Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Published online: 15.05.2025
DOI : 10.15672/hujms.1580654

Research Article

Robust multiple regression based on shrinkage S_n estimator

Lakshmi Raveendran, Sajesh T. Abraham *

Department of Statistics, St. Thomas College (Autonomous), Thrissur, Affiliated to University of Calicut, 680001, Thrissur, Kerala, India

Abstract

Regression analysis is used to model the data statistically. However, data modeling and interpretation are affected by outliers and significant points. Robust regression analysis offers an alternative. In this study, the parameters that define the linear regression problem are estimated using a robust approach. The concept of shrinkage, which has been investigated for outlier detection in multivariate data. A comprehensive simulation analysis is performed to examine the breakdown value of the regression estimator, the affine equivariance, the robustness against contamination, and the efficiency with normal errors. The advantages of the suggested robust estimator in regression are demonstrated by the simulation results and real-world data examples. Simulation and research are conducted using the R software.

Mathematics Subject Classification (2020). 62H86, 62H12, 62F35

Keywords. Data modeling, regression, reweighted estimator, robust shrinkage S_n.

1. Introduction

Regression analysis is commonly used in various academic disciplines, including social sciences, health sciences, engineering, and physical sciences, among others. This method relies primarily on ordinary least squares, which makes it vulnerable to problems, especially in the presence of outliers. Outliers are defined as observations that significantly deviate from the majority of data points in a data set. As a result, robust regression was created as a more reliable and efficient alternative to least squares in scenarios where the data set includes contaminated points. Numerous robust regression algorithms are available, some of which are also resistant to outliers. Unlike classical linear regression, which focuses primarily on estimating unknown regression parameters, robust regression aims to provide reliable estimates even in the presence of data contamination. Consider a linear multiple regression model for a sample of size n as

$$y_i = \alpha + \mathbf{x_i^t}\beta + \epsilon_i, \quad i = 1, 2, 3, ..., n, \tag{1.1}$$

*Corresponding Author.
Email addresses: lakshmi.nss19@gmail.com (L. Raveendran), sajesh.t.abraham@gmail.com (S. T. Abraham)
Received: 07.12.2024; Accepted: 22.04.2025

where α is the unknown intercept, β represents the unknown vector of regression parameters of size $p \times 1$. The error terms ϵ_i follow an independent and identically distributed $i.i.d.$ normal distribution and are independent from $\mathbf{x_i^t}$, the p - dimensional regressor variables. The classical method known as ordinary least squares (OLS) minimizes the sum of squared residuals, and it is the traditional approach for estimating the parameters in Eq. (1.1) as follows:

$$\hat{\beta}_{\text{OLS}} = \min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x_i^t}\beta)^2. \tag{1.2}$$

When the data meet these classical estimators' assumptions, they perform well. The OLS estimator in Eq. (1.2) can be expressed as follows: Let $\mathbf{z} = (\mathbf{x}, y)$ represent the joint variable of the response and carriers. Let μ be the location and Σ be the scatter matrix of $\mathbf{z}$. Partitioning of μ and Σ with respect to $(\mathbf{x}, y)$, we have

$$\mu = \begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_y \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}. \tag{1.3}$$

The empirical mean $\hat{\mu}$ and the empirical covariance matrix $\hat{\Sigma}$ are traditionally used to estimate them. Specifically, the OLS estimators of β and α can be expressed as functions of the $\hat{\mu}, \hat{\Sigma}$ components as follows:

$$\hat{\beta} = \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}, \quad \hat{\alpha} = \hat{\mu}_y - \hat{\beta}^t\hat{\mu}_{\mathbf{x}}. \tag{1.4}$$

The OLS is widely acknowledged for its vulnerabilities, particularly in producing unstable estimates and unreliable predictions in the presence of outliers within the data. A robust alternative to managing outliers is robust regression estimation. Robust methods aim to develop estimators that are resilient to outliers. Efficiency and breakdown point are key metrics to evaluate their performance. The relative efficiency of robust estimators with respect to the OLS is often used to assess their effectiveness. The breakdown point measures an estimator's tolerance to outliers, with the maximum asymptotic breakdown point being $\frac{1}{2}$. In this paper, we explore several well-known and effective robust regression techniques for multiple linear regression models that have contaminated data. In addition, classical normal equations can be represented by a covariance matrix, making it easy to perform multiple regression.

In this paper, we introduce a robust reweighted regression that utilizes the Shrinkage S_n covariance matrix as proposed by [20], while also addressing robust regression estimators. The proposed estimator aims to enhance robustness while maintaining computational efficiency and affine equivariance. The results demonstrate that the shrinkage-based estimator outperforms the classical OLS and some of the other robust estimators, making it a valuable alternative in regression analysis. The core concept is to employ a simulated dataset to compare different methodologies and to depend exclusively on comprehensive empirical simulations to evaluate the characteristics of the proposed estimator.

The paper is organized as follows: Section 2 provides a brief overview of the OLS method, highlighting the importance of robust methodologies and notable robust estimators that have been developed over the years. A novel reweighted regression estimator based on a robust covariance matrix approach is also introduced. In Section 3, the proposed estimator is compared with other robust regression estimators using various simulation methods. Furthermore, the properties of the proposed estimator are evaluated through extensive simulation. Section 4 presents the conclusions of the paper along with an application involving real-world data. Equivariance, breakdown, and robustness properties are studied in Section 5. The sensitivity curve is shown in Section 6 and applications with various data are given in Section 7. The conclusion is given in Section 8.

2. Literature review

The primary aim of robust methods is to establish estimators that demonstrate resilience to outliers. Efficiency and the breakdown point serve as two conventional metrics that are used to assess the efficacy of the robust approaches currently used. The OLS is characterized by the lowest variance among unbiased estimators in scenarios involving normally distributed errors with constant variance without outliers, which results in high efficiency. Therefore, when the error distribution is precisely normal and the data set does not have outliers, the relative efficiency of robust estimates compared to the OLS is frequently considered a crucial metric for evaluating and contrasting the effectiveness of various robust methodologies.

The breakdown point indicates the percentage of outliers that an estimate can withstand, while the asymptotic breakdown point is the limit of the finite sample breakdown point as the sample size approaches infinity. If more than half of the observations are contaminated, distinguishing between contamination and valid data becomes impossible, which makes the maximum asymptotic breakdown point $\frac{1}{2}$ [31]. In comparison, the OLS has an asymptotic breakdown point of 0 and a finite sample breakdown point of $\frac{1}{n}$. In 1887, Edgeworth [12] developed the least absolute deviation (LAD) method, also known as least absolute value (LAV) regression or L_1 estimator, building on Roger Joseph Boscovich's idea. The LAV regression is classified as an L estimator and differs from the traditional least-squares regression by minimizing the sum of the absolute values of the residuals instead of their squares. This makes the LAV regression more robust to outliers as they have less influence on the results. Although LAV is less affected by odd values y compared to OLS, it cannot determine leverage levels [28], and its breakdown point is limited to $\frac{1}{n}$. The M estimator was the next development in this direction. Huber [34] presents the M estimator by replacing the least-squares criterion with a robust residual loss function. It was more efficient than LAD. Since M estimators are resistant to heavy-tailed error distribution and nonconstant error variance, they are very resistant to y outliers with a breakdown point of 0.5. However, the finite sample breakdown point of both LAD and M tends to 0, because of the possibility of leverage points [27]. Due to the vulnerabilities of M-estimators, generalized M-estimators, or GM estimators, were introduced. These estimators were designed to effectively address the challenge of identifying leverage points. However, they did not distinguish between "good" and "bad" leverage points. Moreover, it is important to note that the breakdown point decreases as the data dimension p increases.

The least median of square (LMS) was proposed by [30], which minimizes the median of squared residuals. The slow convergence rate of LMS contributes to its low efficiency, even if it has a high breakdown point. Rousseeuw [31] proposed another L estimator technique, called the least trimmed square (LTS). The trimmed-mean approach is expanded upon. The method is to minimize the sum of trimmed squared residuals. LTS regression is scale- and affine equivariant. Setting q=$\frac{n}{2}$+1 makes sure the estimator has a breakdown of 0.5. The issue is that LTS performs less efficiently than OLS [40]. Rousseeuw and Yohai [32], created a high breakdown value approach, which minimizes the dispersion of residuals with high asymptotic efficiency and improves the convergence rate of the objective function over LTS because LTS and LMS estimators have weak convergence rates. The method has definitely greater efficiency than LTS. Although robust against response outliers, S-estimators are still vulnerable to high leverage points (extreme values in predictor variables), which can distort the estimates. In order to increase efficiency, Croux et al. [10] proposed the generalized S-estimator (GS-estimator), but again there was a constant to define, which depends on sample size and dimension. The most popular MM regression estimator was introduced by [41]. The first phase required a solid and consistent estimate of the regression parameters with a high breakdown point, although not necessarily a high efficiency. The LMS and S-estimates with Huber or bisquare functions

are the most widely used initial estimators in practice. Compared to regular least squares, MM estimation has an efficiency of approximately 95% due to its combination of high breakdown value estimator.

The weighted least squares estimator (REWLSE) presented by [13], which is a reliable and effective method. Under Gaussian errors, the approach simultaneously achieves maximum breakdown point and complete efficiency. The plan is to use hard rejection weights (0 or 1) derived from a first robust estimator. Because of the adaptive cut-off point, which is based on the distribution of the standardized absolute residuals, the method has complete asymptotic efficiency and is asymptotically identical to OLS. The REWLSE performs well with heavy-tailed errors, but it struggles to identify leverage points. Furthermore, it performs poorly in lower-dimensional settings [2]. The REWLSE is competitive with SR [2] in high dimension, but in low dimension the REWLSE shows higher errors. The REWLSE outperforms in heavy-tailed errors, but fails to perform in identifying leverage points. Also, for smaller dimensions, the REWLSE does not perform. With the covariance approach, an alternative way, the OLS estimators could be determined, as defined in equation 1.4. The classical sample estimators' $\hat{\mu}, \hat{\Sigma}$ used in 1.4 are sensitive to the presence of outliers, which is considered a downside of this method.

Robust estimators must be used and were put in place by [9, 25]. They proposed the application of the S-estimator (referred to as technique S) and multivariate M-estimators. A shrinkage-based covariance matrix as an alternative to $\hat{\Sigma}$ was introduced by [2], along with L_1. Median as an option for $\hat{\mu}$ in Eq. (1.4). Consequently, the authors suggested an alternative to the OLS and conducted a comparative study to evaluate the performance of their estimator. In conclusion, some of these robust methods are resistant to response outliers, but are not immune to leverage points or able to discriminate between appropriate and inappropriate leverage. Obtaining maximum breakdown while maintaining good efficiency is a challenge. The best options appear to be the MM-estimator because of their high asymptotic efficiency and high breakdown point. Although some of these methods have a high breakdown point [42], they are computationally challenging in large data sets of high dimensions [15] and [40]. That is why resampling algorithms are used to obtain a number of subsets and then compute the robust regression estimate from a number of initial estimates. However, the property of high decomposition generally requires that the number of elementary sets go to infinity [15]. An improved M-robust regression method for anomaly detection was discussed in [17], regarding dam safety monitoring data, addressing issues such as misjudgment and missed outliers. It introduces an AR factor to handle random variables and optimize the residual calculation model to enhance robustness. They are mainly using improved M-based robust regression for anomaly detection purposes.

A robust alternative to the maximum likelihood estimate (MLE) was proposed by [3], it minimizes the squared difference between the true density $g(x)$ (unknown) and the assumed parametric model $f(x|\theta)$. The proposed method is robust, but can be computationally intensive and sensitive to parameter choices. It may also face challenges in high-dimensional settings, especially when residuals are skewed or difficult to interpret graphically. In this paper, we propose to use robust shrinkage estimators instead of classical estimators in Eq.(1.4) as proposed by [20]. They obtained a shrinkage covariance matrix and evaluated the efficiency, robustness, and sensitivity curve of the proposed estimator in their paper. Here, all these estimators are used to develop the reweighted regression estimator.

3. Proposed method

The principle behind shrinkage estimation lies in the notion of "shrinking" an estimator $\hat{E}$ toward a target estimator $\hat{T}$, which serves to effectively diminish estimation errors.

This approach capitalizes on the fact that while the shrinkage target $\hat{T}$ may exhibit bias, it typically shows lower variance compared to the estimator $\hat{E}$. By carefully calibrating the level of shrinkage, represented by the intensity of the shrinkage η, the resulting shrinkage estimator can surpass $\hat{E}$ in terms of reducing estimation errors, provided that certain general conditions are met [19] as follows:

$$\hat{\Sigma}_{Sh} = (1-\eta)\hat{E} + \eta\hat{T}. \tag{3.1}$$

Using a shrinkage estimator offers a significant benefit in balancing bias and variance. Cabana et al. [5] proposed the shrinkage estimator of the L_1 - median as a robust alternative to location and the shrinkage estimator based on L_1-median is defined as

$$\hat{\boldsymbol{\mu}}_{Sh} = (1-\eta)\hat{\boldsymbol{\mu}}_{MM} + \eta\nu_{\boldsymbol{\mu}}\mathbf{e}, \tag{3.2}$$

where $\nu_{\boldsymbol{\mu}}\mathbf{e}$ is the shrinkage target matrix, $\mathbf{e}$ is a vector of ones with p - dimension and $\hat{\boldsymbol{\mu}}_{MM}$ is the L_1-median from the samples.

The scaling factor $\nu_{\boldsymbol{\mu}}$ and the shrinkage intensity η should be such that they minimize the expected quadratic loss. Lakshmi and Sajesh [20] proposed the shrinkage S_n covariance matrix as a robust alternative to the covariance estimate. S_n covariance of two random variables X and Y be

$$S_n(\mathrm{X},\mathrm{Y}) = 1.4304(\mathrm{med}_i[\mathrm{med}_{j\neq i}\{(x_i - x_j)(y_i - y_j)\}]).$$

Let $\mathbf{X}$ be $n\times p$ matrix with sample size n, number of variables p, and $\mathbf{X_j}(j = 1, 2, \ldots, p)$ be the column of the matrix. The covariance matrix of $\mathbf{X}$ based on S_n would be $\hat{S_n} = S_n(\mathbf{X_i}, \mathbf{X_j})$. Then, the covariance matrix based on shrinkage S_n would be

$$\begin{gathered}\hat{\Sigma}_{Sh} = (1-\eta)\hat{E} + \eta\hat{T},\\ \text{where } \hat{E} = \hat{S_n}.\end{gathered} \tag{3.3}$$

In this paper, we utilize the above-defined location estimate and covariance matrix estimate in Eq. (1.4) and propose a reweighted regression estimator. Consider $\mathbf{z} = (\mathbf{x}, y)$, the joint variable with location and covariance matrix μ, Σ, respectively. The associated squared Mahalanobis distance for each observation $\mathbf{z}_i, i = 1, 2, 3, \ldots, n$, based on $\hat{\mu}_{Sh}$ and $\hat{\Sigma}_{Sh}$ be

$$\mathrm{RD}^2(\mathbf{z}_i) = (\mathbf{z}_i - \hat{\mu}_{Sh})^t\hat{\Sigma}_{Sh}^{-1}(\mathbf{z}_i - \hat{\mu}_{Sh}). \tag{3.4}$$

The weight function based on the robust Mahalanobis distance is $w_i = w(\mathrm{RD}^2(\mathbf{z}_i))$, where a weight of 1 is assigned to the observations $(\mathbf{z}_i)$ with a Mahalanobis distance less than $\frac{\chi^2_{0.95,p}\times\mathrm{median}(\mathrm{RD}^2(\mathbf{z}_i))}{\chi^2_{0.5,p}}$. Thus, the reweighted shrinkage location and S_n covariance matrix is defined as

$$\hat{\mu}^1 = \frac{\sum_{i=1}^n w_i\mathbf{z}_i}{\sum_{i=1}^n w_i}, \quad \hat{\Sigma}^1 = \frac{\sum_{i=1}^n w_i(\mathbf{z}_i - \hat{\mu}^1)(\mathbf{z}_i - \hat{\mu}^1)^t}{\sum_{i=1}^n w_i}. \tag{3.5}$$

The regression estimates, based on one-step reweighted shrinkage are $\hat{\beta}^1$ and $\hat{\alpha}^1$ based on $\hat{\mu}^1$ and $\hat{\Sigma}^1$ be defined as

$$\hat{\beta}^1 = (\hat{\Sigma}^1)_{xx}^{-1}(\hat{\Sigma}^1)_{xy}, \quad \hat{\alpha}^1 = (\hat{\mu}^1)_y - (\hat{\beta}^1)^t(\hat{\mu}^1)_{\mathbf{x}}, \tag{3.6}$$

where $(\hat{\beta}^1, \hat{\alpha}^1)^t$ is our one-step reweighted shrinkage-based regression estimator. The scale estimate of errors based on the regression estimator defined above is given by

$$\hat{\sigma} = (\hat{\Sigma}^1)_{yy} - (\hat{\beta}^1)^t(\hat{\Sigma}^1)_{xx}\hat{\beta}^1.$$

The next step is to reweight by taking into consideration the residuals based on the following one-step reweighted shrinkage based regression estimator

$$r_i = y_i - (\hat{\beta}^1)^t\mathbf{x}_i - \hat{\alpha}^1.$$

The Mahalanobis distance for the above estimator defined residuals is

$$d(r_i) = ((r_i)^t(\hat{\sigma})^{-1}r_i)^{1/2}.$$

Let $wr_i = w(d^2(r_i))$ be the weighting function with respect to the residual Mahalanobis distance, where a weight of 1 is assigned to the residuals with a Mahalanobis distance less than $\chi^2_{1,0.99}$. Define $\mathbf{u}_i$=$(\mathbf{x}_i^t, 1)^t$, then we have

$$\phi^{WShS_n} = ((\hat{\beta}^{WShS_n})^t, \hat{\alpha}^{WShS_n})^t = (\sum_{i=1}^{n} wr_i\mathbf{u}_i\mathbf{u}_i^t)^{-1}\sum_{i=1}^{n} wr_iy_i\mathbf{u}_i. \tag{3.7}$$

Note that Equation (3.7) is the two-step reweighted regression estimator based on shrinkage S_n ($WShS_n$).

4. Simulation

This section presents the results of a simulation study that compares the performance of the proposed $WShS_n$ regression estimator with the OLS and several robust regression techniques previously discussed, namely MM, S, LTS, LMS and the reweighted regression estimator based on the shrinkage comedian (SR). The simulations were conducted using the R software using the following functions: The lmrob.S function of the robustbase package for the S estimator, the rlm function for the MM method of the MASS package, the lmsreg function from the MASS package for the LMS, and the lqs function from the MASS package for the LTS estimator. We utilized the built-in functions in R for our simulation study. Consider the linear model

$$y = \alpha + \mathbf{x}\beta + \epsilon, \tag{4.1}$$

where $\mathbf{x}$ is $n \times p$ matrix, β is $p \times 1$ vector of unknown regression coefficients, α unknown intercept, and ϵ are *i.i.d.* error variable.

The independent variable $\mathbf{x}$ follows a multivariate normal distribution with mean vector $0_{p\times 1}$ and covariance matrix, an identity matrix $p \times p$. For the simulation study, the sample sizes considered in different situations are n = 80, 100, 150, 200, 500, 1000 and the dimensions considered are p = 5, 10, 15, 20, 30. 1000 times each of the simulation scenarios is repeated in our study. We have considered simulation scenarios similar to those found in the literature [1,2,9,13,25,36,42]. In the first scenario, we generate the response variables from a standard normal distribution with $\beta = 0$, $\alpha = 0$, and the errors are considered standard Gaussian. Let the scenario be denoted as NE. In the second scenario to evaluate robustness, normal errors are considered as in NE but with a probability of δ contamination in response and independent variables. Independent variables are taken from $N(\lambda\sqrt{\chi^2_{p,0.99}}, 1)$ and response variables from $N(k\sqrt{\chi^2_{1,0.99}}, 1)$ where λ, k = 0, 0.5, 1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Let the respective scenario be denoted as NEO further in the article. The percentages of contamination considered are δ = 10%, 20%. If $\lambda = 0$ and $k > 0$, we will get y outliers. Similarly, if $\lambda > 0$, $k = 0$, we will get good leverage points, and we will obtain bad leverage points for $\lambda > 0$, $k > 0$. Thus, our choice of λ and k gives data that range from extreme outliers to intermediate outliers. Under the NE criteria, it is clear that the OLS will have maximum efficiency. For analyzing the performance of the estimators, we have considered the efficiency of the estimators as a metric. Let $\phi = (\beta^t, \alpha)^t$ be the joint vector $(p+1)\times 1$ of the regression coefficients. The finite sample efficiency of any robust method R is defined by [13] and given by

$$\text{Eff} = \frac{1/N\sum_{i=1}^{N}||\hat{\phi}_{OLS}^{(i)} - \phi||_2^2}{1/N\sum_{i=1}^{N}||\hat{\phi}_{R}^{(i)} - \phi||_2^2}. \tag{4.2}$$

Table 1. Finite sample efficiency in case of normal errors

n	p	$WShS_n$	MM	S	LTS	LMS	SR
	5	0.90111	0.8901934	0.3133835	0.1687298	0.1870315	0.91122
80	10	0.91998	0.9222617	0.2686664	0.1470201	0.128014	0.92895
	15	0.933981	0.9306423	0.2520925	0.1427853	0.1423768	0.93391
	20	0.91189	0.8765167	0.2664371	0.1724709	0.1269676	0.96774
	30	0.98881	0.8412267	0.273741	0.1836285	0.1050502	0.91334
100	5	0.91198	0.9625692	0.3307681	0.1678611	0.1163469	0.92122
	10	0.95003	0.9477326	0.2681845	0.1338582	0.1469313	0.93895
	15	0.92289	0.9122369	0.2664602	0.1358261	0.1444049	0.93581
	20	0.94478	0.9410851	0.2956971	0.1368254	0.1305699	0.96994
	30	0.92339	0.8474065	0.3050582	0.167848	0.1229417	0.92293
150	5	0.93328	0.9304732	0.3096604	0.11059407	0.14979494	0.891122
	10	0.95001	0.9483747	0.3067942	0.13424255	0.10780639	0.938895
	15	0.94991	0.9346738	0.2526533	0.11922365	0.11141836	0.958391
	20	0.93391	0.9345712	0.2508206	0.11280803	0.10334599	0.967874
	30	0.93443	0.8910862	0.2851218	0.09904998	0.09132823	0.922234
200	5	0.96991	0.9410785	0.3035179	0.11344367	0.12879938	0.89012
	10	0.94411	0.9328904	0.2738827	0.0992274	0.09473856	0.93915
	15	0.94471	0.9434414	0.2659222	0.08857581	0.08025978	0.95691
	20	0.92287	0.9314504	0.2494056	0.08732885	0.09294969	0.93399
	30	0.92278	0.9114167	0.2663687	0.08309036	0.07708978	0.92224
300	5	0.99981	0.9836872	0.3414343	0.10978863	0.11293767	0.90012
	10	0.94441	0.9344473	0.2969611	0.08428847	0.07586824	0.92981
	15	0.96227	0.9637224	0.2685592	0.06195658	0.06536106	0.93441
	20	0.95619	0.9543658	0.2491195	0.05939695	0.05864878	0.97001
	30	0.95671	0.9463761	0.2569738	0.05703498	0.05549416	0.92881
500	5	0.98827	0.9707854	0.352695	0.08741342	0.08629351	0.92199
	10	0.95189	0.9504772	0.3079973	0.06210263	0.06096699	0.93391
	15	0.93891	0.9365855	0.2447813	0.04703221	0.04344129	0.94417
	20	0.95178	0.9402005	0.2720199	0.04044304	0.03248954	0.91887
	30	0.96881	0.9514596	0.2416872	0.03636119	0.03106638	0.93241
1000	5	0.98811	0.947679	0.3242796	0.05436184	0.05822269	0.92289
	10	0.95587	0.9488504	0.3353234	0.03039559	0.03082224	0.94805
	15	0.96111	0.9575694	0.2741748	0.02430555	0.02085405	0.96791
	20	0.94445	0.935774	0.2761304	0.0203216	0.01564593	0.97987
	30	0.95011	0.9417903	0.268742	0.01838522	0.01896355	0.97733

Table 1 presents the simulation results of the relative efficiency for the joint regression estimator ϕ based on shrinkage S_n ($WShS_n$) compared to other robust methods. The results indicate that our proposed method is more efficient than the robust S, LTS, and LMS estimators. Furthermore, $WShS_n$ along with the MM and SR estimators, achieves efficiency values close to one. Table 1 confirms the performance of our proposed estimator in non-contaminated data. The results indicate that, regardless of sample size and dimension, our estimator demonstrates higher relative efficiency compared to other robust methods. Furthermore, its efficiency improves as the dimension increases. In contrast, the LTS and LMS methods consistently perform poorly across all dimensions and sample sizes,

highlighting their limitations for use in non-contaminated datasets. A robust method that matches the performance of OLS in clean data while outperforming it in contaminated cases is ideal, making it highly suitable for practical applications. The efficiency value of our proposed estimator remains consistently close to one, indicating that it performs well in non-contaminated data. To assess the robustness property, that is, the NEO criteria, the mean square error (MSE) of the estimated parameter $\phi = (\beta^t, \alpha)^t$ averaged over the identified simulation runs N. We consider the maximum MSE for different values of k for each λ as

$$\mathrm{MSE}_\lambda(.)_{\max} = \max_{k\epsilon 0,0.5,\ldots,10}\mathrm{MSE}_{\lambda,k}(.).$$

Finally, the maximum MSE is considered; that is, the metric for the assessment of NEO performance is defined as follows:

$$\mathrm{MSE}()_{\max} = \max_{\lambda\epsilon 0,0.5,\ldots,10}\mathrm{MSE}_\lambda(.).$$

Table 2 shows that our proposed method exhibits less $\mathrm{MSE}()_{\max}$ than other methods for contamination of 10% and 20%. In addition, our proposed regression estimator shows a decrease $\mathrm{MSE}()_{\max}$ as the dimension increases. Even for higher contamination in the dataset, our estimator possesses the least compared $\mathrm{MSE}()_{\max}$ to other estimators. The classical method, OLS, is not robust and exhibits a drastically high $\mathrm{MSE}()_{\max}$ value. Among other robust estimators, the MM estimator shows the minimum $\mathrm{MSE}()_{\max}$. The LMS shows the worst performance in terms of $\mathrm{MSE}()_{\max}$ robustness.

Table 2. $\mathrm{MSE}()_{\max}$ of estimates for checking robustness - NEO case

p	δ	OLS	$WShS_n$	MM	S	LTS	LMS	SR
5		9.8874	0.002384	0.095852	0.053662	0.058848	0.059968	0.048566
10		8.9993	0.002662	0.038258	0.027147	0.06561	0.063555	0.029432
15	10%	7.6712	0.002871	0.02485	0.026043	0.078443	0.07844	0.022529
20		6.8201	0.003129	0.021788	0.022529	0.094063	0.093965	0.019781
30		3.1178	0.003704	0.027099	0.019681	0.135806	0.138321	0.015976
5		9.9991	0.024481	0.195169	0.140195	0.122953	0.139308	0.225836
10		9.0001	0.006446	0.091101	0.079506	0.132886	0.130309	0.120671
15	20%	8.5619	0.007887	0.077203	0.059177	0.152587	0.15697	0.079638
20		7.2211	0.007104	0.091241	0.043591	0.233085	0.233583	0.065083
30		4.1871	0.007844	0.064294	0.037972	1.178396	1.555017	0.046544

5. Equivariance and breakdown

Equivariance, breakdown, and robustness properties that determine an estimator's actual usefulness rather than theoretical goodness. Three forms of equivariance are taken into consideration for regression estimators. The regression equivariant is equivalent to adding the coefficients of this linear function to the estimators if we convert the dependent variable by adding a linear function of independent variables. y - equivariant is the estimators transform correctly if the response variable is changed linearly. These two mentioned properties can be summarized as follows:

$$\hat{\phi}_{WShS_n}(\mathbf{x}, yb + \mathbf{x}g + u) = \hat{\phi}_{WShS_n}(\mathbf{x}, y)b + (g^t, u)^t, \tag{5.1}$$

where $b \epsilon \mathbf{R}$ is any non-zero constant, g be any $p \times 1$ vector and $u \epsilon \mathbf{R}$ is any constant. Keeping $\mathbf{x}$ the same and transforming the dependent variable as $yb + \mathbf{x}g + u$, the resulting transformed estimators are $\hat{\beta}^{new}_{WShS_n} = b(\hat{\beta}_{WShS_n}) + g$ and $\hat{\alpha}^{new}_{WShS_n} = b\hat{\alpha}_{WShS_n} + u$.

If the independent variables undergo a linear transformation, the equivalent transformed estimator can be defined for $\mathbf{x}$ - equivariance $\hat{\phi}_{WShS_n}(\mathbf{x}A, y) = ((\hat{\beta}_{WShS_n})^t(A^{-1})^t, \hat{\alpha}_{WShS_n})^t$. Note that if independent variables are transformed using any non-singular matrix $p \times p$ A, then the resulting new regression estimators are $\hat{\beta}^{new}_{WShS_n} = A^{-1}\hat{\beta}_{WShS_n}$ and the intercept remains the same. Since it is impossible to investigate all possible transformations, [26] and [37] suggested that A matrices be randomly generated to check $\mathbf{x}$ - equivariance as $A = TD$, where D is a $p \times p$ diagonal matrix with diagonal entries that are uniformly and independently distributed and T is a random orthogonal matrix.

In this study, we examine the $\mathbf{x}$ - equivariance property of the proposed estimator using the approaches described above. Also, we propose a random non-zero zero b,g and u to check regression and y - equivariance. The approximate affine equivariance of the initial estimates $\hat{\mu}_{Sh}$ and $\hat{\Sigma}_{Sh}$ is demonstrated by extensive simulation by [20]. Our main concern is about the estimator $\hat{\phi}_{WShS_n} = (\hat{\beta}^t_{WShS_n}, \hat{\alpha}_{WShS_n})^t$. We studied the equivariance property of the proposed estimator on transformed data as described above. Similarly to the simulation scenarios NE and NEO, we have considered contamination $\delta = 0\%$, 10%, 20%. $\hat{\phi}_{WShS_n}$ is obtained from the non-transformed data and recorded. The data is then transformed as in the explanations given above and $\hat{\phi}^{new}_{WShS_n}$ is recorded from the transformed data. The MSE is calculated between $\hat{\phi}^{new}_{WShS_n}$ and the estimator should determine if the equivariance property holds. Tables 3 and 4 show $\text{MSE}_\lambda(\hat{\phi}^{new}_{WShS_n})_{\max}$ for each λ. The MSE is minimal even for the higher dimension in the equivariance y and is negligibly low in all contamination. The same pattern as we could see in $\mathbf{x}$ - equivariance, too. The values give us empirical confirmation regarding the equivariance property of the new robust estimator.

Table 3. Affine y - equivariance and regression equivariance $\text{MSE}_\lambda(.)_{\max}$ values

	$p = 5$			$p = 30$		
λ	$\delta = 0\%$	$\delta = 10\%$	$\delta = 20\%$	$\delta = 0\%$	$\delta = 10\%$	$\delta = 20\%$
0	0.01692	0.03357	0.08543	0.0006	0.02763	0.09313
0.5	0.01693	0.03212	0.03269	0.00065	0.00526	0.00292
1	0.01645	0.03259	0.02042	0.00061	0.00541	0.00266
1.5	0.01687	0.03309	0.02099	0.00058	0.00533	0.00272
2	0.01683	0.03275	0.02016	0.00061	0.0058	0.00271
3	0.01767	0.03195	0.02029	0.00061	0.00552	0.00248
4	0.01677	0.03449	0.02045	0.00063	0.00557	0.00294
5	0.01736	0.03307	0.02059	0.00062	0.00558	0.00263
6	0.01725	0.03275	0.02049	0.00064	0.00507	0.00254
7	0.01659	0.03275	0.02065	0.0006	0.00547	0.00262
8	0.01706	0.03267	0.02093	0.00063	0.0052	0.0025
9	0.01662	0.03249	0.02051	0.00059	0.00544	0.00279
10	0.01710	0.03332	0.02015	0.00061	0.0054	0.00284

Table 4. Affine **x** - equivariance $\text{MSE}_\lambda(.)_{\max}$ values

	$p = 5$			$p = 30$		
λ	$\delta = 0\%$	$\delta = 10\%$	$\delta = 20\%$	$\delta = 0\%$	$\delta = 10\%$	$\delta = 20\%$
0	9.55e-27	0.17066	0.35649	0.13403	0.22271	0.12016
0.5	1.51e-6	0.94962	0.36883	0.00089	0.09106	0.26992
1	4.23e-24	0.36592	0.09488	0.00055	0.13569	0.12746
1.5	1.71e-27	0.35441	0.22789	0.00251	0.10637	0.08758
2	3.81e-26	0.08954	0.64404	0.00024	0.10065	0.12487
3	3.81e-26	0.51310	0.56414	0.00656	0.11819	0.14615
4	1.43e-27	0.48495	0.49300	0.00656	0.11531	0.13136
5	2.11e-23	0.73876	1.3392	0.00142	0.11602	0.11201
6	7.41e-25	0.92863	0.44410	0.00194	0.14555	0.16444
7	1.21e-27	0.48099	0.40906	0.00517	0.14801	0.11676
8	0.00016	0.26175	0.19499	0.03388	0.10334	0.12029
9	4.80e-7	0.59238	0.35093	0.00106	0.13357	0.12236
10	1.13e-5	0.34573	0.91354	0.00379	0.19618	0.11513

The maximum percentage of outliers that the estimator can safely accept is measured by the breakdown point. The breakdown point can have a maximum value of 50%. With high contamination levels in mind, simulations such as those done by Sajesh and Srinivasan [37] can be used to investigate the empirical breakdown value. We suggest examining whether error and bias are controlled in these circumstances in order to assess the effectiveness of the suggested estimator $WShS_n$, even if low levels of contamination should be assumed and therefore make these scenarios less relevant in practice. We consider the NEO scenario to check the breakdown, with the percentage of contamination $\delta = 30\%, 40\%, 45\%$. For evaluation, we considered $\text{MSE}()_{\max}$ as defined by

$$\text{MSE}()_{\max} = \max_{\lambda \epsilon 0,0.5,1,\ldots,10} \text{MSE}_\lambda(.)_{\max}.$$

Table 5 shows that our proposed method has the least error compared to $\text{MSE}()_{\max}$ of other estimators, especially for higher contamination like 40% and 45%. In addition, we can see a decrease in $\text{MSE}()_{\max}$ with an increase in dimension. Even for 45% contamination our method has least $\text{MSE}()_{\max}$ which gives us empirical assurance of having a high breakdown property.

Table 5. $\text{MSE}()_{\max}$ table for breakdown property

	$p = 5$			$p = 30$		
Method	$\delta = 30\%$	$\delta = 40\%$	$\delta = 45\%$	$\delta = 30\%$	$\delta = 40\%$	$\delta = 45\%$
OLS	3.43891	4.57897	3.58334	2.4692	4.5482	1.2738
$WShS_n$	0.22585	0.56321	0.96679	0.0113	0.0323	0.0607
MM	0.49268	1.3743	4.0716	3.3505	5.5076	5.5949
LTS	0.29355	0.84809	1.70897	0.5239	0.7636	0.8144
LMS	0.28607	0.88574	1.82799	0.6210	0.8764	0.9119
S	0.28507	0.86094	2.57749	0.1356	0.4699	0.6267
SR	0.3125	0.6782	1.0122	0.02118	0.02442	0.07023

6. Sensitivity curve

The sensitivity curve of an estimator shows how an estimator performs when a small contamination replaces a single observation in the dataset. That is, it measures how the estimator responds to the local effect of a single observation. An estimator with a bounded sensitivity curve has an influence function that remains bounded. In this study, for the evaluation of the sensitivity curve, normal errors are considered, as in the NE case. The independent variables are taken from $N(0_{p\times 1}, \mathbf{I}_{p\times p})$ and the response variables from $N(0,1)$. The contaminant, a single observation in the independent variable.

The dependent variable is taken from $N(\lambda\sqrt{\chi^2_{p,0.99}}, 1)$ and $N(k\sqrt{\chi^2_{1,0.99}}, 1)$ respectively, where λ, k = -10, -8, -6, -4, -2, 0, 2, 4, 6, 8, 10.

$$\mathrm{SC}_n(y) = (n+1)\left(\hat{\theta}_{n+1}(y_1,\ldots,y_n,y) - \hat{\theta}_n(y_1,\ldots,y_n)\right).$$

Here, $\hat{\theta}$ represents the regression coefficient of the respective methods. Then, across the λ, k values, we found the norm of the above-defined difference of the regression coefficients.

In this study, each value is obtained after 1000 simulations. After obtaining the norm values for all values λ, k, we obtain the maximum norm k at different values, λ and it is plotted. The maximum norm for different values of λ and k provides insight into the resistance of robust methods to contamination. Here, we compare our proposed estimator with classical estimators and other robust estimators such as LTS, LMS, MM, and S.

Figure 1 illustrates that our proposed estimator remains more bounded than other methods, regardless of dimension. This empirically confirms the bounded nature of the influence function of our proposed estimator.

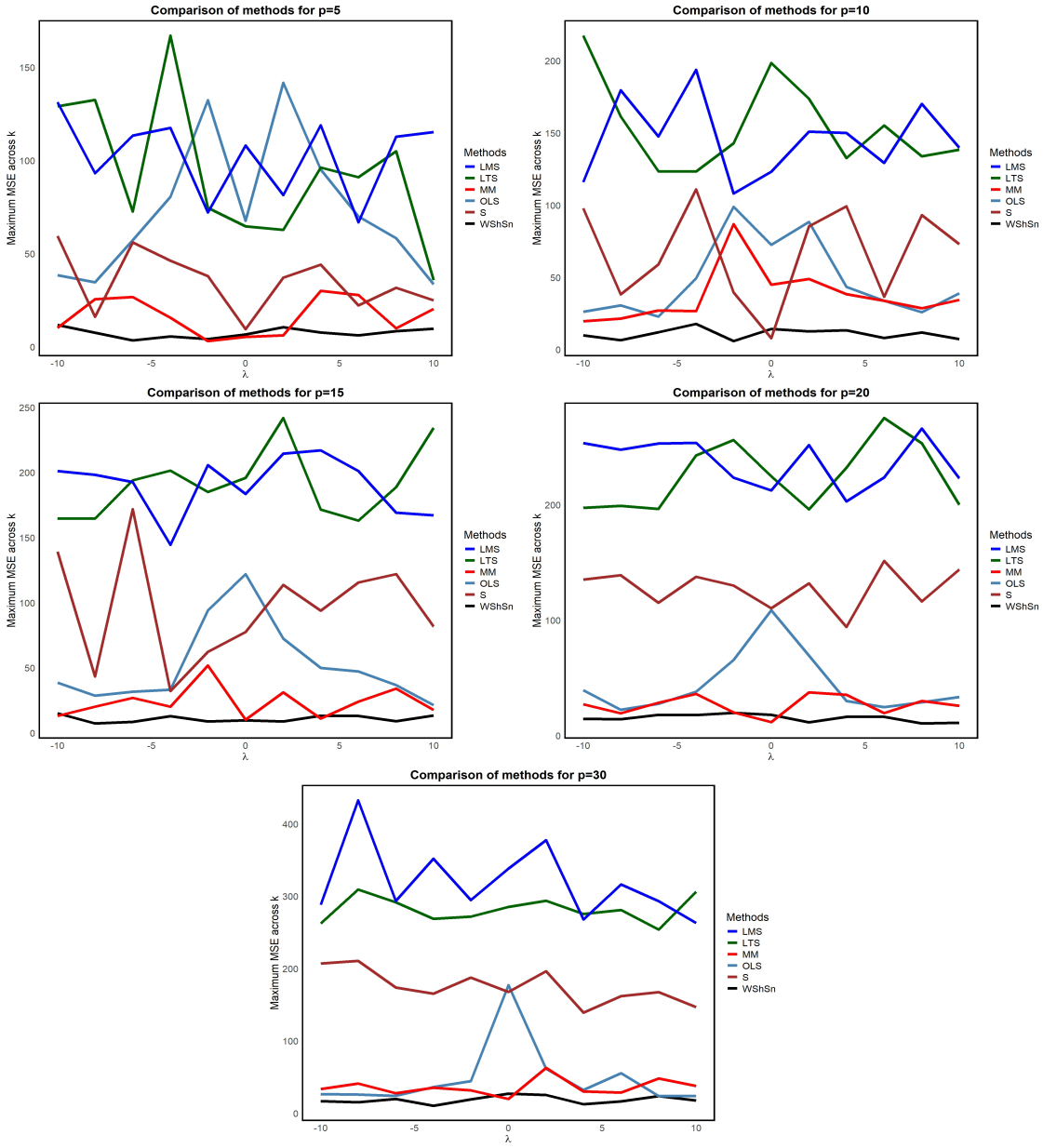

Figure 1. Sensitivity curve of our proposed regression method along with other methods a)$p = 5$ b)$p = 10$ c)$p = 15$ d)$p = 20$ e)$p = 30$

7. Applications

In this section, four real-life data sets are performed, namely, the learning data studied in [18], the mineral data studied in [39], the aircraft data mentioned in [14], and the Belgium phone call data used in [31]. The mean square error (MSE), the mean absolute percentage error (MAPE), the mean absolute deviation (MAD), and the Akaike information criterion (AIC) are used to compare the estimator model in real-life data sets. These measures are defined as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \tag{7.1}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|, \tag{7.2}$$

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i,| \tag{7.3}$$

$$\text{AIC} = n \times ln\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right] + 2p. \tag{7.4}$$

We compare the different estimators by looking at the one that gets the lowest MSE, MAPE, MAD and AIC.

7.1. Application I

The simulation data consisted of student motivation ($\mathbf{x}_1$), learning facilities ($\mathbf{x}_2$), and student learning outcomes in the cognitive domain (y). Motivation and facilities were evaluated using a questionnaire with scores ranging from 0 to 30, while learning outcomes were evaluated using a test, with scores ranging from 0 to 100. Data are tabulated in [18]. 20 respondents were part of the study.

Figure 2 shows that the data contain outliers and observations 7, 14, 15, and 18 detected as outliers. If we closely observe the data, we can confirm these observations as outliers or observations that require close surveillance. For example, if we look at observations 7 and 18, the respective students express high student motivation and a very good learning facility, but the student learning outcome is very low. The low learning outcome of a student who felt good motivation to learn and had good facilities contradicts the explicit statement that the student should be monitored. This discrepancy suggests that, despite favorable learning conditions, other unobserved factors may be influencing their academic performance. These cases alarm researchers with warnings of the need for more research to understand the underlying causes and provide the necessary interventions. The reason for the presence of these outliers varies, such as recording errors, psychological stress, or external distractions, or they might be students who required special attention. We analyze the data with all the estimators used in our comparison study, and the results are shown in Table 6.

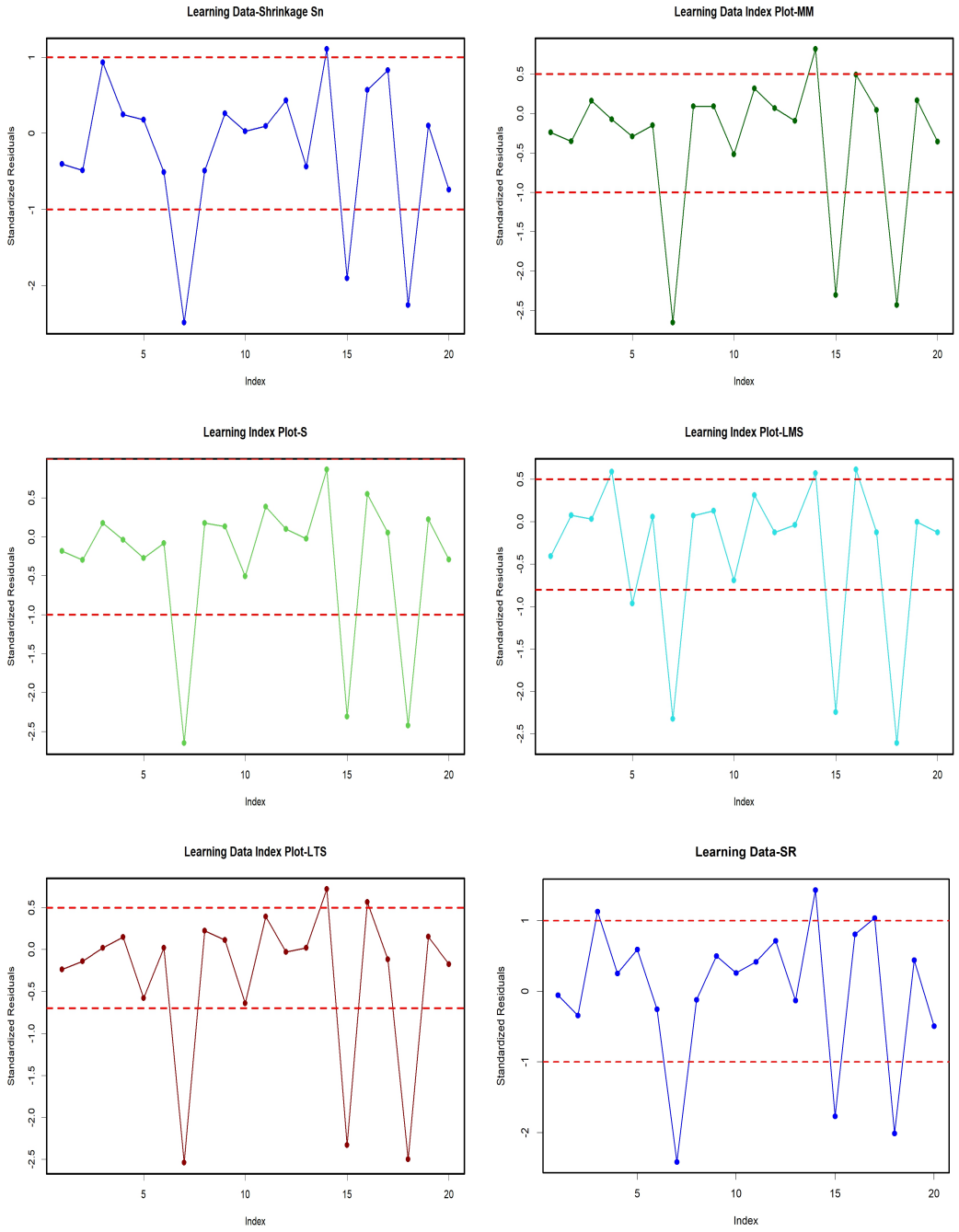

Figure 2. Standardized residuals index plot for outlier detection of $WShS_n$, MM, S, LMS, LTS, SR for the learning data.

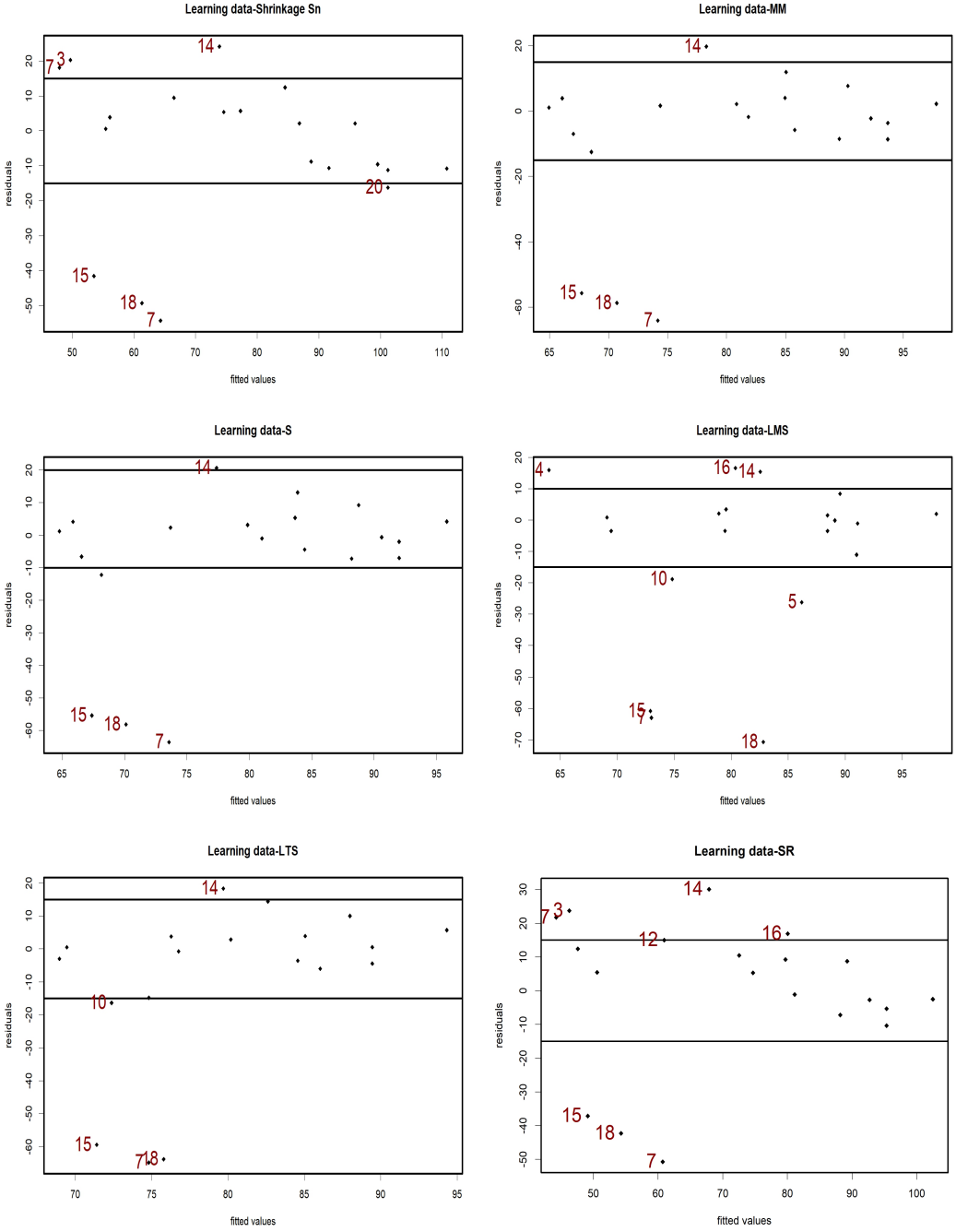

Figure 3. Fitted values versus residuals plot of $WShS_n$, MM, S, LMS, LTS, SR for the learning data.

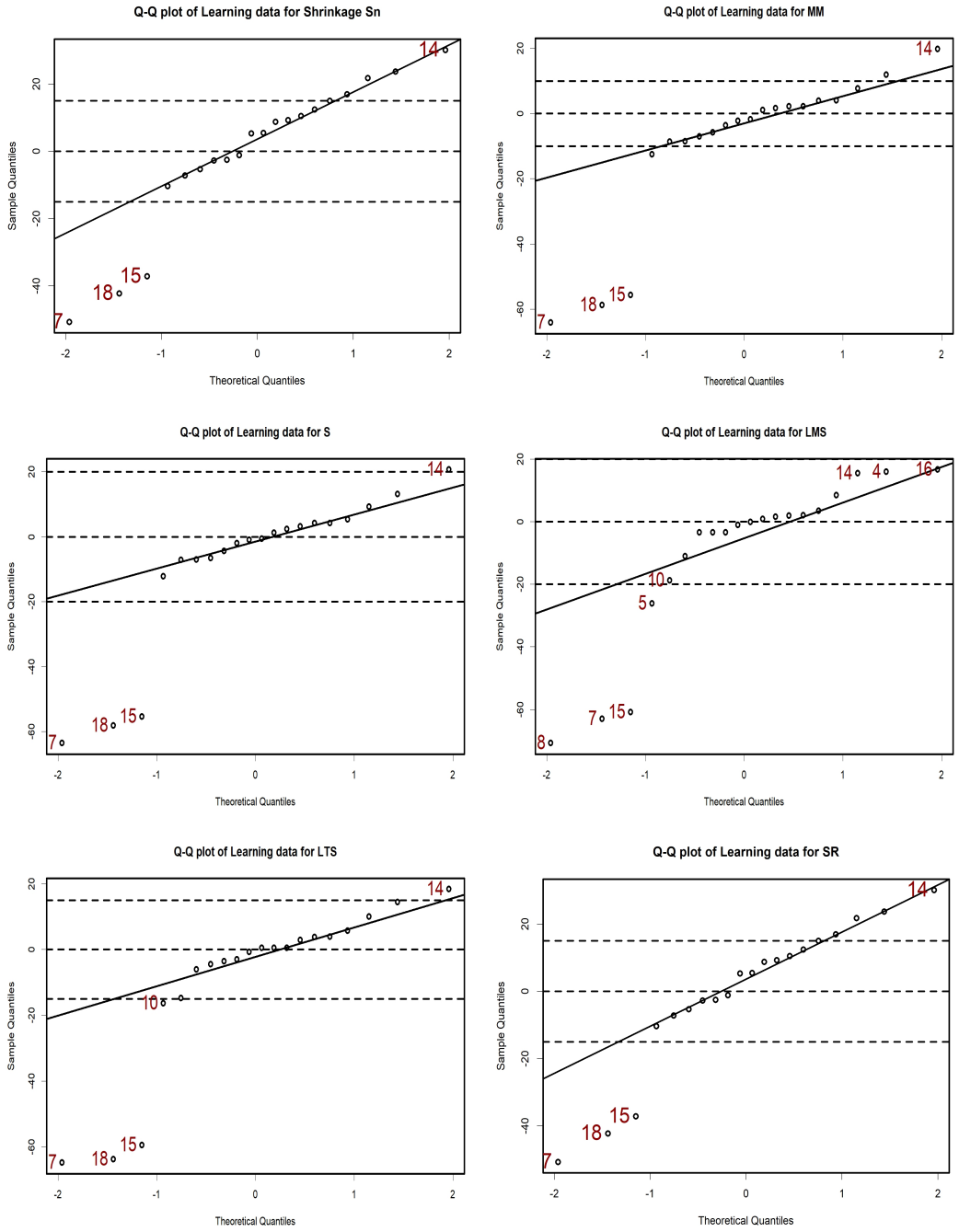

Figure 4. Q-Q plot of residuals of $WShS_n$, MM, S, LMS, LTS, SR for the learning data.

Table 6. The results for the learning data

Method	Intercept	Slope $\mathbf{x}_1$	Slope $\mathbf{x}_2$	RSE	MSE	MAPE	MAD	AIC
$WShS_n$	50.11493	1.91120	1.77704	21.7768	477.2288	0.75461	9.35787	129.91775
MM	39.30157	0.82210	1.12754	24.14265	585.8677	0.86269	14.16605	131.46190
LTS	50.42680	0.97560	0.48780	25.51272	653.8988	0.90765	14.87439	133.65910
LMS	51.63640	1.90910	-0.36360	27.10937	737.9178	0.95359	16.44091	136.07660
S	40.50310	0.76260	1.08140	21.29027	578.4258	0.85499	14.06815	131.20620
OLS	-1.78900	1.91780	2.03000	21.00165	444.0691	0.70329	15.93060	125.91960
SR	-1.37300	1.43300	2.03370	21.02750	445.0811	0.80911	16.78160	126.69180

7.2. Application II

Smith [39] studied the investigation of how various elements of the Golden Grove massive sulfide deposits disperse throughout the lateritic landscape. They conducted measurements of the concentrations (in parts per million) of 22 chemical elements in 53 rock samples from Western Australia. Maronna et al. [27] studied two variables from the data mentioned above in their book. We use the same data as in [27]. The data consists of details of the zinc (*study variable*) and copper (*explanatory variable*) content deposits. We tried to explore the relationship of these two elements.

For model comparison purpose, we consider MSE, MAPE, MAD of the study variables. In addition, we consider AIC for comparison purposes. This particular data set contains outliers [35]. When we look into the fitted versus residual of the OLS, we could see the OLS Q-Q plot exhibits, the line attracted towards point 15, but for robust methods, we could not find such a pattern. Similar false performance of OLS can be found in other datasets that we observed here. Observations 15, 2, 25, 3, and 39 can be detected as extreme lying observations by all methods in the residual versus fitted plot. If we closely look at these observations of behavior in a standardized residual plot, one could find that the smooth pattern of the points is inhibited by the aforementioned observations. The Q-Q plot substantiates the fact that the aforementioned observations cause non-normality in the data.

The observations of data were supposed to be studied in detail, and the reason for exhibiting these kinds of patterns. The values of the coefficient and model parameters are given in Table 7.

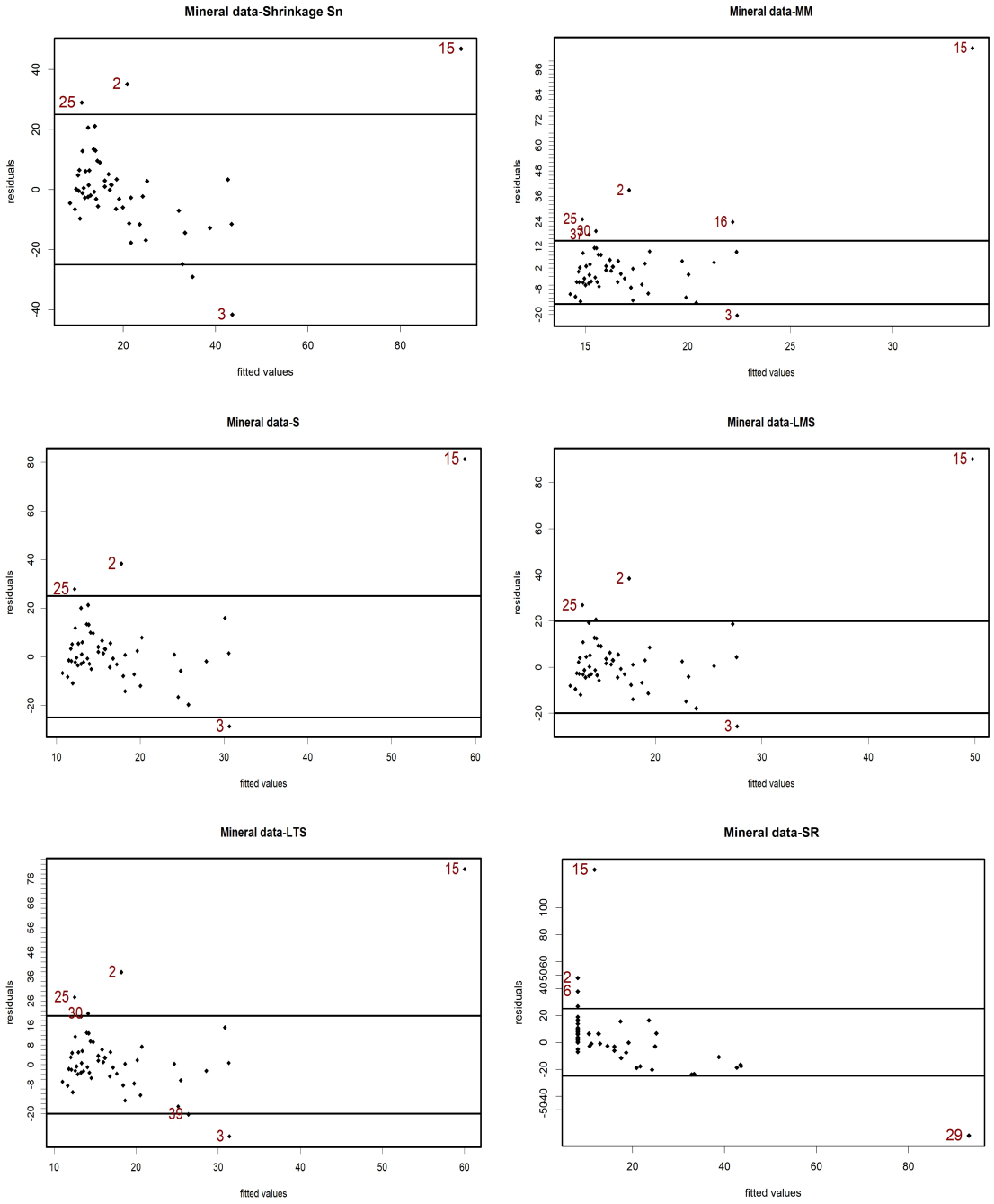

Figure 5. Fitted values versus residuals plot of $WShS_n$, MM, S, LMS, LTS, SR for the mineral data.

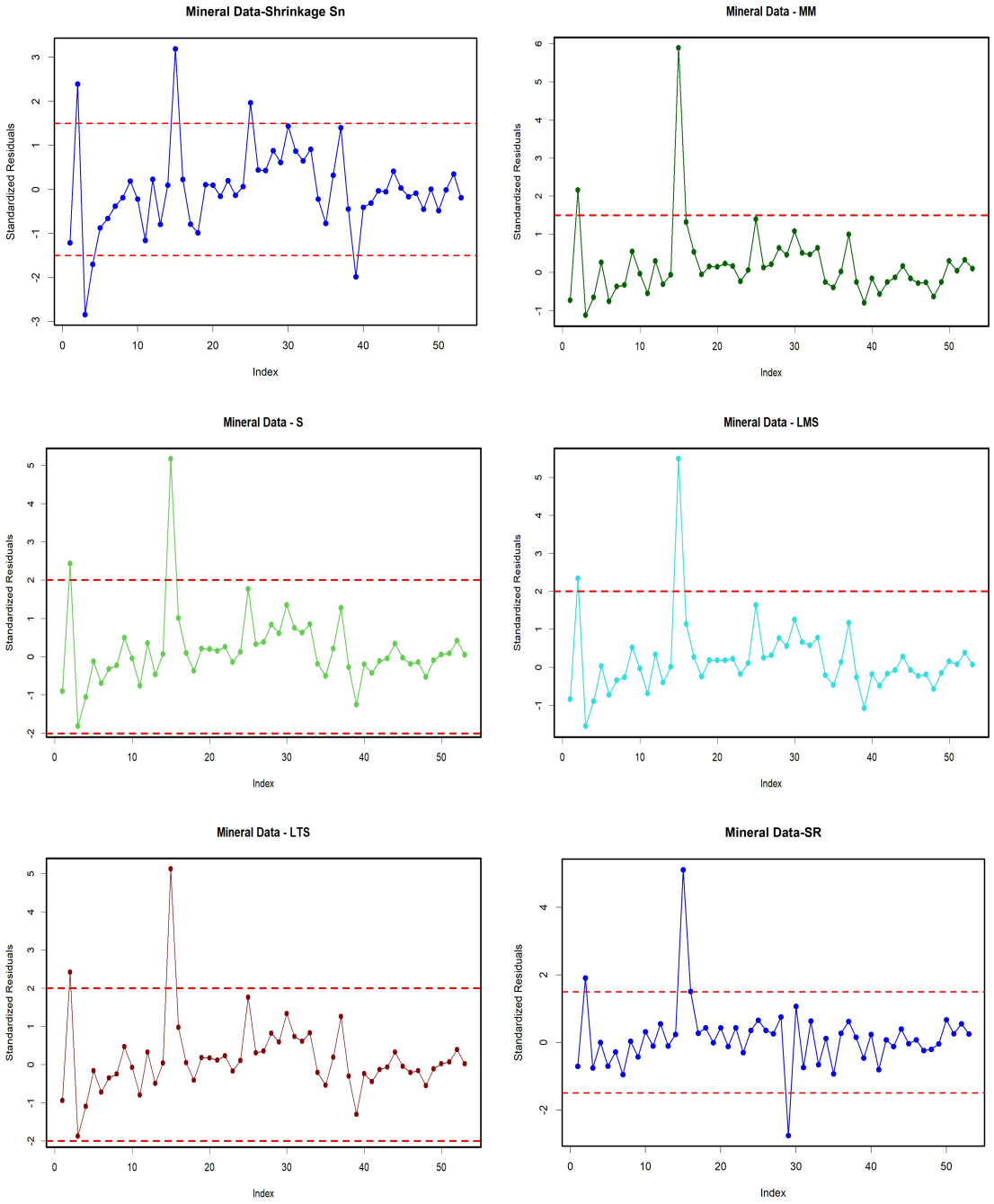

Figure 6. Standardized residuals index plot for outlier detection of $WShS_n$, MM, S, LMS, LTS, SR for the mineral data.

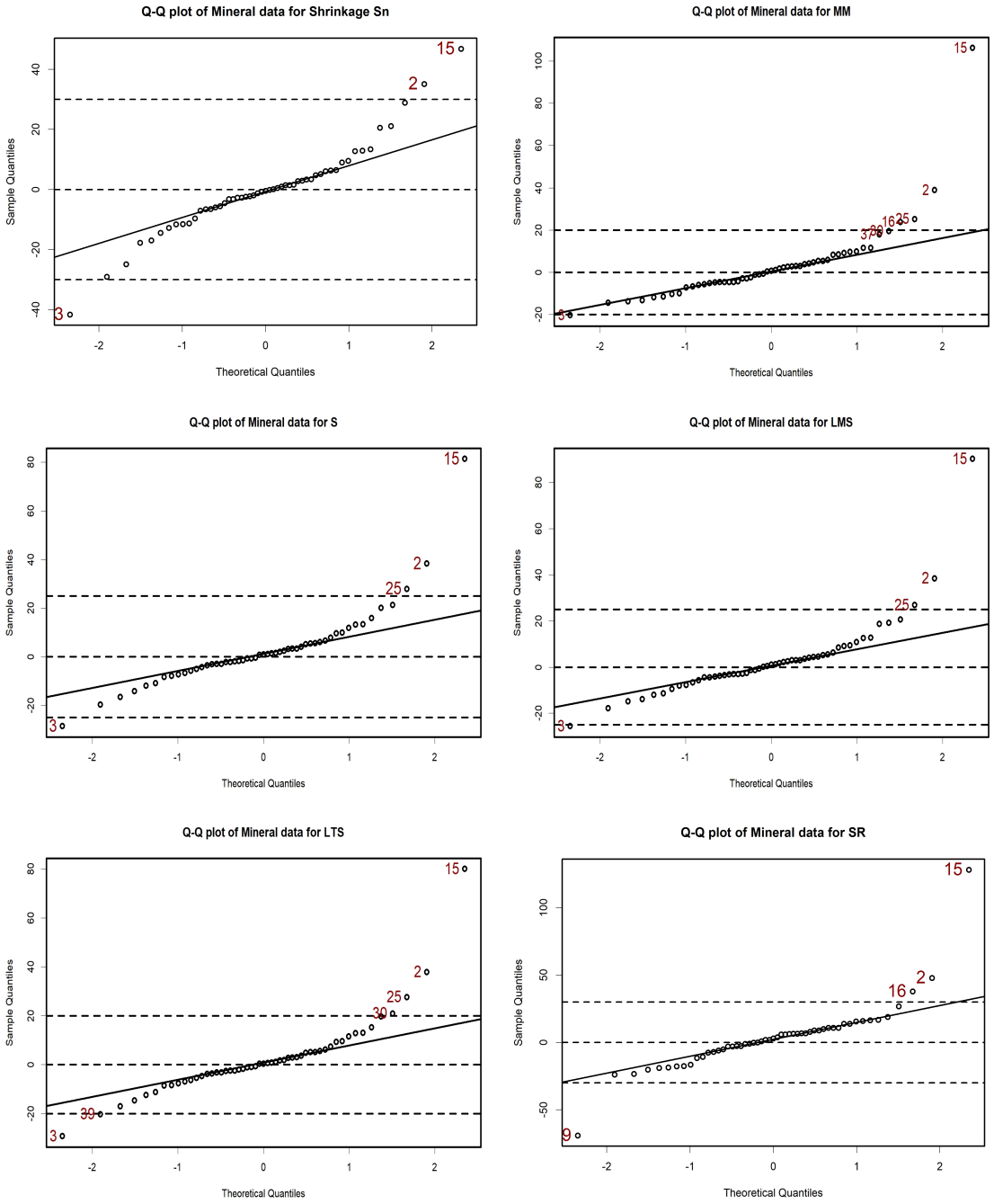

Figure 7. Q-Q plot for residuals of $WShS_n$, MM, S, LMS, LTS, SR for the mineral data.

Table 7. The results for the mineral data

Method	Intercept	Slope	RSE	MSE	MAPE	MAD	AIC
$WShS_n$	10.24283	0.07720	18.03592	331.4453	1.48007	13.0142	301.5835
MM	14.13856	0.031207	18.03535	327.2739	1.017315	9.833758	308.9123
LTS	10.71525	0.07792	15.60709	245.5813	1.038663	9.28958	293.6923
LMS	11.79	0.06	16.45569	272.7899	1.013786	9.443962	299.2612
S	10.41003	0.076318	15.74303	249.8429	1.013658	9.324528	294.6041
OLS	11.4759	0.06313	7.90696	64.51997	0.789537	6.23915	204.0148
SR	7.96063	0.13457	25.07286	604.9256	1.082492	14.42829	343.4706

7.3. Application III

Aircraft data mentioned in [14] consists of 23 observations in five variables. The response variable is cost and there are four explanatory variables, namely, aspect ratio, lift-to-drag ratio, weight of the plane, and maximal thrust. The standardized residual plot shows that some observations deviate and inhibit the smooth pattern of the data. If we look at the fitted versus residual plot, the same observations as in the residual plot remain outlying. Rousseeuw and Driessen [33] quoted some of the data sets that include the aircraft data set and outlying observations in the data sets. Similarly to their findings, we can see that our robust method detects observations 2, 3, 10, 11, 12, 16, 17, 18, 19, 20, and 22 as outliers. All other robust methods are used to detect some of these observations. The regression coefficients and model metrics are presented in Table 8. It can be said that the proposed estimator successfully detects outliers and performs better than the OLS in the presence of outliers.

Table 8. The results for the aircraft data

Method	Intercept		Slope			RSE	MSE	MAPE	MAD	AIC
$WShS_n$	6.0121	-2.0006	1.6189	0.00182	-0.0008	11.443	135.95	0.32531	6.1762	120.98
MM	6.1417	-3.2306	1.6711	0.00192	-0.0009	10.339	111.89	0.34389	5.5301	112.50
LTS	14.4816	-4.6712	1.9305	0.00198	-0.0013	11.643	140.56	0.35657	5.9568	121.75
LMS	5.7789	-2.0064	3.3645	0.00079	-0.0005	17.4118	308.17	0.29575	8.6583	139.81
S	13.3733	-4.0219	1.5413	0.00170	-0.0009	11.8193	144.69	0.33773	5.9442	122.42
OLS	-3.7913	-3.8529	2.4883	0.00349	-0.0019	7.0921	55.298	0.58333	5.6856	100.29
SR	4.62125	-3.2716	1.8726	0.00212	-0.0011	9.65777	98.2725	0.36041	5.40455	113.5181

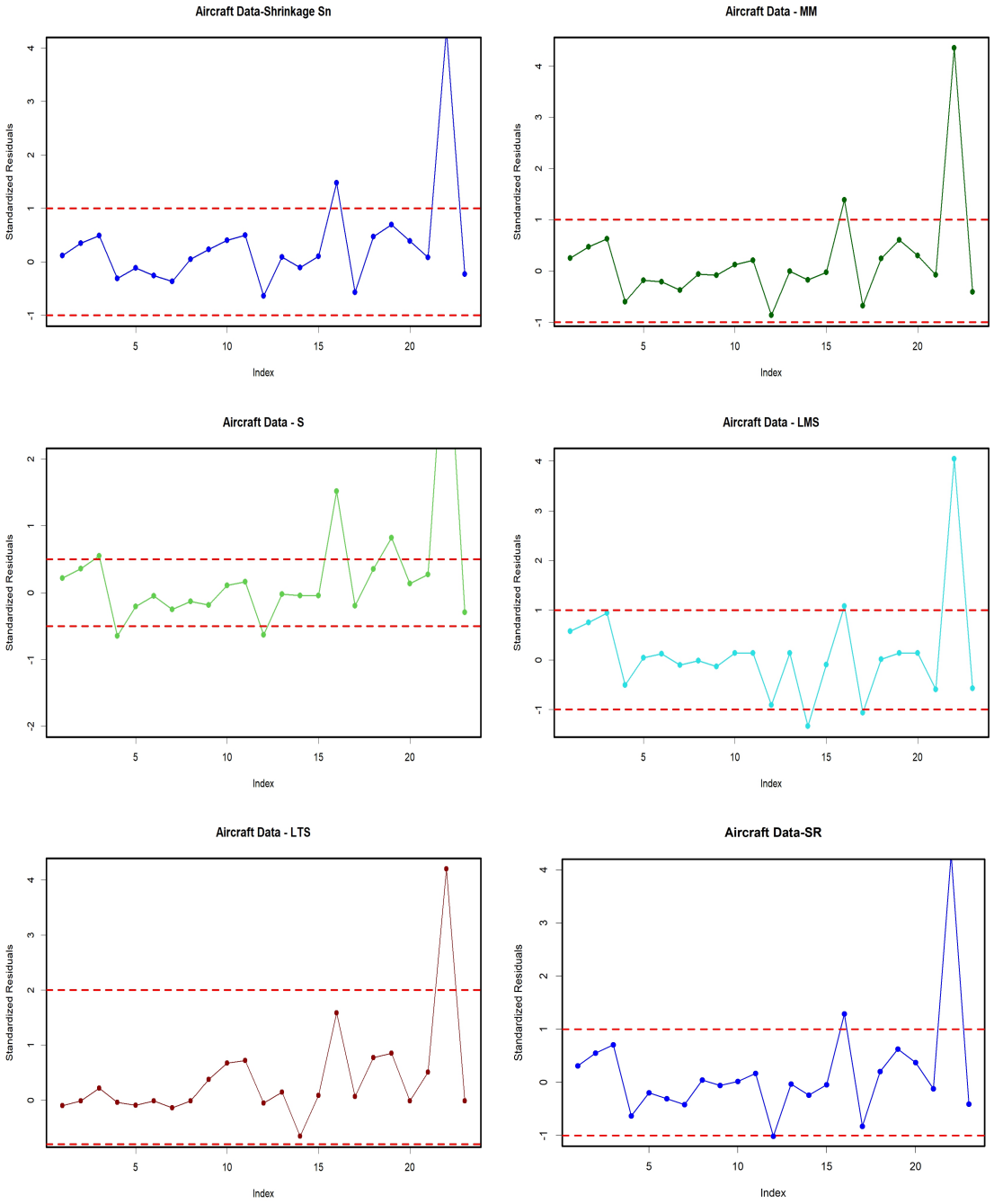

Figure 8. Standardized residuals index plot for outlier detection of $WShS_n$, MM, S, LMS, LTS, SR for the aircraft data.

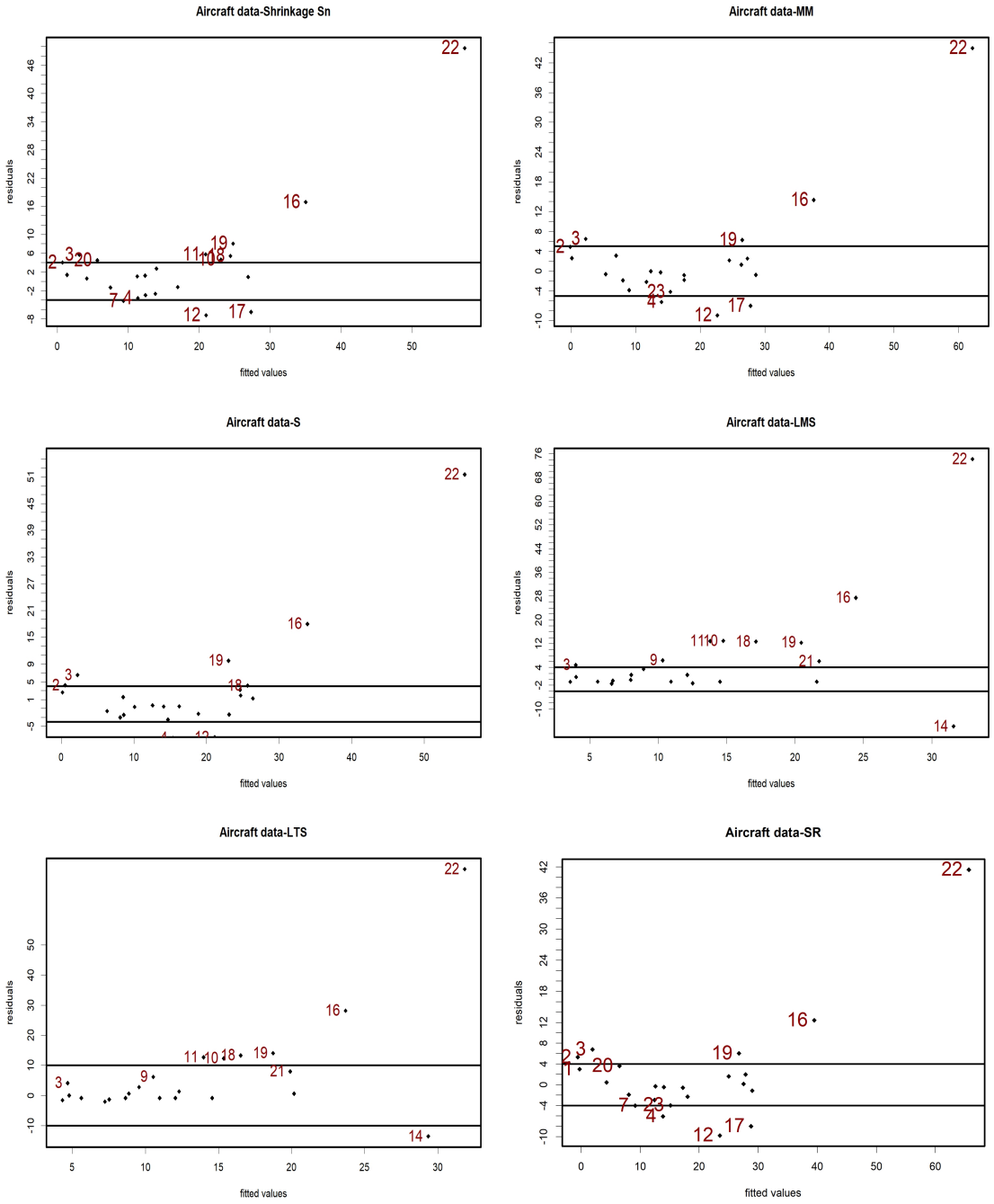

Figure 9. Fitted values versus residuals plot of $WShS_n$, MM, S, LMS, LTS, SR for the aircraft data.

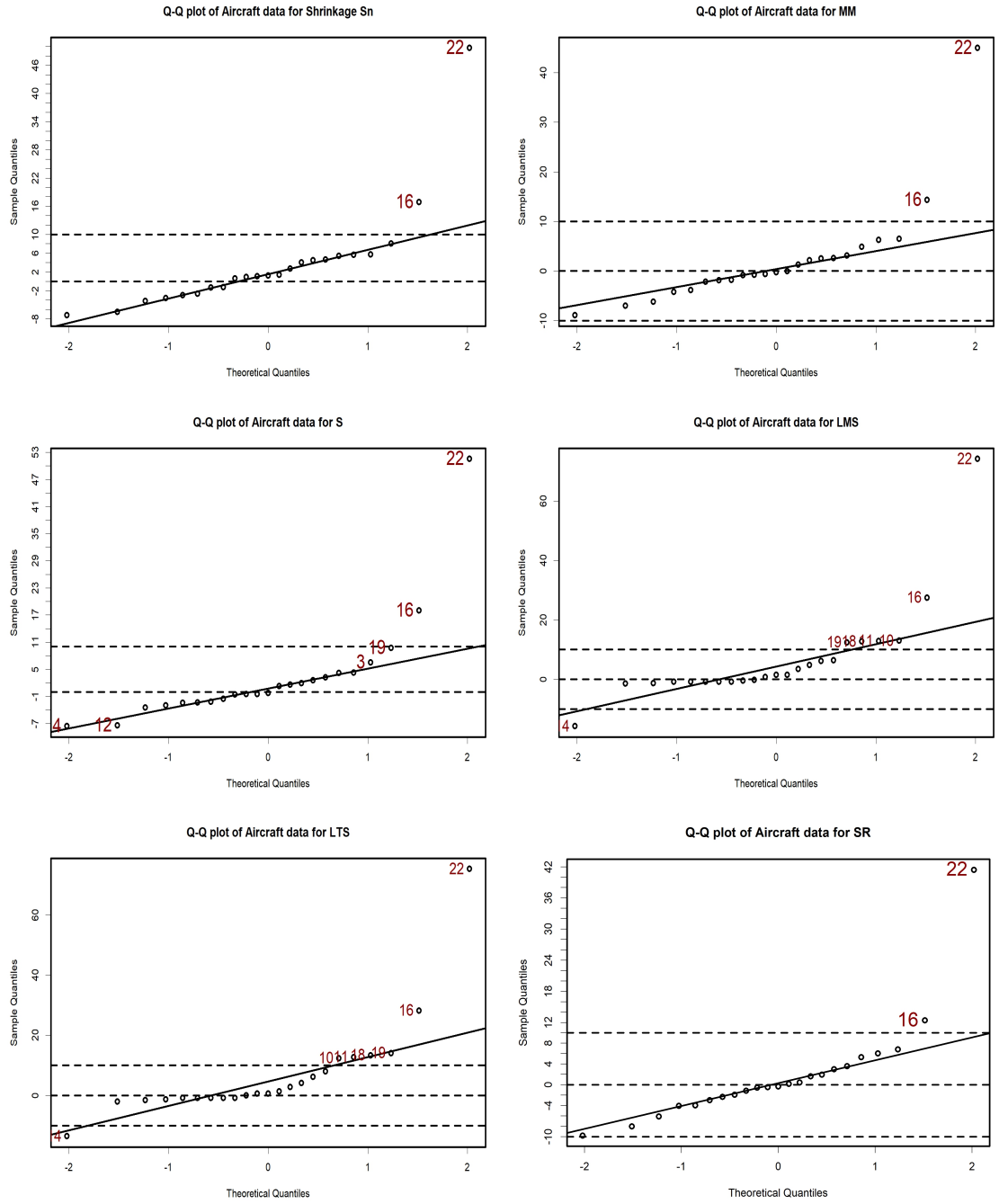

Figure 10. Q-Q plot of residuals of $WShS_n$, MM, S, LMS, LTS, SR for the aircraft data.

7.4. Application IV

The Belgium phone calls data were originally published by the Belgium Statistical Survey and were used in [31]. This data set includes the annual number of international calls made from Belgium between 1950 and 1973. The data set consists of two variables: the year ($\mathbf{x}$) and the number of calls received (y). The data represents simple linear regression data. The OLS provides regression coefficients but when a person without critical thinking looks into the residual plot of the OLS cannot find anything abnormal. If we look

into the residual plots from robust methods and the fitted vs. residual plot, we can easily identify 6 outliers in the y direction and we can find how these observations influence other observations. Our proposed estimator-based graph detects observations 15, 16, 17, 18, 19 and 20 correctly as outliers. Due to the influence of these six outliers, observation 21 was detected as an outlier. The regression coefficients and model metric values are tabulated and given in Table 9.

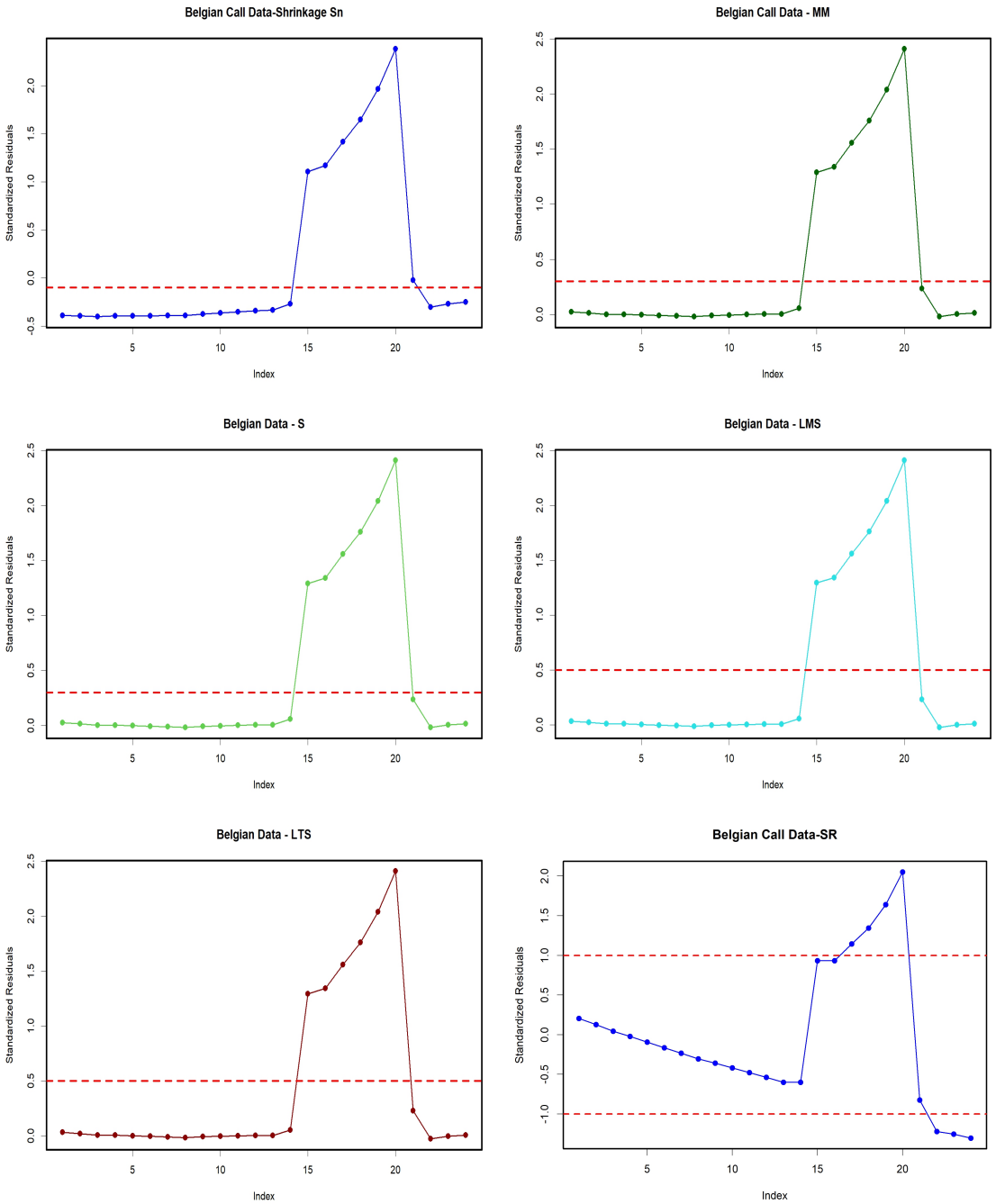

Figure 11. Q-Q plot of residuals of $WShS_n$, MM, S, LMS, LTS, SR for the Belgium phone call data.

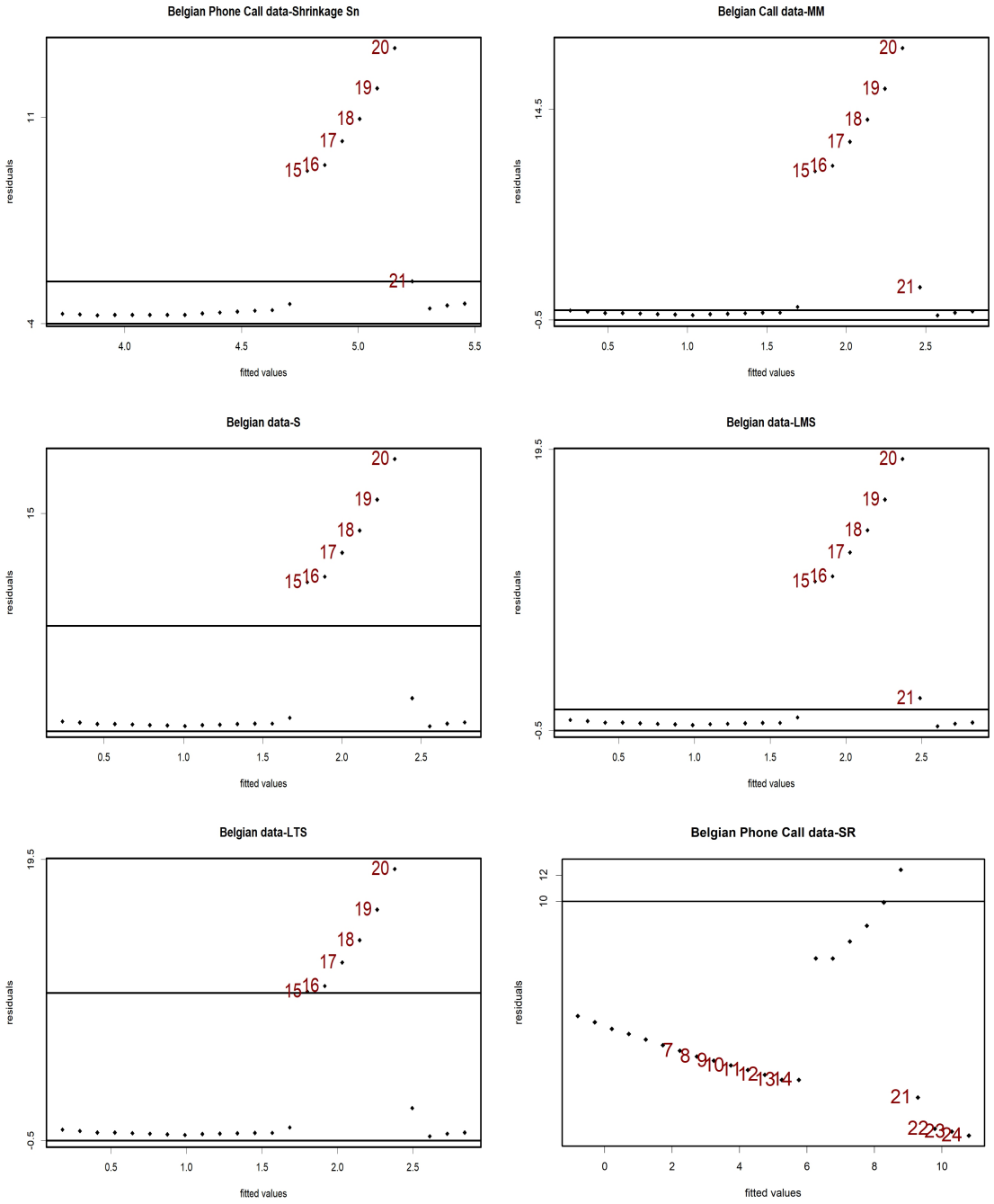

Figure 12. Q-Q plot of residuals of $WShS_n$, MM, S, LMS, LTS, SR for the Belgium phone call data.

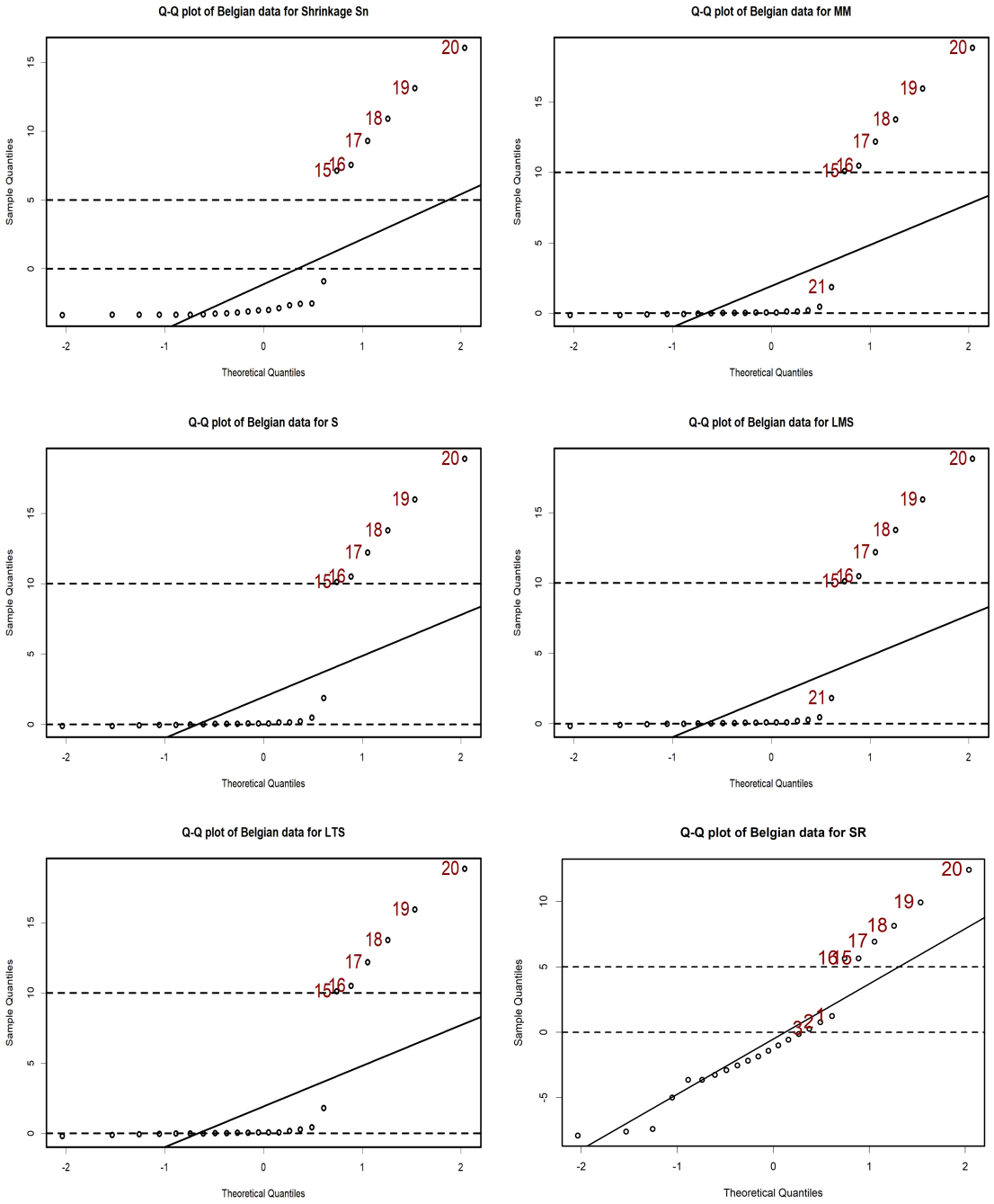

Figure 13. Q-Q plot of residuals of $WShS_n$, MM, S, LMS, LTS, SR for the Belgium phone call data.

Table 9. The results for the Belgium phone call data

Method	Intercept	Slope	RSE	MSE	MAPE	MAD	AIC
$WShS_n$	-5.0119	0.07490	5.7419	37.969	2.7439	4.9275	95.28
MM	-5.2423	0.11009	6.5926	48.462	0.30085	3.5316	101.12
LTS	-5.6162	0.1159	6.5856	48.3706	0.31328	3.5306	101.09
LMS	-5.5947	0.1155	6.5879	48.3989	0.31277	3.5316	101.11
S	-5.2732	0.1102	6.6042	48.6151	0.30198	3.5384	101.21
OLS	-26.006	0.5041	4.8966	28.9765	1.5236	4.2453	88.796
SR	-26.008	0.5042	25.4240	511.7187	10.718	17.0290	157.707

The metrics of different real-life datasets show that our proposed method is capable of withstanding outliers occupying the dataset and gives metric values less than the LTS, LMS and S. Also, the regression coefficients are near other robust methods. The plots show the capability of our method to detect outliers. All this gives assurance regarding the good performance of our proposed method.

8. Conclusion

In multiple regression, the response variable is clearly associated with explanatory variables p, and least squares regression is the preferred method for estimating regression coefficients. It is essential to note that the classical Ordinary Least Squares (OLS) regression is highly sensitive to the presence of outliers. When outliers are present in the dataset, the OLS often produces distorted regression coefficients that do not accurately reflect the true relationship. Furthermore, outliers can significantly inflate the standard errors of the regression coefficients, making them appear less statistically significant and resulting in a misleading assessment of the goodness of fit. It is crucial to address outliers to ensure reliable regression analysis. It is essential for the researcher to conduct a meticulous examination of all aspects rather than simply focusing on the regression coefficient estimates.

A comprehensive assessment of model assumptions, data quality, potential outliers, and the relevance of the results is crucial. Careful observation, thorough diagnostic testing, and domain expertise are essential for developing a reliable and valid regression model. By emphasizing these critical factors, researchers can ensure that their findings are not only accurate, but also meaningful and significant in their area of study.

In this paper, we confidently present our robust regression estimator, $WShS_n$, which is built upon the Shrinkage S_n covariance matrix. Our extensive comparisons demonstrate that $WShS_n$ outperforms other existing robust regression methods. It is crucial to recognize that many available methods fail to deliver satisfactory performance with high-dimensional data. Not all available methods perform well with large datasets or high-dimensional data, and many are not proven to be sufficiently robust against the presence of outliers.

This paper proposes making use of robust location and scatter shrinkage estimators to estimate regression parameters using the concept of shrinkage. The method produces the $WShS_n$ regression estimator. The advantages of the proposed estimator are shown through the simulation study. The simulation study demonstrates that our proposed estimator consistently outperforms other methods in terms of robustness, regardless of higher dimensions, greater contamination, or transformed data. In terms of efficiency, our method shows an advantage over existing approaches like LTS, LMS, S, and MM.

A key feature of $WShS_n$ is that it utilizes all observations, unlike the sub-sample iterations employed in other methods, which adds to its stability. We also applied the new estimator to several real-life datasets and evaluated its performance. The results favor the effectiveness of our proposed method, in line with the findings of the simulation study. Comparisons of real-life models highlight the capabilities of our estimator and reveal how classical OLS can mislead researchers when dealing with data sets containing outliers.

Acknowledgements

We thank the anonymous reviewers for their insightful comments, which significantly helped to improve the manuscript.

Author contributions. All authors have contributed equally to this work.

Conflict of interest statement. No potential conflict of interest was reported by the author(s).

Funding. The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability. The data that support the findings of this study are openly available, with references provided in the relevant sections.

References

[1] J. Agulló, C. Croux and S. Van Aelst, *The multivariate least-trimmed squares estimator*, J. Multivar. Anal. **99**(3), 311-338, 2008.

[2] E. Cabana, R.E. Lillo and H. Laniado, *Robust regression based on shrinkage with application to Living Environment Deprivation*, Stoch. Environ. Res. Risk Assess. **34**(2), 293-310, 2020.

[3] D.W. Scott and Z. Wang, *Robust multiple regression*, Entropy **23**(1), 88, 2021.

[4] X. Liu, E.C. Chi and K. Lange, *A sharper computational tool for regression*, Technometrics **65**(1), 117-126, 2023.

[5] E. Cabana, R.E. Lillo and H. Laniado, *Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators*, Stat. Pap. **62**(2), 1583-1609, 2021.

[6] O.J. Ibidoja, F.P. Shan, J. Sulaiman and M.K.M. Ali, *Robust M-estimators and machine learning algorithms for improving the predictive accuracy of seaweed contaminated big data*, J. Niger. Soc. Phys. Sci. **1137**(1), 1137-1137, 2023.

[7] E. Bas, *Robust fuzzy regression functions approaches*, Inf. Sci. **613**(1), 419-434, 2022.

[8] M. Abonazel and A. Rabie, *The impact of using robust estimations in regression models: An application on the Egyptian economy*, J. Adv. Res. Appl. Math. Stat. **4**(2), 8-16, 2019.

[9] C. Croux, S. Van Aelst and C. Dehon, *Bounded influence regression using high breakdown scatter matrices*, Ann. Inst. Stat. Math. **55**(1), 265-285, 2003.

[10] C. Croux, P.J. Rousseeuw and O. Hössjer, *Generalized S-estimators*, J. Am. Stat. Assoc. **89**(428), 1271-1281, 1994.

[11] V. DeMiguel, L. Garlappi and R. Uppal, *Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?*, Rev. Financ. Stud. **22**(5), 19151953, 2009.
[12] F.Y. Edgeworth, *On observations relating to several quantities*, Hermathena **6**, 279285, 1887.
[13] D. Gervini and V.J. Yohai, *A class of robust and fully efficient regression estimators*, Ann. Stat. **30**(2), 583-616, 2002.
[14] J.B. Gray, *Graphics for regression diagnostics*, Am. Stat. Assoc. Proc. Stat. Comput. Sect. **1985**(1), 102-107, 1985.
[15] D.M. Hawkins and D.J. Olive, *Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm*, J. Am. Stat. Assoc. **97**(457), 136-159, 2002.
[16] M. Falk, *On mad and comedians*, Ann. Inst. Stat. Math. **49**(3), 615-644, 1997.
[17] Z. Han, J. Chen, F. Zhang, Z. Gao, H. Huang and Y. Li, *An efficient online outlier recognition method of dam monitoring data based on improved M-robust regression*, Struct. Health Monit. **22**(1), 581-599, 2022.
[18] P. Jana, D. Rosadi and E.D. Supandi, *Comparison of robust estimation on multiple regression model*, BAREKENG: J. Math. App. **17**(2), 979-988, 2003.
[19] W. James and C. Stein, *Estimation with quadratic loss*, Bayesian Statistics 4, Oxford Univ. Press 4, 361-379, 1992.
[20] R. Lakshmi and T.A. Sajesh, *A robust distance-based approach for detecting multidimensional outliers*, J. Appl. Stat. **1**(1), 1-21, 2024.
[21] H.P. Lopuhaa and P.J. Rousseeuw, *Breakdown points of affine equivariant estimators of multivariate location and covariance matrices*, Ann. Stat. **19**(1), 229-248, 1991.
[22] O. Ledoit and M. Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection*, J. Empir. Finance 10(5), 603-621, 2003b.
[23] O. Ledoit and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, J. Multivar. Anal. **88**(2), 365-411, 2004.
[24] H. Oja, *Multivariate non parametric methods with R: an approach based on spatial signs and ranks*, Springer Sci. Bus. Media **1**(1), 1-1, 2010.
[25] R. Maronna and S. Morgenthaler, *Robust regression through robust covariances*, Commun. Stat. Theory Methods **15**(4), 1347-1365, 1986.
[26] R.A. Maronna and R.H. Zamar, *Robust estimates of location and dispersion for high-dimensional datasets*, Technometrics **44**(4), 307-317, 2002.
[27] R.A. Maronna, R.D. Martin and V.J. Yohai, *Robust Statistics: Theory and Methods*, John Wiley Sons **1**(1), 1-1, 2006.
[28] F. Mosteller and J.W. Tukey, *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley **1**(1), 1-1, 1977.
[29] J. Mottonen, K. Nordhausen and H. Oja, *Asymptotic theory of the spatial median*, Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis **7**(1), 182-194, 2010.
[30] P.J. Rousseeuw, *Least median of squares regression*, J. Am. Stat. Assoc. **79**(388), 871-880, 1984.
[31] P.J. Rousseeuw and A.M. Leroy, *Robust regression and outlier detection*, John Wiley Sons **1**(1), 1-1, 2005.
[32] P. Rousseeuw and V. Yohai, *Robust regression by means of S-estimators*, Robust and Nonlinear Time Series Analysis **1**(1), 256-272, 1984.
[33] P.J. Rousseeuw and K. Van Driessen, *An algorithm for positive-breakdown regression based on concentration steps*, Data Anal. Sci. Mod. Pract. Appl. **1**(1), 335-346, 2000.
[34] P.J. Huber, *Robust estimation of a location parameter*, Ann. Math. Stat. **35**(1), 73101, 1964.
[35] P.J. Huber and E.M. Ronchetti, *Robust statistics*, John Wiley Sons **1**(1), 1-1, 2011.

[36] P.J. Rousseeuw, S. Van Aelst, K. Van Driessen and J.A. Gulló, *Robust multivariate regression*, Technometrics **46**(3), 293-305, 2004.
[37] T.A. Sajesh and M.R. Srinivasan, *Outlier detection for high dimensional data using the Comedian approach*, J. Stat. Comput. Simul. **82**(5), 745-757, 2012.
[38] A.F. Siegel, *Robust regression using repeated medians*, Biometrika **69**(1), 242-244, 1982.
[39] R.E. Smith, N.A. Campbell and R. Litchfield, *Multivariate statistical techniques applied to pisolitic laterite geochemistry at Golden Grove, Western Australia*, J. Geochem. Explor. **22**(13), 193-216, 1984.
[40] A.J. Stromberg, O. Hössjer and D.M. Hawkins, *The least trimmed differences regression estimator and alternatives*, J. Am. Stat. Assoc. **95**(451), 853-864, 2000.
[41] V.J. Yohai, *High breakdown-point and high efficiency robust estimates for regression*, Ann. Stat. **15**(2), 642-656, 1987.
[42] C. Yu and W. Yao, *Robust linear regression: A review and comparison*, Commun. Stat. Simul. Comput. **46**(8), 6261-6282, 2017.

APPENDIX

The median absolute deviation (MAD) is similar to the median, is a reliable estimate of the dispersion for a random variable X. Rousseeuw and Croux in 1993 offered a more effective substitute for MAD with a breakdown of 50% and established S_n, a consistent and unbiased estimator for the function of the relevant population. It is defined as

$$S_n = c\ med_i med_j |x_i - x_j|.$$

Here, the constant ($c = 1.1926$) is chosen as the consistency factor for normal distributions and to make the estimator also unbiased. Also, S_n is location-free that it does not use any location estimator. S_n estimator of scale assures bounded influence function and is more applicable due to its low sensitivity to gross error. A comedian-based robust alternative to the sample covariance between two random variables X and Y was put out by [16], also, he has mentioned the idea of extending S_n scale estimator as a robust alternative to covariance between two random variables, the same is developed here in a multivariate version. S_n covariance of two random variables X and Y be

$$S_n(\mathrm{X}, \mathrm{Y}) = 1.4304(med_i[med_{j\neq i}\{(x_i - x_j)(y_i - y_j)\}]).$$

Let $\mathbf{X}$ be $n \times p$ matrix with sample size n, number of variables p and $\mathbf{X_j}$ $(j = 1, 2, ..., p)$ be the columns of the matrix. Then, the covariance matrix of $\mathbf{X}$ based on S_n would be

$$\hat{S_n} = S_n(\mathbf{X_i}, \mathbf{X_j}).$$

The trace of this equation provides a robust scale estimator is given by

$$\mathrm{trace}(\hat{S_n}) = \sum_{j=1}^{p} S_n(\mathbf{X_j}, \mathbf{X_j}) = \sum_{j=1}^{p} 1.4304.\ S_n^2(\mathbf{X_j}) = \sum_{j=1}^{p} \sigma^2_{\mathbf{X_j}}.$$

Here, $\hat{S_n}$ is an unbiased estimator and a high breakdown estimator [38], but it is not definite positive. Instead of employing conventional methods to ensure positive semidefiniteness of the covariance matrix, our approach involves shrinkage estimation on the empirical covariance matrix $\hat{S_n}$. This estimation leads to a well-conditioned, positive semi-definite matrix, serving as the shrinkage dispersion matrix defined below:

$$\hat{\mathrm{E}}_{\mathrm{Sh}} = (1-\eta)\hat{\mathrm{E}} + \eta\hat{\mathrm{T}}.$$

where $\hat{\mathrm{E}} = \hat{S_n}$. Various options for the shrinkage target $\hat{\mathrm{T}}$ have been proposed in the literature. For example, Leodit and Wolf [22] used a weighted average of the sample covariance matrix and a single-index covariance matrix as a shrinkage target. Leodit and Wolf [22] in their other work chose the shrinkage target as a "constant correlation matrix" with correlations equal to the average of all sample correlations. Using a scaled multiple of the identity matrix as a shrinkage goal, as suggested by [23], ensures a well-conditioned shrinkage covariance matrix even if the sample covariance matrix is not. DeMiguel et al. [11] have introduced an alternative approach to estimating the covariance matrix and its inverse. According to their proposal, a shrinkage estimator can be constructed taking a convex combination of the sample covariance matrix and a scaled shrinkage target. The same approach is executed for sample covariance inverse too in their paper. DeMiguel et al. [11] consider the scaled identity matrix as a target, same as that of [23]. In our procedure, also, we use the shrinkage target $\hat{\mathrm{T}} = \nu_\Sigma \mathrm{I}$. Thus $\hat{\mathrm{E}}_{\mathrm{Sh}} = (1-\eta)\hat{\mathrm{E}} + \eta\hat{\mathrm{T}}$ becomes

$$\hat{\mathrm{E}}_{\mathrm{Sh}} = (1-\eta)\hat{\mathrm{E}} + \eta\nu_\Sigma \mathrm{I},$$

where $\hat{\mathrm{E}} = \hat{S_n}$. We need to estimate η and ν_Σ. The parameters are selected such that minimizing the expected quadratic loss

$$i.e., \min_{\nu_\Sigma,\eta} \mathrm{E}\Big[\|\hat{\Sigma}_{\mathrm{Sh}} - \Sigma\|^2\Big], \text{ s.t. } \hat{\mathrm{E}}_{\mathrm{Sh}} = (1-\eta)\hat{S_n} + \eta\nu_\Sigma \mathrm{I},$$

where$||A||^2 = \mathrm{trace}(AA^T)/p$. Consider the above function to be minimized

$$\begin{aligned}
\mathrm{E}\Big[\|\hat{\Sigma}_{Sh} - \Sigma\|^2\Big] = \\
\mathrm{E}\Big[\|(1-\eta)\hat{S_n} + \eta\nu_\Sigma \mathrm{I} - \Sigma\|^2\Big] = \\
\mathrm{E}\Big[\|(1-\eta)\hat{S_n} + \eta\nu_\Sigma \mathrm{I} - \Sigma + \eta\Sigma - \eta\Sigma\|^2\Big] = \\
(1-\eta)^2\mathrm{E}\Big[\|\hat{S_n} - \Sigma\|^2\Big] + \eta^2\|\nu_\Sigma \mathrm{I} - \Sigma\|^2 + \\
2\mathrm{E}\Big[\langle(1-\eta)(\hat{S_n} - \Sigma), \eta(\nu_\Sigma \mathrm{I} - \Sigma)\rangle\Big].
\end{aligned}$$

Let our associated inner product is $\langle \mathrm{A}_1, \mathrm{A}_2\rangle = \mathrm{trace}(\mathrm{A}_1\mathrm{A}_2)^{\mathrm{T}}/p$. The latter element in the above expression is zero as $\mathrm{E}(\hat{S_n}) = \Sigma$ which is shown above. Therefore, the above minimization expression reduces to:

$$\mathrm{E}\Big[\|\hat{\Sigma}_{Sh} - \Sigma\|^2\Big] = [(1-\eta)^2\mathrm{E}\Big[\|\hat{S_n} - \Sigma\|^2\Big] + \eta^2\|\nu_\Sigma \mathrm{I} - \Sigma\|^2.$$

Parameter ν_Σ presents only on the right side element in above expression. Thus, minimizing the right element gives the optimum value of ν_Σ. Also, $\|\nu_\Sigma \mathrm{I} - \Sigma\|^2 = \nu_\Sigma^2\|\mathrm{I}\|^2 + \|\Sigma\|^2 - 2\nu_\Sigma\langle \mathrm{I}, \Sigma\rangle$. Thus, the first order minimization condition with respect to ν_Σ be

$$\begin{aligned}
2\nu_\Sigma - 2\langle \Sigma, \mathrm{I}\rangle = 0, \\
\nu_\Sigma = \mathrm{trace}(\Sigma)/p.
\end{aligned}$$

Since Σ is unknown, we propose to estimate with $\hat{S}_n$, thus $\nu_\Sigma = \mathrm{trace}(\hat{S_n})/p$. The first order optimal condition of η from equation gives:

$$\eta = \frac{\mathrm{E}\Big[\|\hat{S_n} - \Sigma\|^2\Big]}{\mathrm{E}\Big[\|\hat{S_n} - \nu_\Sigma \mathrm{I}\|^2\Big]}.$$