High dimensional regression coefficient compression model and its application

Liyuan Gao, Yongmei Ding* and Linxin Zhang
Institute of Systems Science, Wuhan University of Science and Technology, Wuhan, Hubei, 430065, China
*Corresponding author’s e-mail: dingyongmei@wust.edu.cn

Abstract. From stock market risk to genetic data survival analysis to energy consumption impact analysis, statistical modeling of high-dimensional regression plays an important role in different fields. Based on the financial data of China for the past 15 years, we select fifteen predictors related to fiscal revenue, design a ten-fold cross-validation algorithm based on the Ridge Regression and Lasso Regression models. Empirical examples show that Lasso Regression is a great way in big data modeling by comparing the cross-validation mean square error and the equation interpretation ability, which achieves the process of coefficient compression.

1. Introduction
Nowadays, high-dimensional data structure is becoming more and more common in many fields, such as finance, medicine, environment, etc, which is accompanied by the dimensionality disaster in the process of statistical modelling. So it is very important to find an effective way as a data-driving modelling method for dimensionality reduction and effective prediction.

As a hot topic with big data science and Artificial Intelligence, high-dimensional data analysis has been attracted by many scholars, and most of them focused on establishing some regression models. In 1998, Fu [1] analyzed the prostate cancer data in his paper, compared bridge regression with ordinary least squares regression, Lasso Regression and Ridge Regression, and found that bridge regression has better effect. Statistical selection is the basis of multivariate statistical models, including non parametric regression. In the process of selecting variables, it is inevitable to ignore the random error. Fan and Li [2] proposed the penalty likelihood method (SCAD) in 2001 to improve the Lasso Regression. Compared with other variable selection methods, this method has better superiority and sufficient accuracy. There are still some shortcomings in the Lasso Regression. Therefore, in Zou [3-4], the elastic network and adaptive Lasso Regression of Lasso Regression were proposed in 2005 and 2006 respectively. His real data and simulation studies show that the elastic network is superior to the Lasso Regression and has similar representation sparsity. It is especially suitable when the explanatory variable p is much larger than the sample size n. In 2007, Wang et al. [5] also proposed relaxed lasso, which solved the convergence speed problem of Lasso Regression in some sparse high-dimensional data, which made the prediction error smaller. Spencer et al. [6] proposed a model improvement of Lasso Regression in the historical data of intelligent house sensors. It adjusted the Lasso Regression coefficient according to the shape of the error curve to improve the prediction accuracy. Mishra et al. [7] discussed and experimented with different weak regressions such as linear regression, Lasso Regression, Ridge Regression, and support vector machine regression by studying the climate characteristics of three coastal regions in India. Domestic scholars have also carried out a
series of useful explorations on high-dimensional regression. In 2017, Yang J. W \[8\] analyzed the stock transaction data, analyzed the cox proportional hazard regression model, the Lasso method and the elastic net method, and then found the group effect property of the elastic net method, which was applicable to many variables and correlations, and required smaller sample size of data.

Ridge Regression and Lasso Regression are also used in dimensionality reduction or regression coefficient compression techniques. Lasso Regression achieves dimensionality reduction through parameter reduction. Ridge Regression is the method that it uses variable constraints to make variable selection. Are these coefficient compression regression methods so good? In order to answer above question, in this paper, we starts from the statistical modeling, and establishes two models of Ridge Regression and Lasso Regression respectively, and designs a cross-validation algorithm. The empirical data is used to cross-validate the mean square error comparison to realize the model comparison and explore the different models. The effects and capabilities of the coefficient compression, and the advantages and disadvantages of different models are judged.

2. Theoretical inquiry

2.1. Ridge Regression

In the least squares estimation, the least squares regression is fitted by minimizing the following functions to estimate the coefficients:

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$  \hspace{1cm} (1)

Ridge Regression is very similar to least squares except that its coefficients are estimated by minimizing a slightly different formula. In particular, ridge estimation coefficient estimate \(\hat{\beta}_R\) is minimized by the following formula:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$  \hspace{1cm} (2)

Among them \(\lambda\) is an adjusting parameter. In addition to seeking the estimator that fits the data better by minimizing RSS, it also contains another term, \(\lambda \sum_{j=1}^{p} \beta_j^2\) called compression penalty term, The two norm of \(\beta_j\), adjusting parameters \(\lambda\) it is used to control the relative influence of these two items on the estimation of regression coefficients. The ridge estimate was proposed by the famous statisticians Hoerl and Kennard in 1970, and they calculated \(\lambda\) as follows:

$$\lambda = \frac{\sigma^2}{\max \alpha_i^2}$$  \hspace{1cm} (3)

The ridge estimate \(\hat{\beta}_R\) is an estimated class of \(\lambda\) that changes as the value of \(\lambda\) changes. If I remember \(\hat{\beta}_R(\lambda)\) by \(\hat{\beta}_R\) the first component, It is the unary function of \(\lambda\). When \(\lambda\) changes on \([0, +\infty]\), the graph of \(\hat{\beta}_R\) is called a ridge map. The ridge method for selecting \(\lambda\) is:

Firstly, the ridge traces of \(\hat{\beta}_R^1(\lambda), \hat{\beta}_R^2(\lambda), \ldots, \hat{\beta}_R^p(\lambda)\) are drawn in the same two-dimensional coordinate system, and then \(\lambda\) values are selected according to their trend of ridge traces change, so that the ridge estimates of P regression coefficients are no longer greatly changed, and are near a fixed level, and the positive and negative symbols of each regression coefficient are required to conform to economic significance. From the expression \(\hat{\beta}(\lambda) = (X^T X + \lambda I)^{-1} X^T y\) of ridge estimation, we can see that the bigger \(\lambda\) is, the bigger the deviation between ridge estimation and least squares estimation is. Therefore, the corresponding RSS will increase with the increase of \(\lambda\) value. Therefore, when selecting ridge estimation, RSS should not increase too much. To sum up the above points, the \(\lambda\) value can be selected according to the specific situation.
2.2. Lasso Regression

In 1996, Robert Tibshirani first proposed Lasso (Least absolute shrinkage and selection operator) regression. This method is compression estimation. By estimating the coefficients, some coefficients are compressed to 0, and variable selection is realized. Compared with Ridge Regression, Lasso Regression overcomes the disadvantage that Ridge Regression can not select variables. The coefficients of Lasso Regression are minimized by solving the following equation:

$$\hat{\beta} = \arg \min_{\beta} \left( \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

The Lasso Regression implements the process of selecting variables by introducing a penalty term \( \sum_{j=1}^{p} |\beta_j| \), a norm of \( \beta_j \). At present, the methods for solving Lasso are mainly divided into two types, namely the coordinate reduction algorithm and the Lasso correction of the minimum angle regression.

2.3. Cross Validation

In order to compare the coefficient compression ability of the four models of Ridge Regression and Lasso Regression better, the k-fold cross-validation method is used to solve the parameters of the model, and the cross-validation mean square error of the two models is calculated. The k-fold cross-validation method is an improvement of the leave-one-out cross-validation method. Firstly, the n observations of the data set are randomly divided into k groups of substantially the same size, or referred to as folds. Secondly, by using the first group as the verification set and then building the model on the remaining k-1 groups, the mean square error MSE_1 can be calculated from the observations of the verification set. To repeat above steps for k times, and to take a different observation group as a verification set at each time. Throughout the process, we will get an estimate of k test errors, namely MSE_1, MSE_2, ..., MSE_k. By averaging these values, we can get a k-fold cross-validation test estimate:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

3. Empirical study of high dimensional data regression analysis

Fiscal revenue is an important indicator of the national economy. It can not only comprehensively reflect China's economic resource allocation, but also evaluate China's overall development level. This paper mainly analyzes the influencing factors of China's fiscal revenue by using the financial data of China in the past 15 years. A total of 15 predictors are selected, as shown in Table 1:

| Variable | Variable Meaning | Variable | Variable Meaning |
|----------|------------------|----------|------------------|
| Y        | General public budget income | X_8      | Consumer Price Index |
| X_1      | Total number of employed persons by three industries | X_9      | Total import and export volume |
| X_2      | Number of State-owned and Collective Workers | X_10     | GDP |
| X_3      | National resident per capital disposable income | X_11     | Total energy consumption Total investment in industry |
| X_4      | Total population | X_12     | Total energy consumption |
| X_5      | Total investment in fixed assets | X_13     | Total investment in industry |
| X_6      | Primary industry | X_14     | General public budget expenditure |
| X_7      | Ratio of output value of tertiary industry to secondary industry | X_15     | Highway mileage |
3.1. Ridge Regression

Figure 1 is a ridge map of the Ridge Regression fitted according to the data. When the ridge of each regression coefficient no longer has a large change, and is near a certain stable value, the value of $\lambda$ can be roughly determined. However, due to the large error of the $\lambda$ value obtained by the ridge map observation, the fitted model is not accurate enough. The ten-fold cross-validation method is used to determine the optimal $\lambda$ value and fit the model; Figure 2 is a cross-validation error map corresponding to different $\lambda$ values. It can be determined from the figure that the $\lambda$ value that minimizes the cross-validation error is 7.63. By taking $\lambda = 7.63$ into the test set of the Ridge Regression model, the mean square error MSE=0.202 of the model can be calculated. Finally, based on all the data variables, the $\lambda = 7.63$ re-fitting Ridge Regression model obtained by cross-validation method is used to calculate the regression coefficient of the model and the standardized model caliber as follows:

$$Y = -1.762 \times 10^{-16} + 0.03580 X_1 + 0.04575 X_2 + 0.04906 X_3 + 0.04148 X_4 + 0.04935 X_5$$
$$+ 0.0489 X_6 + 0.04498 X_7 - 0.01007 X_8 + 0.04430 X_9 + 0.04909 X_{10} + 0.0497 X_{11}$$
$$+ 0.04898 X_{12} + 0.04831 X_{13} + 0.04938 X_{14} + 0.04010 X_{15}$$

(6)

The model fitted by Ridge Regression (6) can be seen that the number of predictors affecting the response variable is still 15, although it plays the role of coefficient compression. This model does not compress its coefficient to zero, and its predictor has the effect of uniform distribution.

3.2. Lasso Regression

Figure 3 shows the path of the Lasso Regression. It can be observed that the rightmost digits are sorted from top to bottom: 12, 1, 9, 10, 4, 6, which indicates that the variable that has an increasing influence on the response variable Y has $X_6, X_{10}, X_{12}$, the less influential variables are $X_1, X_4, X_9$, while the coefficients of other predictors are compressed to zero. That is to say, the lasso correction using the minimum angle regression can play the role of variable selection, and the original 15 variables can be reduced to about 6, which greatly reduces the complexity of the model.
Figure 3. Lasso correction path for Minimum angular regression.

Figure 4. Cross validation error.

Figure 4 is a cross-validation error map corresponding to different $\lambda$ values. From Figure 4 and according to the R software, the $\lambda$ value that minimizes the cross-validation error is determined to be 0.0233. By taking $\lambda = 7.63$ into the test set of the Ridge Regression model, the mean square error $\text{MSE} = 0.00131$ of the model test set can be calculated. Finally, based on the entire data set, the Lasso Regression model was re-fitting using the $\lambda$ value obtained by cross-validation. Finally, the regression coefficient of the model and the standardized model caliber are as follows:

$$Y = -1.455 \times 10^{-16} + 0.006066X_1 + 0.01868X_2 + 0.3979X_6 + 0.3418X_9 + 0.02595X_{12} + 0.1877X_{14}$$  (7)

The model (7) fitted by Lasso Regression, and the most influential ones are $X_6$, $X_{10}$, $X_{14}$. The economic significance is that whenever $X_6$ is increased by one unit, $Y$ increases by an average of 0.3979 units. The primary industry is the lifeline of the country and has the greatest impact on fiscal revenue. When GDP increases by one unit, $Y$ increases by an average of 0.3418 units. Whenever $X_{14}$ is added to a unit, $Y$ increases by an average of 0.1877 units, which means that the higher the consumption, the more fiscal revenue can be promoted. The less affected ones are $X_2$, $X_5$, and $X_{12}$. This means that whenever $X_2$ is increased by one unit, $Y$ increases by an average of 0.006066 units, which indicates that the employment rate of state-owned enterprises also has a certain impact on fiscal revenue; whenever $X_5$ is increased by one unit, $Y$ increases by an average of 0.01868 units; China as an agricultural large countries, their agriculture also has a certain impact on fiscal revenue. When $X_{12}$ is increased by one unit, $Y$ increases by an average of 0.02595 units.

3.3. Model comparison

In this section, we mainly compare the prediction accuracy and model interpretation ability of various regression models. The prediction accuracy is evaluated by the mean square error of the cross-validation test set, and the model interpretation ability is evaluated by the fitted model equation. It can be seen from Table 2 that for the regression analysis of high-dimensional data, the choice of Lasso Regression implementation variables is better than other methods.

| Regression model       | Cross-validation test set mean square error |
|------------------------|--------------------------------------------|
| Ridge Regression       | 0.202                                      |
| Lasso Regression       | 0.00131                                    |

The Lasso Regression model (7) can be derived from the real data, also been deduced the experimental model by Ridge Regression technique, which is expressed by (6). The number of predictors in the model is still 15 and the length of the model is cumbersome. The Lasso Regression model has predictors. The numbers of variables are compressed to six, which is greatly reduced compared to the original 15 predictors, showing a strong compression capability. Combining the
above two factors, Lasso Regression has stronger coefficient compression ability and the mean square error is small, which also shows better prediction effect.

4. Conclusion
Fiscal revenue is an important indicator of the national economy. It can not only comprehensively reflect China's economic resource allocation, but also evaluate China's overall development level. Based on the financial data of China in the past fifteen years, this paper establishes the Ridge Regression and Lasso Regression models respectively, and uses the cross-validation test set and Mean Square Error as the accuracy index of the evaluation model. Compared with the traditional evaluation index and Ridge Regression, the Lasso Regression can repeatedly use the randomly generated subsamples for training and verification, and obtain a more reliable mode.

References
[1] Fu, W. J. J.(1998)Penalized regression: the ridge versus the lasso. Journal of Computational and Graphical Statistics, 7(3): 397–416.
[2] Fan, J. Q., Li, R. Z. (2001) Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association, 96(456): 1348–1360.
[3] Zou, H. (2005) Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, 67(2): 301–320.
[4] Zou, H. (2006) The adaptive LASSO and its oracle properties. Journal of the American Statistical Association, 101(476): 1418–1429.
[5] Wang, H. S., Li, R. Z., Tsai C. L. (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika, 2007, 94(3): 553–568.
[6] Spencer, B., Alfandi, O., Al-Obeidat, F. (2018) A Refinement of Lasso Regression Applied to Temperature Forecasting. Procedia Computer Science, 130: 728–735.
[7] Mishra, S. Mishra, D. Santra, G. H. (2017) Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: An empirical assessment. Journal of King Saud University - Computer and Information Sciences, 2017.
[8] Yang, J. G. (2017) Variable Selection Method in Cox Model and Empirical Study of Stock Market. Zhongnan University of Economics and Law.