Variable selection of yearly high dimension stock market price using ordered homogenous pursuit lasso

Cite as: AIP Conference Proceedings **2266**, 090012 (2020); https://doi.org/10.1063/5.0019161
Published Online: 06 October 2020

Yusrina Andu, Muhammad Hisyam Lee, and Zakariya Yahya Algamal

Variable Selection of Yearly High Dimension Stock Market Price using Ordered Homogenous Pursuit Lasso

Yusrina Andu[1,2,b)], Muhammad Hisyam Lee[1,a)] and Zakariya Yahya Algamal[3,c)]

[1]*Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia*
[2]*Universiti Malaysia Kelantan, 17600, Jeli, Kelantan, Malaysia*
[3]*University of Mosul, Mosul, Iraq*

[a)]Corresponding author: mhl@utm.my
[b)]yusrinaandu@gmail.com
[c)]zakariya.algamal@gmail.com

Abstract. It is noting that the response variable and the explanatory variables are highly correlated in high dimension data. Hence, the selection of informative variables is important in order to achieve a better model interpretation and concomitantly improve the accuracy of the prediction. In this study, the variable selection in stock market price using statistical approach was carried out. It is pertinent since most of the previous study only concerns on the financial interests of the stock market. Therefore, this study considers the homogeneity structure in the highly correlated data on yearly stock market price by applying ordered homogenous pursuit lasso (OHPL) method. The performance results of OHPL were compared with lasso and elastic net. As a result, OHPL a had higher number of selected variables and a better prediction power than of lasso and elastic net. In conclusion, OHPL shows its capability to enhance variable selection while increasing the prediction power of the selected variables than its counterpart.

Keywords: variable selection, high dimension, OHPL, linear regression, homogeneity

INTRODUCTION

Modelling the relationship between the explanatory and response variables using statistical approach is essential for a wide range of studies. In addition, when the number of explanatory variables is larger than of observations, the high dimensionality sample data is assumed to be independent. Thus, in the case of high dimension, selecting a subset of informative variables is of major concern [1]. It is noteworthy that many explanatory variables are added to lower the possibility of the model deviation. Nevertheless, having more explanatory variables may not be relevant as this decreases the accuracy of the prediction and also the model interpretation. Hence, the reduction technique is performed since it is able to select optimal explanatory variables that possess relevant information and improve the statistical models concomitantly. As an advantage, the model would be constructed with a better performance in prediction power and also interpretation [2].

It is noting that the conventional variable selection such as Akaike information criteria (AIC), Mallow's Cp, and Bayesian information criteria (BIC) are impractical in high dimensional due to the high computational time [1, 3, 4]. As a result, modelling high dimension data using these conventional methods may not achieve a better interpretation of the relationship between the explanatory and the response variables. Ordinary least squares (OLS) is widely recognize as the common estimation method in many application studies. Nonetheless, the OLS estimator becomes unreliable in the existence of multicollinearity among the explanatory variables. Moreover, the computation of the OLS estimator could not be carried out if the number of explanatory variables exceeds the number of response variables [5]. Hence, penalized likelihood methods are commonly adapted to overcome these difficulties.

The most known penalized method which is least absolute shrinkage and selection operator (LASSO) was introduced by [6]. LASSO utilised L_1-penalty instead of L_2-penalty, in which the variable selection can be performed by assigning some explanatory variable coefficients to zero. Certainly, LASSO has garnered the attention of many researches particularly in high dimensional studies. On the contrary, it also has some flaws, for instance, the selection of explanatory variables is usually less than the number of observations. In addition, this method attempts to choose only

Proceedings of the 27th National Symposium on Mathematical Sciences (SKSM27)
AIP Conf. Proc. 2266, 090012-1–090012-5; https://doi.org/10.1063/5.0019161
Published by AIP Publishing. 978-0-7354-2029-8/$30.00

one variable among the highly correlated explanatory variables. Besides that, the method also has no oracle properties where the probability of selecting the right set of explanatory variables that have nonzero coefficients converged to one. As well as, if the zero coefficients were known antecedent, it will have a similar means and covariances as the asymptotically normal nonzero coefficients estimators. Notwithstanding, LASSO has become the baseline of many penalized methods and has elucidate several extensions in diverse practical applications.

On the other hand, [7] has then proposed elastic net method to overcome limitation issue in LASSO by combining both L_1-penalty and L_2-penalty. Definitely, both of LASSO and elastic net has the capability to perform variable selection and model estimation concurrently. However, they may lack in variable selection consistency which may lead to poor model performance and lower prediction power. Therefore, a recent approach of homogeneity specifically in the highly correlated variables is proposed. The fundamental of homogeneity was introduced by [8] to divide the regression coefficients into several groups. The classification of the group is pertinent to separate based on regression coefficients values. For example, the regression coefficients values in the same groups are together or close, whereas in the different groups are significantly different to each other. Sparsity occurs when a large number of groups is entirely made up of zero coefficients. This is also referred as a special condition of homogeneity. The successful detection of homogeneity in the model causes the regression model to not only recognized the original structure of the data but also increase the predictive performance.

In a recent study on homogeneity by [4], proposed ordered homogeneity pursuit lasso (OHPL) to improve the limitation of LASSO. This study had successfully improved the prediction power and variable selection in the application of spectroscopic dataset. Likewise, spectroscopic dataset, the highly correlated yearly stock market price also may contain the homogeneity structure in its explanatory variables. Thus, the present study applies OHPL to performed variable selection as well as enhanced the prediction power in the application of yearly stock market price. The comparison of OHPL will be carried out with LASSO and elastic net method.

METHOD

LASSO

The penalized linear regression with LASSO regression coefficients is defined as

$$\hat{\beta}_{PLR}^{LASSO} = argmin_{\beta}\left\{(y-\beta\mathbf{X})^{\mathbf{T}}(\mathbf{y}-\beta\mathbf{X}) + \lambda \sum_{\mathbf{j=1}}^{\mathbf{p}} |\beta_{\mathbf{j}}|\right\} \tag{1}$$

where $\lambda \geq 0$ is a fixed tuning parameter that controls the degree of sparsity in the $\sum_{j=1}^{p} |\beta_j|$ penalty term. LASSO continuously shrinks the regression coefficients towards zero which subsequently improves its prediction accuracy. If the λ is large enough, greater number of coefficients estimates will be exactly zero. Nevertheless, if the λ is approaching zero, this becomes the OLS estimates.

Noteworthy, LASSO has two added advantages as compared to the conventional variable selection methods. It has more stability than subset selection due to the selection process in LASSO is continuous. In addition, they are suitable in computing high dimensional model. Although LASSO is one of the successful variable selection methods for high-dimensional data, however, it may not perform well under certain circumstances. For instance, LASSO will usually choose most n explanatory variables due to the nature of the convex optimization problem. Thus, this becomes an undesirable property for some applications [7].

Elastic Net

As compared to LASSO, elastic net can select more than n variables. To address the limitation in LASSO, elastic net was developed by combining the L_1-norm penalty and the L_2-norm penalty which are from ridge and LASSO, respectively [7]. The ridge penalty is to overcome the highly correlated problem, whereas the LASSO penalty is for the variable selection problem. Elastic net estimator with penalized linear regression is defined as follows

$$PLR_{Elastic}(\beta;\lambda_1;\lambda_2) = (y-\beta\mathbf{X})^{\mathbf{T}}(\mathbf{y}-\beta\mathbf{X}) + \lambda_{\mathbf{1}} \sum_{\mathbf{j=1}}^{\mathbf{p}} |\beta_{\mathbf{j}}| + \lambda_{\mathbf{2}} \sum_{\mathbf{j=1}}^{\mathbf{p}} \beta_{\mathbf{j}}^{\mathbf{2}}. \tag{2}$$

where the two tuning parameters are $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$. The sparsity in the regression coefficients is encourage by the first tuning parameter of λ_1 whereas the grouping effect is encouraged from the second tuning parameter λ_2 in Equation(2). Meanwhile, increasing λ_1 shrinkage value will subsequently decrease the number of explanatory variables selected. On the other hand, there is another tuning parameter which is α, a mixing proportion of the LASSO (L_1) and ridge (L_2) penalties with values of (0, 1). Under the assumption of$\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$, a linear combination of both penalties can be obtained by rewriting Equation (2) as:

$$PLR_{Elastic}(\beta;\alpha) = (y - \beta\mathbf{X})^{\mathbf{T}}(\mathbf{y} - \beta\mathbf{X}) + (\lambda_\mathbf{1}) + (\mathbf{1} - \alpha)\sum_{\mathbf{j=1}}^{\mathbf{p}} |\beta_\mathbf{j}| + \alpha\sum_{\mathbf{j=1}}^{\mathbf{p}} \beta_\mathbf{j}^\mathbf{2}. \tag{3}$$

Equation (3) is a ridge regression if $\alpha = 1$. Contrariwise, it becomes the LASSO regression if $\alpha = 0$ in Equation (3). Noteworthy, when the value of α is between 0 and 1, Equation (3) can shrinks the coefficients as ridge regression and select variables as LASSO [5].The penalized linear regression using elastic net regression coefficient estimates can be obtained as:

$$\hat{\beta}_{PLR}^{Elastic} = argmin_\beta PLR_{Elastic}(\beta;\alpha). \tag{4}$$

Worthy of note that Equation (4) is the naïve elastic net. The solution to this is known as the naïve elastic estimates. According to [7], the purpose of the naïve elastic estimates is to obtain the correct elastic net estimates. Hence, by rescaling the naïve estimators, the elastic net estimator is given by

$$\hat{\beta}^{Elastic} = (1 + \lambda_2)\hat{\beta}_{PLR}^{NaïveElastic} \tag{5}$$

It is remarkably known that elastic net has better performance than of LASSO particularly in encouraging grouping effects on the highly correlated variables. Furthermore, the prediction accuracy is better than of LASSO due to combination of both penalties in elastic net. However, it lacks the oracle properties of selecting the variables consistently which is similar to LASSO [9].

OHPL

Ordered homogeneity pursuit LASSO or OHPL was introduced by [4] to select informative variables that are highly correlated in high dimension data. OHPL can be considered as an improved version of LASSO but with two added advantages. The selection of number of observations is greater than of LASSO and the grouping effect which are homogenous in the response variable can naturally be identified using OHPL. In general, OHPL is a combination of fisher optimal partitions and LASSO method. The fisher optimal partitions functions as the baseline for constructing the selection algorithm intervals. In this step, fisher optimal partitions produces different width of variable intervals. Meanwhile, the next step involves applying LASSO in these intervals to obtain the subset of optimal variable.

To perform variable selection using OHPL, the groups are firstly constructed by using the homogeneity in regression coefficients. Then, the extraction process was carried out on the groups that have the most correlated explanatory variables in each group. This groups of selected variables are known to have the most informative variable and is computed as follows:

$$\mathbf{x}_{\Re_i} = argmax_{i \in Gr_j} |\mathbf{x_i^T y}| \tag{6}$$

where the $\mathbf{x_i}$ is standardized, $\mathbf{y}$ is centered and Gr_j is the group at $j - th$. From Equation (6), LASSO is then applied to these selected group of variables. The final stage involves deriving these group of variables to develop a partial least square model. It is noteworthy that there are three tuning parameters in the OHPL algorithm steps, namely the number of components K, number of groups, g and λ_1 LASSO penalty. The selection of these tuning parameters can be carried out using the cross-validation technique [4].

Performance Assessment

For evaluation purpose, the dataset were divided into two sets. The training set were used to developed the model, choosing tuning parameters and implementing variable selection. Meanwhile the test set is used to assess the model. By using the root mean square error (RMSE), the performance of LASSO, elastic net and OHPL were used as the performance evaluation. RMSE were calculated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y} - y)^2}{n}} \tag{7}$$

where $\hat{y}$ is the predicted response value and n is the number of samples.

Meanwhile, coefficient of determination is another validation method that were used to assess the performance of the methods. The closer the values to 1, indicates the better regression model fits the data. The formula for coefficient of determination is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{8}$$

where $\hat{y}$ is the average of predicted value of response.

RESULT

The stock market price consists of 4362 stock market prices over the period of 30 years from 1987 to 2017, including all sectors in the stock market. Since missing values is not the main concern here, therefore the process of screening and removal were carried out on the missing observations. For instance, those that have at least one missing value and where no recorded data is available, were subsequently omitted from the observations. The finalize data consisted of 901 stock market prices. S&P500 index was the response variable used in this study.

The application of this yearly stock market price in OHPL, elastic net and LASSO were carried out to perform variable selection and model estimation. In order to determine the tuning parameters of λ in both elastic net and LASSO, a grid values between 0.01 and 100000 were produced. The next step involves performing a 5-fold cross-validation, following [7] suggestion. The minimum λ obtained using cross-validation, were applied to perform the analysis. The g were generated as 10 and cross-validation were taken in 5-fold, to obtain the optimal λ. Therefore, the comparison result of LASSO, elastic net and OHPL using yearly stock market price is shown in Table 1.

TABLE 1. Performance result of yearly stock market price

Method	RMSE	R^2	Variables Selected
LASSO	109.40	0.985	38
Elastic Net	95.54	0.991	171
OHPL	71.52	0.980	310

The result for the stock market price also shows the capability of OHPL to improve the accuracy of the original LASSO. This is in a good agreement with [4] by using spectroscopic data. Elastic net selected more variables than of LASSO. Contrariwise, OHPL outperformed the number of selected variables compared to elastic net and LASSO. Unlike elastic net, OHPL have shown the lowest number of RMSE indicating that this method can yield a better model performance as well as better prediction power. On the other hand, LASSO had a limited ability in performing variable selection as well as model performance. This is apparently presented by its higher RMSE and lower number of selected variables value. The result is in agreement with the findings of [4].

From Table 1 result, it shows that OHPL increased the model performance as compared to its counterpart. Moreover, since OHPL is an extension from the original LASSO, therefore the results had proven that it has a higher variable selection power. In addition, the high correlation that exists between the stock market price variables are better represented by OHPL method. Although the coefficient of determination was the lowest in OHPL, nonetheless not much of a difference was observed when compared with elastic net and LASSO. Notwithstanding, the findings in the performance result also showed that the OHPL method can be applied in stock market price, particularly in the study of highly correlated variables as well as for prediction.

CONCLUSION

OHPL which is an improved version of LASSO had effectively show that it had greater selection of number of variables than of LASSO. Moreover, with the homogeneity feature that OHPL possess, it is able to encourage the grouping effect which are homogenous in the application to stock market price. Besides that, OHPL had better prediction accuracy in comparison with LASSO and elastic net. OHPL is more suitable for high dimension data that have similar group structure. Likewise, the explanatory variables in the yearly stock market price have successfully presented that it also owns this property that is applicable to OHPL.

ACKNOWLEDGMENTS

The first author expressed her gratitude to Ministry of Education Malaysia for the scholarship. This study was partially funded by Universiti Teknologi Malaysia under Industry-International Incentive Grant [grant number Q.J130000.3001.01M30].

REFERENCES

[1] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications* (Springer Science & Business Media, 2011).
[2] Z. Y. Algamal, M. H. Lee, A. Al-Fakih, and M. Aziz, SAR and QSAR in Environmental Research **27**, 703–719 (2016).
[3] J. Chen and Z. Chen, Statistica Sinica 555–574 (2012).
[4] Y.-W. Lin, N. Xiao, L.-L. Wang, C.-Q. Li, and Q.-S. Xu, Chemometrics and Intelligent Laboratory Systems **168**, 62–71 (2017).
[5] F. S. Kurnaz, I. Hoffmann, and P. Filzmoser, Chemometrics and Intelligent Laboratory Systems **172**, 211–222 (2018).
[6] R. Tibshirani, Journal of the Royal Statistical Society: Series B (Methodological) **58**, 267–288 (1996).
[7] H. Zou and T. Hastie, Journal of the royal statistical society: series B (statistical methodology) **67**, 301–320 (2005).
[8] Z. T. Ke, J. Fan, and Y. Wu, Journal of the American Statistical Association **110**, 175–194 (2015).
[9] Z. Y. Algamal and M. H. Lee, Computers in biology and medicine **67**, 136–145 (2015).