+k}{2}\right)\right].
\end{aligned}$$

where P and Γ are the gamma and the incomplete gamma functions respectively.

Now, if k is odd it is easy to verify that

$$2^{k/2}\Gamma\left(\frac{v+k}{2}\right) = \sqrt{2}(v+1)(v+3)\cdots(v+k-2),$$

because

$$\Gamma\left(\frac{v+k}{2}\right) = \left(\frac{v+k}{2}-1\right)\times\left(\frac{v+k}{2}-2\right)\times\cdots\times\left(\frac{v+k}{2}-\frac{k-1}{2}\right)\Gamma\left(\frac{v+1}{2}\right)$$

$$= \frac{v+k-2}{2} \times \frac{v+k-4}{2} \times \cdots \times \frac{v+1}{2} \times \Gamma\left(\frac{v+1}{2}\right)$$
$$= (v+1)(v+3)\cdots(v+k-2) \times 2^{-(k-1)/2} \times \Gamma\left(\frac{v+1}{2}\right).$$

Similarly, it is easy to verify that

$$P\left(\frac{c^2}{2}, \frac{v+k}{2}\right) = F_{\chi^2_{v+k}}(c^2),$$

because

$$\begin{aligned} F_{\chi^2_{v+k}}(c^2) &= \int_0^{c^2} \frac{t^{\frac{v+k}{2}-1}e^{-t/2}}{2^{(v+k)/2}\Gamma((v+k)/2)}dt \\ &= \frac{2}{2^{\frac{v+k}{2}}\Gamma((v+k)/2)} \int_0^{c^2} (2y)^{\frac{v+k}{2}-1}e^{-y}dy \\ &= \frac{\int_0^{c^2/2} y^{\frac{v+k}{2}-1}e^{-y}}{\Gamma((v+k)/2)}dy \\ &= P\left(\frac{c^2}{2}, \frac{v+k}{2}\right). \end{aligned}$$

It follows that the expression of the central absolute moment of order k, when k is odd, under elliptical truncation is given by

$$\mu_{k_1,k_2,\ldots,k_v}(U) = \sqrt{2}\frac{(v+k-2)!!}{(v-1)!!}\frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)}\left(F_{\chi^2_{v+k}}(c^2) - F_{\chi^2_{v+k}}(b^2)\right). \quad \text{(A.5)}$$

This result shows that the odd absolute moments of the multivariate standard normal distribution under elliptical truncation are equal to the original moments multiplied by $(F_{\chi^2_{v+k}}(c^2) - F_{\chi^2_{v+k}}(b^2))$, thus generalizing that of Tallis (1963).

It is interesting to notice that, if U is univariate ($v = 1$), equation (A.5) becomes

$$\mu_k(U) = \int_{b<|u|<c} |u|^k d\Phi(u) = \sqrt{\frac{2}{\pi}}(k-1)!!\left(F_{\chi^2_{r+1}}(c^2) - F_{\chi^2_{r+1}}(b^2)\right).$$

If $b \to \infty$ and $a = 0$ we simply recover the usual formula for the central absolute odd moment. Finally, when $k = 1$ we easily find that

$$E(|U|) = \int_{b<|u|<c} |u| d\Phi(u) = 2\left(\phi(b) - \phi(c)\right)$$

where $\phi(c)$ is the density function of the standard normal evaluated at c. If $b = 0$ and $c = \infty$ we get $E(|U|) = \sqrt{2/\pi}$, the usual formula for the expectation of the half normal distribution.

TABLE 3
Bank data: the thirteen explanatory variables

Variable number	Description	Number of zeroes
1	Personal loans	1666
2	Financing and hire-purchase	1529
3	Mortgages	1734
4	Life insurance	1503
5	Share account	435
6	Bond account	987
7	Current account	27
8	Salary deposits	742
9	Debit cards	1030
10	Credit cards	1003
11	Telephone banking	1459
12	Domestic direct debits	426
13	Money transfers	1596

Appendix C: The bank data

There are 1,949 univariate observations on the amount of money made from individual personal banking customers over a year for an Italian bank. Because of the linking of products, it is not straightforward for the bank to attribute the profit to individual sources. The bank made a preliminary classification of its 700 products into 48 macrocategories (macroservices). Among these 48 macrocategories, the 13 most important ones according to the bank are listed in the second column of Table 3 and form our set of explanatory variables. All explanatory variables are discrete, taking values 0, 1, 2, ..., the number of services (inside each macroservice) that each customer has signed up for – number of credit cards, number of domestic direct debits, number of current accounts and so forth. Only x_{11}, telephone banking, is binary, because the bank has just one internet service. Since many customers have not signed up for all services, we also give in Table 3 the number of zeroes for each variable. Translation of the names of the variables is made difficult since Italian banks, or at least the bank in question, charge for many services which come free with a current account at a British bank and so are sometimes not identified, or, even, necessary. (We have no experience of American banks). In the case of joint accounts, the data for the account are entered once for each account holder. Although the individuals may have signed up for different levels of other services, such duplication produces near replicates in x (and, of course, exact replicates of the response).

The revenue of a macroservice sold by the bank is determined not by the products sold inside it, but by the behavior of customers using its products. The data were accordingly selected by the bank using predefined thresholds (for example, customers who had movements of current accounts greater than a certain amount or debts within a certain range) the intention being to identify relatively homogeneous groups of customers with similar behaviour. Our analysis shows that the cluster under study is not as homogenous as the bank hoped, containing as it does six outliers and two subgroups.

The data are available as an Excel file in the supplement to this paper (Riani *et al.*, 2014d). They, together with the routines for monitoring, are included in the FSDA toolbox downloadable from `http://fsda.jrc.ec.europa.eu/` or `http://www.riani.it/MATLAB`.

Supplementary Material

Bank Data

(doi: 10.1214/14-EJS897SUPP; .zip). The supplement provides an Excel file of the Bank Data described in Appendix C and Table 3 of our paper.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Tukey, W. J., and Huber, P. J. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, NJ. MR0331595

Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York. MR1884997

Atkinson, A. C. and Riani, M. (2002). Forward search added variable t tests and the effect of masked outliers on model selection. *Biometrika*, **89**, 939–946. MR1946522

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. MR2758131

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, **26**, 211–246. MR0192611

Cerioli, A., Farcomeni, A., and Riani, M. (2014). Strong consistency and robustness of the Forward Search estimator of multivariate location and scatter. *Journal of Multivariate Analysis*, **126**, 167–183. MR3173089

Croux, C. and Rousseeuw, P. J. (1992). A class of high-breakdown scale estimators based on subranges. *Communications in Statistics – Theory and Methods*, **21**, 1935–1951. MR1173503

Croux, C., Dhaene, G., and Hoorelbeke, D. (2004). Robust standard errors for robust estimators. CES – Discussion paper series OR 0367, Department of Applied Economics, KU Leuven.

Hampel, F. R. (1975). Beyond location parameters: Robust concepts and methods. *Bulletin of the International Statistical Institute*, **46**, 375–382. MR0483172

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.

Hawkins, D. M. and Olive, D. J. (2002). Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm (with discussion). *Journal of the American Statistical Association*, **97**, 136–159. MR1947276

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, **1**, 799–821. MR0356373

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics, Second Edition*. Wiley, New York. MR2488795

Koller, M. and Stahel, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. *Computational Statistics and Data Analysis*, **55**, 2504–2515. MR2787008

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Multivariate Distributions – 1, 2nd Edition*. Wiley, New York.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester. MR2238141

Riani, M., Perrotta, D., and Torti, F. (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, **116**, 17–32. doi:10.1016/j.chemolab.2012.03.017.

Riani, M., Atkinson, A. C., and Perrotta, D. (2014a). The forward search algorithm for very robust regression. (Submitted).

Riani, M., Cerioli, A., and Torti, F. (2014b). On consistency factors and efficiency of robust S-estimators. *TEST*. (In press). doi:10.1007/S11749-014-0357-7.

Riani, M., Atkinson, A. C., and Perrotta, D. (2014c). A parametric framework for the comparison of methods of very robust regression. *Statistical Science*, **29**, 128–143. doi:10.1214/13-STS437.

Riani, M., Cerioli, A., Atkinson, A. C., and Perrotta, D. (2014d). Supplement to "Monitoring robust regression". doi:10.1214/14-EJS897SUPP.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871–880. MR0770281

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York. MR0914792

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. MR0786313

Stigler, S. M. (1989). Francis Galton's account of the invention of correlation. *Statistical Science*, **4**, 73–79. MR1007556

Stigler, S. M. (2010). The changing history of robustness. *The American Statistician*, **64**, 277–281. MR2758558

Tallis, G. M. (1963). Elliptical and radial truncation in normal samples. *Annals of Mathematical Statistics*, **34**, 940–944. MR0152081

Yohai, V. J. (1987). High breakdown-point and high efficiency estimates for regression. *The Annals of Statistics*, **15**, 642–656. MR0888431

Yohai, V. J. and Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, **83**, 406–413. MR0971366

Yohai, V. J. and Zamar, R. H. (1997). Optimal locally robust M-estimates of regression. *Journal of Statistical Planning and Inference*, **64**(2), 309–323. MR1621620