The Dimensional Reduction of Correlation Matrix for Linear Regression Model Selection

[1]HABSHAH MIDI, [2]HASSAN S. URAIBI

[1]Department of Mathematics, Faculty Science, Universiti Putra Malaysia,43400, UPM. Selangor. MALAYSIA
[1,2]Laboratory of Applied and Computational Statistics,Institute for Mathematical Research, University Putra Malaysia, 43400, UPM. Selangor. MALAYSIA
[2]College of Administration & Economics,University of Al-Qadisiyah,P.O Box 244, Diwanyiah, IRAQ
[1]habshah@upm.upm.edu.my, [2]hssn.sami1@gmail.com

Abstract:

This article is concerned with the problem of model selection in high dimensional data. We propose a method that can reduce the time for selecting only the variables which provide important information to the response variable. We call this method the Dimensional Reduction of Correlation Matrix (DRCM). Our proposed procedure based on two steps, whereby in the first step, DRCM attempts to reduce the dimension of correlation matrix by including only those variables that have absolute correlations greater than a threshold value, in the potential model. In the second step , the p-values for the parameter estimates of potential model were computed.. The final regression model only include those variables that are significant. The DRCM is compared with the existing Adaptive Lasso and VIF regression techniques. The result shows that the DRCM is more efficient than the existing methods in terms of reducing the time taken for selecting the best model.

Keyword: Variable selection, Adaptive lasso , VIF Regression, DRCM

1. Introduction

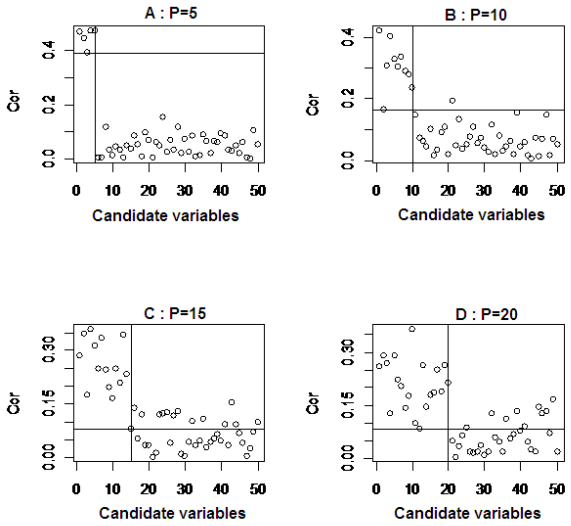
The aims of variable selection is to select the most predictive variables among an enormous number of potential variables. High Dimensional (HD) data formed a major challenge for the statistical practitioners that employ classical statistical methods. A rich literatures have been developed in recent years to overcome the problem of HD data such as Least Absolute Shrinkage Selection Operator (LASSO) (Tibshirani,1996), adaptive LASSO (Zou,2006), elastic net (Zou and Hasti,2005), Least Angle Regression (Tibshirani et al,2004), Dantzig (Candes and Tao, 2007). Lin et al. (2012) proposed Variance Inflation Factor (VIF) regression as a fast algorithm for variable selection. LASSO's method is based on marginal correlation between X's and Y denoted as (R_{XY}) , whereby the first potential variable that enter to be in the candidate set is the one that poses the highest absolute correlation than others.

Fig. 1 exhibits the absolute correlation between each of the candidate variable with Y. Here, we wish to show the problem that will be encountered regarding the correlation values for high dimensional data. The horizontal line in the diagram corresponds to the minimum value of the absolute correlation (MVAC) between each artificial (actual) potential variables(denoted as p) and Y. If there is no problem, then regardless of the number of p, none of the correlation between the noise variable and Y will exceed MVAC. Hereafter, we will see that this is not true for high dimensional data. It can be seen from Fig. 1-A that the correlations between Y and the artificial potential variables lie far from the rest of the correlations, especially, when the dimension of covariate is low. However, by increasing the dimension of p, not only the correlations of the artificial potential variables get closer to the rest of the correlations, but some of the correlation of the noise variables become larger than MVAC. For example, increasing the value of p to 10 make the correlation $|R_{X_{22}Y}|$ to be higher than the $|R_{X_1Y}|$, and $|R_{X_{40}Y}|$, $|R_{X_{11}Y}|$ and $|R_{X_{47}Y}|$ to be very close to

ISBN: 978-960-474-372-8

$|R_{X_1Y}|$. On the other hand, 13 and 9 noise variables appearing for p=15 and 20, respectively which should not be the case when there is no high dimensional problem.

Fig 1: The correlations between X's and Y.



It can be observed from Fig. 1 that by increasing the dimension of the artificial potential covariates, some noise covariates appeared to be highly correlated with Y. In this situation, they are included in the potential candidate set and we call them noise variables. It is important mentioning that some of the noise variables that enter the final model may be larger than the number of actual potential variables. This case is called over-fitting model which usually occurs in HD data and may give a misleading results.

It is important to note that the VIF regression selection technique which is assumed faster than the penalized methods, is not that efficient after all, because it is affected by the choice of two parameters (initial wealth and investment). Lin et al. (2012) pointed out that one may use larger parameters for initial wealth parameter (0.5) and larger investment parameter such as 0.05. By avoiding to select the optimal parameters in the VIF regression technique will consume more time in selecting the best model. On the contrary, the penalize method which choose (λ) in each iteration might not help in reducing the computation time, because the method will select more noise variables in the final model.

The weakness of these two methods is that they did not clearly mention the optimal procedure of choosing the relevant parameters needed in their algorithms. Moreover, they employed a very complex iterative algorithms. Hence we propose a simple DRCM procedure which is based on the reduction of the dimension of the correlation matrix. We expect that this method is more efficient and less time consuming than the adaptive lasso and VIF regression techniques.

2. The Idea of DRCM

By standardizing all variables considered in the Euclidean space will help to reduce the time for calculating the correlations between variables. The correlation between X and Y can be written in terms of cosine $\theta_{X,Y}$

$$\mathrm{Cos}(\theta_{X,Y}) = \frac{< X, Y >}{\|X\|.\|Y\|} = \frac{\mathrm{COV}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

$$= \mathrm{Cor}(X,Y) \qquad (1)$$

where normalization $\|X\|$ and $\|Y\|$ to 1 (unit length) and $\sum_i Y_i = 0$, $\sum_i X_i = 0$, $i = 1, \ldots, m$.

Let $Y = X.\beta +$ □ be the linear regression model where Y is the response variable, X is the matrix of covariates. β is the regression coefficients and □ is the errors. Assuming Y and X are scaled, then $\hat{\beta} = (X^TX)^{-1}X^TY$ $= X^TY$, since $X^TX = I$, and $\frac{1}{n}X^TY = \frac{1}{n}\hat{\beta} = R_{XY}$ is the correlation between X's and Y. Subsequently, the regression estimates can be obtained from the correlation matrix: i.e $|\hat{\beta}| = |R|_{XY}$ where the value of $|R|_{XY}$ is between 0 and 1. When the value of correlation between X and Y equals to 1 and 0, this is an indication of a perfect and no correlation between the variables, respectively. Hence, if all p potential variables are perfectly correlated with Y then $\sum|\hat{\beta}| = \sum|R|_{XY} =$ P, while imperfect correlation indicated by $\sum|\hat{\beta}| = \sum|R|_{XY} = 0$. However, in real practice, these two cases are impossible to occur.

It is important to note that the contribution of each estimated coefficient rely on its numerical value in the $[0,1]$ interval. The choice of the candidate variable to be included in the potential model is very important. The correlation between the

ISBN: 978-960-474-372-8

candidate variable and the response variable should be reasonably high. This decision can be very subjective.

Therefore, a threshold value is needed to identify candidate variables that can be considered in the potential model. This value denoted as M is computed based on the average of absolute correlation between X and Y.

Suppose j_0 be the number of all candidate covariates, the threshold value M is given by $M = \frac{\sum_{i=1}^{j_0} |R_{XY}|}{j_0}$. A variable that has absolute correlation with the response variable which is greater than the threshold value, i.e $|R_{XY}| \geq M$, will be included in the potential model. Consequently the dimension of the correlation matrix is reduced. Subsequently, the parameters of the potential model with j_1 variables are estimated. The final model (in the active model) consists of j_2 variables (covariates) ; those covariates that have P-values less than 0.05.

3. A Simulation Studys

To compare DRCM procedure with adaptive LASSO and VIF regression, we carried out a simulation study similar to Khan(2007) where correlation between the covariates is weak and no outlier in the dataset. We consider d=50 candidate variables and $P = 5, 10, 15 \text{ or } 20$ artificial (real) potential variables.

The candidate variables were generated from $X_j \sim N(0,1)$. The response variable Y is generated using P non-zero covariates, with coefficients (7,6,5,7,6) , repeated one time for $P = 5$, two times for $P = 10$, three times for $P = 15$ and four times for $P = 20$. The error is chosen $\varepsilon_i \sim N(0,1)$. For each case, we generate 1000 simulation data sets. The average number of potential variables, noise variables and time taken to select the final model are presented in Tables 1. A good method is the one that has average potential covariates in the final model which is closer to the number of the generated artificial (actual) potential covariates, has the least value of average noise variables and consume the least time to select the final model.

It is interesting to note that irrespective of the number of candidate variables, all the three methods select the same number of potential covariates in the final model. Nonetheless, the DRCM has the least value of the average number of noise variables and the least time taken to select the final model, followed by the Adaptive Lasso and VIF Regression methods.

4.Conclusion

In this article, we proposed a new method for linear regression model selection. The simulation results suggest that, for the generated Y observations which are based on linear relationship with artificial potential covariates greater than 14 (high dimensional data), will make some noise variables to be included in the final model. The more dimension of the artificial potential covariates being considered, the more noise variables will appear. Although Adaptive lasso performs better than the VIF Regression technique, the time taken to select the final model still can be considered high. The DRCM not only has the smallest average number of noise variables, but also can significantly reduce the time taken to select the final model. Hence, we can conclude that our proposed DRCM is more efficient than the existing VIF Regression and Adaptive Lasso methods.

References

[1] Candes, E. J., and Tao, T. ,The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n, The Annals of Statistics,35,6,2007,2365–2369.

[2] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R, Least Angle Regression (with discussion) , The Annals of Statistics,32,2, 2004, 407–499.

[3] Lin, D., Foster, D. P. and Ungar, L. H., VIF regression: A fast regression algorithm for large data. J. Amer. Statist. Assoc. 106,493,2011, 232–247. [4] Khan. A. Jafar, Aelst. Van Stefan, Zamar. H. Ruben, Building a robust linear model with forward selectionand stepwise procedures, J. Computational Statistics & Data Analysis, 52, 2007,239-248

[5] Tibshirani, R.,Regression shrinkage and selection via the Lasso. J. Roy. Statist.Soc. Ser. B, 58 ,1,1996, 267-288.

[6] ZOU, H. ,The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101.2006, 476,1418-1429.

[7]Zou,H. and Hastie,T, Regularization and variable selection via the elastic net, . J. R. Statist.Soc. Ser. B, 67,2,,2005,301-320.

ISBN: 978-960-474-372-8

Table 1 Average potential , noise variable and time taken for d=50,100,150

P	DRCM			VIF R			Adaptive Lasso		
	avg. of potential	avg. of noise	avg. Time	avg. of potential	avg. of noise	avg. Time	avg. of potential	avg. of noise	avg. Time
	d=50								
5	5	0.07	0.03	5	0.82	0.261	5	0.15	0.075
10	10	0.06	0.04	10	1.35	0.120	10	0.13	0.081
15	15	0.09	0.04	15	3.48	0.204	15	0.14	0.078
20	20	0.02	0.04	20	10.0	0.407	20	0.14	0.082
	d=100								
5	5	0.16	0.08	5	0.97	0.48	5	0.16	0.221
10	10	0.15	0.09	10	1.64	0.14	10	0.18	0.234
15	15	0.14	0.07	15	4.95	0.24	15	0.18	0.176
20	20	0.10	0.08	20	23.4	0.84	20	0.18	0.197
	d=150								
5	5	0.23	0.105	5	0.85	0.633	5	0.165	0.368
10	10	0.19	0.102	10	1.87	0.128	10	0.210	0.345
15	15	0.18	0.117	15	8.34	0.396	15	0.165	0.433
20	20	0.18	0.117	20	37.4	1.670	20	1.700	0.389