B. Supplement to "A Parametric Framework for the Comparison of Methods of Very Robust Regression" by Riani, Atkinson and Perrotta

There are three subsections. In the first we give a second motivating example. The second and third are expanded versions of our analyses of Examples 2 and 3 in the paper. The numbering of the figures continues that of the paper.

B1. Clustered Outliers

We now present a numerical example in which there are appreciable and informative differences between the behaviour of the five estimators. This example is a special case of the simulation results of §5 that serves to illustrate the need for a general structure for the design of these experiments.

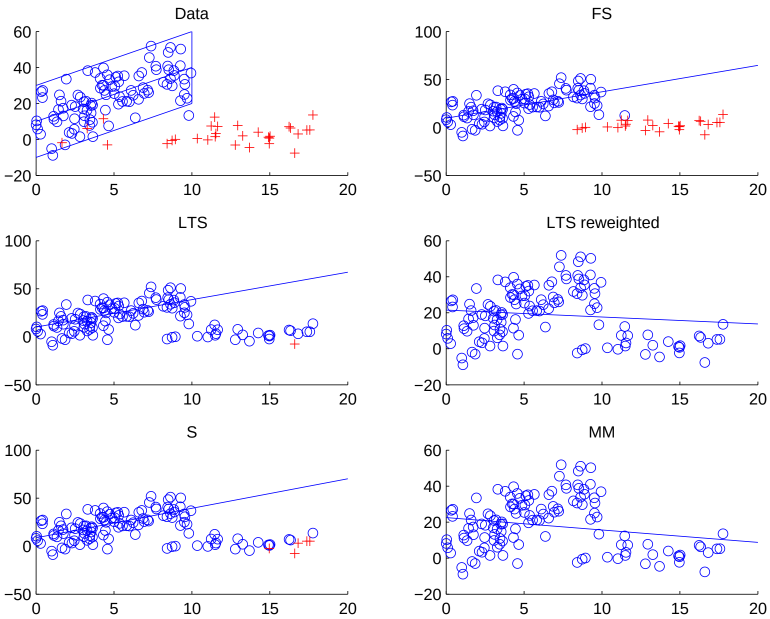

Figure 14: Regression data with clustered outliers simulated from (B1). Reading across: ○ simulated regression data and outliers +; fitted FS line and outliers +; LTS, reweighted LTS, S and MM estimators

There are 100 regression observations from the model $y_i = 10 + 3x_i + 10Z_i$, where $Z_i, \sim \mathcal{N}(0,1)$ and the independent $x_i \sim U(0,10)$. The contamination comes from a bivariate normal distribution, with mean μ and variance Σ, with

$$\mu = \begin{pmatrix} 2.5 \\ 12.5 \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} 20 & 2 \\ 2 & 20 \end{pmatrix}, \tag{B1}$$

where the first component corresponds to the response.

For the moment we consider the results of just two simulations, chosen to illuminate the properties of the different estimation procedures. The top-left hand panel in Figure 14 shows the 100 regression data plotted as circles together with the 30 outliers, plotted as crosses. These form an elliptical group of observations below and to the right of the regression data. The other panels show the robust regression lines and outliers identified by the various robust analyses.

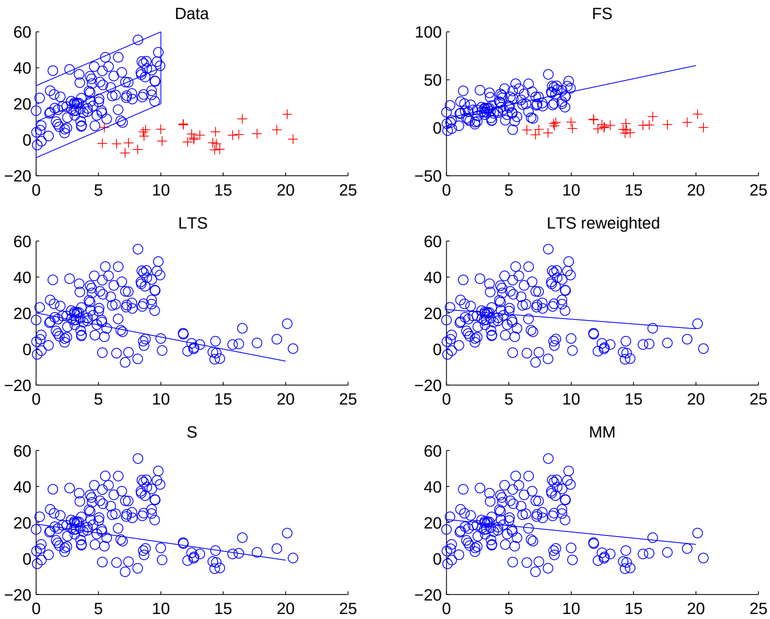

Figure 15: Second set of regression data with clustered outliers simulated from (B1). Reading across: ○ simulated regression data and outliers +; fitted FS line and outliers +; LTS, reweighted LTS, S and MM estimators

The top right-hand panel shows the FS line, which has a slope of 2.76, slightly less than the value of 3 of the data. Virtually all the outliers are identified. The middle panel of the first column of the plot shows the LTS output. Here the line is close to that from the FS with a slope of 2.85. However, the plot shows that only one outlier is identified. This low power for outlier identification arises from the value of consistency factor needed to rescale the estimate of σ^2 when half the data are trimmed and the Bonferroni correction needed to scale the outlier test to have correct size. These procedures do not affect the properties of the LTS parameter estimates. However, they do affect the reweighted least squares fit in the adjacent panel, which has a negative slope of -0.39. Something similar happens with the S estimator in the bottom left-hand panel which has a slope of 3.07. However, the MM modification leads to a negative slope of -0.69.

The second simulation of the same scenario is in Figure 15. Now the FS line has a slope of 2.75 whereas all other methods provide fits with a negative slopes, respectively -1.40, -0.53, -1.00 and -0.69. However, the FS is not uniformly best over repeated simulations. In some simulations it provides a fitted line with a negative slope whereas LTS has a positive slope. In some other simulations all lines can have similar positive, or negative, slopes.

Clearly there is appreciable variability in the estimates produced when the outliers are so close to the main body of the data. In order to discern underlying properties we need to look at the distribution of the estimates of slope, which we will summarize by means and variances. We also need to formalize how far the outliers lie from the main data so that we can study how the properties depend on this distance.

B2. Example 2

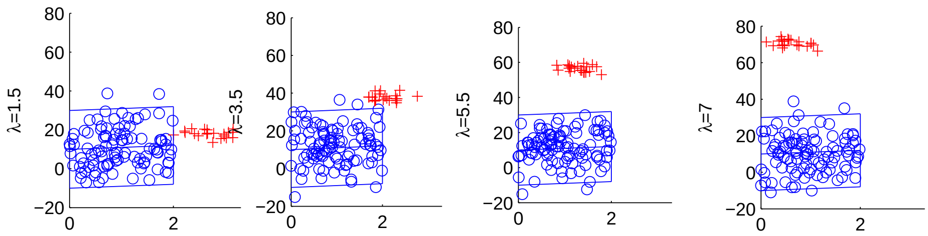

Figure 16: Example 2. Simulated data sets with $n_1 = 100$ and $n_2 = 20$ for four values of λ. As λ increases observations from M_2 become close to those from M_1 and then become remote again. The parallelogram defines the region for the empirical overlapping index

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 16 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.

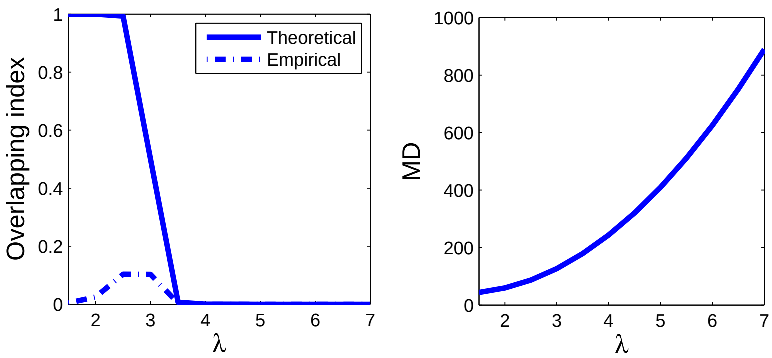

Figure 17: Example 2. Theoretical and empirical overlapping indices and Mahalanobis distance of M_1 from M_2.

The difference between this example and Example 1 is made clear in the plot of the measures of overlap in Figure 17. 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 16 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.

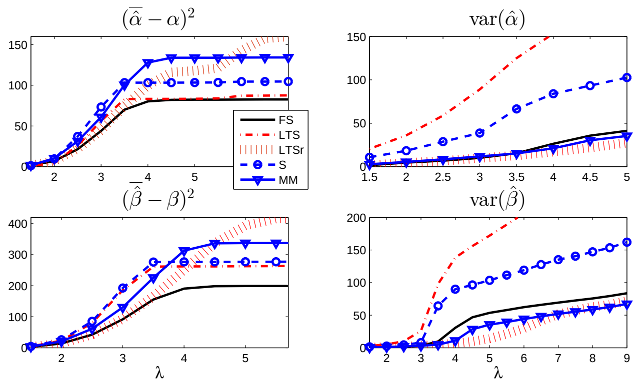

Figure 18: Example 2. Partial sums over Λ of simulated squared bias and variance of the five estimators. Left-hand panels squared bias, right hand panels variance. Top line $\hat{\alpha}$, bottom line $\hat{\beta}$.

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.

To conclude the analysis of the second example we look at the plot of average power in Figure 19. As 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.

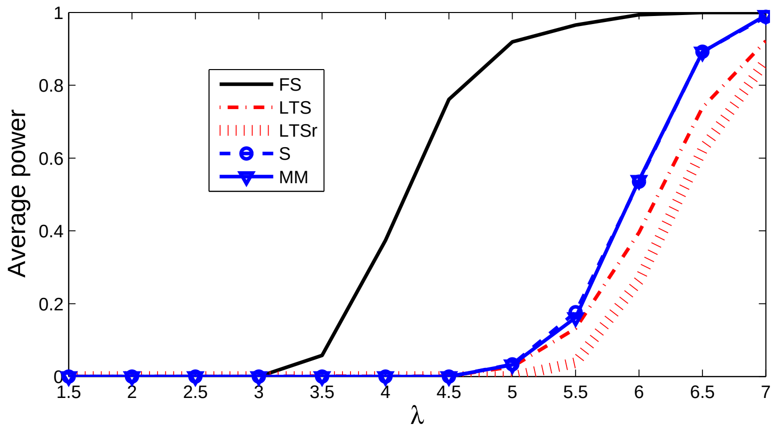

Figure 19: Example 2.Simulated average power of the five procedures over Λ

B3. Example 3

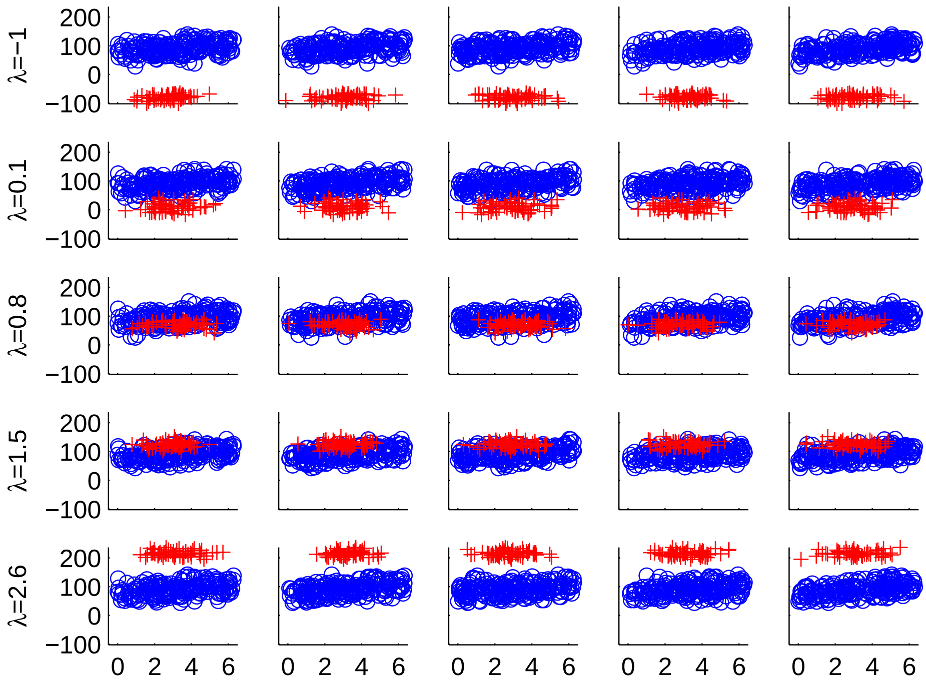

Figure 20: Example 3. Simulated data sets with five explanatory variables, $n_1 = 200$ and $n_2 = 60$ for five values of λ. As λ increases observations from M_2 "pass through" those from M_1, although the centres never coincide. Each row corresponds to one value of λ. The successive columns are for the various x_j, $j = 1, \ldots, 5$

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 20 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 21. 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

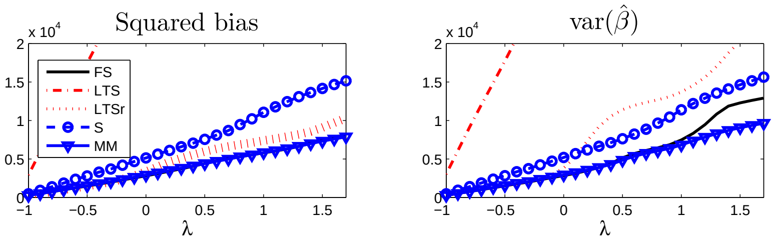

Figure 21: Example 3. Partial sums over Λ of simulated squared bias and variance of the five estimators. Left-hand panel squared bias for $\hat{\beta}$, right hand panel variance.

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.

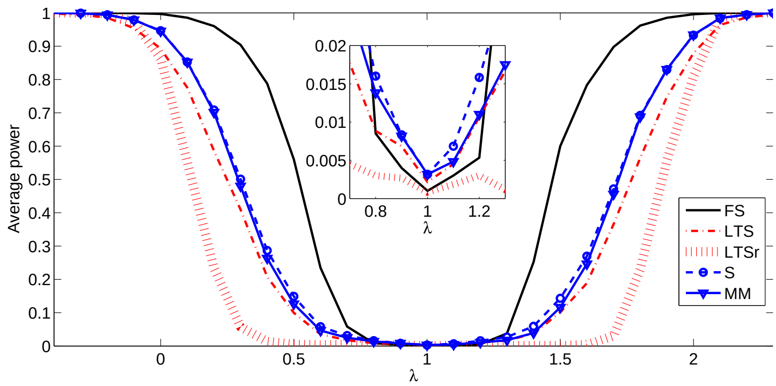

Figure 22: Example 3. Simulated average power of the five procedures over Λ with an inset zoom of the central part of the figure

The plot of average power is in Figure 22. 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.