International Journal of Scientific Research in Mathematical and Statistical Sciences
Volume-5, Issue-5, pp.88-91, October (2018)

Research Paper

E-ISSN: 2348-4519

L_1 Penalized Regression Procedures for Feature Selection

Muthukrishnan.R[1] and Mahalakshmi.P[2*]

[1] Department of Statistics, Bharathiar University, Coimbatore 641 046, Tamil Nadu, India
[2] Department of Statistics, Bharathiar University, Coimbatore 641 046, Tamil Nadu, India

[*]*Corresponding Author: mahastat22@gmail.com, Tel.: +91-8056351592*

Available online at: www.isroset.org

Received: 05/Sept/2018, Accepted:13/Oct/2018, Online: 31/Oct/2018

Abstract— In high dimensional regression analysis, a greater number of independent variables occur in many scientific fields and machine learning applications. To select predictors that are relevant to the response, statistical feature selection should be performed. In the study on variable selection in regression analysis, specifically when there are a greater number of predictor variables or highly correlated variables (or both), traditional method includes forward-backward and mixed stepwise variable selection procedure fails. There is need of alternatives, that is, L_1 penalized regression procedures which provide higher prediction accuracy and computational efficiency. This paper demonstrates such procedures, particularly least absolute shrinkage and selection operator (LASSO) which does shrinkage and variable selection simultaneously and its variants. In case of extreme observations in the data set, robust regression estimators that are adopted in LASSO tolerate outliers with comparatively greater accuracy. In this paper, the performance of these procedures has been analyzed using the performance measure Median Squared Error (MSE) with numerical illustrations.

Keywords— Variable selection, LASSO, Huber, outlier, Robust, R Software

I. INTRODUCTION

Datasets with outliers or heavy-tailed errors are commonly encountered in many scientific field and real-time applications. In regression analysis, those extreme observations may appear in the response variable or in the predictor variables. In this case, the Ordinary Least Square (OLS) estimators fails to produce true value of an estimator. On the other hand, one of the main problems which occur in linear regression is variable selection. Variable selection or feature selection has become widely used as an important task in statistics. Nowadays when it comes to high-dimensional models, penalized estimators are widely considered rather than maximum likelihood estimators. As number of predictor variables increases, the predictive model becomes less effective due to most covariates being inactive in the model. This will cause the problem of over-fitting or under fitting, computations become very complex and also decrease the prediction power due to the noise. The effects of covariates and interpretations would become impossible to understand. So, the need for selecting variables in the predictive model is necessary and hence there are plenty of penalized regression procedures are established in the past few decades to perform feature selection in regression model.

Standard lasso and its variants were developed to reduce the coefficients in the model towards zero exactly. In some cases, it is reasonable to perform feature selection by grouping features. Group lasso proposed by Yuan and Lin (2006) in which coefficients are grouped. This lasso suffered from estimation inefficiency and inconsistency in variable selection in the same way as lasso. To overcome these limitations, Wang and Leng (2008) proposed adaptive group lasso which selects relevant features by adding weight vector in a grouped way. This can find the true consistency and satisfies oracle property.

Nowadays robust variable selection procedures are playing a vital role in the context of regression analysis. Wang et al. (2007) suggested that the lasso penalty to the least absolute deviation (LAD) estimation in robust linear regression. Zou and Yuan (2008) introduced composite quantile regression for a particular case where error variance is infinite. A unified theoretical structure of penalized techniques studied in detail by Negahban et al. (2012). Wang et al. (2013) introduced the exponential squared loss estimation for robust variable selection. Penalized least trimmed square (LTS) procedure was given by Alfons et al. (2013). The penalized Huber's loss for asymptotic contamination was studied by Fan et al. (2016). Lozano et al. (2016) worked on penalized L_2 distance estimation to manage the skewed response variable and variable selection. Qin et al. (2017) studied maximum tangent likelihood estimator (MTE) and its asymptotic properties. It works on variable selection and enjoys the oracle property.

In this paper, Section II briefly recall the various lasso-type methods. Section III demonstrates the performance of various penalty methods with real data. This paper concludes with a discussion in the last section.

II. PENALIZATION METHODS

The lasso and its variants are briefly summarized in this section.

A. LASSO

Standard lasso is performing well when regression error has extreme observations. To obtain a robust estimator, Wang et al. combined the least absolute deviation (LAD) and Lasso penalty to produce LAD-Lasso estimator which is defined as follows

$$\hat{\beta} = \sum_{i=1}^{n}(Y_i - \sum_{j} X_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \quad (1)$$

the sum of squares with a constraint of the form $\sum|\beta_i| \le t$, where $t \ge 0$ is a tuning parameter which controls the amount of shrinkage that remains same for all regression coefficients. Lasso does not only shrink coefficients towards zero but it also provides a selection of the significant covariates. It is known that, the OLS estimator criterion used in lasso regression is very sensitive to outliers.

B. LAD-lasso

Standard lasso is performing well when regression error has extreme observations. To obtain a robust estimator, Wang et al. combined the least absolute deviation (LAD) and Lasso penalty to produce LAD-Lasso estimator which is defined as follows

$$\hat{\beta}^{LAD} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n}|Y_i - X_i^T\beta| + \lambda_n\sum_{j=1}^{p}\hat{w}_j|\beta_j| \quad (2)$$

By using suitable λ_n, LAD-lasso satisfies oracle property. Besides as Zou (2006) showed that by using appropriate λ_n and a weight vector, $\hat{w}_j = (\hat{w}_1,...,\hat{w}_p)$, adaptive LAD-lasso satisfies the oracle property. Moreover, the resulting estimator is not affected by skewed errors since the squared loss is altered to L_1 loss. However, this loss penalizes strongly on small errors. Specifically, when the error is not skewed, it suffers from efficiency over adaptive lasso. In this case, Huber's criterion with lasso is preferable.

C. Huber lasso

The performance of LASSO will be poor if the regression response variable suffers from outliers or if the variable is skewed. Lambert and Zwald (2011) combined Huber's loss function with adaptive lasso penalty, defined by

$$\hat{\beta}_{Hadl} = \min_{\beta} L_\rho(\beta, s) + \lambda\sum_{j=1}^{p}\hat{w}|\beta_j| \quad (3)$$

where weights vector and the Huber's criterion is defined by

$$L_\rho(\beta,s) = \begin{cases} ns + \sum_{i=1}^{n}\rho\left(\frac{Y_i - \sum_{j=1}^{p}\beta_j X_{ij}}{s}\right)s, & if\ s>0, \\ 2M\sum_{i=1}^{n}\left|Y_i - \sum_{j=1}^{p}\beta_j X_{ij}\right|, & if\ s=0, \\ +\infty, & if\ s<0. \end{cases} \quad (4)$$

where s>0 is a scale parameter for the distribution. The criterion $\hat{\beta}_{Hadl}$ is a combination of Huber's loss function and adaptive lasso penalty together. Hence, the resultant estimator tolerates more extremes and filter variables simultaneously. Here, robustness is controlled by the shape parameter M. Huber suggested M as 1.345 to get robustness efficiently for normally distributed data. Generally, Huber's method tolerates more extreme observations in the dataset. But for normally distributed dataset, its efficiency is low.

D. LTS

LTS estimator is defined by adding a penalty parameter λ which leads to the sparse LTS estimator. Combination of Lasso and LTS estimator is defined as

$$\hat{\beta}_{LTS} = \arg\min \frac{1}{h}\sum_{i=1}^{h} r_{(i)}^2(\beta) + \lambda\sum_{j=1}^{p}|\beta_j|, \quad (5)$$

where $r_i^2(\beta) = (Y_i - X_i\beta)^2$, i=1, 2, …, n is the squared residuals and $r_{(1)}^2(\beta) \le ... \le r_{(n)}^2(\beta)$ is their ordered statistics. LTS lasso cannot be computed for high-dimensional data where p > n. It also has a high breakdown point. It performs well when the dataset is contaminated with multiple regression outliers.

E. MTE

In 2017, Qin et. al., introduced penalized MTE for estimation in high-dimensional regression and variable selection. Penalized MTE for variable selection defined as

$$\hat{\beta} = \arg\max_{\beta \in R^d} \left\{ \sum_{i=1}^{n} \ln_t (f(z_i ; \beta)) - n \sum_{j=1}^{d} p_{\lambda_{nj}} \left(\left| \beta_j \right| \right) \right\} \quad (6)$$

where function $\ln_t(.)$ is defined as

$$\ln_t(u) = \begin{cases} \ln(u) & \text{if } u > t, \\ \ln(t) + \sum_{k=1}^{p} \left. \frac{\partial^k \ln(v)}{\partial v^k} \right|_{v=t} \frac{(u-t)^k}{k!} & \text{if } 0 \le u \le t. \end{cases} \quad (7)$$

where $t \ge 0$ is a tuning parameter, $\ln_t(u)$ is a p^{th} order Taylor expansion of $\ln(u)$ for $0 \le u \le t$. MTE-Lasso is defined as,

$$\hat{\beta} = \arg\min_{\beta \in R_d} \left\{ L(\beta) + \lambda_n \sum_{j=1}^{d} \left| \beta_j \right| \right\} \quad (8)$$

where $L(\beta) = -\left(\frac{1}{n} \right) \sum_{i=1}^{n} \ln_t (f(z_i ; \beta))$ is MTE loss function. This penalized MTE performs well in robust estimation and variable selection under high dimensional regression. Also, it enjoys consistency, asymptotic normality and oracle property under fixed dimensional regression.

III. NUMERICAL STUDY

The performance of various penalization procedures has been studied under real data and the results obtained are demonstrated in this section. Penalization methods are applied to Boston housing price data set which is taken from 1970 census. There are totally 506 observations, each having 13 predictor variables namely *crim* (1), *zn* (2), *indus* (3), *chas* (4), *nox* (5), *rm* (6), *age* (7), *dis* (8), *rad* (9), *tax* (10), *ptratio* (11), *black* (12), *lstat* (13) and a dependent variable *mdev*. As the dataset contains outliers, they were detected and removed by using cook's distance. Analysis of this study was carried out by R software. The results such as variables selection, MSE under various procedures by considering with and without outliers are summarized in the following table.

Table 1: Analysis results of Boston house data

Methods	Selected variables		MSE
	With outliers	*Without outliers*	
lasso	1, 2, 4, 5, 6, 8, 9, 10, 11, 12, 13 (**11**)	1, 6, 10, 11, 12, 13 **(6)**	5.81(**4.69***)
lad	1, 2, 6, 8, 10, 11, 12, 13 **(8)**	1, 2, 6,10, 12,13 **(6)**	5.12(**3.71***)
adaptive lad	6, 8, 10, 11, 12, 13 **(6)**	6, 8, 10, 11, 12, 13 **(6)**	4.63(**3.90***)
huber	2, 5, 6, 7, 10, 12, 13 (**7**)	4, 6, 10, 11, 12, 13 **(6)**	4.62(**3.65***)
lts	1, 6, 10, 11, 12, 13 **(6)**	1,6, 01, 11, 12, 13 **(6)**	5.55(**3.92***)
mte	1, 2, 5, 6, 8, 11, 13 **(7)**	1, 6, 8, 9, 11, 13 **(6)**	4.55(**3.64***)

*without outliers

It is observed that both adaptive LAD and LTS methods select the same variables namely *crim, rm, tax, ptratio, black* and *lstat* under with and without outliers. The variables *rm* and *lstat* are most important variables, since all methods selected these two variables under with and without outliers. The variables such as *tax, ptratio, black* was considered the necessary variable for prediction by almost all methods. Further it is noted that, there exist in the multicollinearity among the variables such as (Indus with *nox, dis, tax), (nox* with *age, dis)* and *(age* with *dis)* and *(tax* with *rad)*. Robust procedures automatically eliminate correlated variables and take care of them.

IV. CONCLUSION

The performances of various LASSO penalty methods were studied with Boston housing price data set with and without outliers. From the numerical study, the efficiency of variable selection and accuracy of prediction is also compared with the standard lasso. All the robust procedures perform well when compared with standard lasso by considering the median squared error. Further, it is noted that the robust procedures MTE and Huber lasso performs better when there are extreme observations. It is concluded that the robust procedures perform well even with extreme observations and the presence of multicollinearity among the variables.

REFERENCES

[1] A. Alfons, C. Croux, S. Gelper, "*Sparse least trimmed squares regression for analyzing high-dimensional large data sets*", Annals of Applied Statistics, Vol.7, No.1, pp.226-248, 2013.

[2] A. C. Lozano, N. Meinshausen, E. Yang, "*Minimum distance lasso for robust high-dimensional regression*", Electronic Journal of Statistics, Vol.10, No.1, pp.1296-1340, 2016.

[3] H. Wang, C. Leng, "*A note on adaptive group lasso*", Computational Statistics and Data Analysis, pp.5277-5286, 2008.

[4] H. Wang, G. Li, G. Jiang, "*Robust regression shrinkage and consistent variable selection through the LAD-lasso*", Journal of Business & Economic Statistics, Vol.25, pp.347-355, 2007.

[5] H. Zou, "*The adaptive lasso and its oracle properties*", Journal of the American Statistical Association, Vol.101, No.476, pp.1418-1429, 2006.

[6] H. Zou, M. Yuan, "*Composite quantile regression and the oracle model selection theory*", Annals of Statistics, Vol.36, No.3, pp.1108-1126, 2008.

[7] J. Fan, Q. Li, Y. Wang, "*Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions*". J.R.Statist.Soc. B (Statistical Methodology), In Press, 2016.

[8] K. Knight, W. Fu, "*Asymptotics of lasso-type estimators*", Annals of statistics, Vol.28, pp.1346-1378, 2000.

[9] Lambert-Lacroix, Zwald, "*Robust Regression through the Huber's criterion and adaptive lasso penalty*", Electron. J. Stat., Vol.5, pp.1015-1053, 2011.

[10] M. Mohanadevi, V. Vinothini, "*Accurate Error Prediction of Sugarcane Yeild Using a Regression Model*", International Journal of Computer Science and Engineering, Vol.6, Issue.7, pp.66-71, 2018.

[11] M. Yuan, Y. Lin, "*Model selection and estimation in regression with grouped variables*", J.R.Statist.Soc. B, Vol.68, pp.49-67, 2006.

[12] R Core Team, "*R: A language and environment for statistical computing*", R Foundation for Statistical Computing, Vienna, Austria, 2018.
[13] R. Tibshirani, "*Regression shrinkage and selection via the lasso*", J.R.Statist.Soc. B, Vol.58, pp.267-288, 1996.
[14] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, "*Sparsity and smoothness via the Fused Lasso*", J.R.Statist.Soc. B, Vol.67, pp.91-108, 2005.
[15] S. Li and Y. Qin, "*MTE: Maximum Tangent Likelihood and Other Robust Estimators for High-Dimensional Regression*", R package version 1.0.0.
[16] S. N. Negahban, P. Ravikumar, "*A unified framework for high-dimensional analysis of M-estimators with decomposable regularizes*", Statistical Science, Vol.27, No.4, 538-557, 2010.
[17] X. Wang, Y. Jiang, M. Huang, H. Zhang, "*Robust variable selection with exponential squared loss*", Journal of the American Statistical Association, Vol.108, Issue.502, pp.632-643, 2013.
[18] Y. Qin, Shaobo Li, Yang Li, Yan Yu, "*Penalized maximum tangent likelihood estimation and robust variable selection*", Unpublished, 2017.

AUTHORS PROFILE

R. Muthukrishnan graduated from Manonmaniam Sundaranar University in 2000. Now he is working as an Associate professor in Bharathiar University. His research interests are Robust Statistical Inference, Sampling Techniques, Multivariate Analysis.

P. MAHALAKSHMI pursuing Doctor of Philosophy in Department of Statistics, Bharathiar University. Her research interests are Statistical Inference Particularly in Robust Inference, Multivariate Regression analysis, Statistical Machine learning and R software.

