Robust Regression Estimators In Low Dimensional Data Analysis; A Simulation and Comparison Review

Man Chong Chan, Mehrdad Mohammadi, Yilin Zhu

***Abstract*— In regression analysis, the Ordinary Least Squares Regression might not always be the appropriate model when your data exists extreme outliers or influential points. In such situations, we need an alternative estimator that would not be as much affected by these extreme observations, and Robust Estimator is one such estimator. In this project, 7 methods of Robust Estimation will be discussed, and these methods are Least Absolute Deviations (LAD), Huber M-estimator, Bisquare M-estimator, S-estimation, MM-estimation, Least Trimmed Squares Regression, Least Median Square Regression. Along with general discussion, a simulation study on these estimators were also conducted, and we concluded that Huber M-estimator is more sensitive to the number of outliers, and the Bisquare have a relatively high breakdown point. Meanwhile, MM, S, and LTS have high breakdown points and are more robust when there are more outliers in the data compared to the other estimators.**

I. Introduction

When analyzing real world data, often times, the use of the Ordinary Least Squares Regression might not always be appropriate in solving problems when there exist assumption violations. Whether they are non-normality, heteroscedasticity, extreme outliers or influential points, they all would greatly affect the accuracy of the fit of the least square model. One might try different transformations to eliminate the violations. However, transformation is not the antidote to all assumption violations, and sometimes transformation will not eliminate or even weaken the violations, especially for large amount of contamination. Under these situations, Robust Regression comes into play as an alternative of Ordinary Least Squares Regression.

In this project, we are going to generally discuss and compare the use of different Robust Regression methods in a terms of insensitive estimation methods to outliers and possibly high-leverage points, as well as simulation study on these estimators in both Low Dimension.

II. Robust Regression in Low Dimension

A. Introduction to M-estimation

M-estimation is by far the most popular among all robust estimations and it is essentially an extension of the maximum likelihood estimation. It is unbiased, and also has the the smallest variance possible among all linear unbiased estimators. Below we will start with M-estimates of location and M-estimates of scale, then move on to regression M-estimates as well as a few general numerical algorithms.

1) M-estimates of location: Assume each observation x_i depends on the unknown location parameter μ and also on some random error u_i. We have the M-estimates of location model as follows

$$x_i = \mu + u_i, \quad i = 1, ..., n.$$

We aim to estimate the location μ based on $x_1, ..., x_n$. If we suppose that errors have the same distribution function F and density function f, the maximum likelihood estimation of μ is then given by

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(x_i - \mu),$$

where $\rho \equiv -\log f$. If we further assume that ρ is differentiable and $\psi = \rho'$ [1], the first-order condition would become

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}) = 0.$$

2) M-estimates of scale: The framework of scale M-estimation is the same. The only difference is that now we consider the scale parameter σ and the model will be

$$x_i = \sigma u_i, \quad i = 1, ..., n, \quad \sigma > 0.$$

We could derive the maximum likelihood estimation of σ by

$$\hat{\sigma} = \arg\min_{\sigma} n\log\sigma - \sum_{i=1}^{n} \log f(\frac{x_i}{\sigma}).$$

We still assume the differentiable case. Then the first-order condition implies

$$\frac{1}{n} \sum_{i=1}^{n} \rho_{scale}(\frac{x_i}{\hat{\sigma}}) = 1,$$

where $\rho_{scale}(t) = t\psi(t)$ is different from ρ, and $\psi(t) \equiv \rho'(t) = -\frac{f'(t)}{f(t)}$.

In general, any estimate $\hat{\sigma}$ satisfying $\frac{1}{n}\sum_{i=1}^{n} \rho_{scale}(\frac{x_i}{\hat{\sigma}}) = \delta$ is called a M-estimate of scale, where δ is any positive number.

Recall that the scale parameter does not appear in either LS-estimation or L1-estimation. However, it indeed plays an important role in robust estimation methods. For being less effected by outliers, some M-estimators such as Huber's would trade off equivariance properties ($e.g.$ re-scaling the response should not alter the robustness behavior) if without the involvement of scale parameter. For more information, please refer to page 111 of [5].

3) Regression M-estimates: Now we focus on M-estimation for regression model. Consider a linear regression model with p explanatory variables,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

In this model, one can easily see that $\boldsymbol{x_i}^T\boldsymbol{\beta}$ is actually the location term and ϵ_i the random error term.

Let's consider the M-estimate of location with scale(also known as dispersion) σ, which is unknown and independent of $\boldsymbol{\beta}$. Denote $\boldsymbol{x_i}^T\hat{\boldsymbol{\beta}}$ by $\hat{y}_i(\boldsymbol{\beta})$ and $y_i - \hat{y}_i(\boldsymbol{\beta})$ by $\epsilon_i(\boldsymbol{\beta})$.

[1] This is called ψ-type

Pretty much similar with the procedures we discussed before, $\hat{\beta}$ is determined by

$$\arg\max_{\beta} \frac{1}{\hat{\sigma}^n} \prod_{i=1}^{n} f(\frac{y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}}{\hat{\sigma}}) = \arg\min_{\beta} \sum_{i=1}^{n} \rho(\epsilon_i(\boldsymbol{\beta})),$$

which, based on the assumption that ρ is differentiable, could further lead to

$$\sum_{i=1}^{n} \psi(\frac{\epsilon_i(\boldsymbol{\beta})}{\hat{\sigma}})\boldsymbol{x_i} = 0, \quad \psi \equiv \rho' = -\frac{f'}{f}$$

Commonly used ρ-function will have the following properties:

1. Non-negative: $\rho(\epsilon_i) \geq 0$
2. $\rho(0) = 0$
3. Symmetric: $\rho(-\epsilon_i) = \rho(\epsilon_i)$
4. Monotonic: if $|\epsilon_i| > |\epsilon_j|$, $\rho(\epsilon_i) > \rho(\epsilon_j)$

4) Numerical Algorithm: In this subsection, we will only focus on the algorithm for M-estimation with previously computed scale. It's worth noting that estimating $\hat{\beta}$ and $\hat{\sigma}$ simultaneously can be realized, with a few modifications on the iteratively reweighted least squares(IRWLS). Indeed, this will lead to more robust estimators, such as S-estimator. We will discuss them in the next section.

To compute the scale in advance, a common choice is first implementing the L1-estimation to obtain the residuals ϵ_i, $i = 1, ..., n$, then calculating the median absolute deviation(MAD) of the residuals under L1-estimation

$$\text{MAD}(\boldsymbol{\epsilon}) = \text{Median}(|\boldsymbol{\epsilon} - \text{Median}(\boldsymbol{\epsilon})|).$$

Sometimes "normalized MAD" (MADN) is applied, which is defined as $\text{MAD}/0.675$.

Upon getting the previously computed scale $\hat{\sigma}$, let

$$W(x) = \begin{cases} \frac{\psi(x)}{x} & \text{if } x \neq 0 \\ \psi'(0) & \text{if } x = 0 \end{cases}.$$

be the weighted function.

The procedure, which depends on a tolerance parameter ζ, is

1. Compute an initial L1 estimate $\hat{\beta}_0$ and $\hat{\sigma}$.
2. For $k = 0, 1, 2, ...$:
$\boldsymbol{(a)}$ Given $\hat{\beta}_k$, for $i = 1, ..., n$ compute $\epsilon_{k,i} = y_i - \boldsymbol{x_i}^T\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ and $w_{k,i} = W(\epsilon_{k,i}/\hat{\sigma})$.
$\boldsymbol{(b)}$ Compute $\hat{\beta}_{k+1}$ by solving

$$\sum_{i=1}^{n} w_{k,i}\boldsymbol{x_i}(y_i - \boldsymbol{x_i}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{0}.$$

3. Stop when $\max_i \left(|\epsilon_{k,i} - \epsilon_{k+1,i}|\right)/\hat{\sigma} < \zeta$.

This algorithm converges if $W(x)$ is non-increasing for $x > 0$.

5) Some Classical M-estimators:

• $\boldsymbol{LAD(L1)}$ **:**

The L1 estimate(also called the least absolute deviation or LAD estimate), is defined by

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} |\epsilon_i(\boldsymbol{\beta})|.$$

Here the objective function is simply given by $\rho(x) = |x|$. One may see that the L1 estimate is less affected by the outliers than LS estimate.

• $\boldsymbol{Huber's}$ **:**

The Huber M-Estimator is essentially the combination of Ordinary Least Square(OLS) and LAD, and it gives a unique solution. The objective function is given by

$$\rho(\epsilon) = \begin{cases} \epsilon^2/2 & for\,|\epsilon| \leq k \\ k\,|\epsilon| - k^2/2 & for\,|\epsilon| > k \end{cases}$$

A common choice of k is $k = 1.345\sigma$, and this would give a 95% efficiency.

• $\boldsymbol{Bisquare}$ **:**

Bi-square M-Estimator is a very popular estimator. It is even more robust than Huber Estimator.
The objective function is given by

$$\rho(\epsilon) = \begin{cases} (k^2/6)(1 - [1 - (\epsilon/k)^2]^3) & for\,|\epsilon| \leq k \\ k^2/6 & for\,|\epsilon| > k \end{cases}$$

A common choice of k is $k = 4.685\sigma$.

However, due to the nature of its object function, Bi-square estimator could suffer from converging to local minimum. To avoid that issue, a common cure is to use Huber M-estimate as the starting point of the Bi-square estimate.

B. Estimates Based on a Robust Residual Scale

The estimation approaches we have discussed so far is unreliable when the predictor matrix $\boldsymbol{X}$ is random or contains high leverage points. Taking LS and the L1 estimates for examples, they minimize measures of residual largeness that can be seriously influenced by even a single residual outlier.

Several alternative estimators aim to provide remediation. A more robust alternative is to minimize a scale measure of residuals that is insensitive(or less sensitive) to large values. The three methods presented below follow the same framework that first estimating a robust residual scale $\hat{\sigma}$, then minimizing the scale measure to get $\hat{\beta}$. Formally,

$$\hat{\beta} = \arg\min_{\beta} \hat{\sigma}(\boldsymbol{\epsilon}(\boldsymbol{\beta})),$$

where $\hat{\sigma}$ is the robust estimated scale based on the vector of residual $\boldsymbol{\epsilon}(\boldsymbol{\beta}) = (\epsilon_1(\boldsymbol{\beta}), ..., \epsilon_n(\boldsymbol{\beta}))$.

• $\boldsymbol{S - estimation}$ **:**

S-estimator is a regression estimator that is based on residual scale M-estimation. Just as its name suggested, it is essentially a scale version of M-estimator. It overcomes one of the biggest flaws of M-estimation: it lacks consideration on the data distribution, and use only median as the weighted value. As a estimator based on a robust residual scale, S-estimator is defined by

$$\hat{\beta} = \arg\min_{\beta} \hat{\sigma}(\boldsymbol{\epsilon}(\boldsymbol{\beta})),$$

with specifically designated $\hat{\sigma}$ as we discussed in the second part of section(A), i.e.

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{\epsilon_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) = \delta.$$

We shall briefly introduce the numerical algorithm for S-estimation, which basically follows the framework we presented in part 4) of section(A). However, apart from $\hat{\beta}$, now we will also update $\hat{\sigma}$ at each step, since S-estimation heavily depends on the robust scale estimation. Let

$$\widetilde{W}(x) = \begin{cases} \frac{\rho(x)}{x^2} & \text{if } x \neq 0 \\ \rho''(0) & \text{if } x = 0 \end{cases}.$$

be the weighted function for scale, differing from the former one W.

1. Compute an initial L1 estimate $\hat{\beta}_0$ and $\hat{\sigma}_0$(such as MAD($\boldsymbol{\epsilon}$) or MADN($\boldsymbol{\epsilon}$)).
2. For $k = 0, 1, 2, ...$:
$(\boldsymbol{a})$ Given $\hat{\beta}_k$ and $\hat{\sigma}_k$, for $i = 1, ..., n$ compute $\epsilon_{k,i} = y_i - \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$, $w_{k,i} = W(\epsilon_{k,i})$, and $\tilde{w}_{k,i} = \widetilde{W}(\epsilon_{k,i})$
$(\boldsymbol{b})$ Compute $\hat{\beta}_{k+1}$ and $\hat{\sigma}_{k+1}$ by solving

$$\sum_{i=1}^{n} w_{k,i} \boldsymbol{x_i} (y_i - \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}}) = \mathbf{0}, \quad \text{and}$$

$$\hat{\sigma}_{k+1}^2 = \frac{\hat{\sigma}_k^2}{n\delta} \tilde{w}_{k,i} \epsilon_{k,i}^2$$

3. Stop when $\max_i \left(|\epsilon_{k,i} - \epsilon_{k+1,i}| \right) / \hat{\sigma}_k < \zeta$.
Note that the suggested δ is 0.199.

S-estimators are even more robust than the general M-estimator, especially for contaminated data.

- $\boldsymbol{MM - estimation}$ **:**

MM-estimator is a estimator that is built upon S-estimator and M-estimator. It retains the high breakdown point[2] of the bounded-influence estimator, and also the relatively high efficiency under normality assumption of M-estimator. MM-estimator is defined by

$$\hat{\beta}_{MM} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{\epsilon_i(\boldsymbol{\beta})}{s_{MM}} \right),$$

where s_{MM} is the standard deviation obtained from the residual of S-estimation.

C. Bounded-Influence Regression

- $\boldsymbol{LeastTrimmedSquares(LTS)}$

Least Trimmed Square is one of the most common bounded-influence regression. The idea behind the LTS is to discard a proportion of the largest residuals. Typically, we define the scale estimation for LTS as

$$\hat{\sigma} = \left(\sum_{i=1}^{n} a_i |\boldsymbol{\epsilon}|_{(i)}^2 \right)^{1/2},$$

where a_i's are non-negative constants and call $|\boldsymbol{\epsilon}|_{(1)} \leq ... \leq |\boldsymbol{\epsilon}|_{(n)}$ the ordered absolute values of residuals. By taking $a_1 = ... = a_n = 1$, the LTS estimator is defined as

$$\hat{\beta}_{LTS} = \arg\min_{\beta} \sum_{i=1}^{h} \boldsymbol{\epsilon}_{(i)}^2(\boldsymbol{\beta}).$$

[2]A breakdown point defined as the fraction of data which can be given arbitrary values without making the estimator, arbitrarily too large or too small; The point after which an estimator becomes useless. The higher the BDP is, the more robust is the estimator.

Note that this definition is slightly different from the S-estimator. In particular, $n - h \equiv [n\alpha]$, where $[\cdot]$ denotes the integer part, of the largest absolute residuals are trimmed. The form is called the α-trimmed squares scale where $\alpha \in (0, 1)$. Furthermore, to attain the maximum breakdown point, we may choose $h = [\frac{n+p+1}{2}]$.

- $\boldsymbol{LeastMedianSquare(LMS)}$

Least Median Square is another common bounded-influence regression. Just as its name suggested, instead of minimizing the sum of residuals square like the traditional OLS, Least Median Square aims to minimize the median of residuals square.

$$\hat{\beta}_{LMS} = \arg\min_{\beta} \text{Median}(\boldsymbol{\epsilon}^2(\boldsymbol{\beta}))$$

LMS is highly robust; It achieves Breakdown point = 0.5(the highest BDP possible). However, it has relatively low efficiency: It has at best a relative efficiency of 37% (Rousseeuw and Croux 1993).

III. Simulation and Comparison

We do a simple simulated example with N_1 "good" observations and N_2 "bad" ones. In each of the cases illustrated, $n_1 = 100$ random draws from a bivariate normal distribution with mean $E(X, Y)^T = (0, 0)$ variances equal to 1, and correlation equal to 0.9 are generated. These are shown by the black points in each of the graphs in Figures in appendix. Then, a sample of n_2 " observations were drawn from a bivariate normal but with mean (1.5,1.5) with variances (0.2,0.2) and correlation zero. These are the magenta points on the plots, with $n_2 = (20, 30, 75, 100)$.

A. M-Estimator

Figure [1] compares the objective functions (left), and the corresponding ψ (center) and weight functions (right) for three M-estimators: the least-squares estimator; the Huber estimator; and the Bisquare estimator.

Figure [2] shows the simulated data with an increasing number of "bad" or "outlying" cases. Four lines are shown in each panel: the Huber regression (solid black line); Bi-square fit to all of the data (solid green line); OLS fit to all of the data (magenta line);OLS fit to the "good" data points (dot-dash blue line). If the goal is to match, more or less, the OLS regression fit to the good data, then both the Huber's and the bi-square fit a better line compare to the OLS regression. The Bisquare's regression , however, also does a respectable jobs for $n_2 \geq 30$ comare to the Huber's.

B. S-Estimator and MM-Estimator

In figure [3] the simulated data with an increasing number of "outlying" cases with each estimators line has been plotted. Six lines are shown in each panel: the Huber regression (solid green line); Bi-square fit to all of the data (solid brown line);the S-estimator regression (black line); MM-estimator fit to all of the data (solid red line);OLS fit to all of the data (magenta line);OLS fit to the "good" data points (dot-dash blue line).

The goal is to match to the OLS regression fit to the good data, for $n_2 \leq 20$ the bi-square fit a better line compare to the other regression lines. The S and MM regressions , however, have a better fit as the weight of outliers $n_2 \geq 30$ increases.

C. LTS and LMS

Figure[4] show the simulated data with an increasing number of "outlying" cases with LTS and LMS estimators. AS we can see in the graph with a low number of outlying observations both estimators have similar behavior and do a a good job when the number of outliers are $n_2 \leq 30$.

IV. Discussion and Conclusion

Our short overview for this project was done on theoretical summary of the mostly used robust estimators and some the computational algorithms. We also compared different estimators in each class by running a simulation and graphically comparing the breakdown point of the estimators as the number of outliers increases.

We mainly chose methods that are popularly used and can be found from the existing R packages. R function rlm provides the implementation of M_{Huber} and M_{Tukey} with stating psi function as "Huber" and "Tukey", respectively. LMS, LTS, and S are computed using R function $lqs()$ with the option specified as "lms," "lts", and "S," respectively. In these $lqs()$ computation procedures, re-sampling algorithm is used. R package $robust$ provides the implementation of and the S-estimate is used as an initial estimate via random re-sampling. It is known that using the initial S estimate in two-stage algorithm of MM achieves both high efficiency and robustness.

We concluded that Huber is more sensitive to the number of outliers and the Bisquare have a relatively high breakdown point. MM, S, and LTS have high breakdown points and are more robust when there are more outliers in the data compared to the other estimators. Park et al. (2012) pointed out that MM-estimates cannot detect any outliers when the contamination percentage is equal to and above 30%.
Moreover, it would be worth mentioning that among all those estimators L_1 would give a better understanding of the distribution of the population as we break the analysis to the qunatiles. Thus, it is more robust to outlying points.

Our review could be extended to compare the efficiency and relative efficiency of the above estimators, as well as comparing their Mean Squared Errors. Moreover, one other aspect to compare all these estimators would be to see the relative sensitivity with respect to large outlying points and leverage points.

V. References

[1] Almetwally. E and Almongy. H (2018), "Comparison Between M-Estimation, S-Estmation, and MM-Estimation" *Methods of Robust Estimation with Application and Simulation*

[2] Fox.J. and Weisberg.S. (2018) "An R Companion to Applied Regression, third edition" *SAGE Publications, Inc*

[3] Huber, P. J. (1981), "Robust Statistics" *New York: JohnWiley and Sons*

[4] Park, Y., Kim, D., Kim, S. (2012), "Robust regression using data partitioning and M-estimation" *Communications in Statistics - Simulation and Computation 41:1282–1300.*

[5] Maronna. R, Martin R., Yohai V. (2006), "Robust Statistics, Theory and Methods" *New York: JohnWiley and Sons*

[6] Filzmoser P. and Nordhausen K. (2020), "Robust linear regression for high-dimensional data: An overview" *WIREs Computational Statistics 10.1002/wics.1524.*

VI. Appendix

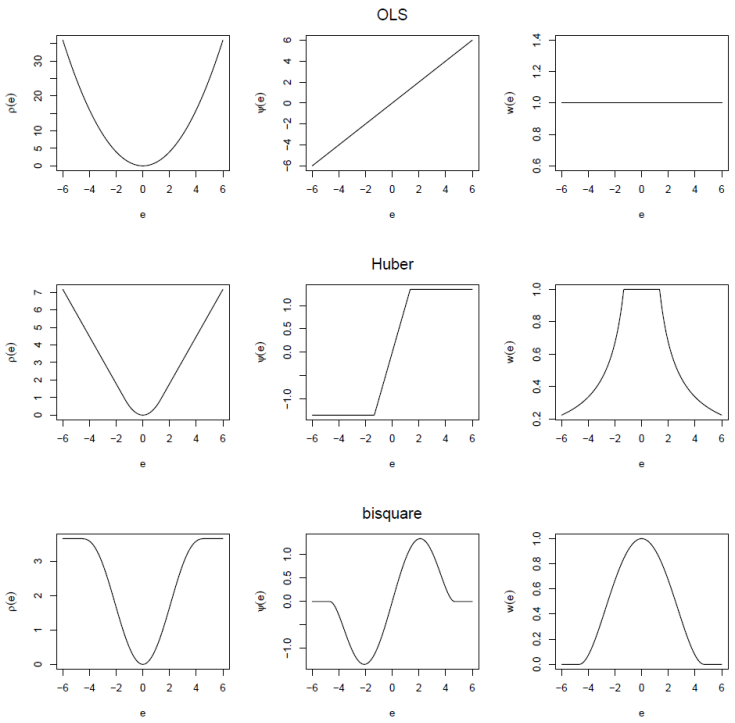

Figure 1. Objective functions (left), and the corresponding ψ (center) and weight functions (right) for three M-estimators

Figure 2. simulated data for M-estimators with an increasing number of outliers.

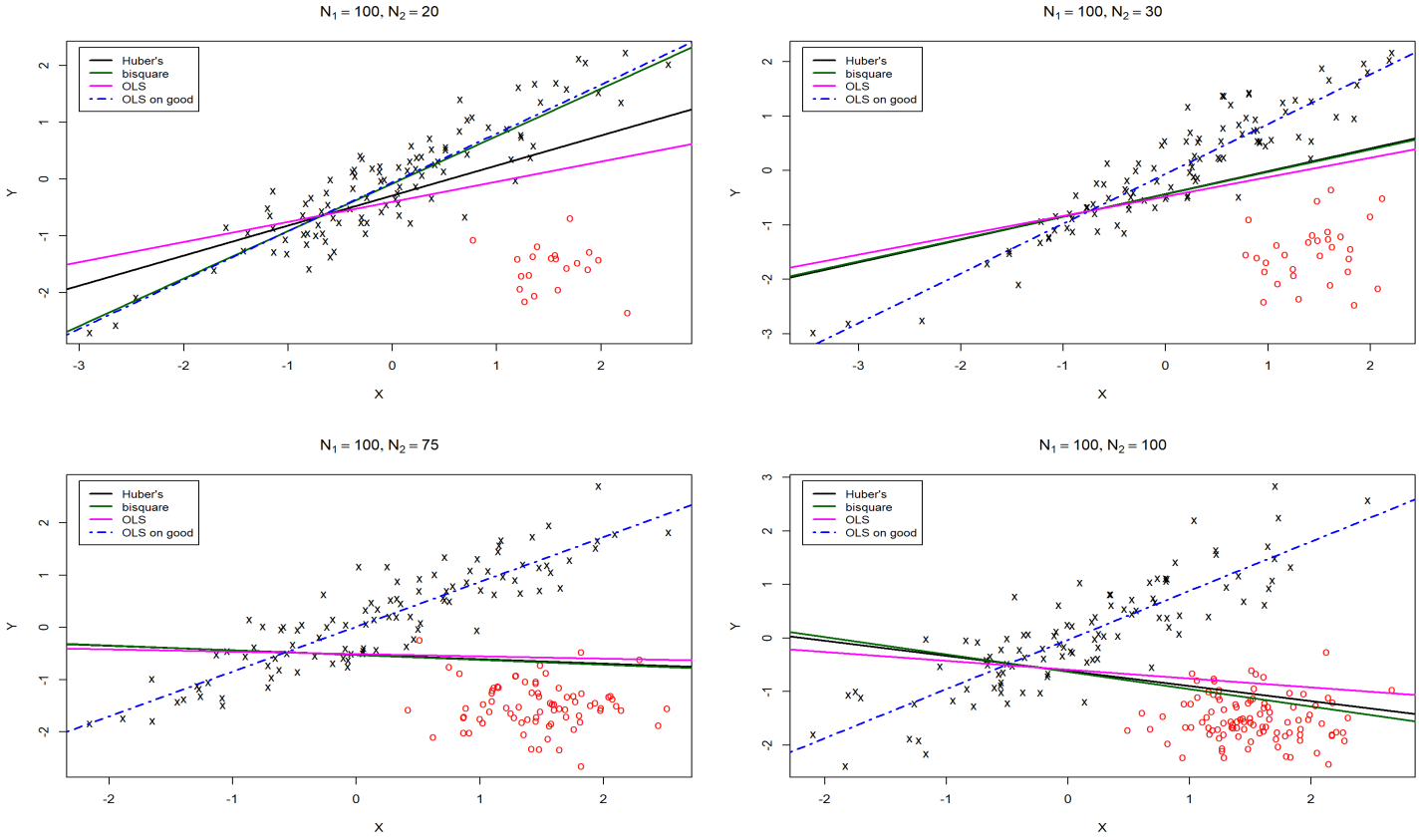

Figure 3. simulated data for S-estimators with an increasing number of outliers.

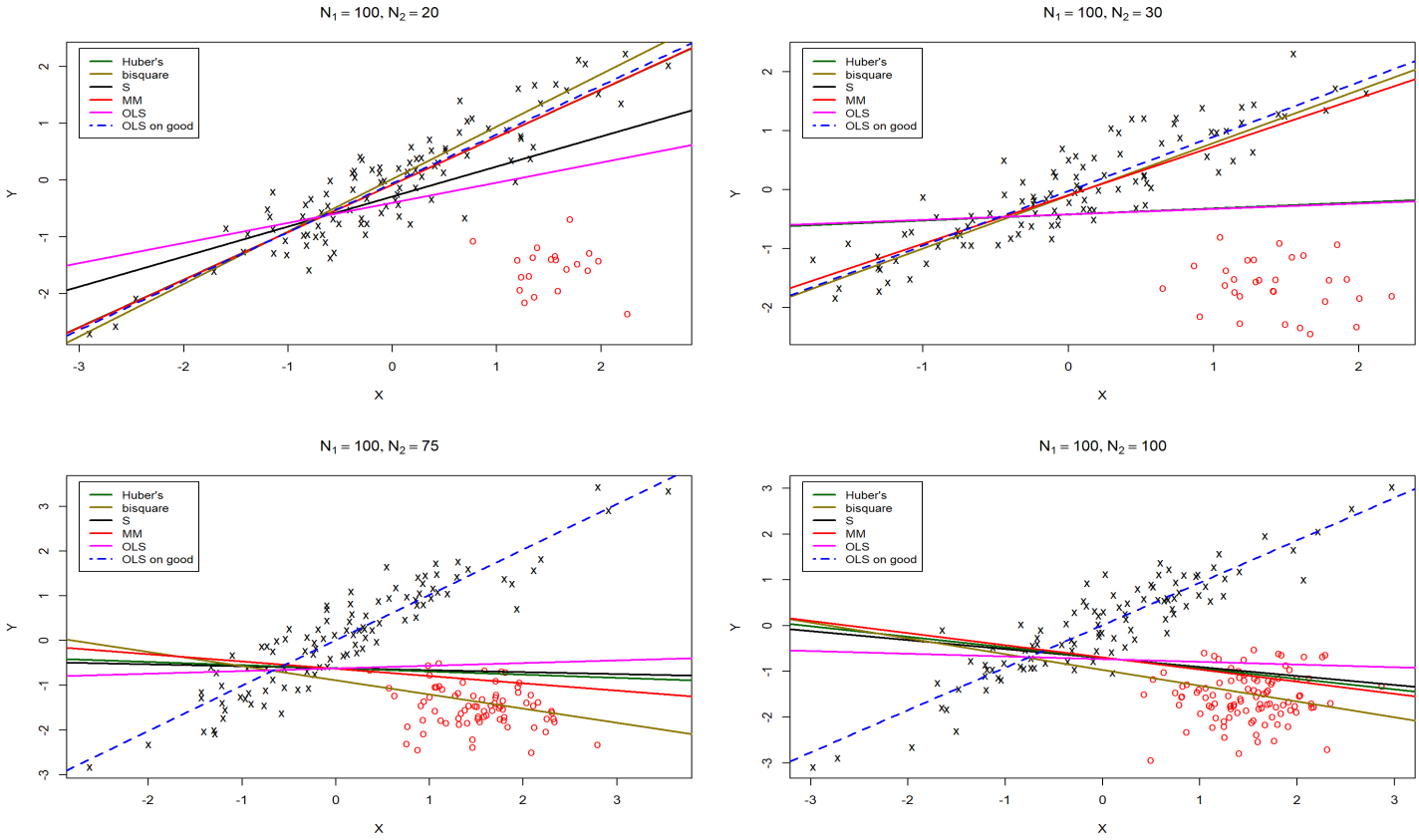

Figure 4. simulated data for LTS-estimators with an increasing number of outliers.

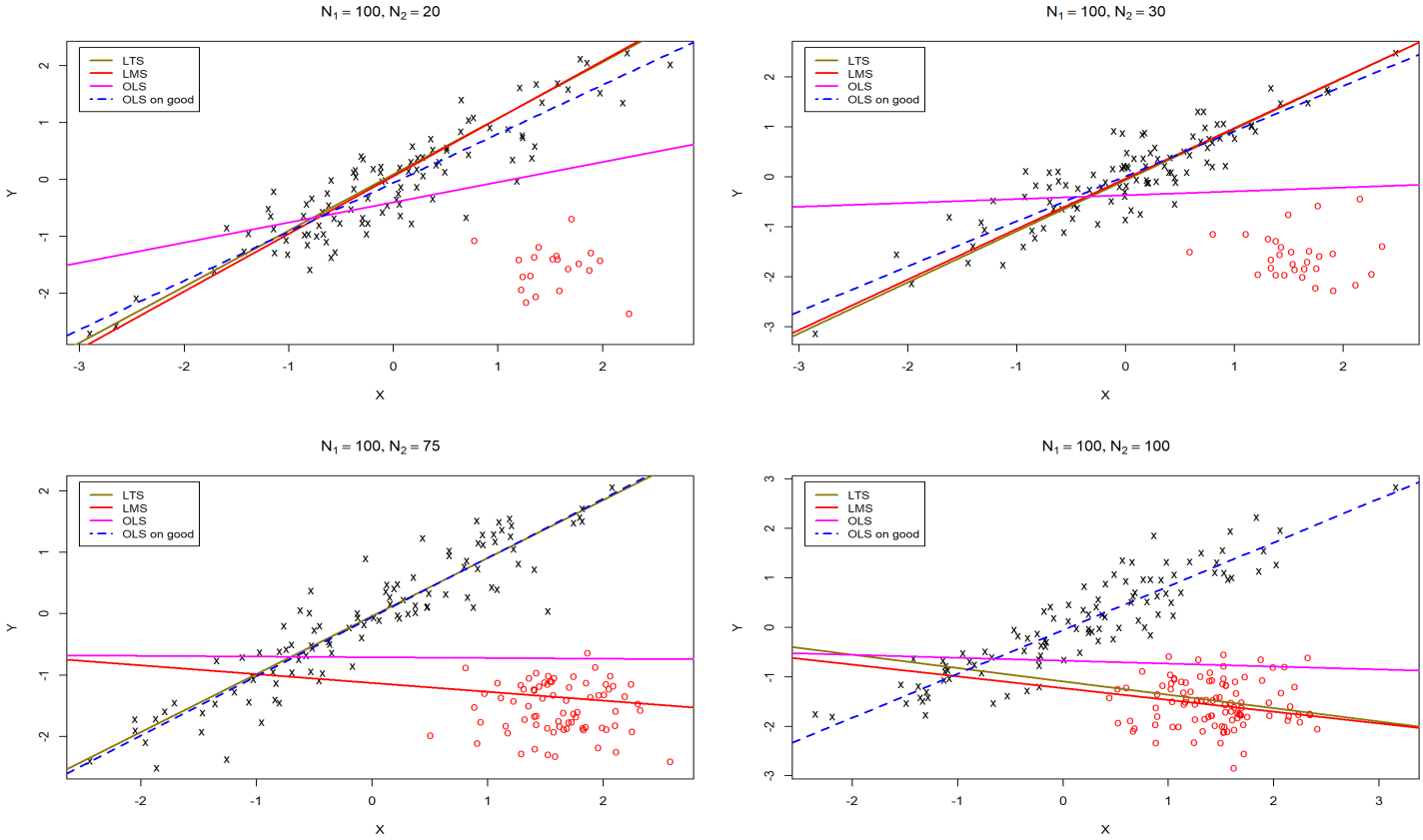

Figure 5. simulated data comparing all estimators.