JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 6, No. 4, October 2022, pp. 844-856

Performance of LAD-LASSO and WLAD-LASSO on High Dimensional Regression in Handling Data Containing Outliers

Septa Dwi Cahya[1], Bagus Sartono[2], Indahwati[3], Evita Purnaningrum[4]
[1,2,3]Department of Statistics, IPB University, Indonesia
[4]Department of Economics, Adi Buana University, Indonesia
septadwicahya@apps.ipb.ac.id[1], bagusco@apps.ipb.ac.id[2], indah.stk@gmail.com[3],
purnaningrum@unipasby.ac.id[4]

Article History:
Received : 03-06-2022
Revised : 19-08-2022
Accepted : 23-08-2022
Online : 08-10-2022

Keywords:
High Dimensional Data;
LAD-LASSO;
Multicollinearity;
Outliers;
WLAD-LASSO.

ABSTRACT

In several research areas, it is common to have a dataset with more explanatory variables than the number of observations, called high-dimensional data. This condition can lead to multicollinearity problem. The least absolute shrinkage and selection operator (LASSO) solves the problem by shrinking the estimated coefficient to zero so that it can simultaneously carry on the variable selection and the parameter estimation. But LASSO performs poorly when the data contains some outliers in the response or explanatory variables. Robust methods have addressed this problem based on the least-absolute-deviation approach, such as LAD-LASSO and WLAD-LASSO. This current research aims to evaluate the performance of the LAD-LASSO and WLAD-LASSO methods on high-dimensional and low-dimensional data containing outliers. To evaluate the performance of these methods, the simulation study was conducted. The simulation study used three scenarios (without outliers, outliers on the response variable (5%, 10%, 15%), outliers both on the response and explanatory variables (5%, 10%, 15%)). We also used the Minimum Regularized Covariance Determinant (MRCD) estimator in calculating the weights on the WLAD-LASSO. The best method from this simulation then will be applied to sembung leaf extract data to identify antioxidant marker compounds in sembung leaf extract. The simulation results show that LAD-LASSO tends to be very tight in selecting, while LASSO tends to be too loose. Meanwhile, WLAD-LASSO is in the middle of those two techniques and performs the best in identifying the important variables correctly. Even the existence of weights cause WLAD-LASSO more robust against the presence of outliers in the response and explanatory variables compared to LAD-LASSO. Furthermore, performance of these methods on high-dimensional data decrease compared to low-dimensional data. The performance of these methods also tends to decrease when the rate of outlier increases. The WLAD-LASSO was then implemented in actual data to find the compound of antioxidant markers in the sembung leaf extract. The compounds/formulas obtained are Umbelliferone, 12-Hydroxyjasmonic Acid, $C_{22}H_{14}N_8O_2$, and Acetyleugenol (with a prediction error is 0.133050). These compounds/formulas can be developed as natural antioxidants and have the potential to be developed as medicinal ingredients.

https://doi.org/10.31764/jtam.v6i4.8968

A. INTRODUCTION

Regression analysis is a statistical method to determine the relationship between the response variable and one or more explanatory variables (Bangdiwala, 2018). Ordinary Least Square (OLS) is usually used to estimate the regression model parameters by minimizing the sum of squared residuals. However, OLS is very sensitive to outliers (Varin, 2021). Least

Absolute Deviation (LAD) is an alternative method for OLS to overcome outliers (Dielman, 2005; Wang, 2013). LAD estimates the regression parameters by minimizing the sum of absolute residuals. LAD gives the same weight to all observations while OLS squares the residuals giving large weight to large residuals. Then LAD is more robust than OLS when the residual distribution is not normal. However, LAD is only robust to outliers in the response variable. LAD is very sensitive to outliers in the explanatory variable (leverage point). The Weighted LAD (WLAD) was introduced to deal with outliers in the explanatory variables (Giloni et al., 2006). In this method, the weight that only depends on the explanatory variable is given to reduce the weight of the leverage point to reduce the effect of outliers on the parameter estimation process (Yang & Li, 2018).

Some research areas (such as biology, signal processing, chemistry, and others) sometimes contain explanatory variables more than observations, known as high-dimensional data ($p \gg n$), so it is necessary to select variables (Wasserman & Roeder, 2009; Lima et al., 2020). In addition, this condition can lead to multicollinearity problem (Zhao et al., 2020). Multicollinearity occurs when several explanatory variables are correlated not only with the response variables but also among explanatory variables (Daoud, 2017). Multicollinearity can be indicated from the columns of the explanatory variable matrix ($\boldsymbol{X}$) that are linearly dependent on each other, so the covariance matrix of the explanatory variable ($\boldsymbol{X}^T\boldsymbol{X}$) is non-invertible. As a result, the variety of parameter estimates becomes large, so the predicted results tend to be unstable and the prediction model obtained is inaccurate (Keith, 2015).

Least Absolute Shrinkage and Selection Operator (LASSO) is a regularization method that minimizes the sum of the residual squares by adding L1-norm constraints (Tibshirani, 1996). LASSO can solve the problem of multicollinearity by shrinking the estimated coefficient close to zero (even exactly zero) to carry on the variable selection and the parameter estimation simultaneously but it is sensitive to outliers (Tibshirani, 2013; Sirimongkolkasem & Drikvandi, 2019). Adaptive LASSO was then introduced using different weights for each regression coefficient on L1 regularization to be more consistent than LASSO in estimating parameters and selecting variables (Zou, 2006; Camponovo, 2022). However, Adaptive LASSO is also based on OLS which is sensitive to outliers in response and explanatory variables (Machkour et al., 2020).

Wang et al. (2007) introduced LAD-LASSO method combining LAD criteria and Adaptive LASSO penalty. They compared LAD-LASSO with LAD based on other traditional variable selection methods, namely Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) in low-dimensional data. As a result, LAD-LASSO is better than LAD-AIC and LAD-BIC. Arslan (2012) introduced WLAD-LASSO method combining WLAD criteria and Adaptive LASSO penalty. Arslan (2012) showed from the simulation study that WLAD-LASSO is robust in low-dimensional data containing outliers. Rahardiantoro & Kurnia (2015) studied a simulation to compare LASSO and LAD-LASSO in high-dimensional data. They obtained that LAD-LASSO is better than LASSO in selecting variables in high-dimensional data containing outliers. Also, Ajeel & Hashem (2020) obtained LAD-LASSO has good performance overcoming outliers in low and high dimensional data.

In this research, we compared the LAD-LASSO and WLAD-LASSO methods on high-dimensional data containing outliers in variable selection and also compared them with LASSO. Then we compared the performance of these methods on low-dimensional data. Comparison of

the performance from these methods is useful for evaluating the best method in the selection of variables. The variable selection method is useful for selecting important variables in the research data, so this method is effective and efficient in determining the explanatory variables that have a significant effect on the response variables (especially if the data contains a lot of explanatory variables). In this research, Minimum Regularized Covariance Determinant (MRCD) estimator was used in computing the weight of WLAD-LASSO rather than the Minimum Covariance Determinant (MCD) estimator. The MCD estimator, used in the simulation study by Arslan (2012), cannot be applied to the case of high-dimensional data such as the simulation study in this paper. The MRCD estimator modifies the MCD estimator for high-dimensional data (Bulut, 2020). This estimator aims to regularize the covariance based on the subset that makes overall determinant the smallest (Boudt et al., 2019).

Furthermore, we applied WLAD-LASSO as the best method that has been evaluated in the previous simulation study to actual data to identify the active compounds that indicate antioxidants in the Sembung leaf extract. The variable selection method applied to the sembung data is useful for selecting important compounds/formulas in the antioxidant content of sembung leaves. These selected compounds/formulas can be developed as natural antioxidants and have the potential to be developed as medicinal ingredients.

B. METHODS

In this paper, we use simulation data and actual data analyzed by R software. Simulation data were obtained from generating data with several levels of outliers on the response variable and the explanatory variable and the level of correlation between the explanatory variables. This simulation refers to research by Arslan (2012), Wahid et al. (2017), and Ajeel & Hashem (2020). However, there are modifications in the level of an outlier on the response and/or the explanatory variable ($\delta = 5\%, 10\%, 15\%$), and the level of correlation between the explanatory variable ($\rho = 0.1, 0.2, 0.3, \dots, 0.9$). Furthermore, simulations were carried out in the case of low-dimensional data ($n = 50, p = 10$) and high-dimensional data ($n = 50, p = 100$). Therefore, we use the Minimum Regularized Covariance Determinant (MRCD) estimator in this study instead of the Minimum Covariance Determinant (MCD) estimator as in Arslan (2012). The MRCD estimator can be used on both high and low dimensional data, while the MCD estimator only can be used on low dimensional data. The following are the steps in a simulation study:

1. Set the number of observations $n = 50$ and the number of explanatory variables $p = 10$
2. Generate a p-dimensional explanatory variable vector on the i-th observation $\boldsymbol{x}_i = \left(x_{1i}, x_{2i}, \dots, x_{pi}\right)^T$ based on a multivariate normal distribution $\boldsymbol{x}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ where covariance matrix $\boldsymbol{\Sigma} = \left(r_{jk}\right)_{p\times p}$; $r_{jk} = \rho^{|j-k|}$; level of correlation between the explanatory variable $\rho = 0.1, 0.2, 0.3, \dots, 0.9$; $i = 1,2,3, \dots, n$; $j, k = 1,2,3, \dots, p$.
3. Set j-th regression parameter as $\beta_j = \begin{cases} 1, & j = 1,2, \dots, 5 \\ 0, & j = 6,7, \dots, p \end{cases}$, so that the linear regression model is $y_i = x_{1i} + \cdots + x_{5i} + \varepsilon_i$ where y_i is a response variable at the i-th observation, x_{ji} is the j-th explanatory variable on the i-th observation, and ε_i is random error on the i-th observation. $i = 1,2,3, \dots, n$; $j = 1,2,3, \dots, p$.
4. Generate the response variable in three scenarios as follows:

a. Scenario 1: no outliers by model $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$ where y_i is a response variable at the i-th observation, $\boldsymbol{x}_i = \left(x_{1i}, x_{2i}, \dots, x_{pi}\right)^T$ is p-dimensional explanatory variable vector on the i-th observation, $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \dots, \beta_p\right)^T$ is a vector of regression parameters, and $\varepsilon_i \sim N(0,1)$ is a random error on the i-th observation.
b. Scenario 2: Outliers on the response variable by model $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$ where y_i is a response variable at the i-th observation, $\boldsymbol{x}_i = \left(x_{1i}, x_{2i}, \dots, x_{pi}\right)^T$ is p-dimensional explanatory variable vector on the i-th observation, $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \dots, \beta_p\right)^T$ is a vector of regression parameters, and ε_i is a random error on the i-th observation. ε_i is generated from: (i) Normal distribution $(1-\delta)N(0,1) + \delta N(25,1)$; (ii) Exponential distribution $(1-\delta)\,[\exp(1) - 1] + \delta \exp(25)$ where the outlier level $\delta = 5\%, 10\%, 15\%$.
c. Scenario 3: Outliers on the response variable and the explanatory variable.
   1) Replace the first 10 rows in the explanatory variable matrix $\boldsymbol{X}$ from normal distribution $N(10,1)$, so that a new explanatory variable matrix is obtained, namely $\boldsymbol{X}^*$.
   2) Generate the response variable by model $\boldsymbol{Y} = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{Y} = (y_1, y_2, \dots, y_n)^T$ is a response variable vector, $\boldsymbol{X}^*$ is an explanatory variable matrix containing $\boldsymbol{x}_i = \left(x_{1i}, x_{2i}, \dots, x_{pi}\right)^T$, and $\boldsymbol{\varepsilon}$ is a random error vector containing ε_i. ε_i is generated from: (i) Normal distribution $(1-\delta)N(0,1) + \delta N(25,1)$; (ii) Exponential distribution $(1-\delta)\,[\exp(1) - 1] + \delta \exp(25)$ where the outlier level $\delta = 5\%, 10\%, 15\%$.

5. Variable selection using LASSO, LAD-LASSO, and WLAD-LASSO methods.
   a. Determine the optimum lambda (λ) based on cross validation.
   b. Estimate the regression parameters as follows:

$$\widehat{\boldsymbol{\beta}}_{LASSO} = argmin_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p}|\beta_j|\right] \quad (1)$$

$$\widehat{\boldsymbol{\beta}}_{LAD-LASSO} = argmin_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n}|y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}| + \lambda \sum_{j=1}^{p}\omega_j|\beta_j|\right] \quad (2)$$

$$\widehat{\boldsymbol{\beta}}_{WLAD-LASSO} = argmin_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} w_i|y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}| + \lambda \sum_{j=1}^{p}\omega_j|\beta_j|\right] \quad (3)$$

   Where y_i is a response variable at the i-th observation, $\boldsymbol{x}_i = \left(x_{1i}, x_{2i}, \dots, x_{pi}\right)^T$ is p-dimensional explanatory variable vector on the i-th observation, $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \dots, \beta_p\right)^T$ is a vector of regression parameters with β_j is j-th regression parameter ($j = 1,2, \dots, p$), the weight w_i is defined as $w_i = min\left\{1, \frac{p}{RD(\boldsymbol{x}_i)}\right\}$ with $RD(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T\widehat{\sum}^{-1}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})$ is a robust mahalanobis distance at i-th observation, $\widehat{\boldsymbol{\mu}}$ is a robust mean vector of observations on each explanatory variable, and $\widehat{\sum}$ is robust covariance matrix of the explanatory variable. $\widehat{\boldsymbol{\mu}}$ and $\widehat{\sum}$ are obtained from the MRCD estimator (Boudt et al., 2019). The weight ω_j is defined as $\omega_j = \frac{1}{|\widehat{\beta}_j|}$ where $\widehat{\beta}_j$ is an estimated parameter from OLS (Zou, 2006).

6. The simulation in steps 2 to 5 is repeated 1000 times.
7. Evaluate the performance of LASSO, LAD-LASSO, and WLAD-LASSO based on No. Correct (NC), No. Incorrect (NC), mean MAD (Mean Absolute Deviation). NC is the average number

of selected significant variables, NI is the average number of selected insignificant variables, and the mean MAD (mMAD) is the average prediction error.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

Where n is the number of observations, y_i is a response variable at the i-th observation, and $\hat{y}_i$ is an estimated response variable at the i-th observation.

8. Steps 1 to 7 are repeated for the number of explanatory variables $p = 100$.

Actual data is a secondary data obtained from analysis results of LC-MS/MS (Liquid Chromatography-Mass Spectrometry/Mass Spectrometry) sembung leaf extract at the Central Laboratory for Biopharmaca Studies IPB in May-June 2021 in collaboration with the Statistics Department of IPB. This data contains 35 observations, 2098 explanatory variables, and one response variable. The observations used were samples of sembung leaf extract. The samples were obtained from the simplicia extraction of sembung leaves with 5 types of solvents, namely water, 30% ethanol, 50% ethanol, 70% ethanol, and 100% ethanol (each type of the solvent was repeated 7 times). The samples were then analyzed using LC-MS/MS that is a combination analysis technique of liquid chromatography and mass spectrometry. Liquid chromatography separates the analyte components in the sample, then proceeds to mass spectrometry for ionization. The ions are separated based on their respective masses, where the physical properties of the ions are measured from the mass-to-charge ratio (m/z) and obtained 2098 mass-to-charge ratios (m/z) which are used as explanatory variables. Each of the 2098 mass-to-charge ratios represents the name of the compound/formula denoted by X1, X2, ..., X2098. Furthermore, antioxidant tests were carried out on samples of sembung leaf extract so that the antioxidant content of each sample was obtained. This antioxidant content is used as a response variable. The following are steps to identify antioxidant compounds in Sembung leaf extract:

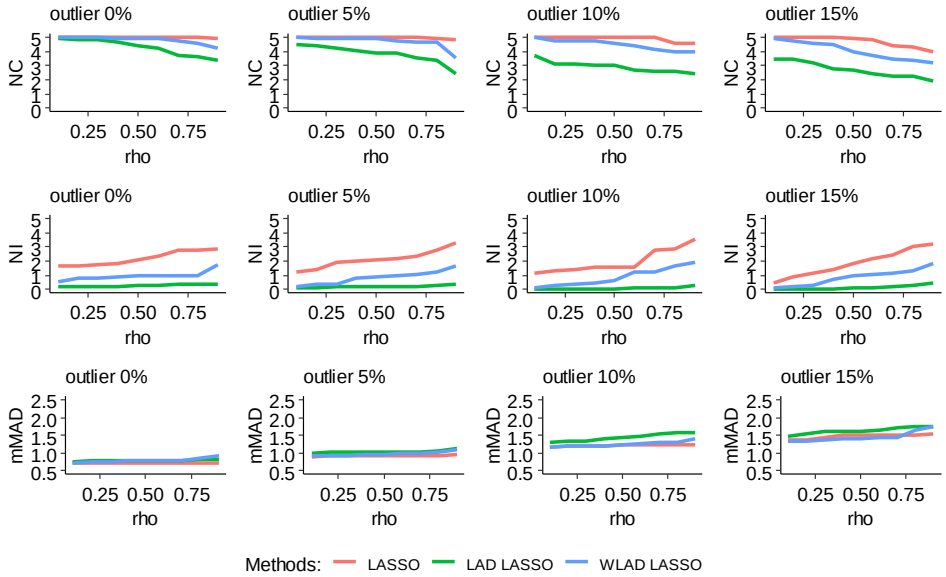
a. Exploration of Sembung leaf extract data
b. The selection of explanatory variables for Sembung leaf extract data uses the best method based on the evaluation of the previous simulation study.
   1) Split the data in to 80% training data and 20% testing data.
   2) Determine the optimum lambda (λ) based on cross validation.
   3) Estimate parameters based on the optimum lambda (λ) obtained.
c. Identify compounds/formulas as antioxidant markers in Sembung leaf extract based on the selected explanatory variables from the method used.

C. RESULT AND DISCUSSION

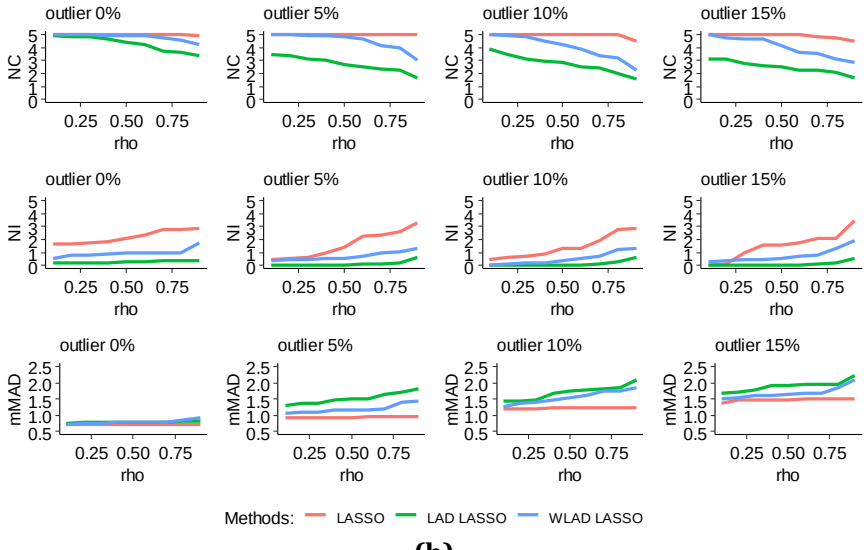
1. Simulation Study of Comparison Robust Methods Performance for Variable Selection

The following figures are the results of LASSO, LAD-LASSO, and WLAD-LASSO simulations on low-dimensional and high-dimensional data. Figure 1 is the result of a simulation on low-dimensional data with a normal distribution error. When there is no outlier (outlier 0%), we can see that LASSO, LAD-LASSO, and WLAD-LASSO performance tend to be similar. The three methods are good in selecting significant variables with low prediction errors (NC tends to be high and mMAD tends to be low at the outlier level 0%). However, when outliers in the response variable exist (outlier 5%, 10%, and 15%) with a normal distribution error (Figure 1a), the significant variables selected for LASSO and WLAD-LASSO tend to be more than LAD-LASSO. The prediction error of WLAD-LASSO and LASSO is lower than LAD-LASSO. The selected

insignificant variable in WLAD-LASSO is less than LASSO and more than LAD-LASSO. These also occur when there are outliers in the response and explanatory variables (Figure 1b). Even the significant variables selected for LAD-LASSO tend to decrease compared to when outliers only on the response variables (Figure 1a). When outliers exist in the response and explanatory variable, LASSO selects insignificant variables more than LAD-LASSO and WLAD-LASSO at the outlier level 5%, 10%, and 15%. The performance of the three methods decreased as the outlier level increased (outlier 5%, 10%, and 15%) on the response and/or the explanatory variables, as shown in Figure 1.



(a)

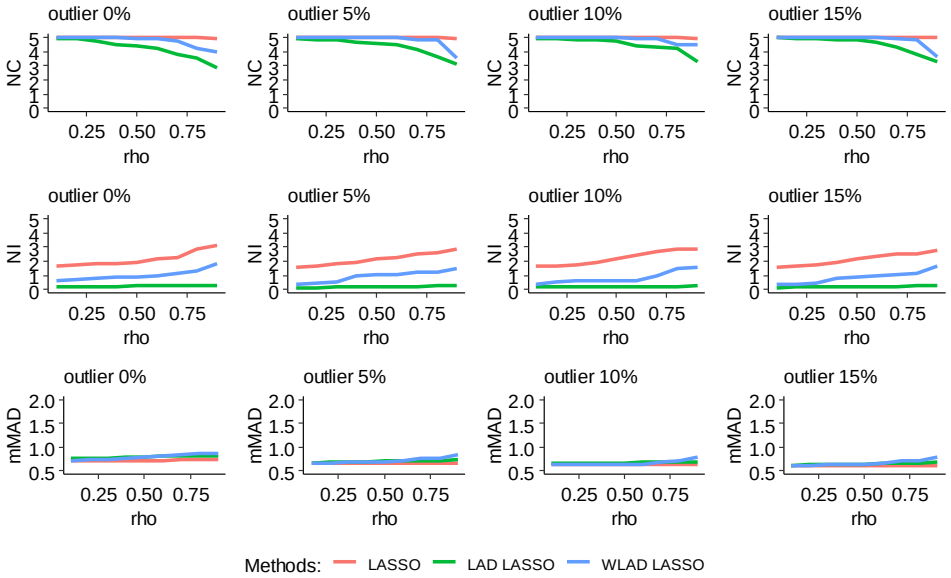


(b)

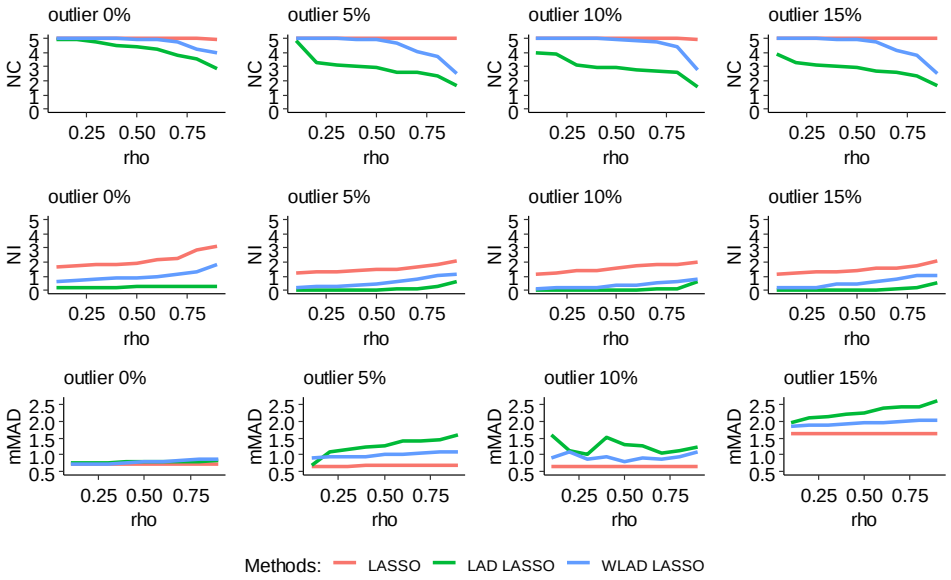
Figure 1. Simulation results $n > p$ $(n = 50, p = 10)$ for normal distribution error which contains outlier in (a) variable Y and (b) variable Y and X

Figure 2 is the result of a simulation on low-dimensional data with an exponential distribution error. When there is no outlier (outlier 0%), LASSO, LAD-LASSO, and WLAD-LASSO are good in selecting significant variables with low prediction errors (NC tends to be high and

mMAD tends to be low at the outlier level 0%). But when outliers exist (outlier 5%, 10%, and 15%) in the response variable (Figure 2a), the significant variables selected for LASSO and WLAD-LASSO tend to be more than LAD-LASSO. The prediction error of WLAD-LASSO and LASSO is lower than LAD-LASSO. The selected insignificant variable in WLAD-LASSO is less than LASSO and more than LAD-LASSO. LAD-LASSO and WLAD-LASSO are based on minimizing the sum of absolute residuals, so they are more robust against outliers in the response variable (H. Wang et al., 2007). Also, they are more robust than LASSO when outliers have an exponential distribution error. These also occur when outliers exist (outlier 5%, 10%, and 15%) in the response and explanatory variables (Figure 2b). Even the weight w_i on WLAD-LASSO makes WLAD-LASSO more robust against outliers in response variables and explanatory variables than LAD-LASSO and LASSO, as shown in Figure 2.



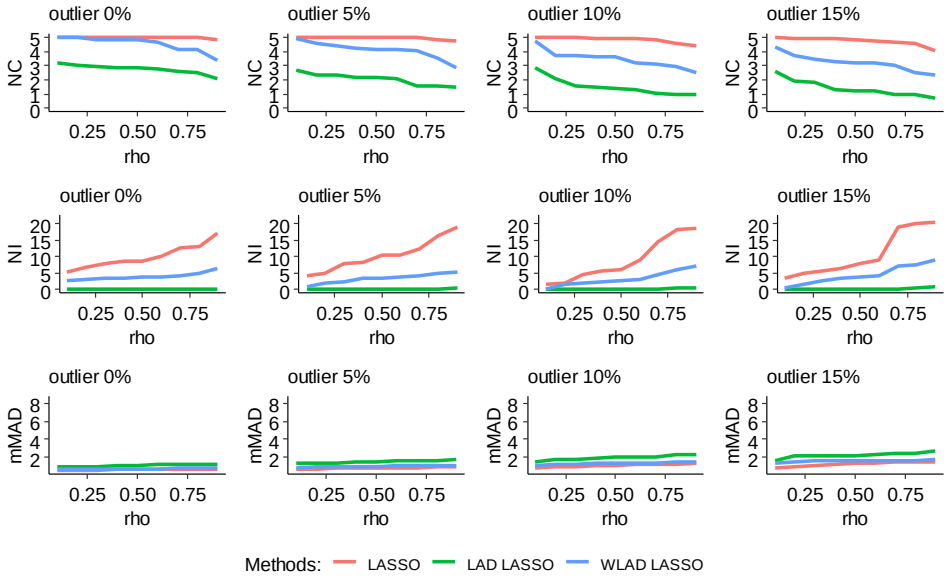
(a)



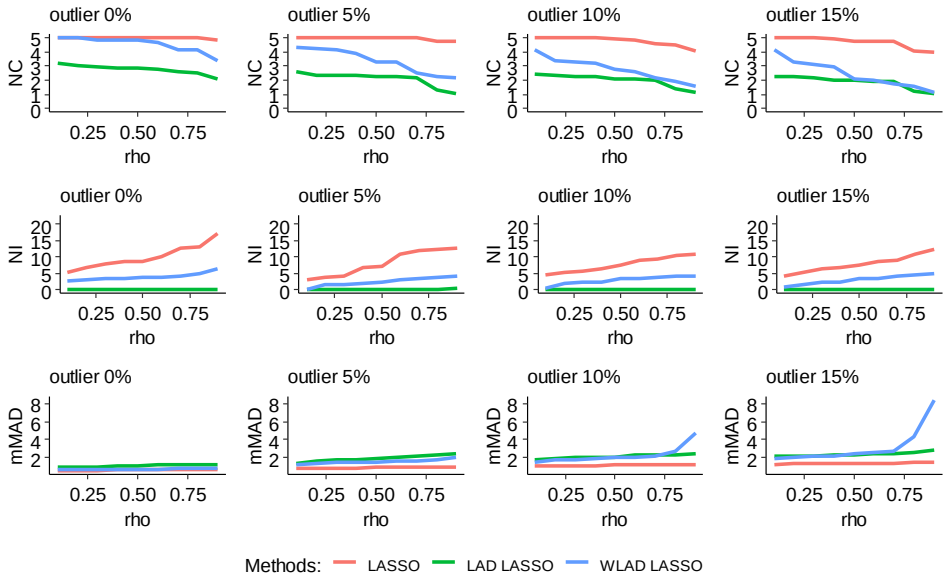
(b)

Figure 2. Simulation results $n > p$ $(n = 50, p = 10)$ for exponential distribution which contains outlier in (a) variable Y and (b) variable Y and X

Figure 3 is the result of a simulation on high-dimensional data with a normal distribution error. Similar to low dimensional data on Figure 1, LASSO, LAD-LASSO, and WLAD-LASSO are good in selecting significant variables with low prediction errors when there is no outlier (NC tends to be high and mMAD tends to be low at the outlier level 0%). But when outliers exist (outlier 5%, 10%, and 15%) in the response variable (Figure 3a), the performance of the three methods decrease as the outlier level increases (NC decreases, while NI and mMAD increases). These also occur when outliers exist (outlier 5%, 10%, and 15%) in the response and explanatory variables (Figure 3b), WLAD-LASSO more robust against outliers in response variables and explanatory variables than LAD-LASSO and LASSO. However, the performance of the three methods decrease compared to low-dimensional data as shown in Figure 3.



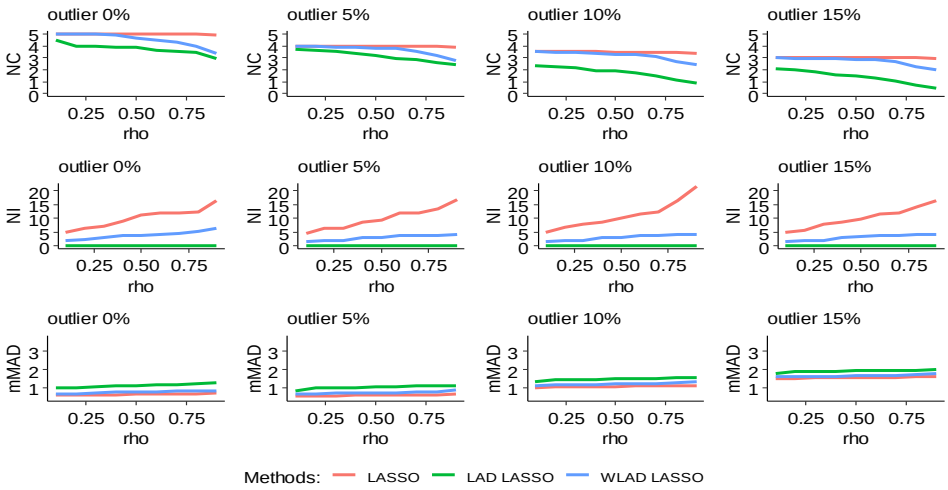
(a)



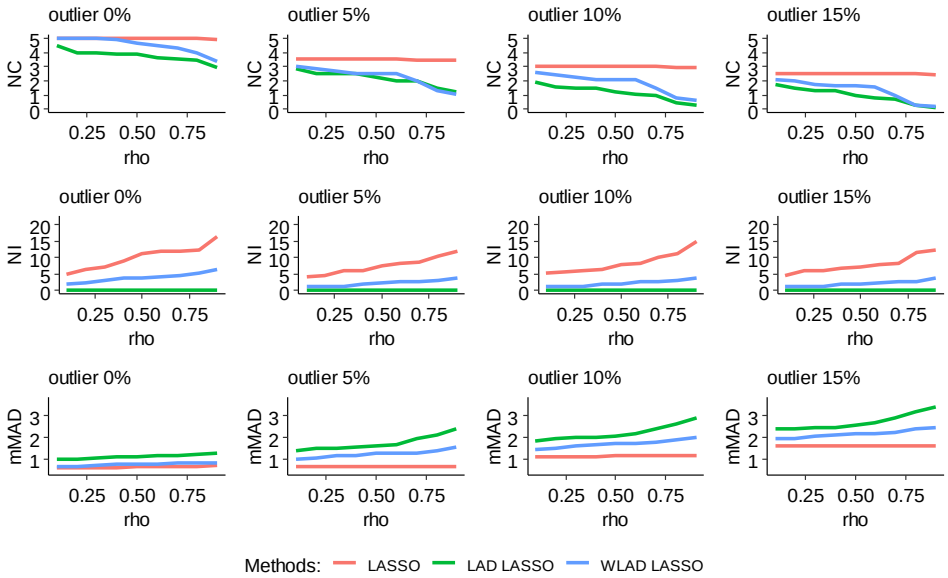
(b)

Figure 3. Simulation results $n < p$ $(n = 50, p = 100)$ for normal distribution which contains outlier in (a) variable Y and (b) variable Y and X

Figure 4 is the simulation result of LASSO, LAD-LASSO, and WLAD-LASSO on high-dimensional data with an exponential distribution error. In this case, these methods are good in selecting significant variables with low prediction errors when there is no outliers (NC tends to be high and mMAD tends to be low at the outlier level 0%). But when outliers exist (outlier 5%, 10%, and 15%) in the response and/or explanatory variables, the performance of the three methods decrease as the outlier level increases (NC decreases, while NI and mMAD increases). Even LASSO tends to select more insignificant variables than LAD-LASSO and WLAD-LASSO. This occurs because LASSO is based on OLS, so the estimation of LASSO parameters is biased when outliers exist or when outliers have an exponential distribution error (LASSO tends to predict 0 coefficients incorrectly). In addition, the MRCD estimator on the weight w_i makes WLAD-LASSO also more robust than LAD-LASSO and LASSO on high-dimensional data (Figure 3 and 4). This estimator aims to regularize the covariance based on the subset that makes overall determinant the smallest (Boudt et al., 2019), as shown in Figure 4.



(a)



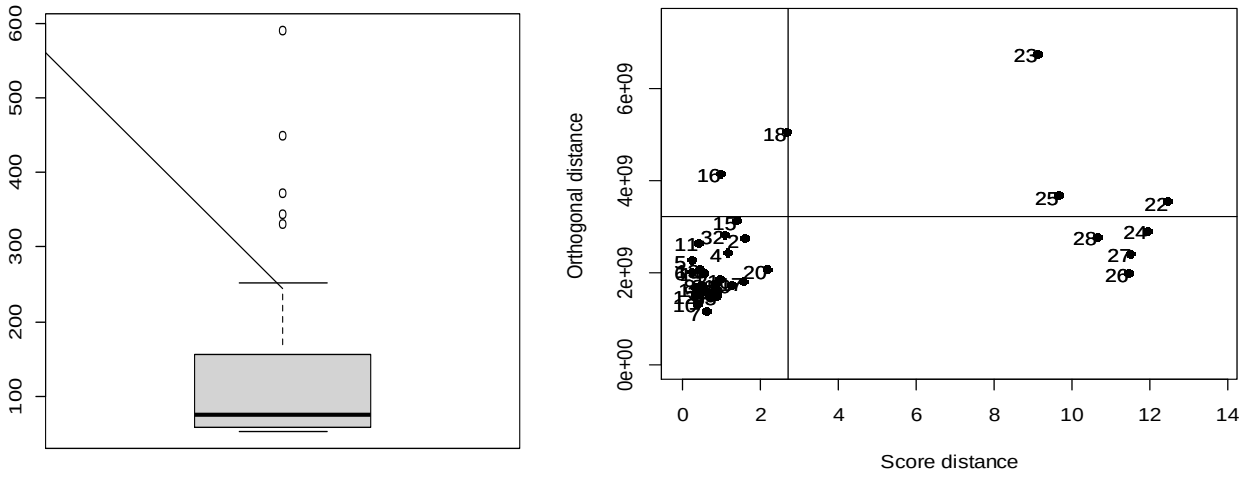
(b)

Figure 4. Simulation results $n < p$ $(n = 50, p = 100)$ for exponential distribution which contains outlier in (a) variable Y and (b) variable Y and X

Furthermore, for the low-dimensional data (Figures 1 and 2) and high-dimensional data (Figures 3 and 4), it can be seen that the performance of LASSO, LAD-LASSO, and WLAD-LASSO decreased when the correlation between explanatory variables (rho) increased. NC indicates these tend to decrease, while NI and mMAD tend to increase when rho increases. It also happens when the outlier level increases. In addition, the performance of the three methods tends to decrease in high-dimensional data compared to low-dimensional data. Then, for low-dimensional and high-dimensional data, LASSO is very loose in selecting variables (significant and insignificant variables are more selected). At the same time, LAD-LASSO is very strict in selecting variables (significant and insignificant variables are slightly selected). In this case, WLAD-LASSO can overcome the weakness of LAD-LASSO which selects very slight significant variables and can overcome the weakness of LASSO which selects very many insignificant variables in the selection of variables. Although the prediction error of LASSO is lower than LAD-LASSO and WLAD-LASSO, the prediction error of WLAD-LASSO is lower than LAD-LASSO.

2. Identification of Antioxidant Marker Compounds in Sembung Leaf Extract Data

Data of Sembung leaf extract contains 35 observations, 2098 explanatory variables as mass-to-charge ratio (m/z) of a compound, and one response variable as the antioxidant of Sembung leaf extract. This data is high dimensional data which causes the covariance matrix of the explanatory variables not to be the full rank matrix. This condition indicates the existence of multicollinearity where the explanatory variables are correlated with each other. Furthermore, about 57% of the explanatory variables weakly correlate negatively with the response variables. Only about 4% of the explanatory variables strongly correlate positively with the response variables. It means that only a few compounds have strong potential as antioxidant markers in Sembung leaf extract. In addition, Figure 5a shows an outlier of about 14.28% in the response variable and Figure 5b shows an outlier of about 25.71% in the explanatory variable (except observations in the 1st quadrant). Detection of outliers on the explanatory variables using the Robust Principal Component Analysis (ROBPCA) method. ROBPCA has been introduced (Hubert et al., 2005; Bulut et al., 2016) and can be applied to both symmetrically distributed data and skewed data (Hubert et al., 2009). This method is a multivariate outlier detection in high-dimensional data, so this method is more efficiently used to detect outliers in many explanatory variables simultaneously than the boxplot (Alaluusua, 2018), as shown in Figure 5.



(a) (b)

Figure 5. Outlier detection (a) Boxplot to detect outliers in response variables (b) Outlier map of ROBPCA to detect outliers in explanatory variables.

Based on the exploration of Sembung leaf extract data, WLAD-LASSO method will be applied to this data. WLAD-LASSO has shown good results in the simulation study of variable selection on high-dimensional data, with an outlier in the response variable and the explanatory variables. This method selected compounds/formulas that are potential antioxidant markers in Sembung leaf extract. From the cross-validation results, the optimum lambda for WLAD-LASSO method is 0.008 with the selected variables shown in Table 1. From Table 1, it can be seen that the compounds/formulas that can be used as antioxidant markers in Sembung leaf extract are Umbelliferone, 12-Hydroxyjasmonic Acid, $C_{22}H_{14}N_8O_2$, and Acetyleugenol. These selected compounds/formulas can be developed as natural antioxidants and have the potential to be developed as medicinal ingredients. Umbelliferone and Acetyleugenol have been known as antioxidant markers. Umbelliferone is a phenol group that shows antioxidant, anti-inflammatory, and anti-hyperglycemic potential (Mazimba, 2017). Acetyleugenol is a derivative of Eugenol which can be used as an antioxidant and anti-inflammatory agent (Leem et al., 2011), as shown in Table 1.

Table 1. The results of selected variables from the WLAD-LASSO method

No	Selected Explanatory Variable	Name of Compound/Formula	MAD
1	X1	Umbelliferone (7-hydroxycoumarin) / $C_9H_6O_3$	0.133050
2	X6	Umbelliferone (7-hydroxycoumarin) / $C_9H_6O_3$	
3	X193	12-Hydroxyjasmonic Acid / $C_{12}H_{18}O_4$	
4	X237	-	
5	X417	$C_{22}H_{14}N_8O_2$	
6	X758	Acetyleugenol / $C_{12}H_{14}O_3$	

D. CONCLUSION AND SUGGESTIONS

Based on the research that has been done, it can be concluded that performance of LASSO, LAD-LASSO, and WLAD-LASSO decreased when the correlation between explanatory variables increased. It also happens when the outlier level increases. In addition, their performances tend to decrease in high-dimensional data compared to low-dimensional data. Then, for low-dimensional and high-dimensional data, LASSO is very loose in selecting variables (significant and insignificant variables are more selected). At the same time, LAD-LASSO is very strict in

selecting variables (significant and insignificant variables are slightly selected). WLAD-LASSO can overcome the weakness of LAD-LASSO, which selects very slight significant variables and can overcome the weakness of LASSO, which selects many insignificant variables in the selection of variables. Although the prediction error of LASSO is lower than LAD-LASSO and WLAD-LASSO, the prediction error of WLAD-LASSO is lower than LAD-LASSO. Because of these results, WLAD-LASSO has been applied to Sembung leaf extract data to identify compounds/formulas as antioxidant markers in Sembung leaf extract. Then, we obtained that the compounds/formulas as antioxidant markers in the Sembung leaf extract are Umbelliferone, 12-Hydroxyjasmonic Acid, $C_{22}H_{14}N_8O_2$, and Acetyleugenol (with a prediction error is 0.133050). For further research, other variable selection methods can be used such as combining the Least Trimmed Square (LTS) weighting with the Elastic Net method. The LTS weighting is based on the residual squared and the absolute residual, so that it can overcome the weakness of OLS based on the residual square and LAD and WLAD based on absolute residual in overcoming outliers. The Elastic Net is also a combination of the Ridge and LASSO methods.

ACKNOWLEDGEMENT

This research was supported by The Ministry of Research and Technology/The National Research and Innovation Agency-Republic of Indonesia under Award Number 1/E1/KP.PTNBH/2021. All content in this research are full authors's responsibility and do not represent The National Research and Innovation Agency's official views.

REFERENCES

Ajeel, S. M., & Hashem, H. A. (2020). Comparison Some Robust Regularization Methods in Linear Regression via Simulation Study. *Academic Journal of Nawroz University*, *9*(2), 244–252. https://doi.org/10.25007/ajnu.v9n2a818

Alaluusua, K. (2018). Outlier detection using robust PCA methods. *Bachelor's Thesis, Aalto University*. https://doi.org/10.13140/RG.2.2.17736.88321

Arslan, O. (2012). Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Computational Statistics and Data Analysis*, *56*(6), 1952–1965. https://doi.org/10.1016/j.csda.2011.11.022

Bangdiwala, S. I. (2018). Regression: multiple linear. *International Journal of Injury Control and Safety Promotion*, *25*(2), 232–236. https://doi.org/10.1080/17457300.2018.1452336

Boudt, K., Rousseeuw, P. J., Vanduffel, S., & Verdonck, T. (2019). The minimum regularized covariance determinant estimator. *Statistics and Computing*, *30*(1), 113–128. https://doi.org/10.1007/s11222-019-09869-x

Bulut, H. (2020). Mahalanobis distance based on minimum regularized covariance determinant estimators for high dimensional data. *Communications in Statistics - Theory and Methods*, *49*(24), 5897–5907. https://doi.org/10.1080/03610926.2020.1719420

Bulut, H., Öner, Y., & Sözen, Ç. (2016). A Proposal for Robpca Algorithm International Journal of Sciences : A Proposal for Robpca Algorithm. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, *29*(2), 119–129. https://www.gssrr.org/index.php/JournalOfBasicAndApplied/article/view/6131

Camponovo, L. (2022). Extended Oracle Properties of Adaptive Lasso Estimators. *Open Journal of Statistics*, *12*(2), 210–215. https://doi.org/10.4236/ojs.2022.122015

Daoud, J. I. (2017). Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series*, *949*(1), 012009. https://doi.org/10.1088/1742-6596/949/1/012009

Dielman, T. E. (2005). Least absolute value regression : recent contributions. *Journal of Statistical Computation and Simulation*, *75*(4), 263–286. https://doi.org/10.1080/00949650420002236 80

Giloni, A., Simonoff, J. S., & Sengupta, B. (2006). Robust weighted LAD regression. *Computational*

Statistics and Data Analysis, *50*(11), 3124–3140. https://doi.org/10.1016/j.csda.2005.06.005

Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, *47*(1), 64–79. https://doi.org/10.1198/004017004000000563

Hubert, M., Rousseeuw, P., & Verdonck, T. (2009). Robust PCA for skewed data and its outlier map. *Computational Statistics and Data Analysis*, *53*(6), 2264–2274. https://doi.org/10.1016/j.csda.2008.05.027

Keith, T. Z. (2015). *Multiple Regression and Beyond: An Introduction to Multiple Regression and Structural Equation Modeling 2nd Edition*. New York:Taylor & Francis.

Leem, H. H., Kim, E. O., Seo, M. J., & Choi, S. W. (2011). Antioxidant and Anti-Inflammatory Activities of Eugenol and Its Derivatives from Clove (Eugenia caryophyllata Thunb.). *Journal Korean Social Food Science Nutrition*, *40*(10), 1361–1370. https://doi.org/10.3746/jkfn.2011.40.10.1361

Lima, E., Davies, P., Kaler, J., Lovatt, F., & Green, M. (2020). Variable selection for inferential models with relatively high-dimensional data: Between method heterogeneity and covariate stability as adjuncts to robust selection. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-64829-0

Machkour, J., Muma, M., Alt, B., & Zoubir, A. M. (2020). A robust adaptive Lasso estimator for the independent contamination model. *Signal Processing*, *174*(2), 1649–1653. https://doi.org/10.1016/j.sigpro.2020.107608

Mazimba, O. (2017). Umbelliferone: Sources, chemistry and bioactivities review. *Bulletin of Faculty of Pharmacy, Cairo University*, *55*(2), 223–232. https://doi.org/10.1016/j.bfopcu.2017.05.001

Rahardiantoro, S., & Kurnia, A. (2015). LAD-LASSO : Simulation Study of Robust Regression in High Dimensional Data. *Indonesian Journal of Statistics*, *18*(2), 105–107. https://journal.ipb.ac.id/index.php/statistika/article/view/16775

Sirimongkolkasem, T., & Drikvandi, R. (2019). On Regularisation Methods for Analysis of High Dimensional Data. *Annals of Data Science*, *6*(4), 737–763. https://doi.org/10.1007/s40745-019-00209-4

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, *7*(1), 1456–1490. https://doi.org/10.1214/13-EJS815

Varin, S. (2021). Comparing the Predictive Performance of OLS and 7 Robust Linear Regression Estimators on a Real and Simulated Datasets. *International Journal of Engineering Applied Sciences and Technology*, *5*(11), 9–23. https://doi.org/10.33564/ijeast.2021.v05i11.002

Wahid, A., Khan, D. M., & Hussain, I. (2017). Robust Adaptive Lasso method for parameter's estimation and variable selection in high-dimensional sparse models. *PLoS ONE*, *12*(8), 1–17. https://doi.org/10.1371/journal.pone.0183518

Wang, H., Li, G., & Jiang, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso. *Journal of Business and Economic Statistics*, *25*(3), 347–355. https://doi.org/10.1198/073500106000000251

Wang, L. (2013). The L1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, *120*(2013), 135–151. https://doi.org/10.1016/j.jmva.2013.04.001

Wasserman, L., & Roeder, K. (2009). High-Dimensional Variable Selection. *Annals of Statistics*, *37*(5A), 2178–2201. https://doi.org/10.1214/08-AOS646

Yang, H., & Li, N. (2018). WLAD-LASSO method for robust estimation and variable selection in Partially Linear Models. *Communications in Statistics - Theory and Methods*, *47*(20), 4958–4976. https://doi.org/10.1080/03610926.2017.1383427

Zhao, N., Xu, Q., Tang, M. L., Jiang, B., Chen, Z., & Wang, H. (2020). High-Dimensional Variable Screening under Multicollinearity. *Stat*, *9*(1), 1–14. https://doi.org/10.1002/sta4.272

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735