A Parametric Framework for the Comparison of Methods of Very Robust Regression

Marco Riani, Anthony C. Atkinson and Domenico Perrotta

April 29, 2013

Abstract

There are several methods for obtaining very robust estimates of regression parameters that asymptotically resist 50% of outliers in the data. Differences in the behaviour of these algorithms depend on the distance between the regression data and the outliers. We introduce a parameter λ that defines a parametric path in the space of models and enables us to study, in a systematic way, the properties of estimators as the groups of data move from being far apart to close together. We examine, as a function of λ, the variance and squared bias of five estimators and we also consider their power when used in the detection of outliers. This systematic approach provides tools for gaining knowledge and better understanding of the properties of robust estimators.

Keywords: distance of outliers; forward search; least trimmed squares; MM estimate; multiple outliers; overlap index; point contamination; regression diagnostics

1 Introduction

Multiple regression is one of the main tools of applied statistics. It has however long been appreciated that ordinary least squares as a method of fitting regression models is exceptionally susceptible to the presence of outliers. Instead, very robust methods, that asymptotically resist 50% of outliers, are to be preferred.

Marco Riani is Professor, Dipartimento di Economia, Università di Parma, Italy (e-mail: mriani@unipr.it, web: http://www.riani.it). Anthony Atkinson is Emeritus Professor, Department of Statistics, The London School of Economics, London WC2A 2AE, UK (e-mail: a.c.atkinson@lse.ac.uk). Domenico Perrotta is Senior Research Scientist, European Commission Joint Research Centre, Ispra, Italy (e-mail: domenico.perrotta@ec.europa.eu).

Our paper presents a systematic, parameterized framework for the non-asymptotic comparison of these methods.

Very robust regression was introduced by Rousseeuw (1984) who developed suggestions of Hampel (1975) that led to the Least Median of Squares (LMS) and Least Trimmed Squares (LTS) algorithms. For some history of more recent developments see Rousseeuw and Van Driessen (2006). More general discussions of robust methods are in Maronna, Martin, and Yohai (2006) and Morgenthaler (2007). We illustrate our methods for the comparison of high-breakdown regression procedures with comparisons of the performance of LTS and other well-established methods, including S and MM estimators, with that of a publicly available algorithm for very robust regression that uses the Forward Search (FS). See Atkinson, Riani, and Cerioli (2010) for a recent discussion of the FS.

Very robust regression estimators share the property that, asymptotically, they have a breakdown point of 50% (see §3.2) as the main data and outliers become infinitely far apart. In order to distinguish between the estimators we study, in a systematic way, their properties as the distance between the two groups of observations decreases. In §4 we introduce a parameterised framework, with parameter λ, for moving the outliers along a trajectory which is initially remote from the main data, but which then passes close to it before again becoming far away. We control whether, at their closest, the two populations share the same centre. We design measures of overlap to calibrate the trajectories.

Numerical results are in §§5 and 6. In §5 we take the outliers from the regression model to have a multivariate normal distribution. This provides a very general scenario for outliers that can range from a seemingly random scatter around the regression plane to points virtually on a line. The special case of point contamination is explored in §6. Boxplots of the estimates from the five methods as λ varies indeed show that, for wide separations, the methods have similar properties. However they differ markedly as the two populations converge. In order to summarise this information we look at cumulative plots, over the range of λ, of the variance and squared bias of the estimators. Another method of comparing robust estimators is by their properties for outlier detection (Cook and Hawkins 1990). In §5 we calculate power curves as a function of λ for the number of outliers detected. Since the curves indicate that the estimators provide tests of varying sizes, we find the size of the outlier tests in §7.

There are two main conclusions. The first is that the parameterised family of departures provides a cogent framework for investigating the behaviour of very robust estimators. The second is that we can clearly establish the properties of the various methods of very robust regression in terms of the bias and variance of estimators and the size and power of outlier tests.

The approach is motivated in the next section by an example in which there is a mixture of two regression lines. Such data arise in the analysis of trade where

different countries or suppliers may report different relationships between value and quantity. Although, in our example, there are only two countries, which makes the data appropriate for a robust analysis assuming one model describes at least half the data, there is no reason why there should not be several suppliers. The comparative robust analysis of the data is in §7.

2 An example: trade data

Our interest in the behaviour of robust regression procedures when the main data and outliers are close together was stimulated by a seemingly simple example with a single explanatory variable. The data, shown in Figure 1, are of a kind discussed by Perrotta, Riani, and Torti (2009) in the detection of fraud in international trade, where false declarations of price are used in tax evasion and money laundering. The result is data which are a mixture of regression lines.

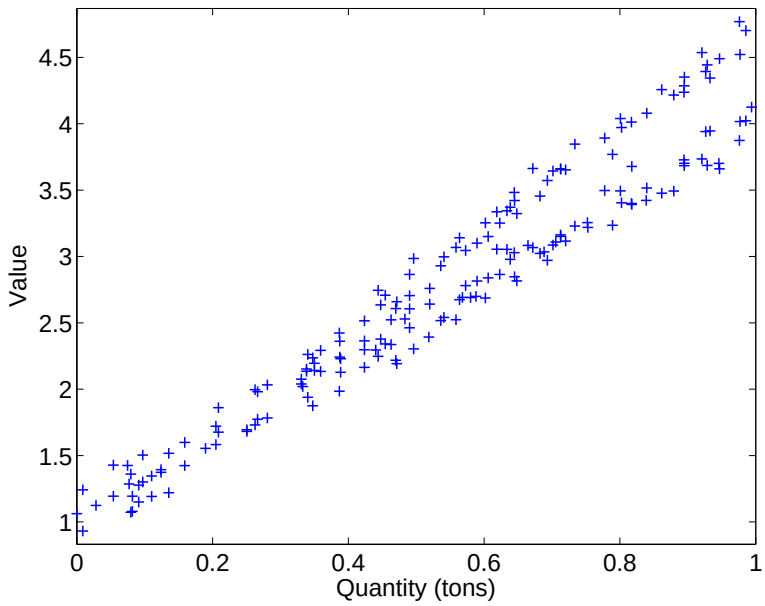

Figure 1: Trade data: a mixture of two regression lines

There are 180 observations in Figure 1 that come from two firms. The structure is of two lines that overlap for lower values; any kind of separation is likely to be impossible. However, the two lines are clearly separate for the higher values of y and x and a robust procedure should respond to this pattern, by downweighting some of the observations in estimation and flagging them as outliers. If the outlier pattern suggests that there is a mixture of regression models, the analysis can move to clusters of regression lines, as in García-Escudero et al. (2010). But

the first stage is the identification of outliers, for which a robust fit is required. Interest in the analysis is not in individual outliers but whether the two lines differ. Accordingly, in §4 we introduce a Bonferroni adjustment to provide, at least theoretically, the desired samplewise size of the outlier test. We return to the analysis of these data in §9.

3 Models, Data, Robustness and Methods

3.1 Outliers and Regression

We consider the usual regression model with random carriers (Huber and Ronchetti 2009, p. 197). The observations are i.i.d. random vectors $(y, x^T) \in \Re^{p+1}$, where $y \in \Re$ and $x \in \Re^p$ satisfy

$$y = x^T\beta + u. \tag{1}$$

The random errors u are distributed independently of the covariates x and β is the $p \times 1$ vector parameter of interest.

In the absence of outliers the least squares estimate $\hat{\beta}$ is the best linear unbiased estimator of β. However even a single outlier can cause $\hat{\beta}$ to be severely biased. Figure 2 of Rousseeuw (1984) is a paradigmatic example in which a cluster of 20 outliers at a remote point in X-space cause the least squares fitted line to pass close to the cluster. The robust line, in that case LMS, completely downweights the outliers and is close to the least squares line for the 30 remaining data points when the outliers have been deleted.

Of course, the outliers are not usually known and the problems of robust estimation and outlier detection are closely related. In robust estimations a fit is found which is close to that without the outliers. The robust fit then allows identification of all important outliers. However, the outliers may be difficult to identify from a non-robust fit since their inclusion can seriously bias the parameter estimates and make the outliers seem less remote. 'Backward' methods of outlier detection that start from a fit to all data and then proceed by eliminating observations that appear to be outlying can therefore fail.

One example for regression is a synthetic dataset due to Hawkins, Bradu, and Kass (1984) with $n = 75$ and three explanatory variables. The figures on p. 95 of Rousseeuw and Leroy (1987) show that the least squares residuals, unlike those from LMS, are not sufficiently large to call attention to the ten outlying observations. Numerous other examples for regression are in Chapters 3 and 4 of Atkinson and Riani (2000); further plots for the Hawkins et al. data are on pp. 72 and 73.

3.2 Maxbias and Breakdown Point

Robustness is concerned with fitting a single model to data which are generated by two, or maybe more, models. We suppose that the larger part of the data, $1-\epsilon$, where $0 < \epsilon < 0.5$, is generated by the model $M_1(\theta_1)$ and the remaining part ϵ of the data is generated by the model $M_2(\theta_2)$. In the absence of outliers, that is when $\epsilon = 0$, an ideal robust estimator would have a variance that achieved the Cramer-Rao lower bound. If the data were contaminated, the estimate would be unbiased. Such estimators do not exist. Maronna, Martin, and Yohai (2006, §3.4) describe some compromises between the two properties.

Robust methods study the properties of methods that fit $M_1(\theta_1)$ in ignorance of knowledge of the form of the outlier generating model $M_2(\theta_2)$, which can be quite general. When $M_1(.)$ is a regression model, $M_2(.)$ is often taken, for example, to distribute observations randomly over a large space, concentrate them in a cluster or to be a second regression model. There is no difficulty in having $M_1(\theta) = M_2(\theta)$ but then we must have $\theta_1 \neq \theta_2$.

With $M_1(\theta_1)$ the usual regression model (1) let $\mathrm{E}\, u^2 = \sigma^2 < \infty$ and $I(x) = \mathrm{E}\, xx^T$. The bias of an estimator $\hat{\beta}$ of β is

$$b(\hat{\beta}) = \{(\hat{\beta} - \beta)^T I(x)(\hat{\beta} - \beta)\}^{0.5}. \tag{2}$$

The bias depends on the estimator, the distribution of y and x, the amount of contamination ϵ and on $M_2(\theta_2)$. In the robustness literature dependence on $M_2(\theta_2)$ is removed by considering estimators that minimize the maximum (asymptotic) bias within a particular class of estimators. The maxbias curve shows how the maximum bias varies with ϵ. The breakdown point of an estimator is the minimum value of ϵ for which $b(\hat{\beta})$ in (2) equals ∞. The estimators we consider all have an asymptotic breakdown point of 50%. An introduction to these ideas for regression is given by Maronna et al. (2006, §5.9). Although 50% is customarily considered to be the maximum possible breakdown value, we briefly discuss this limit in §9.

Unfortunately maxbias curves are only calculable for some estimators and distributions of x. The latter are often assumed to be elliptically symmetrical. A summary of the literature is given by Berrendero and Zamar (2001) who extend results on maxbias curves to regression models with intercepts and to regressors that have Student's t and Cauchy distributions, although without intercepts. Berrendero, Mendes, and Tyler (2007) find maxbias curves for MM estimators, again without intercepts. We describe MM estimators in §3.3.

The theoretical results that are available are asymptotic and do not cover all estimators or models of interest to us. A few numerical results are available for finite samples. Figure 5.14 of Maronna et al. (2006) plots biases of several estimators as a function of a single parameter, the slope of the regression line for the contaminating observations. Figure 3 of García-Escudero et al. (2010) is more in

the spirit of our numerical approach. It shows the simulated bias as a point cluster of outliers moves around a regression line. When the outliers are very close to the line, the bias is negligible, as it is when the outliers are far away and are easily downweighted by, in this case, LTS. Only for intermediate outliers is the bias appreciable.

In our numerical comparisons we study the variance as well as the bias of the estimators. In addition, following the comments in §3.1 about the relationship between robustness and outlier detection, we asses the power of outlier tests using residuals from robustly fitted models.

3.3 Five Methods for Very Robust Regression

We compare and contrast the properties of what are currently considered the five best methods for very robust regression. The algorithms that we use are all publicly available from the Forward Search Data Analysis (FSDA) Matlab toolbox. See Riani, Perrotta, and Torti (2012). In this section we outline the methods that we compare. Full implementation details of the algorithms are in the documentation of the FSDA library. Numerically, all algorithms involve selecting many subsets from the data. An important factor in our ability to conduct as many simulations as were necessary is the efficient sampling of subsets provided in FSDA as described by Torti et al. (2012).

Traditional robust estimators attempt to limit the influence of outliers by replacing the squares of the residuals in least squares estimation of β by a function ρ of the residuals which is bounded. Of the numerous forms that have been suggested for $\rho(.)$ (Andrews et al. 1972, Hampel et al. 1986, Huber and Ronchetti 2009) we use the most popular choice, Tukey's Biweight in which extreme observations are replaced by the value $c^2/6$. See, for example, Rousseeuw and Leroy 1987, (4.31). The M-estimator of scale $\tilde{\sigma}_M$ is the solution to a second equation, for example Rousseeuw and Leroy 1987, (4.30), depending on a second ρ function and a constant K_c. Although the two ρ functions may be different, we again use the biweight. The minimum value of $\tilde{\sigma}_M$ which satisfies this second equation provides the S-estimate of scale ($\tilde{\sigma}_S$) with associated estimate of the vector of regression coefficients ($\tilde{\beta}_S$). K_c and c are related constants which are linked to the breakdown point of the estimator of β. Fixing the breakdown point at 50% gives a value for 1.547 for c and an efficiency for estimation of 28.7% (Rousseeuw and Leroy 1987, pp. 135-143).

The MM-regression estimator is intended to improve the S estimator. The S estimate of scale $\tilde{\sigma}_S$ is used and kept fixed to estimate β, but with a value of K_c giving a higher efficiency. Because of the relationship between K_c and c the hope expressed by Rousseeuw and Leroy (1987, p. 143) is that the MM estimator maintains its high breakdown point for finite samples. Following the recommendation

of Maronna, Martin, and Yohai (2006, p. 126), we take K_c such that the (asymptotic) nominal efficiency is 85%, which gave a high-breakdown estimator in our examples, which included up to 23% of outliers. Small numerical experiments indicate that even slight increases, for example to a nominal efficiency of 87%, result in very low breakdown and estimates similar to those from least squares.

The remaining three estimators of β result from more direct approaches. The forward search (FS) uses least squares to fit subsets of observations of increasing size m to the data, with $p \leq m \leq n$. The forward search for regression was introduced by Atkinson and Riani (2000). A recent general review of forward search methods is Atkinson et al. (2010). For efficient parameter estimation m should increase until all $n - m$ observations not in the subset used for fitting are outliers. The outliers are found by testing at each step of the search. The effect of simultaneous testing can be severe (Atkinson and Riani 2006); the FS algorithm is designed to have size α of declaring an outlier free sample to contain at least one outlier. We perform the outlier test for individual observations at a Bonferronised size $\alpha^* = \alpha/n$, so taking the $1 - \alpha^*$ cutoff value of the reference distribution. In our calculations $\alpha = 0.01$. The automatic algorithm is based on that of Riani, Atkinson, and Cerioli (2009) who used scaled Mahalanobis distances to detect outliers in multivariate normal data. For regression we replace these distances by deletion residuals.

In Least Trimmed Squares (LTS) (Rousseeuw 1984, p. 876) the search is over subsets of size h for which the residual sum of squares from least squares estimates of β is minimized. LTS has an asymptotic breakdown point of 50% when $h = [n/2] + [(p+1)/2]$.

To increase efficiency, reweighted versions of LTS estimators can be computed. These reweighted estimators, denoted LTSr, are computed by giving weight 0 to outlying observations. We then obtain a sample of reduced size $n - k$, possibly outlier free, to which OLS is applied. For comparison of results from LTSr with those from the FS we perform the outlier test at the Bonferronised size α^*.

In FS, LTS and its reweighted version LTSr σ^2 is estimated from subsets formed by hard (0,1) trimming. Consistency factors for the estimators are given by Croux and Rousseeuw (1992, Eq. 6.5) and follow from the results of Tallis (1963) on elliptically trimmed multivariate normal distributions. For LTS we also use the small sample correction of Pison, Van Aelst, and Willems (2002).

4 A Parameterized Family of Departures

As $y_{M2} \sim M_2(\theta_2) \to \infty$ the observations y_{M1} and y_{M2} from the two models become increasingly well separated. Under these conditions the five estimators in our study have similar properties. We are also interested in those data configura-

tions when the observations are not so well separated, so that both y_{M1} and y_{M2} may be used in estimating θ because of overlap between the two samples. Such configurations are highly informative about the differences in properties of robust estimators. We define a finite-sample measure of the overlap of y_{M1} and y_{M2} that is designed to be informative for regression models. In general, the properties of robust estimators depend on the "distance" between the two models. Table 3.1 of Maronna, Martin, and Yohai (2006) is a typical example showing the behaviour of robust estimators as one observation $\rightarrow \infty$. Our proposed distance measure likewise provides a framework for comparisons in the more complicated world of regression procedures.

There is a sample $\mathcal{S}_1$ of n_1 observations from $M_1(\theta_1)$ with distribution $F_1(y_i; x_i, \theta_1)$ conditional on the value of x_i. These values of x_i belong to a design region $\mathcal{X}$. The sample $\mathcal{S}_2$ of n_2 observations from $M_2(\theta_2)$ has conditional expectation $\mathrm{E}(y; x_i, \theta_2)$. Some values of x_i from $\mathcal{S}_2$ may belong to $\mathcal{X}$. We define the indicator

$$I_{i,\gamma} = \begin{cases} = & 1 \text{ if } F_1^{-1}(\gamma/2; x_i, \theta_1) < \mathrm{E}(y; x_i, \theta_2) < F_1^{-1}(1-\gamma/2; x_i, \theta_1) \quad i \in \mathcal{S}_2, x_i \in \mathcal{X} \\ = & 0 \text{ otherwise.} \end{cases} \tag{3}$$

The index is a function of both θ_1 and θ_2 and we examine it over a set of parameter values Θ_1 and Θ_2. For a particular set of parameter values $\theta_{1,k}$ and $\theta_{2,k}$ the overlapping index is defined as

$$O_{\gamma,k} = \sum_i I_{i,\gamma,k} \;\; i \in \mathcal{S}_2. \tag{4}$$

With $M_1(\theta_1)$ normal theory regression, we are therefore counting the total number of observations in $\mathcal{S}_2$ for which $x_i \in \mathcal{X}$, the conditional medians of which lie in a strip around the expectation of $M_1(.)$. As γ decreases, the strip becomes broader in y. If also for all $i \in \mathcal{S}_2, x_i \in \mathcal{X}$, then $O_{\gamma,k} \rightarrow n_2$, the number of observations in $\mathcal{S}_2$.

It is informative to keep θ_1 fixed and to vary θ_2 in a smooth way with a parameter $\lambda \in \mathfrak{R}$. Then we look at a set of indexes

$$\mathcal{O}_\gamma(\lambda) = \{O_{\gamma,k}\} \;\; \theta_1 \in \Theta_1 \text{ and } \theta_{2,k} \in \Theta_2(\lambda). \tag{5}$$

In particular we vary θ_2 linearly using the combination

$$\theta_{2,k} = \lambda_k \theta_2^0 + (1-\lambda_k)\theta_2^1 \quad (-\infty < \lambda_k \in \Lambda < \infty). \tag{6}$$

The set Λ of values considered is problem dependent. With $\theta_2^0 = \theta_1$ the centre of M_2 passes through that of M_1. Other choices of θ_2^0 can produce a trajectory in which the observations y_2 are always outlying. Our examples show how the variance and bias of the parameter estimates change in a smooth way with λ, but in different and informative ways for different estimators.

The contamination M_2 in our examples comes from a multivariate normal distribution. In the Appendix we show how to calculate the probability of intersection between this distribution and a strip around the regression plane. We call this the theoretical overlapping index. Although it ignores $\mathcal{X}$ it does signal cases where y_2 lies close to the regression line, even if remote from $\mathcal{X}$. These observations would then be "good" leverage points, in the sense that they improve the estimates of the regression parameters. For counting vertical outliers we need observations that lie in $\mathcal{X}$. These are signalled by the index defined in (4), which has to be calculated by simulation. We therefore call this the empirical index.

5 The Numerical Effect of Overlap: Normal Contamination

Because of the flexibility of our systematic approach, we can potentially cover a wide range of possibilities. Here we look at three numerical examples with normal contamination. In the next section we consider point contamination. We look at boxplots of the estimates over a suitable Λ and relate these plots to the overlapping indices. We separate out the variance and bias components of the estimates and compare these through cumulative plots over Λ. Finally, we compare the estimators for their power of detecting outlying observations, that is those that come from Model 2. The detection of outliers is particularly important if we require an indication that other methods of data analysis are appropriate.

In our one-variable regression examples M_1 is the regression model $y_i = \alpha + \beta x_i + \epsilon_i$, with the independent $x_i \sim U(a, b)$, these values generated once for all observations and values of λ. The variance of Y is σ_ϵ and overlapping indices were calculated for a strip of width $\pm 2\sigma_\epsilon$ around E(Y).

The expectation of x is $\mu_x = (a + b)/2$. The bivariate normal distribution for M_2 has mean μ and variance Σ given by

$$\mu = \begin{pmatrix} \alpha + \beta(\mu_x + d) \\ \mu_x + d \end{pmatrix} \lambda + \begin{pmatrix} \mu_2 \\ \mu_2 \end{pmatrix} (1 - \lambda) \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 \end{pmatrix}, \tag{7}$$

where the first component corresponds to the response. When $\lambda = 1$ the centres of the two populations are identical when the displacement $d = 0$.

Example 1. We took $n_1 = 100$ with $\alpha = 10, \beta = 3, \sigma_\epsilon = 10, a = 0$ and $b = 10$. For the second population, $n_2 = 30, \sigma_1 = 20, \sigma_2 = 2$ and $\mu_2 = 10$. Also $d = 0$ so the centres coincide at $\lambda = 1$. There were 100 simulations for each value of λ.

Figure 2 shows nine typical simulated data sets. As λ increases from -3 to 4, the centre of M_2 passes through that of M_1, at which point there is almost com-

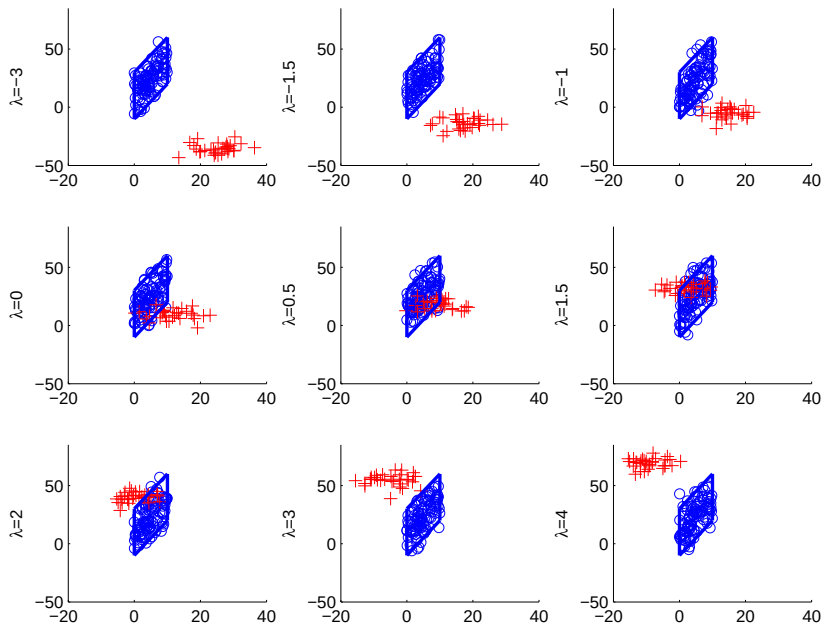

Figure 2: Example 1. Typical simulated data sets with $n_1 = 100$ and $n_2 = 30$ for nine values of λ. As λ increases observations from M_2 become close to those from M_1 and then become remote again. The parallelogram defines the region for the empirical overlapping index

plete overlapping of the observations from the two populations. That the overlap is not complete is shown by the plots of the indices in the upper panel of Figure 3, the maxima of which are less than one. The theoretical index is slightly higher than the empirical index as there is some probability of observations falling within the band of y values that are not in $\mathcal{X}$. On the other hand, the plot of the squared Mahalanobis distance from the mean of M_2 to that of M_1 has a minimum of zero, showing identity of the two centres.

We now consider the effect of these data configurations on the estimation of β. The left-hand panels of Figure 4 show boxplots, from 100 simulations, of the values of the five estimators for a series of values of λ, together with a typical data configuration for each. For $\lambda = -3$ observations from M_2 lie below and to the right of those from M_1. If these outliers are not identified the slope of the line is decreased. The boxplots all show some simulations where such estimates occur. LTS has the highest variance amongst the estimators in the main part of the boxplot, that is for the estimates when all outlying observations are rejected, with S the second most variable. For $\lambda = -1$, LTSr and MM are most affected by the outliers. The value $\lambda = 1$ corresponds to virtually complete overlap of the two groups. All methods, on average, give estimates that are biased downwards.

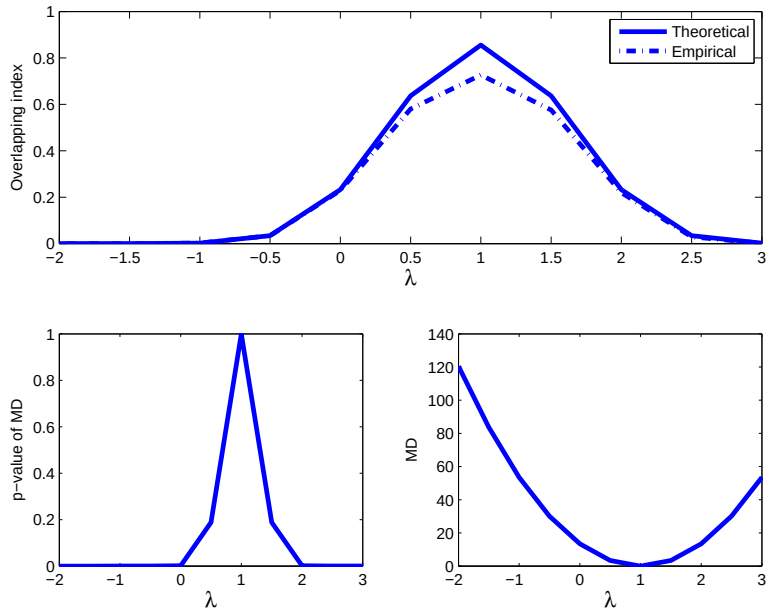

Figure 3: Example 1. Upper panel: theoretical and empirical overlapping indices for the data in Figure 2, showing maxima at $\lambda = 1$. Lower panel: squared Mahalanobis distance of M_1 from M_2 (right) and corresponding p-values (left)

However, those for LTS and S are both more variable and more biased. In the last panel, for $\lambda = 3$, the outliers are not as well separated as they are in panel 1. LTSr now has appreciable negative bias, due to the inclusion of outliers in the reweighting stage.

Figure 5 provides a powerful summary of the results on the variance and bias of the estimates of α and β as λ varies. The left-hand panels show the partial sums of the squared bias over Λ and the right hand panels show the partial sums of the variances. The values for α are in the top row and those for β in the bottom row.

The plots illustrate the trade off between bias and variance for some of the estimators. For values of λ up to three or so, LTS and S have the highest variances and the lowest biases and have very similar properties. Over the same range LTSr and MM have high biases and low variances. The effect of the modification of LTS to LTSr and S to MM has, in general, been to reduce variance at the cost of an increase in bias. The bias values for FS are in between those of these two groups, but closer to the lower pair of values, especially for estimation of β. The variance of FS is close, and ultimately less than, the low values for LTSr and MM.

The bottom right panel of Figure 4 shows that for $\lambda = 3$, the outliers are becoming distinct from y_1. As λ increases further the two groups become increasingly distinct, an effect that is evident in Figure 5. For the extreme values of λ, the

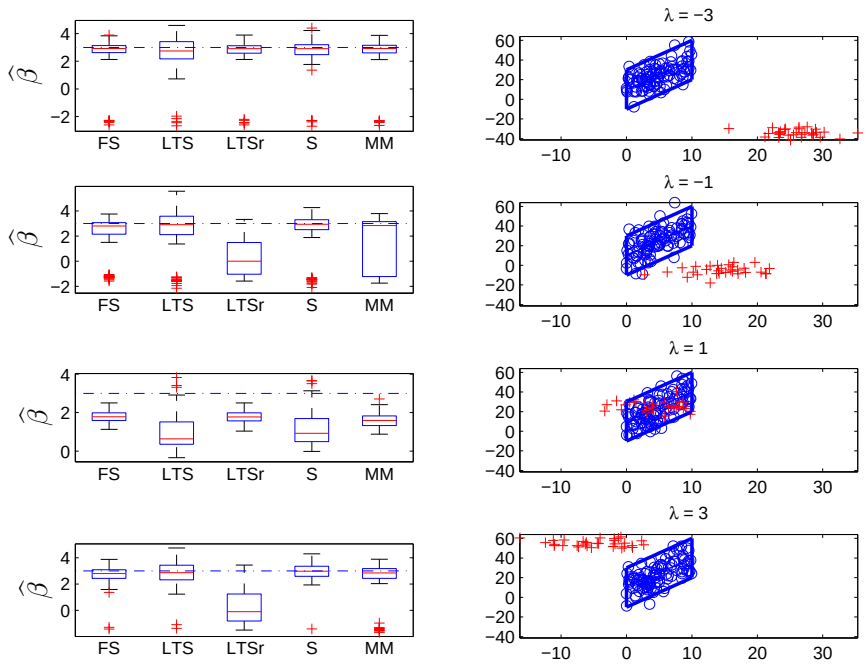

Figure 4: Example 1. Four simulated data sets for $\lambda = -3, -1, 1$ and 3. Left-hand panels: boxplots, from 100 simulations, of estimates of β (dotted line: $\beta_1 = 3$) for FS, LTS, LTSr S and MM estimators. Right-hand panels: typical simulations for these four values of λ

horizontal value of the summed squared bias for all estimators shows that the bias is zero. The two populations are sufficiently far apart that the asymptotics defining high breakdown apply. This is achieved for slightly less separation by MM than LTSr. The plots of partial sums of variances, on the other hand, increase steadily, since the estimators are always subject to the effect of the random variability in the observations. The sums of variances for S and, particularly, LTS are, however, increasing more rapidly at the ends of the region than those for the other three methods, a result in line with the rows of boxplots for $\lambda = \pm 3$ in Figure 4.

These plots illustrate the differing performance of the five estimators. Since this is a paper about robust statistics we also looked at plots in which the variance of the estimators was replaced by the average median absolute deviation from the median. These plots were close to those of the variances shown here.

In addition to good parameter estimates we would also like our estimate to signal the presence of outliers if the model fitted to the data is incorrect. Accordingly, we calculated the average power, that is the average number of observations correctly detected as being contaminated, which is the average number of detected observations from M_2. In testing for the presence of outliers we used a test of

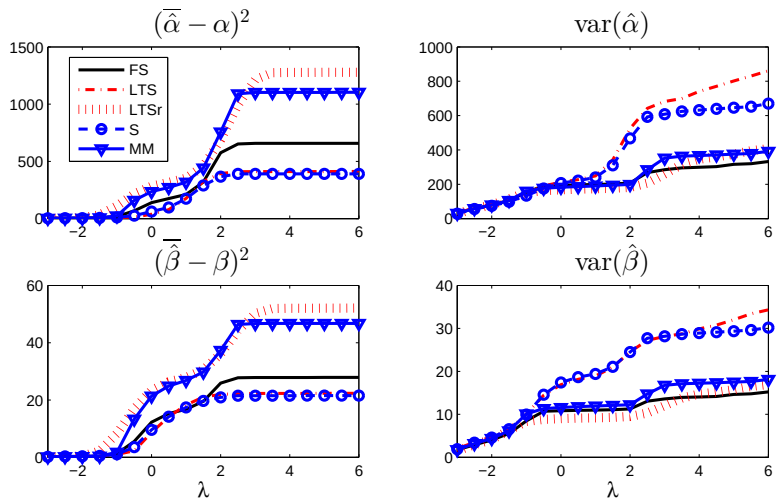

Figure 5: Example 1. Partial sums over Λ of simulated squared bias and variance of the five estimators. Left-hand panels squared bias, right hand panels variance. Top line $\hat{\alpha}$, bottom line $\hat{\beta}$.

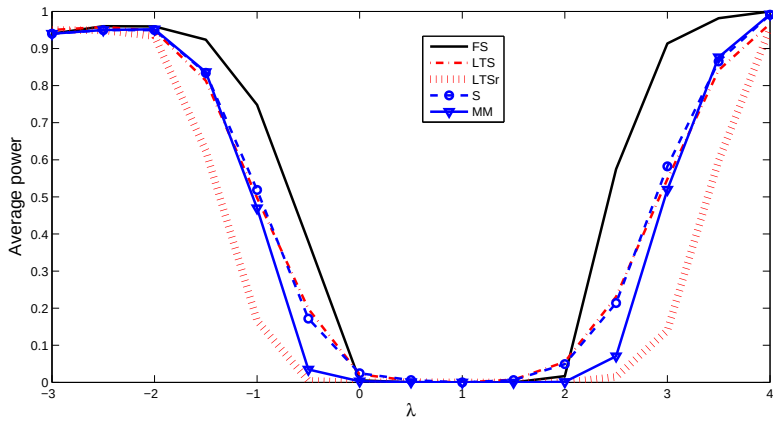

Figure 6: Example 1. Simulated average power of the five procedures over Λ

Bonferronised size α^*. The results are in Figure 6. Outliers are not detected for central values of λ as the parameter estimates are sufficiently corrupted by observations from M_2 that no observations appear outlying. As the means of the two populations move apart, the number of outliers detected increases. Over most of the range FS has the highest power and LTSr the lowest. The other three estimates lie between these extremes with MM having lower power for values of λ near zero. As with any power curves calculated for tests whose exact sizes are not known, we need to calibrate these findings against the size of the tests (see §7).

Example 2. In the interests of space we present only a part of our results,

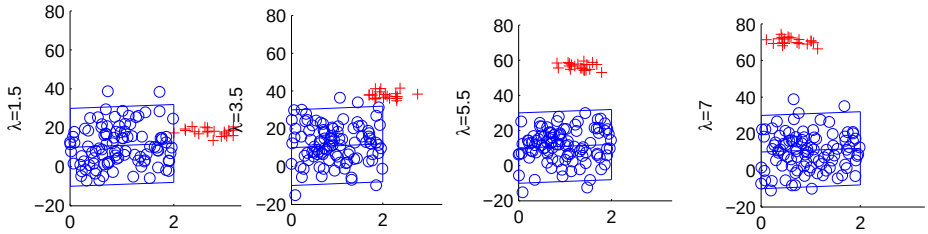

Figure 7: Example 2. Simulated data sets with $n_1 = 100$ and $n_2 = 20$ for four values of λ. As λ increases observations from M_2 become close to those from M_1 and then become remote again. The parallelogram defines the region for the empirical overlapping index

leaving the remainder for the online supplement.

We stay with a single explanatory variable but now choose a trajectory for λ such that $\theta_2^0 \neq \theta_1$, so that most of the observations y_2 are outlying. The parameter values for population 1 were $b = 2$ and $\beta = 1$. For population 2, $\Sigma = \text{diag}(4, 0.1), \mu_2 = 3.4$ and $d = 2$ so that the centres no longer coincided. Also $n_2 = 20$. Figure 7 shows scatterplots of typical samples for four values of λ. In the first, for $\lambda = 1.5$, there is a set of horizontal outliers, which can be expected not appreciably to affect the estimate of slope. As λ increases the observations from M_2 rise above those from M_1, generating increasingly remote vertical outliers.

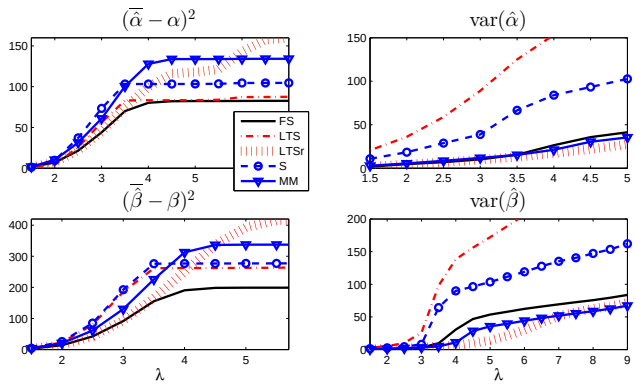

Figure 8: Example 2. Partial sums over Λ of simulated squared bias and variance of the five estimators. Left-hand panels squared bias, right hand panels variance. Top line $\hat{\alpha}$, bottom line $\hat{\beta}$.

The behaviour of the five estimators for this new situation is summarized in the partial sum plots of Figure 8. The plots of variances are simply interpreted: S and LTS have high variance for both α and β over the whole range of λ with MM

and LTSr having low values which are slightly less than that of FS.

The comparison of biases is less straightforward. The scatterplots of Figure 7 suggest that the two populations should be adequately separated by the time $\lambda = 4$. For lower values of λ, S and LTS have similar higher biases for β. The biases for α do not show much difference for lower values of λ. In the right-hand halves of the plots in Figure 8, with $\lambda > 4$, the two populations are more separated. The plots of bias show that S and LTS provide unbiased estimates (horizontal plots) for smaller values of λ than does MM. The LTSr estimates are not unbiased, even for the largest values of λ. The FS has excellent properties; it has the lowest bias for both parameters and a variance which is close to those from MM and LTSr.

The plot of average power for this example in Figure 19 of Riani, Atkinson, and Perrotta (2013) leads to similar conclusions to those for Example 1 in Figure 6. FS has the highest power and LTSr the lowest, but now the difference between FS and the other rules is much greater. S and MM have indistinguishable performances with LTS closer to that of LTSr.

Example 3. The third example had five explanatory variables ($p = 6$), independently uniformly distributed on $(0, 2\sqrt{10})$ with regression parameters $\beta = 5$ for all variables and $n_1 = 200$. For population 2, $\Sigma = \text{diag}(100, I_5)$, $\mu_2 = 3$, $d = 2$ and $n_2 = 60$.

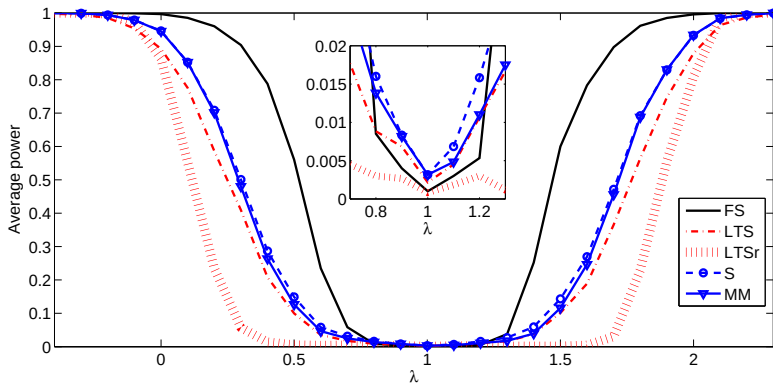

Figure 9: Example 3. Simulated average power of the five procedures over Λ with an inset zoom of the central part of the figure

This is a larger example, with $n_1 = 200$ and $n_2 = 60$. As λ increases from -1 to 2.6 the outliers "rise through" the central observations. Since $d \neq 0$, the centres of the two distributions are never identical. Unlike our other two examples, this one does not include outliers at leverage points, so that the differences in behaviour of the methods are, to some extent, reduced.

With five explanatory variables the major contribution to the mean squared error of the parameter estimates comes from β, so we only consider these values,

which are plotted in Figure 21 of Riani, Atkinson, and Perrotta (2013). With independent x_j the bias and variance are the sums of those for the individual components. LTS behaves surprisingly poorly, with the uniformly highest bias and variance . LTSr and S have medium behaviour for both properties, with the order reversed for bias and variance, while MM and FS have the same, lowest values for bias and similar values for variance until $\lambda = 1$ when that for FS increases, although staying below that for S. Unlike the other two examples, the relative behaviour of the estimators is little affected by the value of λ, a reflection of the stability of the outlier pattern over Λ. Of course, the magnitude of the outliers is largest for extreme values, but leverage points are not introduced or removed.

The plot of average power is in Figure 9. As in the other plots of average power, FS has the highest power and LTSr the least. The other three estimators have very similar properties to each other. However, in assessing power we need to be sure that we are comparing tests with similar sizes. The zoom in the centre of the plot for values of λ close to one shows that we are not, with FS and LTSr having the smallest values. For accurate comparisons we need to scale the other three tests downwards, which will reduce the curves below the plotted values. However, even when $\lambda = 1$, outliers are still present and, since $d \neq 0$, we are not looking at the null distribution of the test statistics. We consider null distributions and the resulting size of tests in §7.

6 The Numerical Effect of Overlap: Point Contamination

Point contamination plays an important role in the theory of robust estimation, for example in finding conditions of maximum bias in regression (Martin, Yohai, and Zamar 1989; Berrendero and Zamar 2001). Accordingly, we extend our simulations to such contamination. Although it is a special case of (7) as $\Sigma \to 0$, there are new features.

The first feature is the response of the FS algorithm to several identical observations. As the search progresses observations are not only added to the subset used in estimation, but remote observations are deleted. If several of the identical observations are included, those outside the point contamination will seem remote and the search will collapse, since the fitted model will be singular. If such singularity occurs, we identify all identical observations and force them to enter at the end of the search. As the figures show, in some cases this has a powerful beneficial effect on the estimates. The second feature is that the overlapping index now has the value of either zero or one.

We took 100 x values between 0 and 1, with the normally distributed values

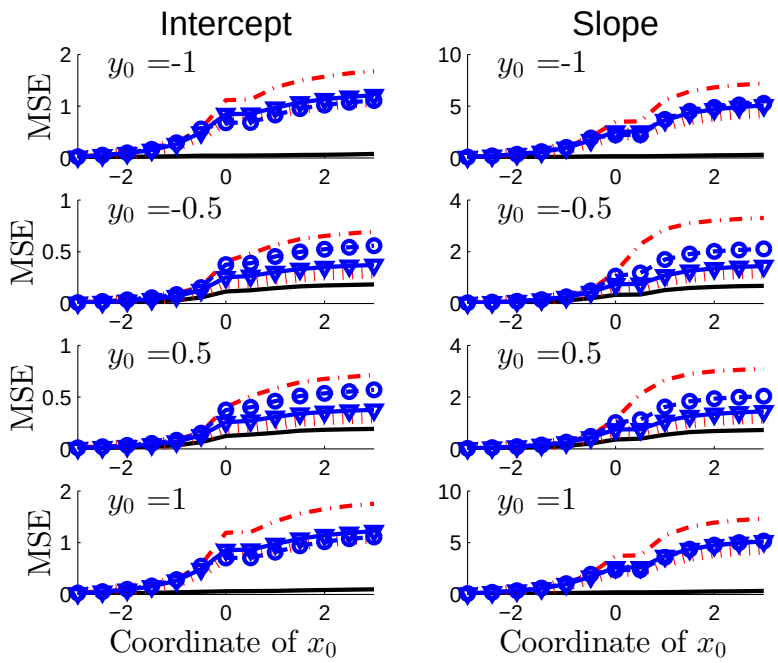

Figure 10: Point contamination at (x_0, y_0). Partial sums of mean squared errors of estimates of α and β for five values of y_0 as x_0 varies from -3 to 3

of y such that approximately 95% lay between -0.5 and 0.5. We add 30 identical contaminating observations at (x_0, y_0) where both the vertical and horizontal directions of contamination range from -3 to 3.

Figure 10 shows plots of the partial sums of the mean squared error of the estimates of the intercept and slope for four values of y_0 over a fine grid of values of x_0 from -3 to 3. The most notable features are the poor performance of LTS and the good performance of FS. This is particularly striking in the more extreme vertical contaminations, $y_0 = \pm 1$, where the FS estimates are virtually unaffected by the thirty outliers.

In Figure 11 we look at the same quantities as y_0 varies for four fixed values of x_0: -1, -0.5, 0.5 and 1. Recall that the values of x range from 0 to 1, so these plots are not symmetrical around $x_0 = 0$. The most striking feature is the excellent performance of the FS, which is by far the best except when the contamination passes through the centre of $\mathcal{X}$; even then it is slightly better than MM and S. LTS behaves particularly poorly when $x_0 \in \mathcal{X}$, but is uniformly poorest. S and MM are similar, and slightly better than LTSr.

The use of point contamination allows sharp comparison of the algorithms for very robust regression. In the more diffuse situations of §5 the plots of the power curves, such as those of Figure 9, help to strengthen the comparisons.

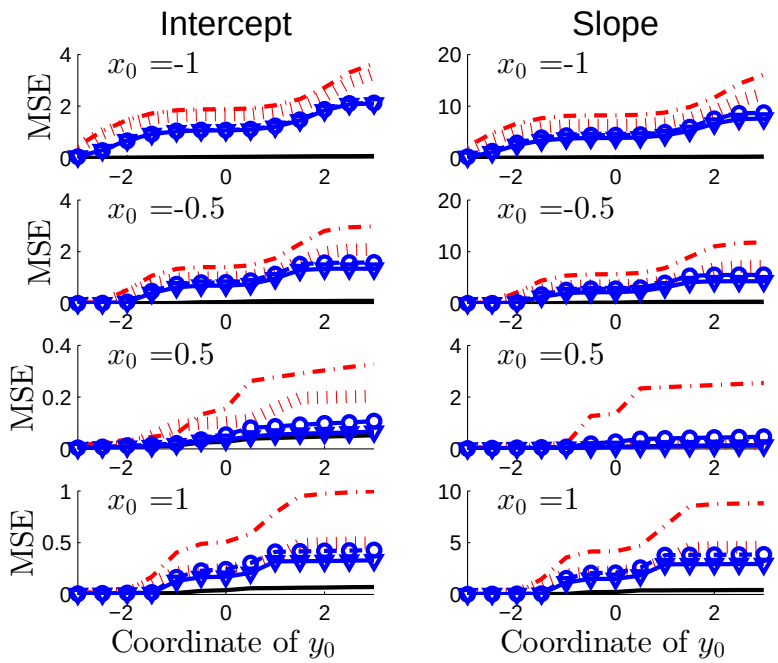

Figure 11: Point contamination at (x_0, y_0). Partial sums of mean squared errors of estimates of α and β for five values of x_0 as y_0 varies from -3 to 3

With this two dimensional model for contamination it is possible to explore the properties of the estimators over a grid of values for (x_0, y_0). With higher dimensional problems, such as Example 3, we will again need to construct a trajectory Λ along which the point contamination moves.

7 Size Comparisons

In order to establish the size of the outlier tests we ran simulations for sample sizes n from 100 to 1,000 for several different dimensions of problems. The results for $p = 6$ and 11 are in Figure 12. In the simulations the samples were allowed to grow with n, so that samples for larger values of n contained those for smaller, leading to smoother curves. Both the response and the explanatory variables were simulated from independent standard normal distributions, with all regression coefficients set to one. Since all methods are affine equivariant, these arbitrary choices do not affect the results. For each value of n we present the average of 10,000 simulations, in which we counted the number of samples declared as containing at least one outlier, with the tests conducted at the 1% Bonferronised level.

The figure shows that, for three out of the five rules, the sizes are very far from

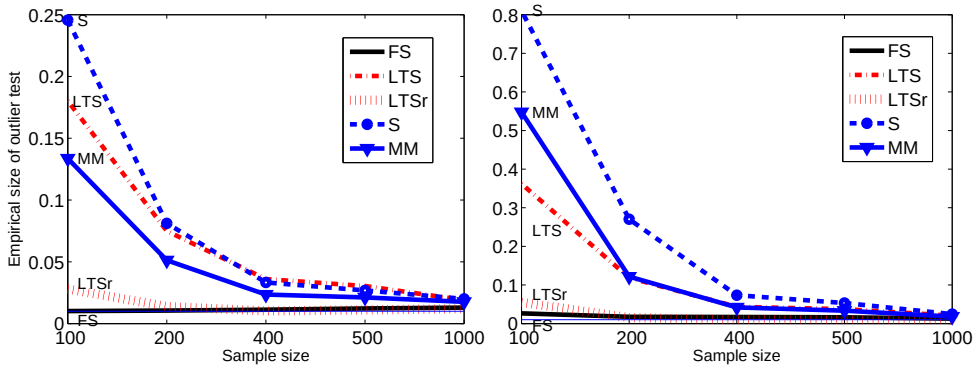

Figure 12: Size of nominally 1% Bonferronized outlier tests for, left-hand panel, $p = 6$ and for $p = 11$. Note the different vertical scales in the two panels

the nominal value of 1%. For $n = 100$ the sizes For MM, LTS and S when $p = 6$ range between 0.13 and 0.25. For $p = 11$ the range for these rules is 0.36 to 0.81. The sizes decrease with n, but are even so still around 2% for these rules when $n = 1,000$. The size for LTSr is closer to nominal, being around 3% and 6% for $n = 100$ and decreasing rapidly with n. Only FS has a size around 1% for both values of p and all n.

These calculations of size show that FS is correctly ordered as having highest power. The curves, such as those in Figure 9, for LTSr do not need appreciable adjustment for size. However size adjustment for MM, LTS and S may well lead to procedures with less power than LTSr.

A simple method of adjusting power for size is a normal, or logistic, plot of the power curves, as in Figure 8.12 of Atkinson (1985), when the slope of the curve indicates power and the intercept size. Although such a comparison would be possible here, our purpose is not to establish the exact properties of outlier tests. Rather we are concerned with introducing a general framework for the comparison of methods for very robust regression.

8 Trade data again

We began our discussion of very robust regression in §2 with the trade data plotted in Figure 1. We now conclude with a plot of the fitted lines and of the outliers identified by the five methods we have been comparing. The results are in Figure 13 where the top-left hand panel repeats the plot of the data. The other panels show that FS, LTS and S all provide fits to the lower of the two lines evident for the higher values of value and quantity. The other two methods, reweighted LTS and MM provide fitted lines which lie more between the groups. Only FS indi-

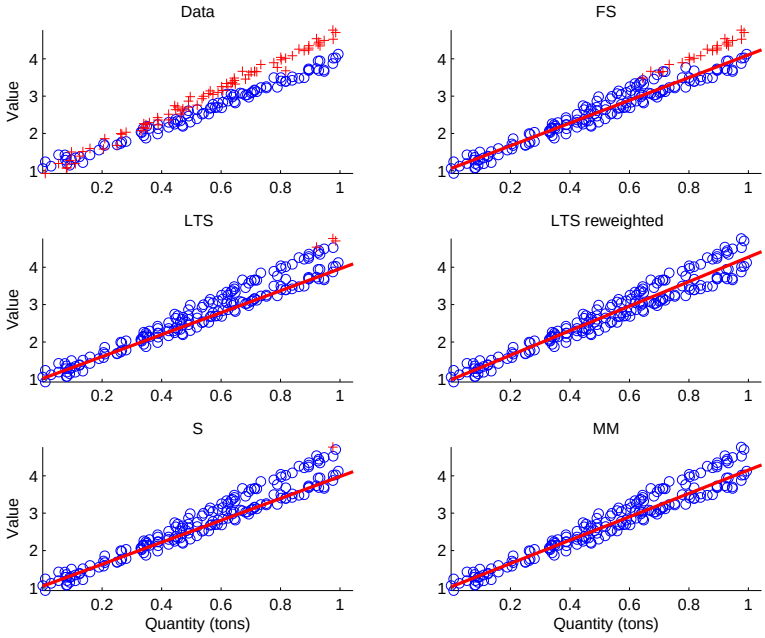

Figure 13: Trade data: results of five very robust analyses. Reading across: ○ simulated regression data and outliers +; fitted FS line and outliers +; LTS, reweighted LTS, S and MM estimators

cates that there are a large number of outliers which might perhaps be modelled separately. These results are in line with the conclusions to be expected from the simulation results of earlier sections, particularly the low power of the outlier tests for all except FS. However, the power comparisons combined with the size calculations of §7, show that we cannot change the level of the tests without damaging the size of the test when there no outliers and so identifying far too may outliers in the null case.

9 Discussion

The largest contrast between estimators is shown in the figures for point contamination of §6. The relatively poor behaviour of LTS recalls the impression of Cook and Hawkins (1990, §6) that the related MVD method for multivariate data finds "Outliers Everywhere". The superior performance of FS comes from the data-dependent flexibility of the number of observations included in the final fit.

Several authors, for example Cook, Hawkins, and Weisberg (1993) and Hawkins and Olive (2002), have commented on the persistence of the effects of the initial estimator, even asymptotically. The FS escapes such persistence because,

although the subset used in fitting grows in size, observations can be deleted as well as added. This provides the algorithmic flexibility that leads to such good performance in §6. In addition, the flexibility of the FS combined with the plotting of diagnostic measures makes possible the detection of subpopulations in the data, not just the point contamination of §6. An example of cluster detection is shown in Figure 10 of Atkinson and Riani (2007).

There is also some theoretical explanation for the relative behaviour of the other estimators. In particular, the MM estimator is intended to improve the efficiency of the S estimator and, indeed, this estimator has a lower variance in Examples 1 and 2. But this is achieved at the cost of having higher bias than the S estimator. The same is true for the comparison of LTSr and LTS. For those values of (x_0, y_0) in §6 for which $x_0 \in \mathcal{X}$, so that there are no leverage points to introduce serious biases, LTSr and MM are respectively an improvement on LTS and S.

We have illustrated the use of our framework for comparing FS with methods designed to have a breakdown of 50%. Of course, the framework can be used for comparisons with breakdown levels more likely to be used in practice, such as 70 or 80%. The properties of FS, since they do not depend on a specified breakdown level, will not be changed.

Acknowledgments The work on this paper was jointly supported by the project MIUR PRIN *MISURA - Multivariate models for risk assessment*, by the JRC Institutional Work Programme 2007–2013 of the SITAFS Research Action, and by the OLAF-JRC project *Automated Monitoring Tool on External Trade*. Much of it was completed at the Isaac Newton Institute for Mathematical Sciences in Cambridge, England, during the 2011 programme on the Design and Analysis of Experiments.

We are most grateful to the referees whose thoughtful and detailed comments led to improvement and clarification of our paper.

Supplementary Material

Supplement to "A parametric framework for the comparison of methods of very robust regression" (DOI:?????). Riani, Atkinson, and Perrotta (2013) includes further analyses of data. The first is a second motivating example; the other two are expanded versions of our analyses of Examples 2 and 3 in the paper. This material is also available at http://www.riani.it/pub/RAP13supp.html, together with further, dynamic graphics and links to the programs used to generate the results in our paper.

References

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

Atkinson, A. C. (1985). *Plots, Transformations, and Regression*. Oxford: Oxford University Press.

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

Atkinson, A. C. and M. Riani (2006). Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics 15*, 460–476.

Atkinson, A. C. and M. Riani (2007). Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis 52*, 272–285. doi:10.1016/j.csda.2006.12.034.

Atkinson, A. C., M. Riani, and A. Cerioli (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.

Berrendero, J. R., B. V. M. Mendes, and D. E. Tyler (2007). On the maximum bias functions of MM-estimates and constrained M-estimates of regression. *Annals of Statistics 35*, 13–40.

Berrendero, J. R. and A. H. Zamar (2001). Maximum bias curves for robust regression with non-elliptical regressors. *Annals of Statistics 29*, 224–251.

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

Cook, R. D., D. M. Hawkins, and S. Weisberg (1993). Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. *Statistics and Probability Letters 16*, 213–218.

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

García-Escudero, L. A., A. Gordaliza, A. Mayo-Iscar, and R. San Martin (2010). Robust clusterwise linear regression through trimming. *Computational Statistics and Data Analysis 54*, 3057–3069. doi:10.1016/j.csda.2009.07.002.

Hampel, F., E. M. Ronchetti, P. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics*. New York: Wiley.

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

Hawkins, D. M., D. Bradu, and G. V. Kass (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics 26*, 197–208.

Hawkins, D. M. and D. J. Olive (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.

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

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

Martin, R. D., V. J. Yohai, and A. H. Zamar (1989). Min-max bias robust regression. *Annals of Statistics 17*, 1608–1630.

Morgenthaler, S. (2007). A survey of robust statistics. *Statistical Methods and Applications 15*, 271–293. Erratum: 16, 171–172.

Perrotta, D., M. Riani, and F. Torti (2009). New robust dynamic plots for regression mixture detection. *Advances in Data Analysis and Classification 3*, 263–279. doi:10.1007/s11634-009-0050-y.

Pison, G., S. Van Aelst, and G. Willems (2002). Small sample corrections for LTS and MCD. *Metrika 55*, 111–123. doi: 10.1007/s001840200191.

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

Riani, M., A. C. Atkinson, and D. Perrotta (2013). Supplement to “A parametric framework for the comparison of methods of very robust regression”.

Riani, M., D. Perrotta, and F. Torti (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.

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

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

Rousseeuw, P. J. and K. Van Driessen (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery 12*, 29–45.

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

Torti, F., D. Perrotta, A. C. Atkinson, and M. Riani (2012). Benchmark testing of algorithms for very robust regression: FS, LMS and LTS. *Computational Statistics and Data Analysis 56*, 2501–2512. doi:10.1016/j.csda.2012.02.003.

Appendix A: The Theoretical Overlapping Index

The response and the explanatory variables lie in a space of dimension $p+1$. Let these variables be w. Then the regression plane can be written as $b^T w - c = 0$. The equation of the normal to the plane through a point w_0 on the plane is

$$z_1 = w_0 + bd, \tag{A1}$$

where the scalar d is the distance from the plane. The outlying observations, including the response, have a multivariate normal distribution. Let these be $W \sim \mathcal{N}(\mu, \Sigma)$. We require the probability that W lies on one side of the plane. To obtain this rotate W to a set of variables Z with z_1 (A1) the normal to the plane. Integrating out the other p variables shows that the required probability comes from the marginal distribution of $Z_1 \sim \mathcal{N}(b^T\mu, b^T\Sigma b)$. Let the distance in the z_1 direction from μ to the plane be $d(c)$. Then, from (A1), at the plane $b^T w = c = b^T\mu + b^T b d(c)$, so that

$$d(c) = (c - b^T\mu)/b^T b. \tag{A2}$$

Since the distance $d(c)$ in the z_1 direction has been rescaled by the factor $1/b^T b$ the required probability is

$$\Pr\,(b^T W > c) = \Pr\,(Z_1 > c - b^T\mu) = \Phi\{d(c) b^T b/(b^T\Sigma b)^{0.5}\} = \Psi(c), \text{ say}, \tag{A3}$$

where Φ is the cdf of the (univariate) standard normal distribution. We require this probability in terms of the regression model, which we now write as $y = \alpha + \beta^T x$. Then

$$b^T = (1 \quad -\beta^T), \quad w^T = (y \quad x^T) \quad \text{and} \quad c = \alpha.$$

Finally, we require the probability that W lies between two planes. For any x the required strip around this model is $y \pm 2\sigma_\epsilon$. The two planes then are defined by constants $c^+ = \alpha + 2\sigma_\epsilon$ and $c^- = \alpha - 2\sigma_\epsilon$. From (A3) the required probability is $\Psi(c^+) - \Psi(c^-)$.