Forum Statistika dan Komputasi : *Indonesian Journal of Statistics* available online at:
ISSN : 0853-8115 *journal.ipb.ac.id/index.php/statistika*
Vol. 18 No.2, Oktober 2015, p: 105-107

LAD-LASSO: SIMULATION STUDY OF ROBUST REGRESSION IN HIGH DIMENSIONAL DATA

Septian Rahardiantoro[1], Anang Kurnia[1]

[1]Department of Statistics, Bogor Agricultural University
E-mail : rahardiantoro.stk@gmail.com and anangk@apps.ipb.ac.id

ABSTRACT

The common issues in regression, there are a lot of cases in the condition number of predictor variables more than number of observations ($p \gg n$) called high dimensional data. The classical problem always lies in this case, that is multicolinearity. It would be worse when the datasets subject to heavy-tailed errors or outliers that may appear in the responses and/or the predictors. As this reason, Wang et al in 2007 developed combined methods from Least Absolute Deviation (LAD) regression that is useful for robust regression, and also LASSO that is popular choice for shrinkage estimation and variable selection, becoming LAD-LASSO. Extensive simulation studies demonstrate satisfactory using LAD-LASSO in high dimensional datasets that lies outliers better than using LASSO.

Keywords: high dimensional data, LAD-LASSO, robust regression

INTRODUCTION

The classical multiple linear regression problem follows the model $\boldsymbol{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i;\ i = 1,2,\dots,n,$ with $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)'$ $p-$dimensional regression covariates, a response $\boldsymbol{y}_i$, and $\boldsymbol{\beta} = \left(\beta_1, \beta_2 \dots, \beta_p\right)'$ the associated regression coefficients where assumes errors $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2)$. Estimation of regression parameter, $\boldsymbol{\beta}$ could be using ordinary least squre (OLS) that minimize the sum square of error. The formula follows $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$, implies assume $\boldsymbol{X'X}$ is a nonsingular matrix, with matrix covariates $\boldsymbol{X}_{n\times p}$ and response vector $\boldsymbol{y}_{n\times 1}$.

Common issues in the certain background, there are a lot of regression cases in the condition number of predictor variables more than number of observations ($p \gg n$). When $\boldsymbol{X}$ is full rank ($p \leq n$), the exploration of causal relationship could be accomplished using classical multiple regression above. But when the number of predictors is large compared to the number of observations, $\boldsymbol{X}$ is likely not full rank, that means $\boldsymbol{X'X}$ become singular and the regression approach is no longer feasible (i.e., because of multicolinearity) [1]. LASSO regression [2], is a penalized regression method that is so popular choice for handling this conditions. It is so useful for shrinkage estimation and variable selection.

The worst condition of datasets for regression problem is when they subject to heavy-tailed errors or outliers that may appear in the responses and/or the predictors. In such a situation, it is well known that the traditional OLS may fail to produce a reliable estimator, and the least absolute deviation (LAD) estimator can be very useful. Wang et al (2007) [3] developed the combined method from LAD and LASSO regression. The basic idea is to combine the usual LAD criterion and the LASSO-type penalty together to produce the LAD-LASSO method.

Simulation study have been developed to see the LASSO and LAD-LASSO processes for handling high-dimensional data containts outliers in a lot of scenarios. The simulation using R software and some of R packages.

LAD-LASSO

Consider the linear regression model above, moreover assume that $\beta_j \neq 0$ for $j \leq p_0$ and $\beta_j = 0$ for $j > p_0$ for some $p_0 \geq 0$. Thus the correct model has p_0 significant and $(p - p_0)$ insignificant regression variables. Usually, the unknown prameters of classical regression model can

be estimated by minimizing the OLS criterion, $\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$. Furthermore, to shrink unnecessary coefficients to 0, Tibshirani (1996) [2] proposed the following LASSO criterion

$$LASSO = \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + n\lambda\sum_{j=1}^{p}|\beta_j|,$$

where $\lambda > 0$ is the tuning parameter. Then the LASSO formula have been modified by Fan and Li 2001 [4] for avoiding the bias,

$$LASSO^* = \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + n\sum_{j=1}^{p}\lambda_j|\beta_j|,$$

As a result, $LASSO^*$ is able to produce sparse solutions more effectively than $LASSO$. To obtain a robust $LASSO$-type estimator, the modification of $LASSO^*$ into the following LAD-$LASSO$ criterion:

$$\text{LAD-}LASSO = Q(\beta) = \sum_{i=1}^{n}|\boldsymbol{y}_i - \boldsymbol{x}i'\boldsymbol{\beta} + n\lambda j=1p\beta j,$$

As can be seen, the LAD-$LASSO$ criterion combines the LAD criterion and the lasso penalty, and hence the resulting estimator is expected to be robust against outliers and also to enjoy a sparse representation.

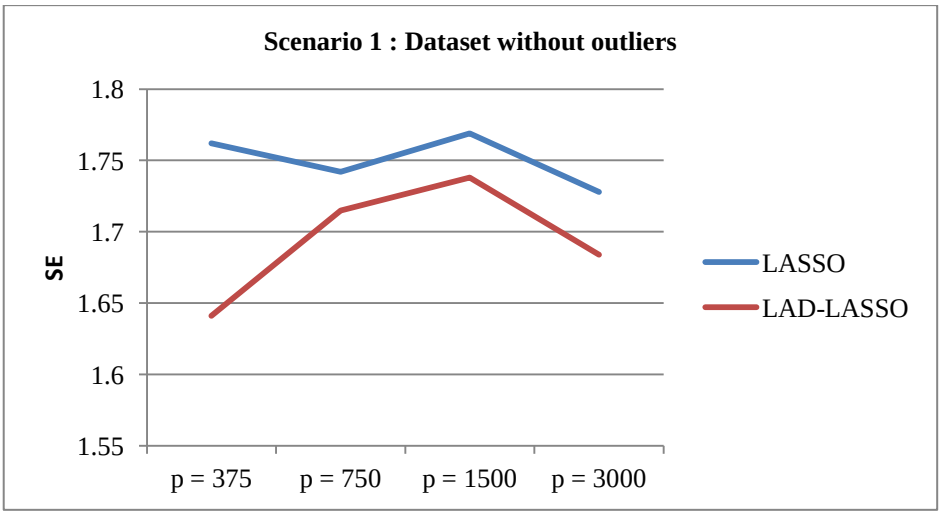
SIMULATION STUDY

The simulation in this research using R software that would evaluate the standard errors performance. Using R package, flare for sparse linear regression, the simulation set in $n = 100$, and vary p from 375 to 3000 as shown in Table 1. The datasets generated independently with each row of the design matrix from a p-dimensional normal distribution $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{jk} = 0.5^{|j-k|}$ [5]. Then the response vector generated follows $\boldsymbol{y}_i = 3\boldsymbol{x}_{i1} + 2\boldsymbol{x}_{i2} + 1.5\boldsymbol{x}_{i4} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$ is independently generated from $N(0,1)$. The scenario before generated without effects of outliers.

Next scenarios generated using the effects of outliers by replacing the distributions of errors that generated from some heavy-tailed distributions, in this research using the standard t-distribution with 5 df (t_5). For comparison purpose, all of scenarios to be evaluated using $LASSO$ and LAD-$LASSO$.

Table 1. Average of standard errors

Scenario 1 : Dataset without outliers				
Method	$p = 375$	$p = 750$	$p = 1500$	$p = 3000$
LASSO	1.762	1.742	1.769	1.728
LAD-LASSO	1.641	1.715	1.738	1.684
Scenario 2 : Dataset with outliers				
Errors : standard t-distribution with 5 df (t_5)				
Method	$p = 375$	$p = 750$	$p = 1500$	$p = 3000$
LASSO	1.514	1.653	1.623	1.638
LAD-LASSO	1.503	1.588	1.636	1.617



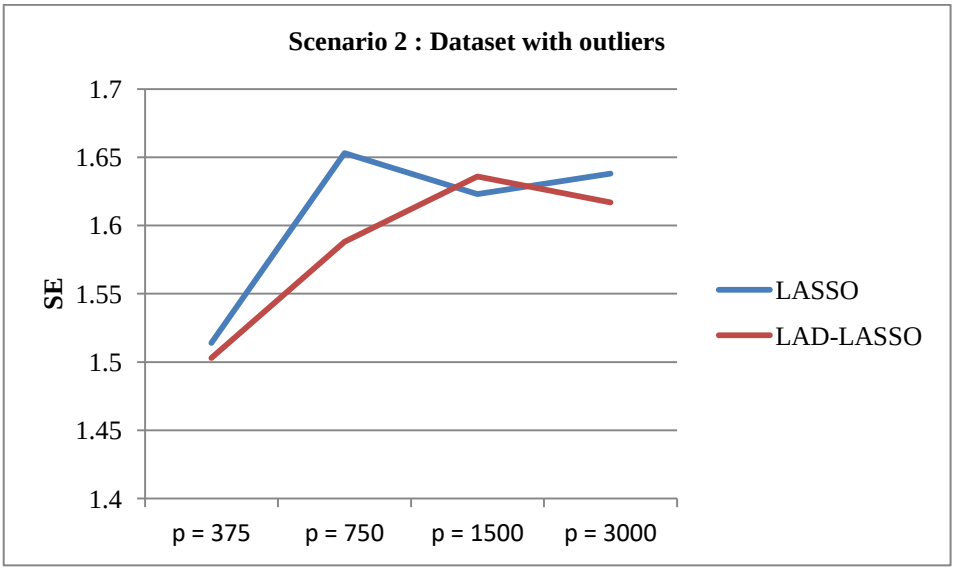


Figure 1. Graph of SE on scenario 1 and 2

From the table 1 above, it is shown that in the datasets without outliers, the better performance is from LAD-LASSO that have standard errors minimum. It is also happen almost in the datasets with outliers, the performance of LAD-LASSO is better than LASSO.

CONCLUSION

In the conditions of high-dimensional datasets contains outlier, the LAD-LASSO result the better solution (smaller standard errors) than LASSO. The concept is from combination of LAD and LASSO. And it could be as a suggestion for some researchers when handling high-dimensional datasets with $p \gg n$, and also contains outliers.

REFERENCES

[1] Myers, RH. "*Classical and Modern Regression with Applications Second Edition"*. Boston: PWS-Kent, 1990.
[2] Tibshirani, R. (1996). "Regression Shrinkage and Selection via the LASSO", *Journal of the Royal Statistics Society Series B*, 58, 267-288.
[3] Wang, H. Li,G. Jiang, G. (2007). "Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso", *JBES asa v.2007.*
[4] Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.
[5] Li, X. Zhao, T. Yuan, X. *et al* (2015). "An R Package flare for High Dimensional Linear Regression and Precision Matrix Estimation", *R publication.*