

Study on the performance of Robust LASSO in determining important variables data with outliers

ROCHYATI^{1*}, KUSMAN SADIK¹, BAGUS SARTONO¹, EVITA PURNANINGRUM²

¹Department of Statistics, IPB University, Bogor, Indonesia

²Department of Management, PGRI Adi Buana University, Surabaya, Indonesia

Abstract. A variable selection method is required to deal with regression models with many variables, and LASSO has been the most widely used methodology. However, as several authors have noted, LASSO is sensitive to outliers in the data. For this reason, the Robust-LASSO approach was introduced by applying some weighting schemes for each sample in the data. This research presented a comparative study of the three weighting schemes in Robust LASSO, namely Huber-LASSO, Tukey-LASSO, and Welsch-LASSO. The study did a rich simulation containing many scenarios with various characteristics on the covariance structures of the explanatory variable, the types of outliers, the number of outliers, the location of active variables, and the number of variables. The study then found that Tukey-LASSO outperformed Huber-LASSO and Welsch-LASSO in identifying significant variables. The Robust LASSO performance generally decreased as the covariances among explanatory variables increased and the data dimension increased. Exploration of sembung leaf extract data shows that the data is high dimensional data which contains outliers of about 14,28% on the response variable and about 25,71% on the explanatory variables. Based on the research, the number of variables selected using the Tukey-LASSO method was nine compounds, Huber-LASSO and Welsch-LASSO were eight compounds, and LASSO 13 compounds. The Tukey-LASSO prediction accuracy is superior to the other three methods.

Keywords: high dimensional regression, Huber, Tukey, variable selection, Welsch

INTRODUCTION

Regression analysis is a statistical analysis that aims to investigate the relationship between the response variable and the explanatory variable. The Least Square Method (LSM) is usually used to estimate the regression model parameters by minimizing the sum of the residual squares [1]. Under certain conditions, the least squares method could be more satisfactory. One of the things that causes the least squares method to be unsatisfactory is if the data has a vast number of explanatory variables, even far exceeding the observations or also known as high dimensional data. High-dimensional data appears when the object of observation is rare or requires high costs. Some examples include gene expression data and spectroscopic data. In data like this, the problem of double collinearity between explanatory variables cannot be avoided. The estimation results have a significant variance if the least squares estimation method is used. As a result, the estimation results tend to be unstable, and the prediction model needs to be

more accurate. The more explanatory variables used, the more complex the model is to interpret, so it is necessary to select the variables [2].

LASSO (Least Absolute Shrinkage and Selection Operator) and its derivatives have been widely considered for high-dimensional data analysis [3]. LASSO is a shrinking technique for several regression coefficients close to zero to precisely zero so that it can automatically select the regression variables. LASSO works by adding constraints to the least squares method. However, as several authors have noted, LASSO is sensitive to outliers. Therefore, it is necessary to have a robust LASSO method.

LAD LASSO is a robust LASSO method that combines LAD criteria with LASSO penalties [4]. However, this method has lower efficiency than OLS when there are no outliers in the response variables [5]. Then the Huber-LASSO method was developed, which is a robust LASSO method by adding the Huber criteria. This method is superior to LASSO and LAD-LASSO [5]. However, Huber-LASSO is only robust against outliers in the response. Therefore, in 2018, [5] introduced the robust LASSO method with the Tukey double weight criterion called Tukey-LASSO, where the method is a LASSO method that is robust not only on the response variable but also on the explanatory variable.

*Corresponding Author:
bagusco@apps.ipb.ac.id

Received: 8 June 2022 | Revised: 4 September 2022 |
Accepted: 6 January 2023

Therefore, this research aims to 1) examine the Robust LASSO method in identifying significant variables in simulation data containing various characteristics of covariance, types of outliers, number of outliers, location of active variables, and number of variables and 2) assess the application of the best method obtained from point 1) on the Sembung leaf extract data.

METHODOLOGY

Data

In this study, simulated and actual data were used. The simulation data in this study were analyzed by the LASSO and Robust LASSO methods (Huber-LASSO, Tukey-LASSO, and Welsch-LASSO). Then the number of selected variables (variable selection), the number of active variables identified correctly (number correct), the number of noise variables correctly identified (number incorrect), the percentage of occurrences when the model contains the correctly fitted variables only, and the average Mean Square Prediction Error (MSPE) were seen. Meanwhile, the simulation data used were low and high-dimensional data, with $p = 10$; 200 consisted of five informative variables, which had ordered and unordered indices, and the rest were zero variables. Based on previous research, we choose the number of observations $n = 100$, the proportion of outliers = 0% (without outliers); 3%; 6%; 9%; 12%; 15%, and the correlation level $\rho = 0.1$; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9. In addition, there were two types of outliers: outliers that were only on the response variable and outliers that were not only on the response variable but also on the explanatory variable. Meanwhile, the repetition was done 1000 times.

Meanwhile, the actual data in this study were obtained from experimental results by the team of the Center for Biopharmaceutical Studies, LPPM, IPB, in collaboration with the Statistics Department of IPB in May-June 2021. These data resulted from LC-MS (Liquid Chromatograph-Mass Spectrometry) analysis to see which compounds significantly affected the formation of antioxidants in Sembung leaf extract. The data consisted of 35 observations and 2,098 explanatory variables.

In this study, the explanatory variable (X) was mass spectrometry data, and the response variable (y) was the level of the antioxidant content of Sembung leaves. Then, the data also had additional information regarding molecular mass, formula, and compound predictions but should have been included in the research analysis.

Data Analysis Procedure

Simulation Study

The simulations were carried out to examine the effect of the contamination level of outliers and several types

of correlation in the Robust LASSO model. The steps carried out in the simulation study are as follows:

1. Setting the number of observations n as much as 100 and the number of explanatory variables p as much as 10
2. Generating a value on the explanatory variable x through a multivariate normal distribution with a mean of 0 dan $\text{cov}(x_{ij}, x_{il}) = \rho^{|j-l|}$, where $j, l = 1, 2, \dots, p$ and $i = 1, 2, \dots, n$ as well as ρ determined by 0.1, 0.2, 0.3, 0.4 0.5, 0.6, 0.7, 0.8 and 0.9.
3. Deciding which model to use, i.e., $y_i = x_{1i} + x_{2i} + x_{3i} + x_{4i} + x_{5i}$ or determining the parameters for the population in the regression model with

$$\beta_t = \begin{cases} 1, & t = 1, 2, \dots, 5 \\ 0, & \text{others} \end{cases}$$
4. Generating the response variable y with three scenarios:
 - a. Scenario 1: response variable data y without outliers in the form of $y_i = x_i \beta + \varepsilon_i$ where $\varepsilon_i \sim N(0, 1)$.
 - b. Scenario 2: response variable data y containing outliers in the form $y_i = x_i \beta + \varepsilon_i$, with spread error $(1 - k)N(0, 1) + kN(3, 1)$, where k is the percentage of outliers in the specified data of 3%, 6%, 9%, 12%, and 15%
 - c. Scenario 3: data with outliers on the response variable y and the explanatory variable X
 - 1) Replacing ten rows in the explanatory variable data with values derived from the distribution $N(10, 1)$. Next, the X matrix is denoted by X^* .
 - 2) Response variable data containing outliers in the form of $Y = X\beta + \varepsilon$, with spread error $(1 - k)N(0, 1) + kN(3, 1)$, where k is the percentage of outliers in the specified data of 3%, 6%, 9%, 12%, and 15%
5. Processing data using LASSO, Huber-LASSO, Tukey-LASSO, and Welsch-LASSO methods to obtain estimators for parameters β .

$$\begin{aligned} \beta_{\text{robust-LASSO}} &= \arg \min_{\beta} \left[\sum_{i=1}^n \rho \left(\frac{y_i - X_i^T \beta}{\sigma} \right)^2 + \lambda \sum_{j=1}^p \hat{\omega} |\beta_j| \right] \end{aligned}$$

where ρ is the weight of the Huber, Tukey, or Welsch methods. The data analysis was then followed by calculating some statistics, including the number of the explanatory variables that were correctly identified as active variables (no correct), the number of the explanatory variables that were incorrectly identified as active variables (no incorrect), and the average MSPE.

$$MSPE = n^{-1} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

6. Select the best parameter estimation method in the model based on the values of no correct, no incorrect, and correctly fitted.
7. Repeating the process 1 to 6 with the number of observations $n = 100$ and the number of explanatory variables $p = 200$

Applied Studies

This study was conducted by applying the best Robust LASSO estimator method based on the simulation study results on actual data. Variables affecting the percentage of Sembung plant content were also identified in the applied study. The stages carried out in this study are as follows:

1. Doing data preparation
 2. Identifying the response variables and explanatory variables used
 3. Doing data exploration
 4. Standardizing the variable so that it has a mean of 0 and a variance of 1
 5. Dividing the data into 80% training data and 20% testing data
 6. Determining the optimal lambda value for each method. The optimal lambda is obtained based on the Minimum Mean Squared Error Cross Validation (MSE CV). MSE CV is calculated based on the following equation:
- $$MSE\ CV = \frac{1}{k} \sum_{k=1}^K (y_i - \hat{y}_{-k}(x_i))^2$$
7. Applying the Robust LASSO method using the MTE package on the R software introduced by [6]
 8. Identifying which compounds have a significant effect on the formation of antioxidants in Sembung plants

RESULTS AND DISCUSSION

Scenario Study on Simulation Data

This paper complements what [5] did. This study compares the Lasso, Adaptive Lasso, Huber-LASSO, sparse LTS, Rlars, and Tukey-LASSO methods with $\rho = 0.5$, and the percentage of outliers is 10%. In contrast, this study compares Huber-LASSO, Tukey-LASSO, and Welsch-LASSO with various outliers and characteristics. Multiple degrees of correlation between covariate variables. In this study, it is not only using the first five variables that are adjacent but also jumping. In addition, this study uses Matlab software for the Tukey-LASSO method and R software for all forms.

This study examined the performance of Robust LASSO for selecting variables on data containing outliers. Meanwhile, the percentages of outliers used were 0%, 3%, 6%, 12%, and 15%, applied to data with 100 observations, and the number of explanatory variables used in this simulation was 10 and 200. On the other hand, the location of the outliers was not only on the response variable but also on the explanatory variable. The outlier used in the explanatory variable was the bad leverage point outlier. Then, in the simulation, active explanatory variables were also set, which had sequential indices with one another. The design employed in this scenario consisted of five explanatory variables set as active explanatory variables. The model used in the simulation process was $y_i = x_{1i} + x_{2i} + x_{3i} + x_{4i} + x_{5i} + \varepsilon_i$. Each simulation scenario was analyzed using the Huber-LASSO, Tukey-LASSO, Welsch-LASSO, and LASSO methods with 1000 replications. Then, the performance was evaluated by looking at the number of active variables correctly identified (number correct), the number of noise variables correctly identified (number incorrect),

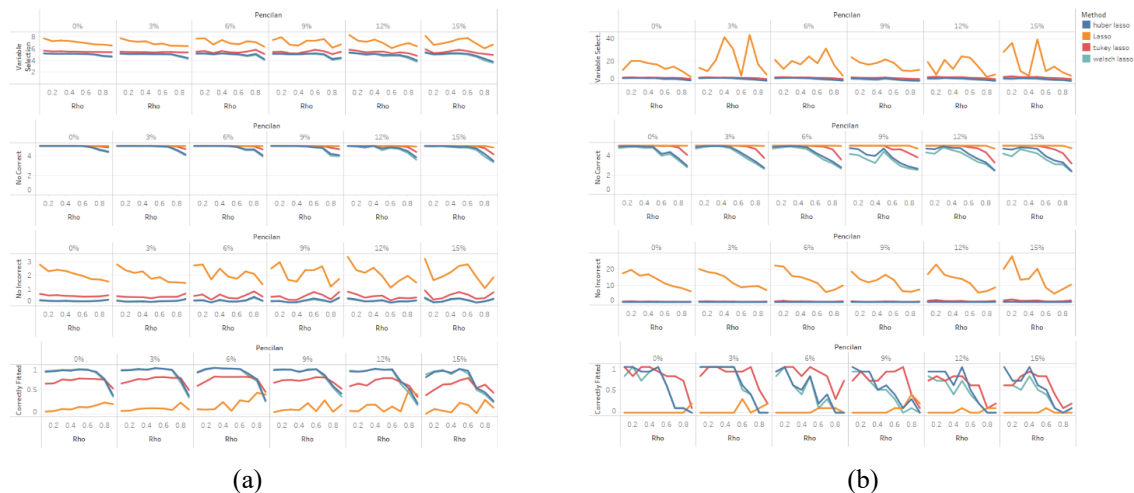


Figure 1. Performance of four simulation methods for selecting variables for outlier scenarios only on response variables and active explanatory variables having sequential indices, for (a) $p=10$; $n=100$, (b) $p=200$; $n=100$.

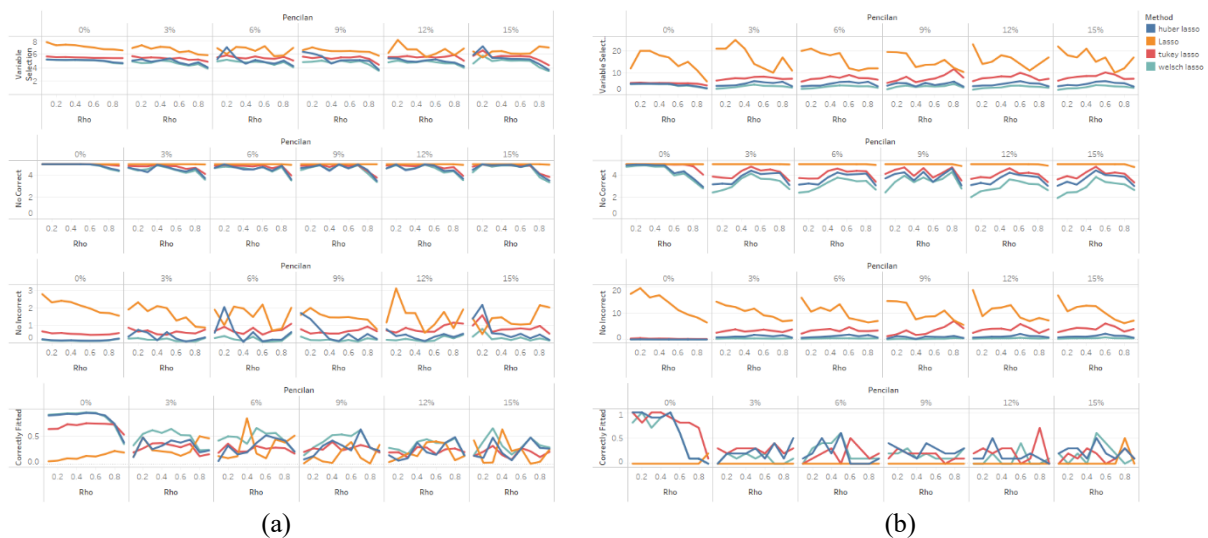


Figure 2. The performance of the four simulation methods of selecting variables for the outlier scenario on the response variable and the explanatory variable, and the active explanatory variable having sequential indices, for (a) $p=10$; $n=100$, (b) $p=200$; $n=100$.

and the percentage of occurrences when the model contained only the correctly fitted variables.

Figure 1(a) shows the performance plot of the four simulation methods of selecting variables for the outlier scenario only on the response variable and the active explanatory variable having sequential indices, for the number of explanatory variables of 10 and the number of observations of 100. From Figure 1(a), it can be seen that The number of active variables was correctly identified, which means that from 5 active variables, how many were identified as active variables (number correct) in 1000 repetitions of the Huber-LASSO, Tukey-LASSO, and Welsch-LASSO methods was almost the same and decreased along with the increasing correlation between explanatory variables. Meanwhile, the number of noise variables was correctly identified, which means that from the remaining $p-5$ variables, how many were identified that he identified as zero (number incorrect), the Huber-LASSO, Tukey-LASSO, and Welsch-LASSO methods looked almost the same but not on LASSO. The LASSO method tended to have more incorrect values than the other three methods. It indicates that LASSO was not very selective, so it tended to choose many unimportant variables to be included in the model, as shown in the variable selection graph. In LASSO, the selection of variables also tended to be loose because the selected variables were many but wrong.

Meanwhile, the Huber-LASSO, Tukey-LASSO, and Welsch-LASSO methods had reasonable solutions, not too loose and not too tight, since they tended to be careful in selecting variables. Still, the essential variables sometimes needed to be selected. Then, for the percentage of events when the model contained only the correctly fitted variables, it can be seen that the percentage of Robust LASSO tended to be superior

to LASSO. It shows that Robust LASSO was better than LASSO in the correctly fitted.

Figure 1(b) depicts the performance plot of the four simulation variables selection methods for the outlier scenario only on the response variable and the active explanatory variable that had sequential indices, for the number of explanatory variables of 200 and the number of observations of 100. From Figure 1(b), it can be seen that the number correct for 1000 repetitions of the Huber-LASSO, Tukey-LASSO, and Welsch-LASSO methods was almost the same and decreased along with the increasing correlation between explanatory variables. Meanwhile, for number incorrect, it was more apparent that the Huber-LASSO, Tukey-LASSO, and Welsch-LASSO methods looked the same and were more robust than the LASSO method, which had a selection of variables that tended to choose many variables but was wrong. For correctly fitted, LASSO methods tended to approach zero, and Robust LASSO methods decreased as the correlation between explanatory variables increased. It indicates that the Robust LASSO methods were better than the LASSO method. Among the Robust LASSO methods, the Tukey-LASSO method tended to be superior to other Robust LASSO methods.

Next, a simulation was carried out where the number of explanatory variables was 10, and the number of observations was 100; outliers were not only on the response variable but also on the explanatory variable; the active explanatory variable had a sequential index. The simulation results can be seen in Figure 2(a). Figure 2(a) displays that the existence of an explanatory variable outlier greatly affected all methods.

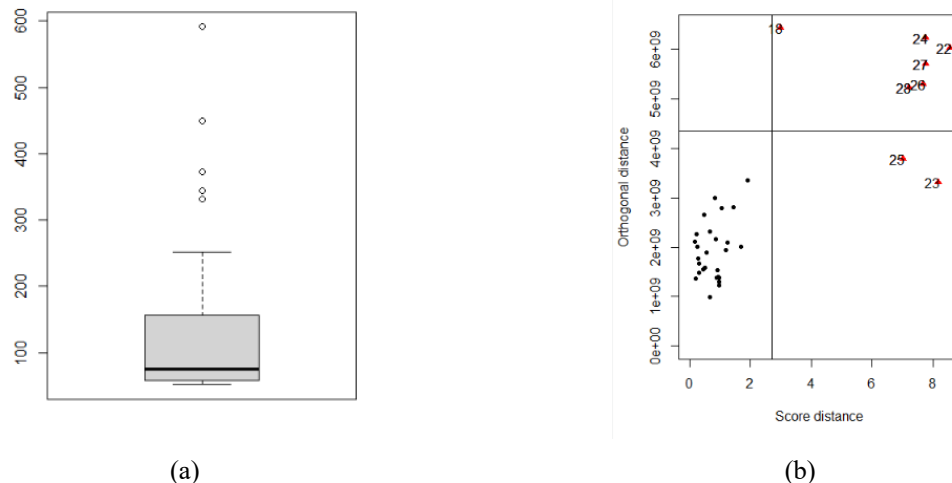


Figure 3. Detection of outliers in Sembung leaf extract data in (a) the response variables, (b) the explanatory variable.

in the explanatory variable greatly affected the four methods. From Figure 2(b), the value of the number correct for 1000 repetitions of the Tukey-LASSO method tended to be superior to the Huber-LASSO and Welsch-LASSO methods. Meanwhile, for the number incorrect, it can be seen that the Huber-LASSO, Tukey-LASSO, and Welsch-LASSO methods were the same and worked well when there were no outliers.

Meanwhile, when there were outliers, the Huber-LASSO and Welsch-LASSO methods tended to be better than Tukey-LASSO. However, the three methods were still better than the LASSO method. It signifies that the Robust LASSO method was superior to the LASSO method in selecting variables. In addition, in Figure 2(b), it is also seen that the correctly fitted LASSO method also tended to approach zero. It indicates that the LASSO method was not better than the Robust LASSO method. In the Robust LASSO methods, Tukey-LASSO tended to be superior to the other two methods when the correlation between explanatory variables was higher.

Study on Actual Data

Before analyzing the data on Sembung leaf extract using the explanatory variable, namely mass spectrometry, and the response variable, namely antioxidant, outliers were detected employing the ROBPCA method. The ROBPCA method is a method for detecting and classifying outliers [7]. ROBPCA produces a diagnostic plot that displays and classifies the outliers in high-dimensional data. Outliers were detected from the data with 35 observations and 2098 explanatory variables, as shown in Figure 3. The boxplot in Figure 3(a) illustrates an outlier in the response variables, with the proportion of outliers on the response variable being around 14.28%. Figure 3(b) illustrates an outlier in the explanatory variables,

with the proportion of outliers on the explanatory variable being around 25.71%.

Meanwhile, in ROBPCA, it can be seen that the Orthogonal Distance (OD) value was significant, while the Score Distance (SD) value was small, smaller than the cut-off point of 2.716203. Thus, it was concluded that no observations were classified as orthogonal outliers. Numbers 18, 22, 24, 26, 27, and 28 were lousy leverage points because OD and SD were also large. Meanwhile, numbers 25 and 23 were observations classified as good leverage points since OD was small, but SD was large.

The simulation results show that for selecting variables in high-dimensional data and data containing outliers, the Robust LASSO method was quite good compared to LASSO. Then, the best Robust LASSO method was applied, Tukey-LASSO, in this case, on Sembung leaf extract data. The first thing to do was find the optimum lambda with 10-folds for the Huber-LASSO, Tukey-LASSO, Welsch-LASSO, and LASSO, respectively, 0.002, 0.005, 0.002, and 0.06. With the optimum lambda, further analysis of the Sembung leaf extract data was carried out with the results presented in Table 1.

Based on Table 1, it can be seen that the number of selected variables in the Huber-LASSO and Welsch-LASSO methods is the same. In addition, it can be seen that the LASSO method has the most significant number of selected variables and the MSPE among the other three methods. Umberliferone (7-hydroxycoumarin) is a phenol compound with antioxidant, anti-inflammatory, and antihyperglycemic potential [8]. In addition, there is the compound Quercetin, one of the most efficient flavonoid antioxidants [9]. This is under research conducted by [10], which said that the compounds contained in the

Table 1. The results of the method

Method	Variable Selection	Compound/Formula	MSPE
Tukey-LASSO	9	Umberlliferone (7-hydroxycoumarin)/C ₉ H ₆ O ₃ , Quercetin/C ₁₅ H ₁₀ O ₇ , Austroinulin/C ₂₀ H ₃₄ O ₃ , Gurjunene/C ₁₅ H ₂₄ , C ₅ H ₁₁ NO ₂ , C ₇ H ₇ NO ₂ , C ₈ H ₁₆ N ₂ O ₉ , C ₁₂ H ₁₄ O ₃	23.907
Huber-LASSO	8	Umberlliferone (7-hydroxycoumarin)/C ₉ H ₆ O ₃ , Aromadendrene oxide/C ₁₅ H ₂₄ O, Gurjunene/C ₁₅ H ₂₄ , C ₅ H ₁₁ NO ₂ , C ₇ H ₇ NO ₂ , C ₈ H ₁₆ N ₂ O ₉ , C ₁₂ H ₁₄ O ₃	23.915
Welsch-LASSO	8	Umberlliferone (7-hydroxycoumarin)/C ₉ H ₆ O ₃ , Aromadendrene oxide/C ₁₅ H ₂₄ O, Gurjunene/C ₁₅ H ₂₄ , C ₅ H ₁₁ NO ₂ , C ₇ H ₇ NO ₂ , C ₈ H ₁₆ N ₂ O ₉ , C ₁₂ H ₁₄ O ₃	23.919
LASSO	13	C ₈ H ₁₆ N ₂ O ₉ , C ₁₂ H ₁₄ O ₃ , C ₁₂ H ₁₆ O ₃ , C ₂₂ H ₁₁ NO ₁₀ , C ₁₃ H ₂₀ N ₈ O ₅ , C ₁₄ H ₆ O ₄ , C ₂₀ H ₂₈ O ₄ , C ₈ H ₁₅ NO ₄	24.447

leaves of Sembung, namely flavonoids. Flavonoid compounds belong to the class of polyphenolic compounds that have been proven to have antioxidant activity.

CONCLUSION

In low-dimensional and high-dimensional data, the LASSO method tends to be loose in selecting variables, so many chosen variables need to be corrected. In contrast, the Huber-LASSO, Tukey-LASSO, and Welsch-LASSO methods tend to be careful in selecting variables. Still, essential variables sometimes need to be selected. Among the Huber-LASSO, Tukey-LASSO, Welsch-LASSO, and LASSO methods, Tukey-LASSO outperforms the three methods regarding variable selection and prediction accuracy. This is indicated by the value of the model selection criteria, which is better, and the error is smaller than the other methods. Furthermore, the method was applied to Sembung leaf extract data to identify compounds/formulas as antioxidant markers in Sembung leaf extract. Based on the analysis of the data, it was found that outliers were found in the data. Of the four methods applied, the best prediction accuracy is obtained by the Tukey-LASSO method because it has a smaller error than the other three methods.

ACKNOWLEDGMENT

Research reported in this publication was supported by the Ministry of Research and Technology/National Agency of Research and Innovation - the Republic of Indonesia under Award Number 1/E1/KP.PTNBH/2021. The content is solely the authors' responsibility and does not necessarily represent the official views of the National Agency of Research and Innovation. I would also like to thank Indonesia Endowment Fund for Education (LPDP)

from the Ministry of Finance Republic Indonesia for granting the scholarship and supporting this research.

REFERENCES

- [1] Varin S. 2021. Comparing the predictive performance of OLS and 7 robust linear regression estimators on a real and simulated datasets. *Int. J. Eng. Appl. Sci. Technol.* **5** (11) 9-23. DOI:10.33564/ijeast.2021.v05i11.002.
- [2] 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. *Sci. Rep.* **10** (1) 8002. DOI:10.1038/s41598-020-64829-0.
- [3] Kim, Y.; Hao, J.; Mallavarapu, T.; Park, J.; Kang, M. 2019. Hi-LASSO: High-Dimensional LASSO. *IEEE Access* **7** 44562–44573. DOI:10.1109/access.2019.2909071
- [4] Li, Q.; Wang, L. 2020. Robust change point detection method via adaptive LAD-LASSO. *Stat Pap.* **61** (1). DOI:10.1007/s00362-017-0927-3.
- [5] Chang, L.; Roberts, S.; Welsh, A. 2018. Robust lasso regression using Tukey's biweight criterion. *Technometrics* **60** (1) 36–47. DOI:10.1080/00401706.2017.1305299.
- [6] Li, S.; Qin, Y. 2022. Maximum tangent likelihood estimation for linear regression. CRAN 1–9. <https://github.com/shaobo-li/MTE>
- [7] Hubert, M.; Rousseeuw, P.J.; Vanden Branden, K. 2005. ROBPCA: A new approach to robust principal component analysis. *Technometrics*. **47** (1) 64-79. DOI:10.1198/004017004000000563.
- [8] Mazimba O. 2017. Umbelliferone: Sources, chemistry and bioactivities review. *Bull. Fac. Pharmacy, Cairo Univ.* **55** (2) 223-232. DOI:10.1016/j.bfopcu.2017.05.001.

- [9] Ozgen, S.; Kilinc, O.K.; Selamoğlu, Z. 2016. Antioxidant activity of quercetin: a mechanistic review. *Turkish J. Agric. - Food Sci. Technol.* **4** (12) 1134-1138. doi:10.24925/turjaf.v4i12.1134-1138.1069.
- [10] Ruhardi, A.; Sahumena, M.H. 2021. Identifikasi Senyawa Flavonoid Daun Sembung (*Blumea balsamifera* L .) [Identification of Flavonoid Compounds in Sembung Leaves (*Blumea balsamifera* L.)]. *J. Syifa Sci. Clin. Res.* **3** 29–36. DOI: 10.37311/jsscr.v3i1.9925.