Advances and Applications in Mathematical Sciences
Volume 21, Issue 10, August 2022, Pages 6103-6115

FEATURE SELECTION THROUGH ROBUST LASSO PROCEDURES IN PREDICTIVE MODELLING

R. MUTHUKRISHNAN[1] and C. K. JAMES[*]

Bharathiar University
Coimbatore-641046, India

Abstract

Feature selection plays an important role in the construction of predictive models by selecting a subset of important features. Least absolute shrinkage and selection operator (LASSO) is one of the widely used techniques which does shrinkage and variable selection simultaneously. It's often utilised to make the model easy to understand. This procedure is mainly based on ordinary least square principle with penalty. Least square principle is very sensitive to outliers, hence this procedure gives unreliable results when extreme observations are present in the data. Moreover, the conventional approaches of feature selection are not resistant to the presence of outliers. In this context, robust statistics helps to play as an alternative. Robust statistics are used to describe the structure by best fitting the majority of data and also to identify deviating data points. This paper explores LASSO type robust procedures which have been developed in the past, such as Least Absolute Deviation (LAD), adaptive, Huber procedures. The efficiency of these procedures have been studied under real environment and by comparing the error measures such as RMSE, MAPE in the context of model fitting. It is concluded that the adaptive procedures perform well in the context of prediction accuracy and variable selection. With the help of robust statistical procedures, it is now possible to develop a variety of automated feature selection algorithms because of increasing availability of fast computing.

1. Introduction

Data containing outliers is one of the most typical challenges we encounter in various scientific domains and real-time applications which can cause several complications in regression analysis. Outliers can exist either in dependent variables or covariates (predictor variables). Because of the presence of outliers, traditional approaches such as Ordinary Least Squares

2020 Mathematics Subject Classification: 62J07, 62F35

Keywords: Feature selection - LASSO - LAD - RMSE - Outlier.
[*]Corresponding author; E-mail: jamesck74@gmail.com
Received February 17, 2022; Accepted April 4, 2022

(OLS) methods fail to determine the real value of the estimate and also the interpretation becomes more complex if we are following the traditional methods. This pay a route for feature selection methods or penalised estimator approaches. Feature selection/variable selection is becoming more essential in statistics which plays a significant part in statistical analysis owing to the complexity and dimensionality of data. The key advantage of the feature selection approach is that it identifies the essential variables while also assisting in a parameter estimate. The loss function plus a penalty function helps to choose the most essential variables from a bigger range of variables. Over fitting occurs when the number of variables rises, making computation more complex. As the number of variables increases, interpretation becomes more difficult, resulting in decreasing prediction power. This piques the researchers' interest in approaches for simultaneous variable selection and parameter estimation. In the last several decades, various shrinkage regression algorithms for variable selection in regression methods have been proposed.

The most popular shrinkage methods/penalized methods for variable selection is Least Absolute Shrinkage and Selection Operator (LASSO) introduced by (Tibshirani, [5]) which makes the coefficients in the model exactly towards zero. The LASSO is operated by a loss function of the OLS technique and a penalty factor (also known as the L1 penalty). Later, it was shown that LASSO does not meet the oracle property, prompting (Zou, [8]) to create adaptive lasso in 2006, in which adaptive weights are used to shrink distinct coefficients in the L1 penalty.

The presence of extreme observations or outliers in the data renders both approaches ineffective for executing feature selection procedures. This prompted the researchers to consider developing robust variable selection methods, which are resistant to outlier or extreme observation. (Rosset and Zhu, [4]) and (H. Wang et al., [6]) took the initial step in this direction by combining the Huber as a loss function with the lasso penalty factor to create Huber-LASSO, which can withstand outliers in the dependent variables and Least Absolute Deviation (LAD) as a loss function with same procedure to overcome the outliers in the dependent variables. In 2008, (Zou and Yuan, [9]) proposed the notion of composite quantile regression. (Lambert-Lacroix and Zwald, [2]) developed Huber Adaptive lasso, which can also resist outliers

in response variables. (X. Wang et al., [7]) proposed the Exponential Squared LASSO for feature selection, and (Qin et al., [3]) studied the Maximum Tangent Likelihood Estimator (MTE) and its asymptotic properties in 2017, which aids in variable selection methods, and many researchers are working to develop more robust feature selection methods that will be useful for statistical analysis. In this study, we will compare all of the key conventional LASSO and their adaptive versions. The rest of the paper is organised as follows. Section 2 briefly defines the LASSO-type approaches. Section 3 shows that the performance of the various LASSO-type algorithms on real data, and the work concludes with a conclusion and provides scope for further research.

2. Shrinkage Methods

To explain the regression regularisation approaches, we begin with the usual model for multiple linear regression. Let the data be $(x_1, y_1), \ldots, (x_n, y_n)$, let the design matrix be $X = (x_I^T, \ldots, x_n^T)^T$ and the general linear model be defined as

$$Y = X\beta + \varepsilon \tag{1}$$

Here $\beta = (\beta_1, \beta_p)^T$ represent the regression coefficients $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(0, \sigma^2 I_n)$ known as error terms, x_k termed as the repressors for observation x_k and $k = 1, \ldots, n$ and $y = (y_1, \ldots, y_n)^T$. The OLS estimates β by minimizing the error sum of squares, i.e. $\hat{\beta}_{OLS} = \min_{\beta} \{(Y - X\beta)^T (Y - X\beta)\}$.

In general, OLS yields unbiased estimators with high variances, and lowering the variance by a little amount can enhance prediction accuracy.

2.1 LASSO

A method is used to reduce the variance of the estimate and to accomplish variable selection, named LASSO was proposed by (Tibshirani, [5]) as a novel approach for linear model estimation. It is a regression shrinkage method that is frequently used in models with such a large number of variables although little observations. This approach minimises the residual sum of

squares subject to the sum of the absolute value of the coefficients being less than a constant, which is same as minimizing the sum of squares with a constant $\Sigma|\beta_j| \le s$, and brings some of the β coefficients shrunk to zero. The method applies L1 regularisation to the objective under optimization by imposing a penalty. This penalty is the total of the absolute values of the coefficients and decides which coefficients and how much to shrink. The LASSO estimate is given by

$$\beta_{Lasso} = \min_{\beta}\left(\sum_{i=1}^{n}\left(y_i - \sum\nolimits_j \beta_j X_{ij}\right)^2 + \lambda \sum\nolimits_j |\beta_j|\right) \tag{2}$$

where λ is termed as the shrinkage parameter.

2.2 Adaptive LASSO

The LASSO proposed by Tibshirani does not satisfy the oracle property, in order to overcome this problem Zou (2006) introduced a weight function to each β coefficients and it is defined as

$$\beta_{AdaptiveLASSO} = \min_{\beta}\left(\sum_{i=1}^{n}\left(y_i - \sum\nolimits_j \beta_j X_{ij}\right)^2 + \lambda \sum\nolimits_{j=1}^{p} \hat{w}_j |\beta_j|\right) \tag{3}$$

where $W_j (j = 1, \ldots, p)$ are the weight functions, which can be estimated by $w_j = \dfrac{1}{|\hat{\beta}_j|^{\gamma}}$, where γ is a positive constant and $\hat{\beta}_j$'s are the initial estimates of β coefficients.

2.3 LAD-LASSO

Wang et al. [6] combined the LAD along with LASSO penalty to obtain the robust estimator, namely LAD-LASSO estimator defined as

$$\beta_{LAD\,LASSO} = \min_{\beta}\left(\sum_{i=1}^{n}\left(\left| y_i - \sum\nolimits_{j=}^{p} \beta_j x_{ij} \right|\right) + \lambda \sum\nolimits_{j=1}^{P} |\beta_j|\right) \tag{4}$$

The resulting estimator should be resistant to outliers and have a sparse representation.

2.4 Adaptive LAD-LASSO

For a robust estimate and to have a consistent variable selection, Zou's Adaptive LASSO method for variable selection is paired with LAD regression in the situation of heavy tailed errors and is given by

$$\beta_{AdapiveLAD} = \min_{\beta}\left(\sum_{i=1}^{n}\left(\left| y_i - \sum_{j=}^{p}\beta_j x_{ij}\right|\right) + \lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|\right) \tag{5}$$

where $\hat{w}_j = (w_1, \ldots, w_p)$ is known as weight vector.

2.5 Huber LASSO

When there are outliers in the regression response variable, LASSO performance may diminish. Rosset and Zhu [4] and Wang et al. [6] were the first to attempt to solve this problem. Rosset and Zhu [4] use Huber's criterion as a loss function for the LASSO penalty. They employ Huber's loss function with a fixed M and a penalty equal to the L1 penalty defined as

$$\beta_{HuberLASSO} = \min_{\beta}\sum_{i=1}^{n}H_M(y_i - x_i^T\beta) + \lambda\sum_{i=1}^{p}|\beta| \tag{6}$$

Where $H_M(t) = \begin{cases} t^2 & \text{if } |t| \le M \\ 2M|t| - M^2 & \text{if } |t| > M \end{cases}$

2.6 Adaptive Huber LASSO

The Huber's Criteria with Adaptive LASSO was proposed by Lambert-Lacroix and Zwald [2], this combines Huber's objective function and the Adaptive LASSO penalty stated by

$$\beta_{AdaptiveHuberLASSO} = \min_{\beta}\sum_{i=1}^{n}H_M(y_i - x_i^T\beta) + \lambda\sum_{i=1}^{p}\hat{w}_j|\beta_j| \tag{7}$$

where $\hat{w}_j = (w_1, w_p)$ is known as weight function and the Huber's function is defined by

$$L_H(\alpha, \beta, s) = \begin{cases} ns + \sum_{i=1}^{n} H_M(Y_i - \alpha - X_i^T\beta) & \text{if } s > 0 \\ 2M + \sum_{i=1}^{n} | y_i - \alpha - X_i^T\beta | & \text{if } s = 0 \\ +\infty & \text{if } s < 0 \end{cases}$$

where, $s > 0$ is the scale parameter for the function. For small and large residuals the loss function changes from quadratic to linear which helps in penalizing outliers.

2.7 Maximum Tangent Likelihood Estimation (MTE)

Qin et al. [3] proposed MTE defined as $\hat{\beta} = \max_{\beta} \sum_{i=1}^{n} \ln_t(f(z_i : \beta))$, where z_i represents the dependent and independent variables, and the normal distribution with a mean of zero is represented by f. However, even if only a small fraction of the data is contaminated, the performance of such an estimator generally falls dramatically. When model assumptions are valid, the robust statistical process should perform almost ideally, and it should continue perform well when the assumptions are broken. For variable selection, the penalised maximum tangent likelihood estimate (penalised MTE) is used and is defined as

$$\beta = \arg\max_{\beta \in R^d} \left\{ \sum_{i=1}^{n} \ln_t(f(z_i : \beta)) - n \sum_{j=1}^{d} p\lambda_{nj}(| \beta_j |) \right\} \tag{8}$$

where the $\ln_t(\cdot)$ is as follows

$$\ln_t(u) = \begin{cases} \ln(u) & \text{if } u > t \\ \ln(t) + \sum_{k=1}^{p} \frac{\partial^k \ln(v)}{\partial v^k} |_{v=t} \frac{(u-t)^k}{k!} & \text{if } 0 \le u \le t \end{cases} \tag{9}$$

In this case, $t \ge 0$ is a tuning parameter (u) is just a p^{th} order Taylor expansion of $\ln(u)$ for $0 \le u < t$.

2.8 Adaptive MTE

The Zou's Adaptive LASSO for consistent variable selection is combined with MTE regression in the MTE-LASSO criteria.

$$\beta = \arg\max_{\beta \in R^d} \left\{ L(\beta) + \lambda_n \sum_{j=1}^{d} W_j(|\ \beta_j\ |) \right\} \tag{10}$$

Where $L(\beta)$ is the MTE loss function which is defined in (7) and (8), λ_n is the regularization parameter of L1 penalty and Where $W_j = (w_1, \ldots, w_p)$ is known as weight vector.

3. Experimental Results

In this section, the performance of various LASSO type feature selection procedures has been studied on real environment. The real data set contains outliers, they were detected and removed by using cook distance (Cook, 2000) and the analysis has been carried out using *R* software. The results such as number of variable selected, Mean Absolute Error (MAE), Median Absolute Error (MDAE), Mean Absolute Percent Error (MAPE), Root Mean Squared Error (RMSE) under various procedures by considering with and without outliers are summarized in the following tables.

3.1 Diabetes data set

The Diabetes data initially used by Efron et al. (2004). The data contains 442 observations and nine covariates and a dependent variable. We standardize all the variables before doing the analysis. The Figure 1 – (a) gives the information of correlation plot and Figure 1 – (b) gives number of variables chosen by each method. The Table (1) gives error values and variables chosen by each method under with and without outliers.

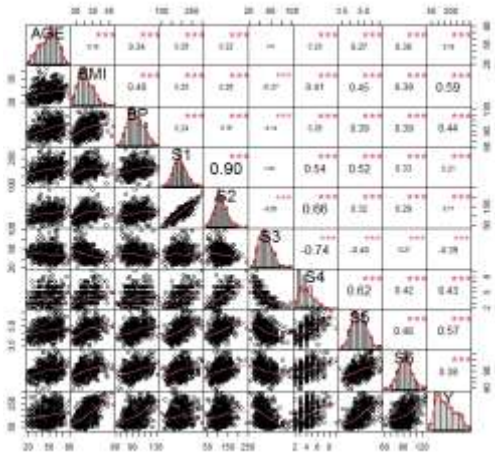

Figure 1-(a). Correlation Plot.

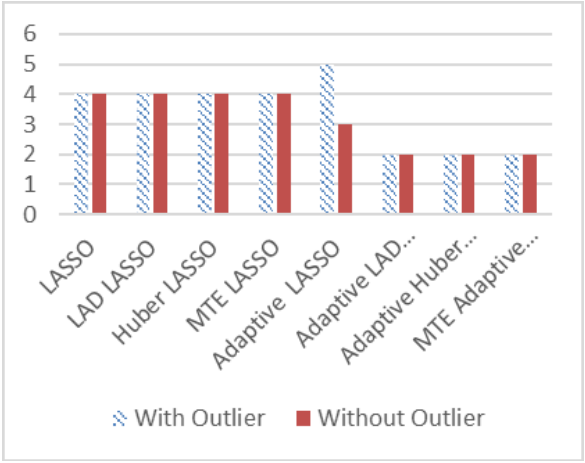

Figure 1-(b) No. of variables chosen by each method.

All LASSO algorithms selects same variables under with and without outlier study except adaptive LASSO. The other adaptive methods except adaptive LASSO method selects less number of variables when compared with other standard LASSO methods. The error value is also minimum for adaptive methods.

Table 1. Error values, variables selected by each method under with and without outlier (.) without outlier.

Methods	MAE	MDAE	RMSE	MAPE	Variables Selected	No. of Variables Selected
LASSO	0.564 (0.579)	0.515 (0.510)	0.696 (0.668)	1.142 (1.470)	Bmi, bp, s3, s5 (Bmi, bp, s3, s5)	4 (4)
LAD LASSO	0.561 (0.546)	0.489 (0.526)	0.693 (0.635)	1.287 (1.563)	Bmi, bp, s3, s5 (Bmi, bp, s3, s5)	4 (4)
Huber LASSO	0.558 (0.556)	0.500 (0.509)	0.692 (0.645)	1.221 (1.524)	Bmi, bp, s3, s5 (Bmi, bp, s3, s5)	4 (4)
MTE LASSO	0.564 (0.583)	0.516 (0.503)	0.696 (0.673)	1.146 (1.458)	Bmi, bp, s3, s5 (Bmi, bp, s3, s5)	4 (4)
Adaptive LASSO	0.560 (0.545)	0.506 (0.503)	0.694 (0.630)	1.222 (1.664)	Bmi, bp, s1, s2, s5 (Bmi, s2, s5)	5 (3)
Adaptive LAD Lasso	0.580 (0.532)	0.466 (0.488)	0.717 (0.633)	1.259 (1.576)	Bmi, s5 (Bmi, s5)	2 (2)
Adaptive Huber LASSO	0.575 (0.532)	0.464 (0.502)	0.708 (0.622)	1.206 (1.689)	Bmi, s5 (Bmi, s5)	2 (2)
MTE Adaptive LASSO	0.574 (0.546)	0.502 (0.508)	0.705 (0.632)	1.151 (1.66)	Bmi, s5 (Bmi, s5)	2 (2)

From the results, it is concluded that adaptive Huber LASSO algorithm produces less prediction errors and also select less number of variables when compared with standard LASSO, and adaptive LASSO algorithms for feature selection.

3.2 Boston Housing dataset

This dataset contains information collected by the U.S Census Service

concerning housing in the area of Boston Mass and the dataset is available at the ml bench package in *R*-software. The data set contains 506 observations and 15 covariates and a dependent variable CMDEV (Corrected median value of owner-occupied homes per 1000 US $). The objective of the data set is to predict the value of prices of the houses in Boston region using the given features such as crime rate, accessibility to highway, number of rooms, distance of employment centres, nitric oxides concentration etc. We standardize all the variables before doing the analysis. The Figure 2 – (a) gives the information of correlation plot and Figure 2 – (b) gives number of variables chosen by each method. The Table (2) gives variables chosen by each method and gives the various error measurements.

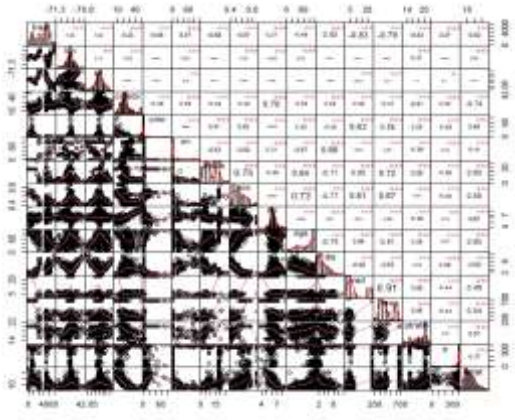

Figure 2 (a). Correlation Plot.

The LAD- LASSO selects all variables in without outlier case and Huber method selects almost same number of variables in both cases. The number of variables selected by the Adaptive method is less compared with other LASSO methods and also rm, ptratio, istat are the variables mostly selected by all methods.

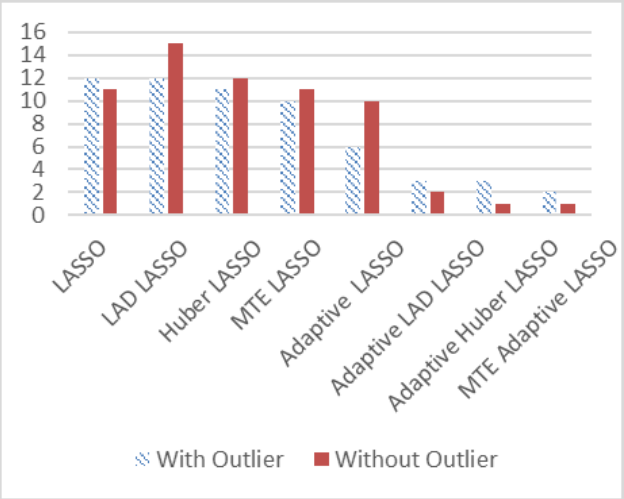

Figure 2 - (b). No. of variables chosen by each method.

Table 2. Error values, variables selected by each method under with and without outlier (.) without outlier.

Methods	MAE	MDAE	RMSE	MAPE	Variables Selected	No. of Variables Selected
LASSO	0.402 (0.125)	0.302 (0.122)	0.595 (0.145)	4.261 (0.782)	Tract, ion, lat, crim, zn, nox, rm, dis, tax, ptratio, b, istat (ion, crim, zn, indus, nox, rm, age, dis, tax, ptratio, b, istat)	12(11)
LAD LASSO	0.373 (0.239)	0.266 (0.176)	0.585 (0.316)	4.063 (1.192)	Tract, ion, crim, zn, nox, rm, age, dis, tax, ptratio, b, istat (Tract, ion, lat, crim, zn, indus, nox, rm, age, dis, rad, tax, ptratio, b, istat)	12(15)
Huber LASSO	0.390 (0.237)	0.290 (0.188)	0.599 (0.315)	3.964 (1.168)	Tract, ion, crim, nox, rm, age, dis, tax, ptratio, b, istat (Tract, ion, crim, zn, indus, rm, age, dis, tax, ptratio, b, istat)	11(12)
MTE LASSO	0.403 (0.302)	0.306 (0.244)	0.595 (0.390)	4.230 (1.310)	Tract, ion, crim, nox, rm, dis, tax, ptratio, b, istat (ion, crim, zn, indus, rm, age, dis, tax, ptratio, b, istat)	10(11)

Adaptive LASSO	0.442 (0.126)	0.361 (0.119)	0.624 (0.148)	5.270 (0.808)	nox, rm, dis, tax, tratio, istat (ion, crim, zn, rm, age, dis, tax, ptratio, b, istat)	6(10)
Adaptive LAD LASSO	0.386 (0.123)	0.275 (0.113)	0.600 (0.147)	4.720 (0.758)	rm, ptratio, istat (rm, age)	3 (2)
Adaptive Huber LASSO	0.410 (0.123)	0.299 (0.114)	0.611 (0.143)	4.812 (0.773)	rm, ptratio, istat (rm)	3(1)
MTE Adaptive LASSO	0.457 (0.128)	0.332 (0.139)	0.652 (0.149)	5.048 (0.789)	rm, istat (rm)	2(1)

From the table it is observed that under with outlier case the various error values of LAD and Huber is very less and also the predictive capacity is good for both methods, and in Adaptive case except Adaptive LASSO all other Adaptive algorithms selects only less number of variable and also the error measurements and predictive capacity is better for both Adaptive (LAD and Huber) methods.

4. Conclusion

Feature selection approach is a method which helps in identifying the essential variables from a larger set of variables. Feature selection is becoming highly significant in statistics, and it plays an important role in statistical analysis. LASSO methods are popular among them which helps in shrinking some variables exactly to zero. LASSO methods fails if the data contains outlier so we come across some robust LASSO techniques and we compared the robust LASSO methods and its adaptive version with the standard LASSO. In this study it is clear that Adaptive methods selects less number of variables when compared with all other LASSO methods. The predictive capacity and various other error measurements of Adaptive methods is good by choosing a less number of variables from a larger set of variables. So the adaptive procedures outperform when compared with the other ordinary LASSO methods, especially Adaptive Huber LASSO and Adaptive LAD-LASSO can resist to outlier to a large amount. Moreover, the study shows that the robust procedures, namely LAD and Huber algorithms outperform over the other algorithms in the context of feature selection. A

further study in the adaptive based methods is useful to get an improved version of the adaptive method or else a better robust loss function with penalty factor could able to get better results in feature selection methods.

References

[1] R. D. Cook, Detection of influential observation in linear regression, Technometrics, 42(1) (2000), 65-68. https://doi.org/10.1080/00401706.2000.10485981

[2] S. Lambert-Lacroix and L. Zwald, Robust regression through the Huber's criterion and adaptive lasso penalty, Electronic Journal of Statistics 5 (2011), 1015-1053. https://doi.org/10.1214/11-EJS635

[3] Y. Li, S. Li, Y. Qin and Y. Yu, Penalized maximum tangent likelihood estimation and robust variable selection, (2017). ArXiv Preprint ArXiv:1708.05439

[4] S. Rosset and J. Zhu, Piecewise linear regularized solution paths, Annals of Statistics, 35(3) (2007), 1012-1030. https://doi.org/10.1214/009053606000001370

[5] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, In Source: Journal of the Royal Statistical Society, Series B (Statistical Methodological) 58(1) (1996).

[6] H. Wang, G. Li and G. Jiang, Robust regression shrinkage and consistent variable selection through the LAD-lasso, Journal of Business and Economic Statistics 25(3) (2007), 347-355. https://doi.org/10.1198/073500106000000251

[7] X. Wang, Y. Jiang M. Huang and H. Zhang, Robust variable selection with exponential squared loss, Journal of the American Statistical Association 108(502) (2013), 632-643. https://doi.org/10.1080/01621459.2013.766613

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

[9] H. Zou and M. Yuan, Composite quantile regression and the oracle model selection theory, Annals of Statistics 36(3) (2008), 1108-1126. https://doi.org/10.1214/07-AOS507