Calibrated Very Robust Regression

Marco Riani
Dipartimento di Economia, Università di Parma, Italy[*]
Anthony C. Atkinson
The London School of Economics, London WC2A 2AE, UK[†]
and Domenico Perrotta
European Commission, Joint Research Centre, Ispra, Italy[‡]

December 14, 2011

Abstract

The behaviour of algorithms for very robust regression depends 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

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.

[*]e-mail: mriani@unipr.it
[†]e-mail: a.c.atkinson@lse.ac.uk
[‡]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 §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 §2 we introduce a parameterised framework, 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 §4. 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 points virtually on a line to a seemingly random scatter around the regression plane. 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 §4 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 §5.

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 FS is the preferred method of very robust regression, a conclusion particularly strongly supported by the power and size of test curves. Among our concluding comments in §6 we mention procedures for insightful methods of data analysis, even if no group of observations is in the majority.

2 Models, Data and Robustness

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, $1-\Delta$, where $0 < \Delta < 0.5$, is generated by the model $M_1(\theta_1)$ and the remaining part Δ of the data is generated by the model $M_2(\theta_2)$. In the absence of outliers, that is when $\Delta = 0$, 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. Figure 2 of Rousseeuw (1984) is a paradigmatic example. There is no difficulty in having $M_1(\theta) = M_2(\theta)$ but then we must have $\theta_1 \neq \theta_2$.

The breakdown point of an estimator $\tilde{\theta}$ is the largest value of Δ for which $\tilde{\theta}$ provides information about θ, remaining bounded and bounded away from the edge of the parameter space as the observations take any values (see Maronna et al. 2006, §3.2). The very robust regression procedures we compare have asymptotic breakdown points of 50%. This is customarily considered to be the maximum possible value, a point we return to in §6.

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 configurations when the observations are not so 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 methods. 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 $\to \infty$. Our proposed distance measure likewise provides a framework for comparison of regression procedures.

There is a sample $\mathcal{S}_1$ of n_1 observations from $M_1(\theta_1)$ with distribution $F_1(y_i; x_i, \theta_1)$. 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 expectation $\mathrm{E}(y; x_i, \theta_2)$. Some values of x_i from $\mathcal{S}_2$ may belong to $\mathcal{X}$. Let

$$y(\gamma) = F_1^{-1}(\gamma; x_i, \theta_1).$$

Then we let the indicator

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

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

$$O_{\gamma,k} = \sum_i I_{i,\gamma,k} \;\; i \in \mathcal{S}_2. \tag{2}$$

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 expectations of which lie in a strip around the expectation of $M_1(.)$. As γ decreases, the strip becomes broader. If also for all $i \in \mathcal{S}_2, x_i \in \mathcal{X}$, then $O_{\gamma,k} \to n_2$, 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

$$\mathcal{O}_\gamma(\lambda) = \{O_{\gamma,k}\} \;\; \theta_1 \in \Theta_1 \text{ and } \theta_{2,k} \in \Theta_2(\lambda). \tag{3}$$

In particular we vary θ_2 linearly using the combination

$$\theta_{2,k} = \lambda_k \theta_1^0 + (1-\lambda_k)\theta_2^0 \quad (-\infty < \lambda_k \in \Lambda < \infty). \tag{4}$$

The set Λ of values considered is problem dependent. With $\theta_1^0 = \theta_1$ the centre of M_2 passes through that of M_1. Other choices of θ_1^0 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 (2), which has to be calculated by simulation. We therefore call this the empirical index.

3 Five Methods for Very Robust Regression

We compare and contrast the properties of five methods for very robust regression. The algorithms that we use are all publicly available from the Forward

Search Data Analysis (FSDA) Matlab toolbox at `www.riani.it/MATLAB`. 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.

We consider the usual regression model with random carriers (a point we come back to in §6) where we observe i.i.d. random vectors $(y_i,\ x_i^T) \in \Re^{p+1}$, $i = 1, \ldots, n$, where $y_i \in \Re$ and $x_i \in \Re^p$ satisfy

$$y_i = x_i^T \beta + u_i \qquad i = 1, \ldots, n. \tag{5}$$

The u_i are random errors independent from the covariates (x_i) which have common variance equal to σ^2 and β is the $p \times 1$ vector parameter of interest. Given an estimator of β, say $\hat{\beta}$, the residuals are defined as

$$r_i(\hat{\beta}) = y_i - x_i^T \hat{\beta}.$$

Traditional robust estimators attempt to limit the influence of outliers by replacing the square of the residuals in the estimation of β by a function ρ of the residuals which is bounded. The regression M-estimate of β is the value that minimizes the objective function

$$\sum_{i=1}^{n} \rho\{r_i(\beta)/\sigma\}. \tag{6}$$

Of the numerous form that have been suggested for $\rho(.)$ (Andrews et al. 1972, Hampel et al. 1986, Huber and Ronchetti 2009) perhaps the most popular choice is Tukey's Biweight function

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if } |x| \leq c \\ \frac{c^2}{6} & \text{if } |x| > c, \end{cases} \tag{7}$$

where c is a crucial tuning constant.

In equation (6) it is assumed that σ is known, yielding the estimate $\tilde{\beta}_M(\sigma)$. Otherwise, an M-estimator of scale $\tilde{\sigma}_M$ is defined as the solution to the equation

$$\frac{1}{n} \sum_{i=1}^{n} \rho\{r_i(\beta)/\sigma\} = K_c, \tag{8}$$

where both β and σ are iteratively jointly estimated. Although the ρ function used to obtain the scale estimator is not necessarily the same as that in (6), we again use

the biweight (7). If we take the minimum value of $\tilde{\sigma}_M$ which satisfies equation (8), we obtain the S-estimate of scale ($\tilde{\sigma}_S$) and the 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. For an asymptotic breakdown point of 50% for the S-estimator it is necessary that $2K_c = \rho(c)$. The value of K_c is determined by the requirement of a consistent estimator of scale when the observations are normally distributed. 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 in (8) 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 of 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.

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. 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 residuals.

In Least Trimmed Squares (LTS) (Rousseeuw 1984, p. 876) the search is over subsets finding $\tilde{\beta}_{\text{LTS}}$ to

$$\text{minimize } SS_{\text{T}}\{\hat{\beta}(h)\} = \sum_{i=1}^{h} r_i^2\{\hat{\beta}(h)\}, \tag{9}$$

where $\hat{\beta}(h)$ is the LS estimate of β for a subset of size h. With $h = [n/2] + [(p + 1)/2]$. LTS has an asymptotic breakdown point of 50%.

To increase efficiency, reweighted versions of LTS estimators can be computed. These reweighted estimators, denoted LTSr, are computed by giving weight 0 to observations which (9) suggests are outliers. 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 a Bonferronised size $\alpha^* = \alpha/n$, so taking the $1 - \alpha^*$ cutoff value of the reference distribution. In our calculations $\alpha = 0.01$.

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 follow from the results of Tallis (1963) on elliptically trimmed multivariate normal distributions.

4 The Numerical Effect of Overlap

Because of the flexibility of our systematic approach, we can potentially cover a wide range of possibilities. Here we look at three numerical examples, two with one explanatory variable and one with five variables. 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, such as those sketched in §6, 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

$$\mu = \begin{pmatrix} \alpha + \beta(\mu_x + d) \\ \mu_x + d \end{pmatrix} \lambda + \begin{pmatrix} \mu_2 \\ \mu_2 \end{pmatrix} (1-\lambda) \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 \end{pmatrix}, \tag{10}$$

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

In our first example 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.

In the second example only a few of the parameters were changed. In population 1 $b = 2$ 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 $n_2 = 20$.

The third example had five explanatory variables ($p = 6$), but the structure can also be explained in this framework. These variables were independently uni-

formly 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 $n_2 = 60$.

We start with a one-variable example in which $d = 0$, so that the two populations have the same mean when $\lambda = 1$. Figure 1 shows nine 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 complete 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 2, 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 3 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, more often for LTS than for the other estimators. More importantly, LTS has the highest variance amongst the estimators in the main part of the boxplot, that is when all outliers are rejected, with S the second most variable. FS, LTSr and MM have similar behaviour. For $\lambda = -1$, LTSr and MM are most affected by the outliers with S slightly more stable than FS. The value $\lambda = 1$ corresponds to virtually complete overlap of the two groups. All methods, on average, give estimates that are biased downwards. 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. To a lesser extent MM is also more prone to include outliers than the other methods, of which LTS has the highest variance.

Figure 4 provides a powerful summary of the results on the variance and bias of the estimates of β and also of α 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. Again this pair of estimators have very similar properties. 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 3 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 4. For the extreme values of λ, the 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 3.

These plots illustrate the differing performance of the five estimators. 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, that is the average number of observations from M_2. In testing for the presence of outliers we again used a test of Bonferronised size α^*. The results are in Figure 5. 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. This we do in §5.

As a second example we stay with a single explanatory variable but now choose a trajectory for λ such that $\theta_1^0 \neq \theta_1$, so that most of the observations y_2 are outlying. Figure 6 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. (In reading this plot, note the rescaling of x in the third and fourth panels.)

The difference between this example and Example 1 is made clear in the plot of the measures of overlap in the upper panels of Figure 8. The theoretical overlapping index has a value close to one as, for lower values of λ, observations from M_2 have a high probability of lying inside the strip around M_1. However, the empirical index has a lower value since, as the first panel of Figure 6 showed, few of these observations fall within $\mathcal{X}$. For larger values of λ both indices have values

close to zero. Since the centres of the two populations are never identical, the minimum value of the squared Mahalanobis distance is greater than zero.

The behaviour of the five estimators for this new situation is summarized in the partial sum plots of Figure 7. The scatterplots of Figure 6 suggest that the two populations should be adequately separated by the time $\lambda = 4$. For the moment we ignore FS and focus on these lower values of λ for which S and LTS have similar biases for β, but these are now the highest. The biases for α do not show much difference for lower values of λ. The plots of variances are much simpler to interpret: S and LTS have high variance for both α and β over the whole range of λ with MM and LTSr having low values. These two estimators also have similar low biases for β and much the same bias as all other rules for α.

This discussion is for that part of Λ for which the two populations are relatively close. The right-hand halves of the plots of bias show that S and LTS provide unbiased estimates for smaller values of λ than do MM. Since it will not be known when analysing data, how far away any contaminants are from the main population, it is sensible to choose an estimator that behaves well over all of Λ. In this example the FS again has excellent properties; it has the lowest bias for both parameters and a variance which is close to those from MM and LTSr. However, the plots of bias show that the values for LTSr do not become zero (remember these are partial sums that are plotted) even when the populations are very well separated.

To conclude the analysis of the second example we look at the plot of average power in the lower half of Figure 8. As in Figure 5, FS has the highest power and LTSr the lowest, but now the difference between LTS and the other rules is much greater. S and MM have indistinguishable performances with LTS closer to that of LTSr.

Our final example, which we treat more briefly, has five explanatory variables ($p = 6$). Typical scatterplots of y against each x are shown in Figure 9 for this 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, a feature more visible in the coloured pdf version of the paper. However, 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.

We summarize the behaviour of the five estimators in the partial sum of variance and bias plots of Figure 10. 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. With independent x_i the bias and variance are the sums of those for the individual components. The most obvious feature of the plot is the poor behaviour of LTS. LTSr and S have medium behaviour for bias and variance, with the order reversed for the two properties, 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 11. 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 $\lambda = 1$ the outliers are still present and, since $d \neq 0$, we are not exactly looking at the null distribution of the test statistics. We consider null distributions and the resulting size of tests in the next section.

5 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, so that larger values of n contained smaller ones, 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 the nominal value. 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 11, 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.

6 Discussion

In §3 we stipulated that the carriers be random. In defence Huber and Ronchetti (2009, p. 197) comment that "Some authors have made unqualified, sweeping claims that their favorite estimates have a breakdown point approaching 1/2 in large samples. Such claims may apply to random carriers, but do not hold in designed situations". They then use an argument based on a saturated D-optimum design with p support points (for example Atkinson, Donev, and Tobias 2007, p. 222) to show that the breakdown point is, at best, $1/(2p)$. They encourage the joint study of robustness and experimental design. An example is Müller (1997).

Our comparisons show that FS has the best performance of the five estimators, particularly, but not only, when we look at the comparisons of power. In making these comparisons we have used Bonferroni levels to determine outliers. Even so we do not avoid the impression given by LTS for large p and small n of finding "Outliers Everywhere" as Cook and Hawkins (1990, §6) did in their investigation of the related MVD method for multivariate data. We think that the superior performance of FS comes from the data-dependent flexibility of the number of observations included in the final fit. All algorithms use fits to subsets of p observations as, at least part, of the estimation method. In LTS all subsets are of size p. In LTSr subsets of size p determine the final subset of size $n - k$ (§3). S estimators are also found starting from p observations. In our examples, with up to 23% outliers we took K_c in the second stage to give an asymptotic nominal efficiency of 85%. 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. 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 behaviour is apparent in random start forward searches for both regression and multivariate data. Figure 10 of Atkinson and Riani (2007) shows the trajectories of 300 searches starting from randomly selected subsets of size p. For a single population, perhaps with some outliers, the subsets become the same (the trajectories converge) in the last 1/3 of

the search. If there are several populations, some trajectories are initially attracted to each, and families of trajectories reveal the populations. Such plots indicate that we do not have to have at least half of the observations coming from one population. Huber and Ronchetti (2009, p. 198) discuss methods when very high contamination or mixture models are an issue and suggest a straight data analytic approach through projection pursuit. We would argue that, in the regression context, the FS is a natural method for disentangling mixtures and revealing outliers, even when no population gives rise to a majority of the data.

Acknowledgment This paper was completed at the Isaac Newton Institute for Mathematical Sciences in Cambridge, England, during the 2011 programme on the Design and Analysis of Experiments.

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., A. N. Donev, and R. D. Tobias (2007). *Optimum Experimental Designs, with SAS*. 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.

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.

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. 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.

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

Müller, C. H. (1997). *Robust Planning and Analysis of Experiments*. Lecture Notes in Statistics 124. Berlin: Springer-Verlag.

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.

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.

7 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 $\Psi(c^+) - \Psi(c^-)$.

Figures

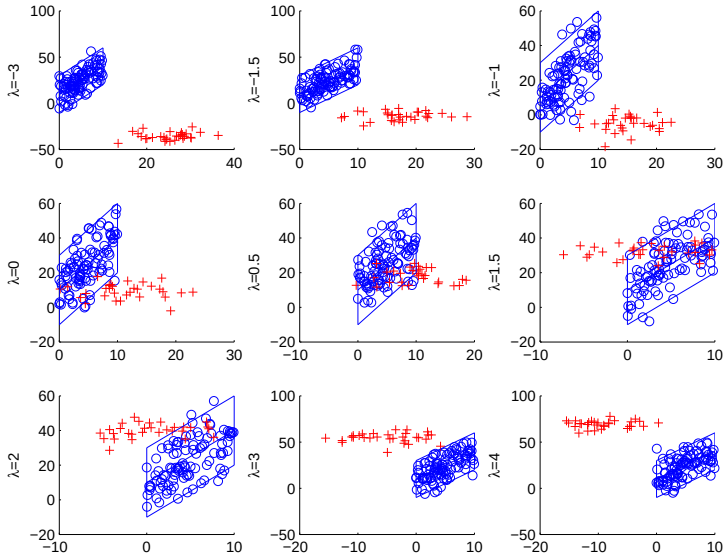

Figure 1: Example 1. Simulated data sets with $n_1 = 100$ and $n_2 = 30$ for nine values of λ. As λ increases observations from M_2 become close to those from M_1 and then become remote again. The parallelogram defines the region for the empirical overlapping index

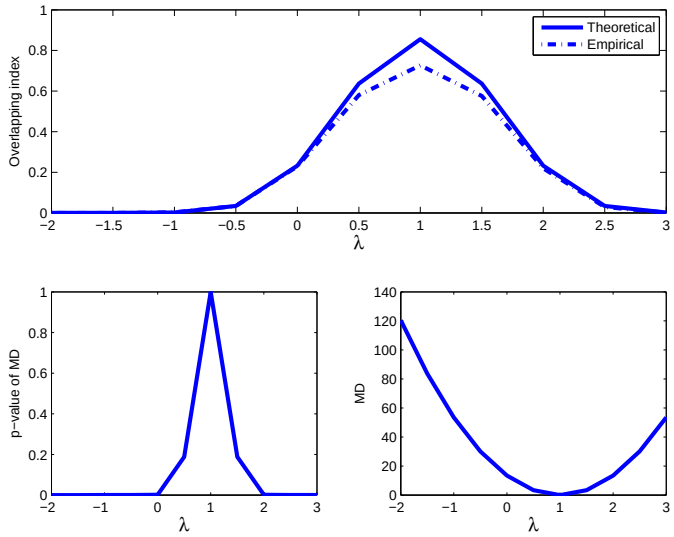

Figure 2: Example 1. Upper panel: theoretical and empirical overlapping indices for the data in Figure 1, showing maxima at $\lambda = 1$. Lower panel: squared Mahalanobis distance of M_1 from M_2 and corresponding percentage points

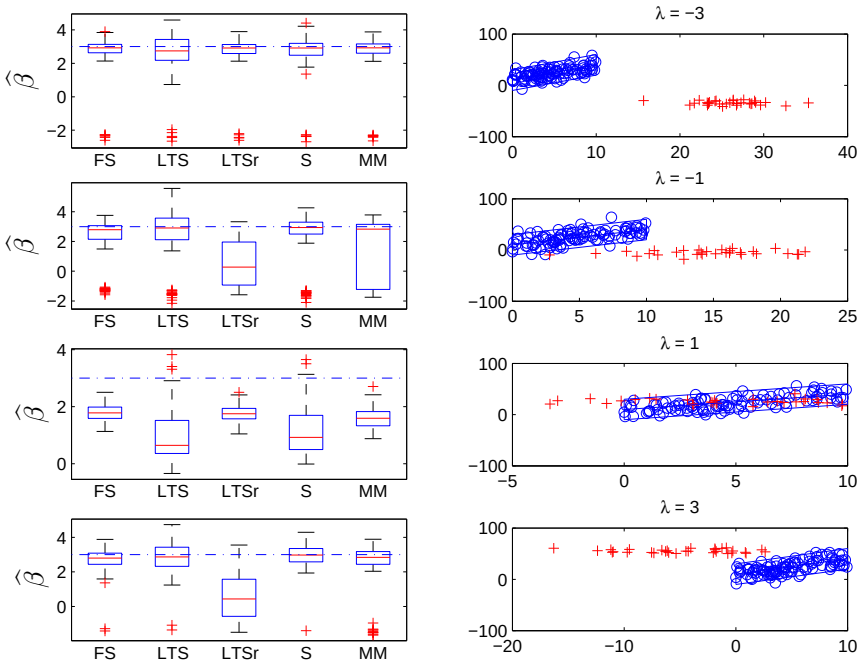

Figure 3: Example 1. Four simulated data sets for $\lambda = -3, -1, 1$ and 3. Left-hand panels: boxplots, from 100 simulations, of estimates of β (dotted line: $\beta_1 = 3$) for FS, LTS, LTSr S and MM estimators. Right-hand panels: typical simulations for these four values of λ

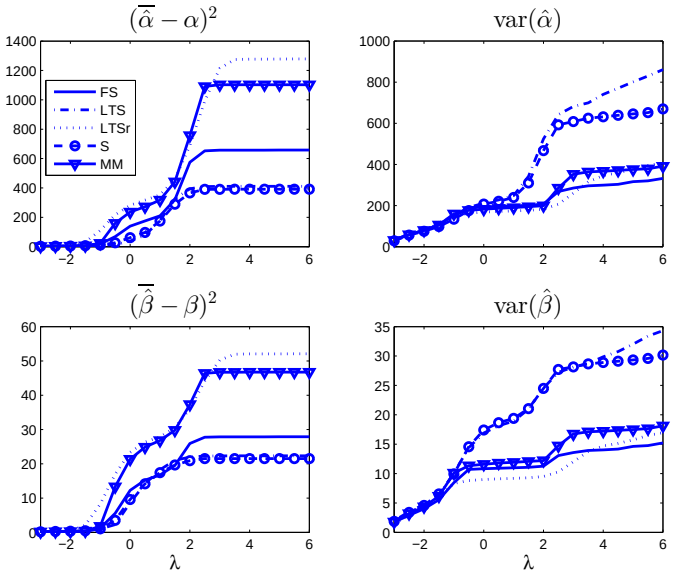

Figure 4: Example 1. Partial sums over Λ of simulated squared bias and variance of the five estimators. Left-hand panels squared bias, right hand panels variance. Top line $\hat{\alpha}$, bottom line $\hat{\beta}$.

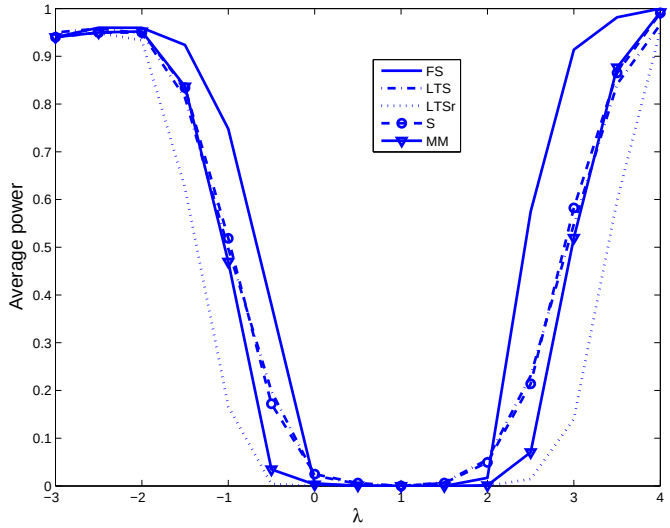

Figure 5: Example 1. Simulated average power of the five procedures over Λ

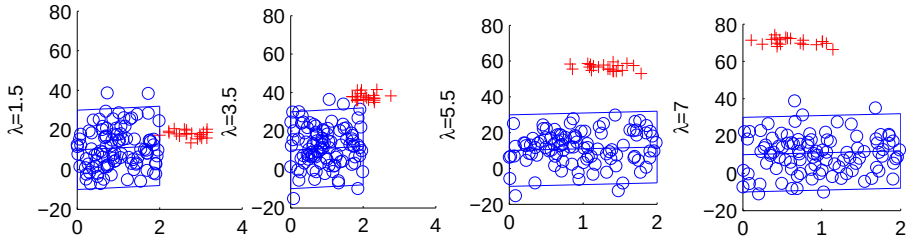

Figure 6: Example 2. Simulated data sets with $n_1 = 100$ and $n_2 = 20$ for four values of λ. As λ increases observations from M_2 become close to those from M_1 and then become remote again. The parallelogram defines the region for the empirical overlapping index

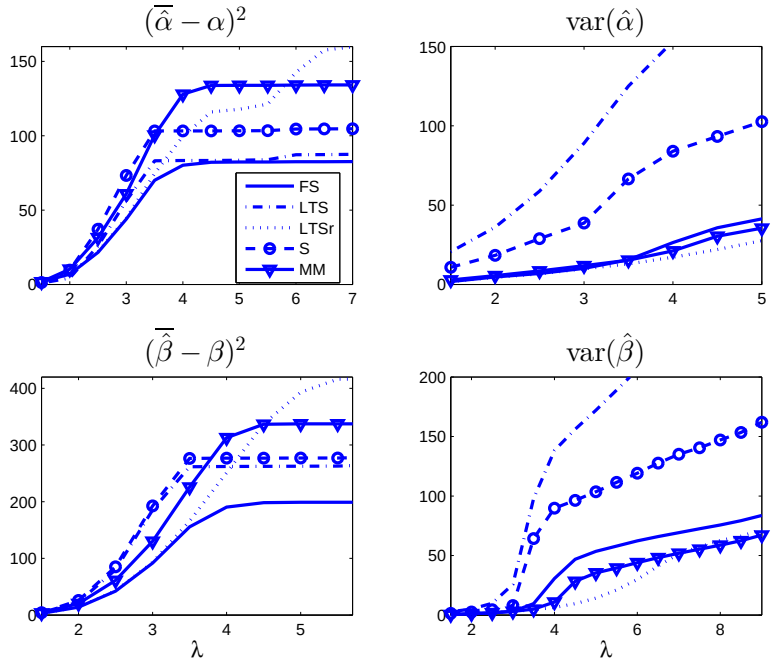

Figure 7: Example 2. Partial sums over Λ of simulated squared bias and variance of the five estimators. Left-hand panels squared bias, right hand panels variance. Top line $\hat{\alpha}$, bottom line $\hat{\beta}$.

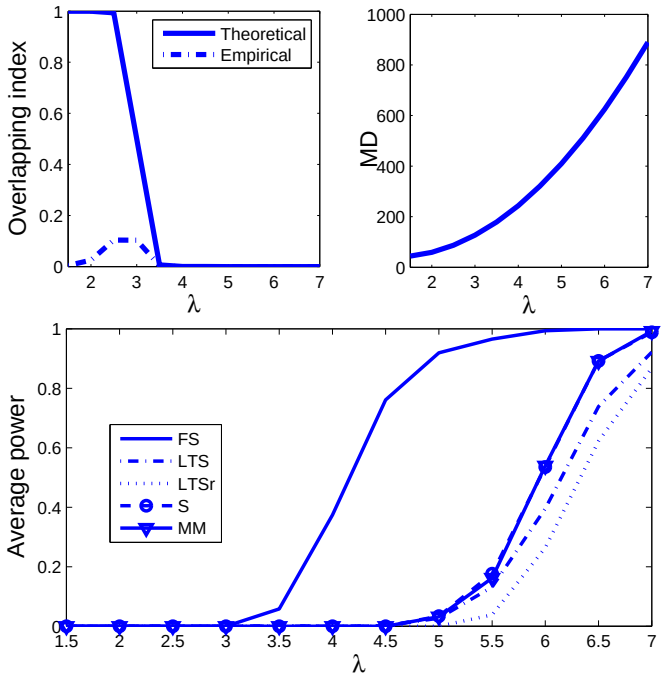

Figure 8: Example 2. Top panels: theoretical and empirical overlapping indices and Mahalanobis distance of M_1 from M_2. Bottom panel: simulated average power of the five procedures over Λ

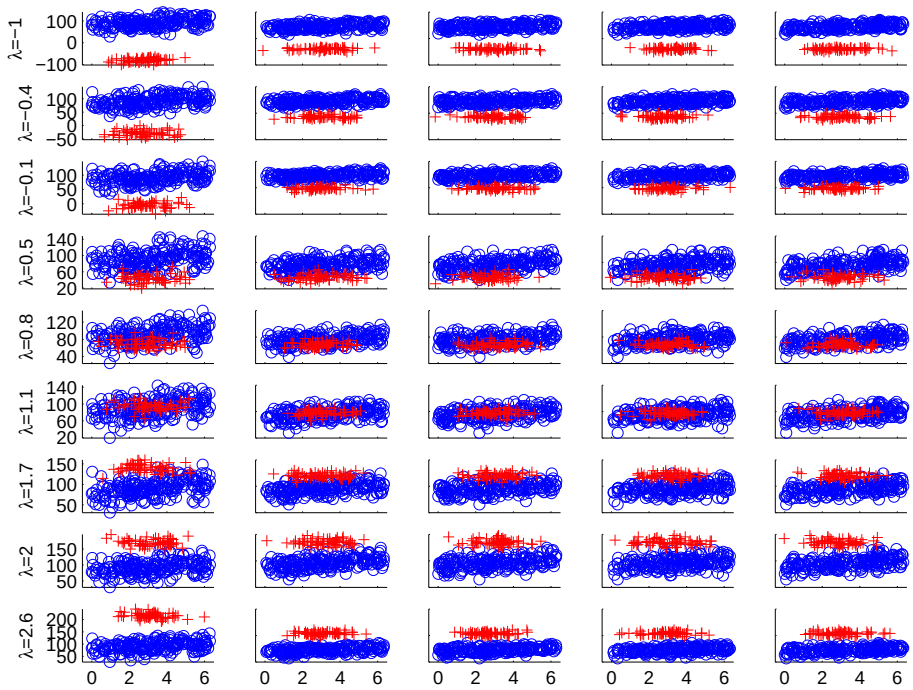

Figure 9: Example 3. Simulated data sets with five explanatory variables, $n_1 = 200$ and $n_2 = 60$ for nine values of λ. As λ increases observations from M_2 "pass through" those from M_1, although the centres never coincide.

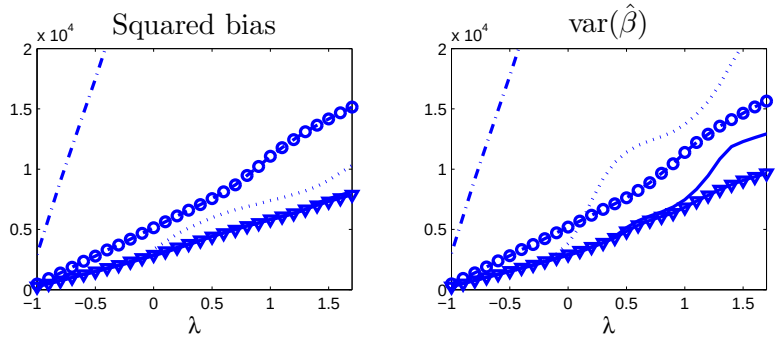

Figure 10: Example 3. Partial sums over Λ of simulated squared bias and variance of the five estimators. Left-hand panel squared bias for $\hat{\beta}$, right hand panel variance.

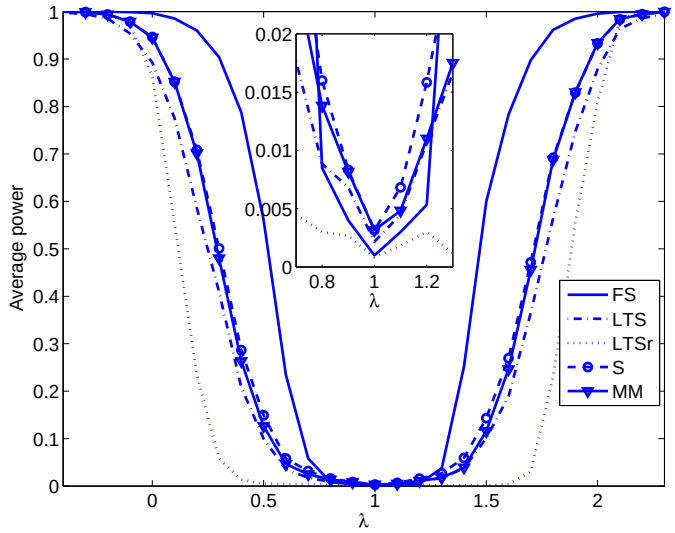

Figure 11: Example 3. Simulated average power of the five procedures over Λ with an inset zoom of the central part of the figure

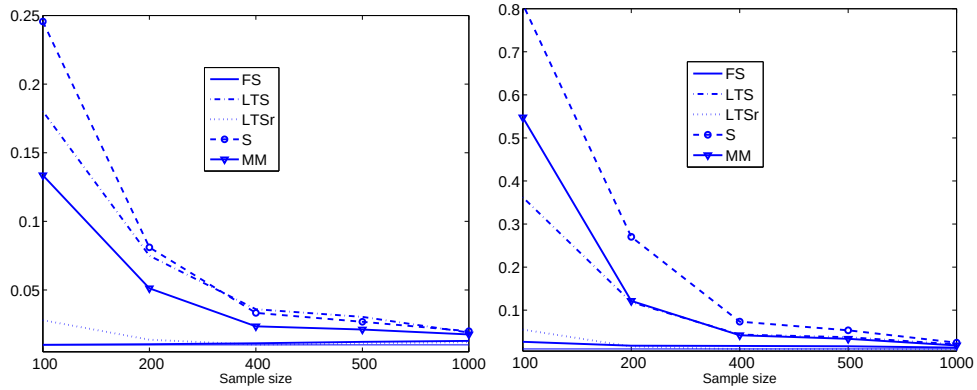

Figure 12: Size of nominally 1% Bonferronized outlier tests for, left-hand panel, $p = 6$ and for $p = 11$. Note the different vertical scales in the two panels