

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

NAAS Rating (2026): 4.49

Maths 2026; 11(6): 51-55

© 2026 Stats & Maths

<https://www.mathsjournal.com>

Received: 15-04-2026

Accepted: 18-05-2026

Muthukrishnan R

Professor and Head, Department
of Statistics, Bharathiar
University, Coimbatore, Tamil
Nadu, India

Dheepika G

Research Scholar, Department of
Statistics, Bharathiar
University, Coimbatore, Tamil
Nadu, India

A robust framework for chemometric data analysis using principal component regression

Muthukrishnan R and Dheepika G

DOI: <https://www.doi.org/10.22271/math.2026.v11.i6a.2337>

Abstract

In spectral datasets, especially high-dimensional data, multicollinearity among predictors and the presence of outliers significantly affect the performance of classical principal component regression (PCR). To address these issues, Robust Principal Component Regression (RPCR) considered more reliable in the presence of multicollinearity, outliers and high dimensionality in data. Robust PCR methods are examined by combining Robust Principal Component Analysis (RPCA) with robust regression techniques, including M, Least Median of Squares (LMS), Least Trimmed Squares (LTS), S, MM, and Quantile regression. Robust regression analysis delivers a substitute for least squares regression analysis, which fails to fulfil the fundamental assumption due to data characteristics. The experimental study has been carried out to study the performance of these procedures in a real environment by using the various error metrics, such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), along with the coefficient of determination (R^2). According to the error measures and R^2 , this study concludes that the RPCR based on robust PCA combined with MM-estimator performs well when compared with the other estimation methods.

Keywords: Principal component regression, robust principal component regression, robust regression, multicollinearity, outliers

1. Introduction

Chemometric datasets often show multicollinearity because the number of predictors is very large, and many variables contain similar or overlapping information, and several predictors become highly correlated with each other, leading to multicollinearity. High-dimensional spectral data often contains outliers among the variables; it significantly affects the classical regression techniques. Regression techniques are mainly used to deal with investigating the relationship between a response and a bunch of predictor variables for analysing and modelling the multi-factor data. In regression analysis, Ordinary Least Squares (OLS) fails when there is multicollinearity or outliers, the number of predictors exceeds the number of observations ($p > n$) and its assumptions are violated. In such situations, robust estimators are used to obtain more stable and reliable parameter estimates. Because OLS estimates are strongly influenced by extreme observations, robust estimators perform better since they reduce the influence of outliers and provide more stable estimates.

Principal component regression is an approach in which PCA is applied to the predictors, followed by regression on a selected subset of orthogonal components. (PCR) is a commonly used method to handle multicollinearity and to predict the parameter of a regression model, but it remains sensitive to non-typical observations. To overcome the issues, this study performs Robust Principal Component Regression (RPCR), which works well when compared to classical PCR. Robust regression is applied to obtain more reliable estimates in the presence of outliers. Robust Principal Component Regression is obtained by integrating robust principal component analysis with robust regression techniques. This study is to comparatively evaluate RPCR methods based on M-Estimator (M), Least Median of Squares (LMS), Least Trimmed Squares (LTS), S-Estimator (S), MM-Estimator (MM), and Quantile Regression.

Corresponding Author:

Muthukrishnan R

Professor and Head, Department
of Statistics, Bharathiar
University, Coimbatore, Tamil
Nadu, India

The primary objective of this study is to evaluate and compare various robust regression techniques in order to determine the most effective estimator for high-dimensional spectral data affected by both multicollinearity and outliers.

The rest of the paper is organized as follows. In section 2, provides a brief overview of classical principal component regression along with the robust principal component regression approach. It also outlines the robust regression methods applied in this study, including M, LMS, LTS, S, MM, and quantile regression. In section 3, presents the evaluation of these techniques under real-world conditions and summarizes the findings, while the final section offers the overall conclusions of the study.

2. Methodology

The conventional regression procedure, namely, classical Principal Component Regression (PCR), and the robust procedures, Robust Principal Component Regression (RPCR), are a combination of robust Principal Component Analysis (RPCA) with robust regression methods such as M-Estimator (M), Least Median of Squares (LMS), Least Trimmed Squares (LTS), S-Estimator (S), MM-Estimator (MM), and Quantile Regression, which are briefly discussed in this section.

Principal Component Regression

Principal Component Regression (PCR) is a multivariate regression method developed from Principal Component Analysis (PCA) introduced by Pearson^[13] and extended by Hotelling^[3]. PCR employs PCA to reduce high-dimensional, correlated predictor variables into a smaller number of uncorrelated principal components (PCs). PCR is used to handle multicollinearity in high-dimensional datasets, reduce dimensionality when predictors are too many ($p > n$), improve prediction accuracy and model robustness, and simplify complex data for easier interpretation.

Let X be the predictor matrix and y be the response vector. Standardise the predictor variables (X) to have a zero mean and unit variance, ensuring they are on the same scale.

$$X_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j}, j = 1, 2, \dots, p \quad (1)$$

Apply Principal Component Analysis (PCA) on the predictor matrix (X) to transform the original correlated variables into a new set of uncorrelated components (principal components). Then compute the covariance (or correlation matrix) and also eigen decomposition.

$$S = \frac{1}{n-1} (X^*)' X^* \quad (2)$$

$$S = P \Lambda P' \quad (3)$$

where $P'P = I_p$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

The Principal components are obtained as

$Z = X^* P$ and $Z'Z = (n-1) \Lambda$. Hence the principal component are orthogonal (uncorrelated). Choose the top k principal components corresponding to the largest eigenvalues $Z_k = X^* P_k$, $k < p$. Then the proportion of variance explained is,

$$k = \frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^p \lambda_j} \quad (4)$$

Select k such that the cumulative variance explain most of the variance in the data, it also determined using cross-validation

or by checking eigenvalues. Fit a linear regression model OLS using the k selected components as predictors and the target variable (Y). The estimator is

$$Y = Z_k \alpha + \varepsilon, \varepsilon \sim N(0, \sigma^2 I) \quad (5)$$

$$\hat{\alpha} = (Z_k' Z_k)^{-1} Z_k' y \quad (6)$$

Principal Component Regression as a method to reduce multicollinearity by regressing the response variable on principal components instead of the original correlated predictors. Principal Component Regression (PCR) reduces multicollinearity but has some drawbacks. It is sensitive to outliers, chooses components based on predictor variance rather than their relevance to the response, and often produces components that are difficult to interpret. Selecting the right number of components is challenging; it assumes linear relationships and may overlook important low-variance predictors. Massy^[11] introduced Principal Component Regression as a technique to handle multicollinearity by transforming correlated predictors into a smaller set of uncorrelated principal components and then performing regression on those components; however, PCR uses classical PCA, which is sensitive to outliers. Keithley *et al.*^[9] work on PCR, which is applied in chemometrics and spectral calibration models, and performance is affected by noisy data. These limitations have been overcome by the development of robust PCR methods to improve performance and interpretability. Robust PCR methods identify or reduce the influence of outliers, producing more reliable principal components and regression estimates.

Robust Principal Component Regression

Robust PCR is an advanced form of PCR that integrates robust PCA with regression to handle outlier-prone data, and it is designed for datasets where classical PCR fails due to outliers or non-normal errors. Hubert *et al.*^[6] introduced robust PCA, which is resistant to outliers. Serneels & Verdonck^[17] proposed robust principal component regression. Todorov & Filzmoser^[19] Applied robust PCR for chemometric datasets. RPCR and PCR follow the same overall framework of dimension reduction and then regression, but the key change happens at the point where PCR becomes robust PCR is when we replace classical PCA with robust PCA and OLS regression with robust regression methods. This modification ensures that unusual observations do not influence the model significantly, making robust PCR more reliable and stable for real-world datasets where noise and outliers are present. Robust PCR is a two-way approach; first apply Robust Principal Component Analysis (ROBPCA) to extract outlier-resistant Principal Components (PCs)^[2, 6]. Then select the top robust PCs that explain major clean variance while ignoring abnormal values and, secondly, fit a robust regression model (e.g., M-estimator, LMS-estimator, LTS-estimator, S-estimator, MM-estimator or quantile regression) using these robust PCs to predict the response variable.

Robust Regression

Robust regression provides an alternative to classical regression techniques when data do not satisfy key assumptions such as normality or the absence of outliers. Its primary aim is to overcome the limitations of traditional methods. When the data follow a normal distribution and are free from outliers, robust regression yields results similar to

those obtained using conventional approaches. However, unlike OLS, which is highly affected by extreme values, robust methods limit their impact. As a result, these approaches produce more reliable parameter estimates and improved model performance, particularly in practical datasets that often contain abnormal data points. This leads to more accurate parameter estimates and better model performance in real-world datasets. In this section, various robust regression techniques including M-estimators, Least Median of Squares (LMS), Least Trimmed Squares (LTS), S-estimators, MM-estimators, and Quantile Regression are discussed.

M-estimator

The M-estimator was introduced by Huber [4] as a generalisation of least squares using robust loss functions. Here the M stands for “maximum likelihood type”. The M-estimator is used to estimate parameters by minimising a robust loss function instead of squared residuals. The breakdown point of the M-estimator is close to zero; it reduces the influence of large residuals using a ψ -function. M-estimators are highly efficient under normal errors. However, they are not robust to leverage points. They are suitable when data contain mild outliers but no extreme leverage points. The M-estimator is defined as

$$\widehat{\beta}_M = \arg \min_{\beta} \sum_{i=1}^n \rho \left(\frac{r_i(\beta)}{\sigma} \right) \quad (7)$$

Where $\rho(\cdot)$ is a robust loss function (e.g., Huber function) introduced for the effect of outliers, ρ is a positive definite, symmetric function with zero as its unique minimum and σ is a scale estimate. An algorithm was developed by Susanti *et al.* [18]. It provides robustness while maintaining efficiency under normal conditions, and it is sensitive to high-leverage points and does not have a high breakdown point. It requires selection of a tuning parameter, which affects performance. M-estimators have some constraints, such as dependence on tuning parameters, sensitivity to high leverage points and a relatively low breakdown point.

Least Median of Squares (LMS)

The Least Median of Squares (LMS) estimator is a robust regression technique introduced by Rousseeuw [14]. LMS is designed to estimate regression parameters in the presence of severe outliers. In this method the sum is replaced by the median in the method of least squares. Unlike ordinary least squares, LMS minimises the median of squared residuals, making it highly resistant to outliers. It is mainly used as an initial robust estimator or for outlier detection. It is highly resistant to both vertical outliers and leverage points. It has very low asymptotic efficiency under normal errors and a non-smooth objective function. The LMS estimator is defined as

$$\widehat{\beta}_{LMS} = \arg \min_{\beta} \text{median} (r_i^2(\beta)) \quad (8)$$

Where $r_i^2(\beta)$ is a median. The LMS estimator provides a high breakdown point (up to 50%), making it highly robust to outliers, but it suffers from low efficiency and high computational cost.

Least Trimmed Squares (LTS)

The Least Trimmed Squares (LTS) is a robust regression technique to improve the efficiency of LMS while maintaining robustness. It estimates regression parameters by

minimising the sum of the smallest h squared residuals. Large residuals corresponding to outliers are trimmed from the objective function. This method has an objective function of the lowest trimmed squared residuals as

$$\sum_{i=1}^h r_i^2(\beta) \quad (9)$$

Where $r_i^2(\beta) \leq \dots \leq r_n^2(\beta)$ are the ordered residuals and h is the trimming parameter. The LTS estimator is defined as,

$$\widehat{\beta}_{LTS} = \arg \min_{\beta} \sum_{i=1}^h r_i^2(\beta) \quad (10)$$

The computation of LTS estimator uses an algorithm called FAST-LTS, it was introduced by Rousseeuw and Van Driessen [15]. LTS can be calculated for α , the trimming proportion tending to 50%. LTS attains a high breakdown point at $h = \left(\frac{n}{2}\right) + \left[\frac{(p+1)}{2}\right]$. LTS estimator provides a high breakdown point with better efficiency than LMS, making it suitable for outlier-contaminated data, but it still involves high computational cost and parameter selection issues. Despite challenges such as computational cost and the need to select an appropriate trimming proportion.

S-Estimator (S)

The S-estimator was introduced by Rousseeuw and Yohai [16] to focus on robustness through scale minimization. It estimates regression parameters by minimizing a robust estimate of scale of the residuals. S-estimators combine the flexibility and asymptotic features of M-estimators to provide a simple high breakdown regression estimator. Their Influence function is bounded and resistant to leverage points. It has low statistical efficiency under normality. It is also used as an initial estimator for other estimators. The S-estimator $\widehat{\theta} = \min_{\theta} s(r_1(\theta), r_2(\theta), \dots, r_n(\theta))$ is defined by minimizing the scale estimate $\widehat{\sigma}$

$$\widehat{\sigma} = s(r_1(\widehat{\theta}), r_2(\widehat{\theta}), \dots, r_n(\widehat{\theta})) \quad (11)$$

The estimator of scale may be defined by the function ρ , For any sample $r_1, \dots, r_n$ of real numbers, we define the scale estimates $s(r_1, r_2, \dots, r_n)$ as a solution of

$$b = \frac{1}{n} \sum_{i=1}^n \rho \left(\frac{r_i(\beta)}{s} \right) \quad (12)$$

where “ s ” is the robust scale estimate and b is a tuning constant. S-estimators provide a high breakdown point and strong resistance to outliers, but they generally have low statistical efficiency under normal error distributions and higher computational complexity. The name “S-estimator” stems from their reliance on scale-based calculations.

MM-Estimator (MM)

The MM estimator is a robust regression technique introduced by Yohai [20] to combine high robustness with high efficiency. It is used to estimate regression parameters in the presence of outliers and leverage points. MM-estimation combines the robustness of S-estimators with the efficiency of M-estimators. This estimator is a modification of the M-estimator, designed to provide both high robustness and high efficiency. The construction begins with a high-breakdown initial estimator, such as an S-estimator. Its asymptotic efficiency can be tuned to exceed 95%. The influence function is bounded and redescending, reducing the effect of extreme outliers. The MM estimator is defined as

$$\hat{\beta}_{MM} = \arg \min_{\beta} \sum_{i=1}^n \rho \left(\frac{r_i(\beta)}{s} \right) \quad (13)$$

Where $r_i(\beta)$ are residuals, s is a robust scale estimate based on the initial robust estimator, and ρ is a loss function like Huber, Hampel or Tukey's biweight. MM-estimators combine the high breakdown point of S-estimators and the high efficiency of M-estimators, making them one of the most reliable robust regression methods, though they involve complex computation and tuning.

Quantile Regression

Quantile regression was introduced by Koenker and Bassett^[10], it models conditional quantiles of the response variable rather than the conditional mean. It estimates regression parameters by minimizing an asymmetric absolute loss function. Median regression ($\tau = 0.5$) provides robustness against outliers in the response variable. It does not require any normality, distributional or homoscedasticity assumptions. It is sensitive to leverage points and does not achieve a high breakdown point. It is useful for studying distributional heterogeneity. The quantile regression estimator is defined as,

$$\hat{\beta}_{\tau} = \arg \min_{\beta} \sum_{i=1}^n \rho_{\tau}(y_i - x_i^T \beta) \quad (14)$$

Where $\rho_{\tau}(u) = u(\tau - I(u < 0))$ is the loss function. Quantile regression provides a robust alternative to OLS by modeling different parts of the response distribution, making it effective under outliers and heteroscedasticity, though it can be computationally intensive.

3. Experimental Result

This section presents the results of experimental studies carried out under a real environment and compares the performance of the various regression procedures, and thus obtained results were also discussed. The numerical analysis was carried out using two real-world case studies, both characterized by the presence of multicollinearity and outliers. The calculated performance metrics MSE, RMSE, and R^2 for the various regression estimators are compiled and presented in Table 1 and Table 2.

The presence of multicollinearity in the real datasets has been identified by computing the condition index (CI) or condition number, which is used to measure the overall multicollinearity of the variables proposed by Belsley *et al.*^[1]. The condition index was computed using the singular values obtained from the singular value decomposition of the standardised design matrix, ensuring numerical stability. The condition index shows how severely predictors in a regression model are linearly dependent. The condition index, which by definition exceeds unity, increases with the degree of linear dependence among explanatory variables. Conventionally, CI values less than 10 are interpreted as no multicollinearity, values in the range 10 to 30 indicate moderate multicollinearity, and values greater than 30 are taken as evidence of strong multicollinearity; CI values greater than 100 are interpreted as severe multicollinearity.

The outliers in the real datasets were detected by using the Robust PCA Outlier Map (SD-OD plot)^[5], and the analysis has been carried out using R software. The SD-OD plot displays score distance and orthogonal distance obtained based on the robust PCA score. On the horizontal axis, mark for each data point its score distance; observations whose scores exceed the cut-off value are outlying within the PCA subspace. On the vertical axis, mark for each data point

its orthogonal distance; it measures how far a data point lies from the PCA subspace. The observations beyond the cut-off limits are identified as outliers or leverage points.

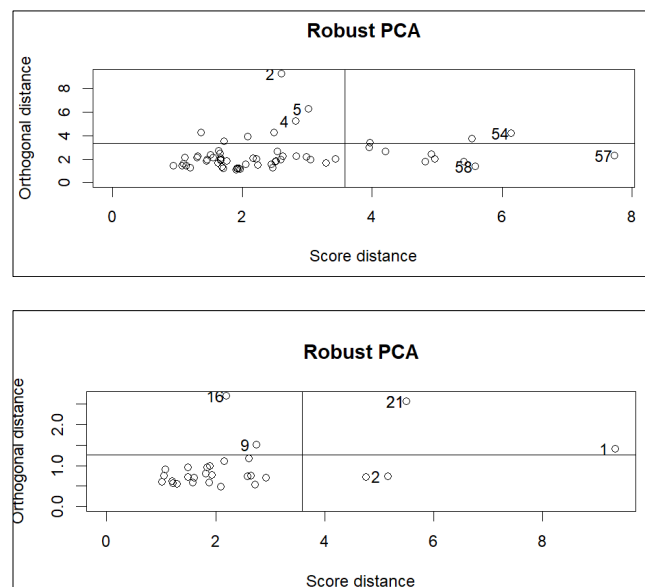


Fig 1: Diagnostic plot Gasoline Data Figure 2: Diagnostic plot Yarn Data

Case 1: Gasoline Dataset. Kalivas^[8] introduced the gasoline data. The data come from a study that looked at how gasoline quality is represented by its octane number. Modelling it through NIR spectral measurements allows efficient fuel quality monitoring. The data frame contains 60 samples with 401 NIR spectral measurements of gasoline data. The dependent variable was the octane number, and the independent variable was the spectra, which were recorded over the wavelength range of 900-1700 nm at 2 nm intervals, resulting in 401 spectral variables. Six outliers are found in this dataset by using the robust PCA outlier plot, which is shown in the Figure 1. Since the condition index of this gasoline data set was greater than 100, the explanatory variables are highly correlated. As a result, the dataset has severe multicollinearity.

Case 2: Yarn Dataset. The data come from a study that looked at how yarn density reflects an important physical property of the material; modelling it using NIR spectral measurements enables efficient quality assessment of yarn samples. The data is available in the R-package "pls". This dataset has 28 observations. The data contain an independent variable, as the spectra were recorded over the 268 wavelengths. Five outliers are found in this dataset, which is shown in the Figure 2. Since the condition index of the independent variables is greater than 100, the explanatory variables are highly correlated; there is an indication of severe multicollinearity.

Table 1: Computed error measures and R^2 value under various regression methods (Gasoline Data)

Model	MSE	RMSE	R^2
PCR	0.0419	0.2046	0.9818
RPCR-M	0.0410	0.2024	0.9822
RPCR-LMS	0.0697	0.2641	0.9697
RPCR-LTS	0.0429	0.2071	0.9814
RPCR-S	0.0531	0.2303	0.9770
RPCR-MM	0.0409	0.2023	0.9822
RPCR-QUANTILE	0.0441	0.2099	0.9809

Table 2: Computed error measures and R^2 value under various regression methods (Yarn Data)

Model	MSE	RMSE	R^2
PCR	0.5297	0.7278	0.9992
RPCR-M	0.5267	0.7257	0.9992
RPCR-LMS	1.2736	1.1286	0.9982
RPCR-LTS	0.6402	0.8001	0.9991
RPCR-S	0.7241	0.8510	0.9990
RPCR-MM	0.5222	0.7226	0.9993
RPCR-QUANTILE	0.5479	0.7402	0.9992

The results obtained in Table 1 & Table 2 show that the error values for different estimators are slightly different from each other. Also, the RPCR based on MM regression has the smallest MSE and RMSE and a higher R^2 value than all others. Hence, RPCR is more efficient than the other estimators in the case of datasets having indications of multicollinearity and outliers.

4. Conclusion

Regression analysis is a fundamental statistical learning method that plays an important role across a wide range of research fields. However, the commonly used principal component regression approach may not be adequate when the data violate underlying model assumptions. In such situations, alternative methods are required to develop a more reliable and effective model for the given dataset. In chemometric data analysis, PCR is commonly used because it effectively handles multicollinearity and high-dimensional data. However, when the dataset contains outliers, the performance of PCR deteriorates, as classical PCA and OLS regression are highly sensitive to extreme observations. To overcome this limitation, robust PCR is preferred. By incorporating robust PCA and robust regression techniques, it reduces the influence of outliers and provides more stable and reliable results. Therefore, even if chemometric data contains outliers, strong multicollinearity, high dimensionality, or nonlinearity, robust PCR can still perform effectively and produce accurate predictions, making it more suitable for real-world chemometric applications. On the basis of error and R^2 value, the study concluded that the robust principal component regression based on MM gives a better model for the data when the datasets have multicollinearity and outliers. Further, this regression procedure can be beneficial to researchers who work on chemometrics and machine learning techniques by considering factors such as multicollinearity, outliers and high dimensionality.

References

1. Belsley DA, Kuh E, Welsch RE. Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons; 2005.
2. Filzmoser P. Robust principal component regression. In: Computer data analysis and modeling. Robust and computer intensive methods. Minsk: Belarusian State University; 2001. p. 132-137.
3. Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology. 1933;24(6):417-441.
4. Huber PJ. Robust estimation of a location parameter. In: Breakthroughs in Statistics: Methodology and Distribution. New York: Springer; 1992. p. 492-518.
5. Hubert M, Verboven S. A robust PCR method for high-dimensional regressors. Journal of Chemometrics. 2003;17(8-9):438-452.
6. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: A new approach to robust principal component analysis. Technometrics. 2005;47(1):64-79.
7. Jolliffe IT. A note on the use of principal components in regression. Journal of the Royal Statistical Society Series C. 1982;31(3):300-303.
8. Kalivas JH. Two data sets of near infrared spectra. Chemometrics and Intelligent Laboratory Systems. 1997;37(2):255-259.
9. Keithley RB, Wightman RM, Heien ML. Multivariate concentration determination using principal component regression with residual analysis. Trends in Analytical Chemistry. 2009;28(9):1127-1136.
10. Koenker R, Bassett G Jr. Regression quantiles. Econometrica. 1978;46(1):33-50.
11. Massy WF. Principal components regression in exploratory statistical research. Journal of the American Statistical Association. 1965;60:234-246.
12. Muthukrishnan R, Ramakrishnan K. Effect of classical and robust regression estimators in the context of high-dimensional data with multicollinearity and outliers. Reliability: Theory and Applications. 2024;19(1):335-341.
13. Pearson K. Principal components analysis. Philosophical Magazine. 1901;6(2):559-572.
14. Rousseeuw PJ. Least median of squares regression. Journal of the American Statistical Association. 1984;79(388):871-880.
15. Rousseeuw PJ, Van Driessen K. Computing LTS regression for large data sets. Data Mining and Knowledge Discovery. 2006;12(1):29-45.
16. Rousseeuw PJ, Yohai VJ. Robust regression by means of S-estimators. In: Robust and Nonlinear Time Series Analysis: Proceedings of a Workshop Organized by the Sonderforschungsbereich 123 "Stochastische Mathematische Modelle", Heidelberg, 1983. Berlin: Springer; 1984. p. 256-272.
17. Serneels S, Verdonck T. Principal component analysis for data containing outliers and missing elements. Computational Statistics and Data Analysis. 2008;52(3):1712-1727.
18. Susanti Y, Pratiwi H, Sulistijowati S, Liana T. M-estimation, S-estimation, and MM-estimation in robust regression. International Journal of Pure and Applied Mathematics. 2014;91(3):349-360.
19. Todorov V, Filzmoser P. An object-oriented framework for robust multivariate analysis. Journal of Statistical Software. 2010;32(1):1-47.
20. Yohai VJ. High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics. 1987;15(2):642-656.