16th European Symposium on Computer Aided Process Engineering
and 9th International Symposium on Process Systems Engineering
W. Marquardt, C. Pantelides (Editors)

A performance comparison of some high breakdown robust estimators for nonlinear parameter estimation

Eduardo L.T. Conceição[a] and António A.T.G. Portugal[a]

[a]CEM group, Department of Chemical Engineering, University of Coimbra,
Pólo II, Pinhal de Marrocos, 3030–290 COIMBRA, Portugal

While the inevitable occurrence of departures from the assumptions made beforehand can damage least squares reliability, robust estimators will resist them. A number of alternative robust regression estimators have been suggested in the literature over the last three decades, but little is known about their small-sample performance in the context of nonlinear regression models. A simulation study comparing four such estimators together with the usual least squares estimator is presented. It is found that the MM- and τ-estimators are quite efficient when the proportion of outliers in data is not too large.

Keywords: high breakdown point, robust regression, nonlinear regression, Monte Carlo

1. INTRODUCTION

In the nonlinear regression model, one observes the response variable y obeying the model

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + e_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{x}$ is a vector of explanatory variables, $\boldsymbol{\theta}$ is a vector of unknown true parameters to be estimated, and e is the measurement error. Define the residuals corresponding to $\boldsymbol{\theta}$ as $r_i(\boldsymbol{\theta}) = y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta})$. It is common to consider the errors e_i as independent and identically distributed random variables with zero mean and variance σ_e^2, which follow a specified type of distributions.

The goal of each possible estimator is to draw reliable estimates of the parameters from data and additionally protect against departures from statistical model assumptions made beforehand because in practice it is very unlikely that the model assumptions hold perfectly. They may include the presence of outlying observations and other departures from the imposed model distribution. Of course, it is recognized that neither classical least squares (LS) nor, more generally, maximum likelihood methodology is satisfactory as far as the robustness requirement is concerned, since it depends heavily on the assertion that the actual error process follows exactly the distribution assumed. For this reason, a vast amount of literature in *robust* alternative techniques was developed over the last 30 years.

A measure of robustness frequently used in the literature is the *breakdown point* (BP) which is, roughly speaking, the smallest proportion of contaminated data which leads to

unreliable model parameters. Thus, a regression estimator with high breakdown point (HBP) is capable of handling *multiple* outliers, even if they are grouped. Besides, it is also important to know the uncertainty (bias and sampling variability) of point estimates on "clean data". This is assessed by the statistical efficiency criterion calculated as the ratio of the mean squared error (MSE) in a least squares estimate to the actual MSE of a (robust) estimate, computed at the Gaussian (normal) distribution. Unfortunately, HBP estimators tend to have low efficiency.

Rousseeuw [1] proposed the first high breakdown regression estimator, the least median of squares (LMS), but its very low asymptotic efficiency is a well known drawback. The same author [1] suggested the least trimmed squares (LTS) estimator which is more efficient than the LMS estimator. Since then, several methods have been proposed which combine good asymptotic efficiency with HBP. Among them are the three-stage MM-estimator starting with initial HBP regression estimates of Yohai [2] and the τ-estimator of Yohai and Zamar [3].

Little is known about the *small-sample* properties of these estimators in the context of nonlinear regression models. Thus, the main purpose of this article is to investigate their small-sample performance by means of a Monte Carlo simulation study based on real data sets. The simulation design considers the effects of proportion of outliers in data and different error distributions. Another goal is to compare the use of the LMS and LTS estimators as the initial HBP estimator in the MM-estimator.

The remainder of the paper is organized as follows. Section 2 defines the LMS, LTS, MM-, and τ-estimates. In Section 3 we summarize the basic aspects of the simulation study. The different estimators are then compared in Section 4.

2. DEFINITIONS OF ROBUST ESTIMATORS

Least median of squares (LMS) Rousseeuw [1] proposed the first regression estimate with the highest possible BP of 1/2, by minimizing the median of squared errors, that is

$$\hat{\boldsymbol{\theta}}_{\text{LMS}} = \arg\min_{\boldsymbol{\theta}} \operatorname*{med}_i r_i^2(\boldsymbol{\theta}), \tag{2}$$

where $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ and med denotes the median.

Least trimmed squares (LTS) The LTS estimate is defined as [1]

$$\hat{\boldsymbol{\theta}}_{\text{LTS}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{h} r_{(i)}^2(\boldsymbol{\theta}), \quad n/2 \leqslant h \leqslant n, \tag{3}$$

where $r_{(i)}^2(\boldsymbol{\theta})$ is the ith squared residual sorted from smallest to largest and h is the number of these terms which are included in the summation called the *coverage* of the estimator. Therefore, the $n - h$ "trimmed" observations that correspond to the largest residuals do not directly affect the estimator.

Let $\alpha = 1 - h/n$ be the amount of trimming called the trimming proportion, with $0 \leqslant \alpha \leqslant 1/2$. The maximal BP for LTS equals 1/2 and is obtained by choosing α close to 1/2. However, one may expect a tradeoff between a high value for α and a loss in efficiency. Thus, the choice of α (or equivalently h) determines the overall performance of

the LTS estimator, and some effort is required to tune this parameter. Hence, it has been suggested that lower values for α (the most commonly suggested values are 0.25 and 0.1) will give a good compromise between robustness and efficiency.
Note that the objective function of the LTS estimator is nonconvex and not differentiable the same happening for the LMS estimator. Consequently, these optimization problems cannot be solved by standard derivative-based methods.

The MM-estimator This method proposed by Yohai [2] involves the following steps as suggested by Stromberg [4]:
1. Compute an initial HBP estimate $\hat{\boldsymbol{\theta}}_{\mathrm{HBP}}$ (we use LMS as well as LTS) and obtain the corresponding residuals $r_i(\hat{\boldsymbol{\theta}}_{\mathrm{HBP}})$.
2. Next, calculate the robust residual scale estimate s_n given by the solution of the following equation

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{r_i(\hat{\boldsymbol{\theta}}_{\mathrm{HBP}})}{s_n}\right)=b \quad \text{with} \quad \rho_0(u)=\rho(u/k_0) \quad \text{and} \quad b/\rho_0(\infty)=0.5, \tag{4}$$

where $k_0 = 0.212$ and ρ is the Hampel loss function defined as follows

$$\rho(u)=\begin{cases}\frac{u^2}{2} & \text{for } |u|<a\\ a\left(|u|-\frac{a}{2}\right) & \text{for } a\leqslant|u|<b\\ ab-\frac{a^2}{2}+(c-b)\frac{a}{2}\left[1-\left(\frac{c-|u|}{c-b}\right)^2\right] & \text{for } b\leqslant|u|\leqslant c\\ ab-\frac{a^2}{2}+(c-b)\frac{a}{2} & \text{for } |u|>c,\end{cases} \tag{5}$$

where $a = 1.5$, $b = 3.5$, and $c = 8$. (Scale estimators measure dispersion and are used to standardize residuals.)
3. Obtain the LS estimate $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$. Then find the M-estimates [5] $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}_1$ to minimize

$$Q(\boldsymbol{\theta})=\sum_{i=1}^{n}\rho_1\left(\frac{r_i(\boldsymbol{\theta})}{s_n}\right) \quad \text{with} \quad \rho_1(u)=\rho(u/k_1), \tag{6}$$

which satisfies $Q(\hat{\boldsymbol{\theta}}_0)\leqslant Q(\hat{\boldsymbol{\theta}}_{\mathrm{HBP}})$ and $Q(\hat{\boldsymbol{\theta}}_1)\leqslant Q(\hat{\boldsymbol{\theta}}_{\mathrm{LS}})$, respectively, where k_1 is chosen as 0.9014 to achieve 95% asymptotic efficiency at the Gaussian distribution. This means that a *local minimum* can be used. The final MM-estimate is then $\hat{\boldsymbol{\theta}}_{\mathrm{MM}}=\min(\hat{\boldsymbol{\theta}}_0,\hat{\boldsymbol{\theta}}_1)$. The basic idea is that this estimate inherits the HBP of the initial estimate and simultaneously improves the efficiency with the M-estimator at step 3.

$\boldsymbol{\tau}$-estimator Yohai and Zamar [3] proposed another HBP estimator with high efficiency. The τ-estimates are defined by minimizing a robust scale of the residuals given by

$$\tau(\boldsymbol{\theta},s_n)=s_n\sqrt{\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{r_i(\boldsymbol{\theta})}{s_n}\right)} \tag{7a}$$

subject to the constraint

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{r_i(\boldsymbol{\theta})}{s_n}\right)=b, \tag{7b}$$

where $b/\rho_0(\infty) = 0.5$ and s_n is an M-estimator of scale implicitly defined by equation (7b). The choice of ρ_0 regulates the robustness, whereas the choice for ρ_1 can be tuned to give good asymptotic efficiency under the Gaussian model. Yohai and Zamar [3] and Tabatabai and Argyros [6] used the ρ function

$$\rho_c(u) = \begin{cases} \frac{u^2}{2}\left(1 - \frac{u^2}{c^2} + \frac{u^4}{3c^4}\right) & \text{for } |u| \leqslant c \\ \frac{c^2}{6} & \text{for } |u| > c. \end{cases} \tag{8}$$

They recommended $\rho_0 = \rho_{c_0}$ with $c_0 = 1.56$ and $\rho_1 = \rho_{c_1}$ with $c_1 = 1.608$. In this case, the τ-estimator's BP is 0.5 and its asymptotic efficiency at the Gaussian distribution is 95%.

Note that the above ρ function does not have a continuous second derivative, which might result in outcomes far from optimality using standard optimization algorithms.

3. SIMULATION STUDY

Description of the test model: oxidation of propylene We consider for the oxidation of propylene the model that involves rate constants with Arrhenius temperature dependence analyzed in Watts [7]

$$-r_{\mathrm{C_3H_6}} = \frac{k_{\mathrm{a}} k_{\mathrm{r}} c_{\mathrm{O_2}}^{0.5} c_{\mathrm{C_3H_6}}}{k_{\mathrm{a}} c_{\mathrm{O_2}}^{0.5} + n k_{\mathrm{r}} c_{\mathrm{C_3H_6}}}, \tag{9}$$

where $r_{\mathrm{C_3H_6}}$ denotes the rate of propylene disappearance, k_{a} and k_{r} denote the rate constants of adsorption of oxygen and oxidation of propylene, respectively, c denotes concentration, and $n =$ (moles oxygen required)/(mole propylene reacted) is the stoichiometric number. To reduce correlation between kinetic parameters in the Arrhenius expression for a rate reaction we used the reparametrization reported in Lohmann et al. [8] resulting in $\boldsymbol{\theta} = (\ln k_{\mathrm{a}}(350\ ^\circ\mathrm{C}), \ln k_{\mathrm{a}}(390\ ^\circ\mathrm{C}), \ln k_{\mathrm{r}}(350\ ^\circ\mathrm{C}), \ln k_{\mathrm{r}}(390\ ^\circ\mathrm{C}))$ as the vector of parameters to be estimated.

Experimental Simulation is conducted to compare the small-sample behavior of the estimates described in the former section. More precisely, we compare LS, LMS, LTS(α) for $\alpha = 0.1$, 0.25, and 0.5, τ-, and MM-estimators starting with three different HBP initial estimates—LMS, LTS(0.25), and LTS(0.5). Each sample contains 66 observations $(\boldsymbol{x}_i, y_i)$ in which $\boldsymbol{x}_i$ is taken from the experimental data. We have taken the set of LS estimates of the experimental data as the true parameters to generate predictions of the measured quantities y_i according to model (1). The error terms e_i are generated from five different distributions: Gaussian $\mathrm{N}(0, \sigma_e^2)$, Cauchy, Skew-Normal [9], and two "scale" contaminated Gaussians $0.9\mathrm{N}(0, \sigma_e^2) + 0.1\mathrm{N}(0, (2\sigma_e)^2)$ and $0.7\mathrm{N}(0, \sigma_e^2) + 0.3\mathrm{N}(0, (5\sigma_e)^2)$—denoted as $\mathrm{CN}(0.1, 2)$ and $\mathrm{CN}(0.3, 5)$, respectively. Two proportions of outliers in data are considered in each simulated data set, namely 10% (small contamination) and 30% (high contamination). Certain observations, chosen at random, are modified to be "bad" data points by shifting upwards the response variable by $5\sigma_e$ for 10% bad data and $10\sigma_e$ for 30% bad data. Here, σ_e may be estimated from the data as $\hat{\sigma}_e = \sqrt{\sum_{i=1}^{n} r_i(\hat{\boldsymbol{\theta}}_{\mathrm{LS}})/(n - n_p)}$ with n_p being the number of parameters in model (1). The number of Monte Carlo replications

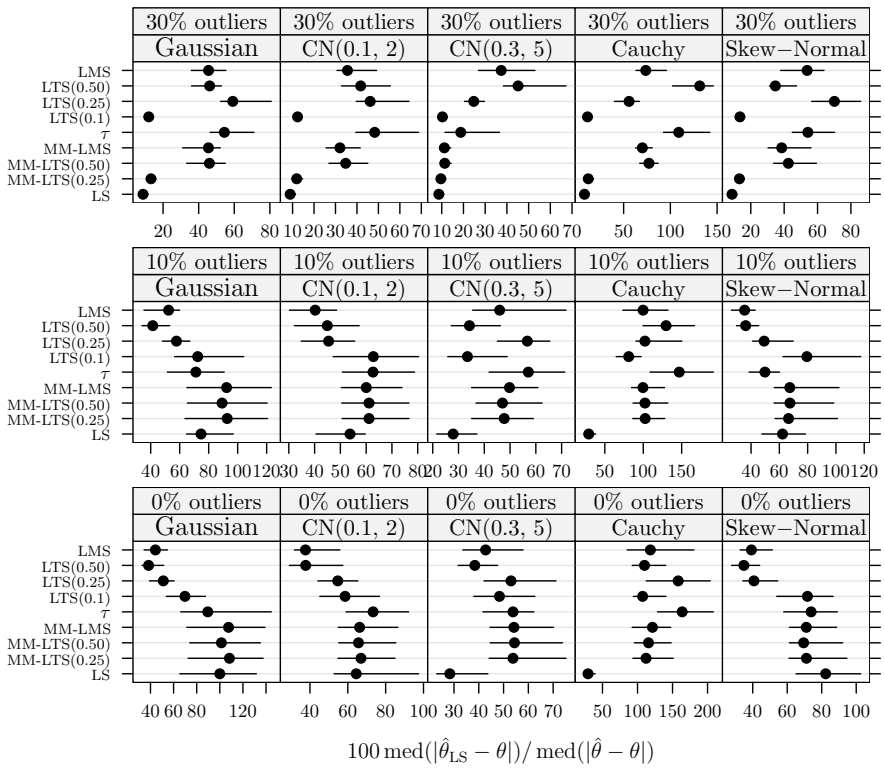

Figure 1. Efficiencies of the competing estimates for $\ln k_{\mathrm{a}}(390\ ^{\circ}\mathrm{C})$. The efficiency criterion has been normalized by its value for the LS estimate obtained only under Gaussian error (contamination 0%). Each circle shows the value of efficiency, whereas the darker line segment shows the bootstrap [15] percentile 95% confidence interval obtained with 999 bootstrap replications.

is 100. We used the criterion $\mathrm{med}(|\hat{\theta} - \theta|)$ [10], a robust analog of the MSE, to evaluate the performance of an estimator.

Computing the estimates The major computational difficulty with the estimates considered in this paper is that they cannot be calculated by standard optimization algorithms. We therefore adopted the improved version by Lee et al. [11] of the differential evolution (DE) algorithm proposed by Storn and Price [12] for all the regression estimators. This method is a stochastic global search heuristic that applies to bound constrained problems. Note that if the univariate scale estimator is computed from (7b), then by plugging s_n into (7a) a τ-estimate can be obtained by solving an unconstrained minimization

problem. A convenient procedure to obtain the solution for both the scale estimators (4) and (7b) is the algorithm of Brent [13] that does not require derivatives. The final stage of the MM-estimator uses the L-BFGS-B algorithm [14].

4. MONTE CARLO RESULTS

Fig. 1 displays results concerning the performance of the robust estimators. For shortness, we only report the simulation results for the $\ln k_{\mathrm{a}}(390\ ^{\circ}\mathrm{C})$ parameter.

As expected, for departures from Gaussian distributed data (especially for $\mathrm{CN}(0.3, 5)$ and Cauchy), we clearly see the advantage of the robust methods over the classical. This becomes even more visible for the outlier contamination scenarios. Generally, no significant difference could be found between the MM-estimates computed with the LMS estimator and when using the LTS estimator, except when the outlier proportion is 30% for which LTS(0.25) is worst. Furthermore, we also note that for (uncontaminated) Gaussian distributed errors the loss in efficiency of both τ- and MM-estimates with respect to least squares is rather small or barely distinguishable.

For null or small contamination levels, we can observe that the τ- and MM-estimates show an overall best behavior, albeit quite close to the LTS(0.1) estimator. On the other hand we can see that, in general terms, the LMS and LTS(0.5) estimates are the worst, followed by LTS(0.25). For high contamination, essentially these estimators behave the opposite way compared to small fractions of contamination. Note that among HBP estimates, LTS(0.1) and MM- clearly lose.

These simulation results support the use of the MM- or τ-estimator as a valuable alternative to the existing classical methods in the practical applications for which the proportion of outliers in data is not too large.

REFERENCES

1. P.J. Rousseeuw, J. Am. Stat. Assoc. 79 (1984) 871.
2. V.J. Yohai, Ann. Stat. 15 (1987) 642.
3. V.J. Yohai and R.H. Zamar, J. Am. Stat. Assoc. 83 (1988) 406.
4. A.J. Stromberg, J. Am. Stat. Assoc. 88 (1993) 237.
5. P.J. Huber, Robust Statistics, Wiley, New York, 1981.
6. M.A. Tabatabai and I.K. Argyros, Appl. Math. Comput. 58 (1993) 85.
7. D.G. Watts, Can. J. Chem. Eng. 72 (1994) 701.
8. T. Lohmann, H.G. Bock, and J.P Schlöder, Ind. Eng. Chem. Res. 31 (1992) 54.
9. A. Azzalini, Scand. J. Stat. 12 (1985) 171.
10. J. You, Comput. Stat. Data Anal. 30 (1999) 205.
11. M.H. Lee, C. Han, and K.S. Chang, Ind. Eng. Chem. Res. 38 (1999) 4825.
12. R. Storn and K. Price, J. Glob. Optim. 11 (1997) 341.
13. R.P. Brent, Algorithms for Minimization without Derivatives, Prentice-Hall, Englewood Cliffs, NJ, 1973; reissued by Dover Publications, Mineaola, NY, 2002.
14. C. Zhu, R.H. Byrd, P. Lu, and J. Nocedal, ACM Trans. Math. Softw. 23 (1997) 550.
15. R. Wehrens, H. Putter, and L.M.C. Buydens, Chemom. Intell. Lab. Syst. 54 (2000) 35.