Abstract. This paper analyzes the influence factors to explain the cross-section of expected returns. Our paper proposes the PCA-LASSO model and its reduced model, comparing them with the results of traditional linear regression model and LASSO regression model. The results show that the proposed PCA-LASSO model and its reduced model are superior to the existing two methods in the prediction of the stock returns of the industry. For practical applications, the research method proposed in this paper has wide applicability and the research conclusion can provide important reference for the securities investors.

Introduction

As the stock returns changing over time, it is affected by the large number of complex unknown factors in the economic development reality. What factors are related to the stock returns, which is a core issue in stock research, but also one of the issues that investors are most concerned about. For investors, the first is to reduce the risk of loss caused by the uncertainty of stock price. We must first understand the factors that affect the stock price, that is, to understand the influence factors of stock returns.

The study of foreign in securities began in the 1950s. In 1952, Markowitz [1] published the Portfolio selection in the Journal of Finance, which opened a new direction for modern securities investment. Based on the Markowitz model, the CAPM was developed by Mossin, Limmer and Sharpe (1964) [2]. The model gives the linear relationship between the returns and the market system risk. However, the empirical results show the CAPM model can’t explain the various phenomena that appear in the US stock market. Fama and MacBeth (1973) [3] used the cross-sectional sequence of stock excess returns to test the β values of these stocks and concluded. On the basis of predecessors, Fama and French [4-6] proposed a well-known three-factor model that gives relationship between the stocks or portfolio excess returns and the three factors. However, the model is still flawed, and it does not explicitly point out how these factors are associated with potential micro factors. Fama and French (2015) [7] form a five-factor model based on the three-factor model, which improves the goodness of the regression model. Harvey (2016) [8] proposed the problem of traditional quantitative selection factor method, and introduced the multiple test method to try to explain the problem of forecasting the cross section of earnings. And the paper mentions some research on data snooping and variable selection in predictive regressions, like Foster, Smith, and Whaley (1997)[9], Cooper and Gulen (2006)[10], and Lynch and Vital-Ahuja (2012)[11]. In terms of research methods, there are several ways to analyze stock returns in research methods. But the regression analysis is often used by domestic and foreign scholars. However, the traditional multivariate linear regression model is susceptible to extreme values, and the multiple collinearity between indexes is not fully taken into account, which may affect the validity of the model estimation.

In summary, studies mainly aim at exploring the factors that affect the stock returns. The research methods are relatively simple. However, compared with the traditional investment analysis technology, the LASSO model is a compression estimation method, which by constructing a penalty function to obtain a more refined model, making it compresses some coefficients to zero. But if LASSO regression is directly applied to the original data, the selected explanatory variables may still have a degree of multiple collinearity.
Therefore, this paper will propose a new PCA-LASSO model, which is a fusion of principal component regression and LASSO compression method, realizing the complete solution to the multi-collinearity problem in the estimation process and the quick choice of the obvious influencing factors. And we further put forward the reduced model of PCA-LASSO, which on the one hand to optimize the prediction accuracy of the model, on the other hand more clearly affect the returns of the important factors. The research of this paper not only focuses on the important factors that affect the returns, but also pay more attention to the improvement of the prediction accuracy of the stock returns.

Research Methodology

This section will show the four theoretical models used in this article.

(1) Multiple linear regression model

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i, \quad i = 1, \ldots, n. \]  

The dependent variable Y is affected by k explanatory variables. The sample size is n, and the regression coefficient \( \hat{\beta}_{OLS} \) is obtained based on the least squares method. In the actual study, one of the biggest problems faced by the multiple linear regression models is multi-collinearity. It is complicated to use the stepwise regression method for variable selection. In the study of industry stock returns, due to the small sample size, it faces with a serious lack of freedom of the problem.

(2) LASSO model

In 1996, LASSO (the absolute shrinkage and selection operator) method is proposed by Tibshirani. The sum of the absolute values of the model coefficients is used as a penalty term to compress the model coefficients under which the residual sum of the residuals is minimized. In this process, some unimportant regression coefficients can be sequentially compressed to zero, thus in the process of parameter estimation to achieve variable selection, and ultimately get a more refined model. Specifically, the LASSO method minimizes the sum of squares of residuals by constraining \( \sum_{j=1}^{k} |\hat{\beta}_j| \) less than a certain value. The coefficients are estimated as:

\[
\hat{\beta}_{LASSO} = \arg \min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} \beta_j x_{ji})^2, \quad \text{s.t.} \quad \sum_{j=1}^{k} |\beta_j| \leq t. 
\]  

The LASSO method can complete the selection of important explanatory variables in the process of parameter estimation to obtain the optimal model, which reduces the multiple collinearity of the model to a certain extent, but it is not completely eliminated.

(3) PCA-LASSO model

The observed data matrix of k explanatory variables is \([X_{n \times k}]\), and the observed vector of the dependent variable is \([Y_{n \times 1}]\). The PCA-LASSO model can be implemented in two steps: The first step, we have the principal component analysis of the explanatory variables and get the main component matrix and the factor load matrix. That is the following formula.

\[
[PC_{n \times k}] = [X_{n \times k}] \cdot [A_{k \times k}] ,
\]  

where \([PC_{n \times k}]\) is the principal component matrix, and contains k principal component column vectors. \([A_{k \times k}]\) is the factor load matrix. In the second step, the k principal components obtained in the first step are LASSO regression using the dependent variable Y, and the optimal compression degree is determined based on the cross-validation method, so as to select the influence of the dependent variable Principal component, the following principal component regression model based on LASSO is obtained.

\[
[Y_{n \times 1}] = [PC_{n \times k}] \cdot [\hat{\beta}^\text{pcaLASS}_{k \times 1}].
\]  

Among them, the partial regression coefficients in \(\hat{\beta}^\text{pcaLASS}_{k \times 1}\) are compressed to zero. Then, we realize the automatic selection of important principal components.
In the estimation of the PCA-LASSO regression model described above, the effect of each explanatory variable on the dependent variable can be accurately measured. We can obtain the following form:

\[ Y_{n \times 1} = [X_{n \times k}] \cdot [A_{k \times k}] \cdot [\hat{\beta}_{k \times 1}^{pcaLASS}] = [X_{n \times k}] \cdot [\phi_{k \times 1}] \cdot [\hat{\beta}_{k \times 1}^{pcaLASS}], \]  \hspace{1cm} (5)

Where the regression coefficient vector is \( \phi_{k \times 1} = [A_{k \times k}] \cdot [\hat{\beta}_{k \times 1}^{pcaLASS}] \), the degree of influence of each explanatory variable on the dependent variable is measured.

(4) Reduced model of PCA-LASSO model

It is noted that the above mentioned final PCA-LASSO model (5) where the regression coefficient vector \( \phi_{k \times 1} \) generally does not have a regression coefficient completely equal to zero, that is, all explanatory variables are involved in the interpretation of the dependent variable and prediction. We can get an optimal streamlined model based on the PCA-LASSO model. The specific approach is as follows:

(a) All the explanatory variables are arranged in descending order according to the order of the absolute values of the regression coefficients corresponding to \( \phi_{k \times 1} \).

(b) Using the dependent variable \( Y \) and the most important explanatory variable and its corresponding regression coefficient in \( \phi_{k \times 1} \), the first regression equation is constructed. The prediction error RMSE is calculated by CV method in the training data.

(c) In the model, we introduce a more important variable and its regression coefficient. We will get a total of \( k \) regression equations and calculate the RMSE for each equation in the training data set using CV method.

(d) In the \( k \) regression equations, we select the smallest RMSE model as a simplified model of PCA-LASSO.

In general, the PCA-LASSO model based on the above steps will have higher prediction accuracy, and the reduced model also completes the selection of important explanatory variables.

Data

According to the classification standard of the SFC industry, the information technology industry, a total of 161 listed companies is the research object. We select the second quarter, the third quarter of the financial data of 2016 for the study sample. Among them, the second quarter of 2016 (2016.06) data will be used to build the model and the third quarter of 2016 (2016.09) data will be used for out-sample forecasts. The data comes from financial data terminal iFinD.

Variable Selection

The variables of this paper are divided into the following six aspects: profitability, development ability, debt paying ability, operating capacity, cash flow capacity and valuation ability. There are a total of 42 secondary indicators shown in Table 1.

<table>
<thead>
<tr>
<th>First-level indicators</th>
<th>secondary indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profitability</td>
<td>X1</td>
</tr>
<tr>
<td>Development ability</td>
<td>X2</td>
</tr>
<tr>
<td>Debt paying ability</td>
<td>X3</td>
</tr>
<tr>
<td>Operating capacity</td>
<td>X4</td>
</tr>
</tbody>
</table>

Table 1. Indicator description.
Data preprocessing

Before the model analysis of the data, in order to eliminate the impact of different indicators and dimension units on the analysis results, we first standardized (normalized) the data of all the indicators. In this paper, the Z-score normalization method is adopted, that is $Z_i = \frac{X_i - \mu}{\sigma}$, where $\mu$ is the mean of the sample data and $\sigma$ the standard deviation of the sample data.

Empirical analysis

Several commonly used methods of studying the stock returns are analyzed, and the characteristics and possible problems of these methods are analyzed. On this basis, we propose our PCA-LASSO model and the simplified model of PCA-LASSO. A comparative analysis of the empirical results of the different models is given in the later section.

Estimation of parameter in sample

(1) For the multiple linear regression model, the following equation is obtained (the result is only two decimal places). The vast majority of variables are not significant, significant indicators are $X_{101}$, $X_{201}$, $X_{305}$ and $X_{308}$. Because of the serious multi-collinearity in the model, the result is not reliable, and only the regression equation is used as a reference.

$$\text{Ret} = 0.3X_{101^*} + 0.15X_{102} \ldots - 0.01X_{602} - 0.08X_{603} - 0.38^{***} \quad (6)$$

(2) The LASSO regression is established directly on the original data, and the 42 financial indexes are reduced dimension. We use CV interactive verification method to determine the optimal degree of compression. The MSE of the interactive verification is the smallest when $s = 0.08$, and the optimal estimation equation is as follows:

$$\text{Ret} = -0.03X_{207} - 0.01X_{305} - 0.03X_{311} - 0.01X_{406} + 0.01X_{504} - 0.07X_{603} \quad (7)$$

(3) Based on the PCA-LASSO model, the principal component analysis is carried out by using the PCA-LASSO model, and 42 principal component vectors and factor load matrices are obtained. These principal components are obtained from the original 42 observation variables in the orthogonal space projection, and did not lose the original information, but each principal component contains more concentrated information. And then we use LASSO regression to select the principal component that has important influence on the dependent variable. Here, we omit the specific principal component and factor load matrix values, using LASSO method to compress the 42 principal components of the trajectory shown in Figure 1. Figure 2 shows the relationship between the prediction error and the compression degree $s$ based on the CV method. The horizontal axis represents the degree of compression index $s$ and the value is between 0 and 1. When calculated by $s = 0.31$, the CV-MSE achieves a minimum value of 0.0684.
Based on the results of the optimal principal component regression, the regression results of returns and the principal components are obtained. In $\hat{\beta}_{kx1}^{pca\text{LASSO}}$, many regression coefficients are compressed to zero, thus completing the choice of the principal component. Table 2 shows the principal component whose coefficient is not zero.

<table>
<thead>
<tr>
<th>PC1</th>
<th>PC5</th>
<th>PC6</th>
<th>PC8</th>
<th>PC11</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.01080608</td>
<td>-0.00862424</td>
<td>0.049909092</td>
<td>-0.011080063</td>
<td>0.015123798</td>
</tr>
<tr>
<td>PC12</td>
<td>PC17</td>
<td>PC19</td>
<td>PC23</td>
<td>PC30</td>
</tr>
<tr>
<td>-0.017831577</td>
<td>-0.027978516</td>
<td>0.007090934</td>
<td>0.013672322</td>
<td>-0.023287705</td>
</tr>
</tbody>
</table>

As can be seen from Table 2 above, there are 10 main components that have an important effect on returns and the other principal component coefficients are pressed to zero. Based on the factor load matrix and the optimal LASSO regression coefficient, we can get the coefficient vector $\hat{\phi}_{kx1}$, which directly measures the influence of 42 indexes on Y. We sort these 42 coefficients according to the absolute value of the regression coefficients. The results are shown in Table 3. It can be seen from Table 3 that the top ten financial indicators with large impact on the returns are X101, X103, X311, X305, X409, X406, X308, X306 and X312.

<table>
<thead>
<tr>
<th>X101</th>
<th>X103</th>
<th>X311</th>
<th>X105</th>
<th>X305</th>
<th>X409</th>
<th>X406</th>
<th>X308</th>
<th>X306</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1164</td>
<td>-0.0978</td>
<td>-0.0614</td>
<td>0.0597</td>
<td>0.05743</td>
<td>-0.0499</td>
<td>-0.0486</td>
<td>-0.048</td>
<td>0.0457</td>
</tr>
<tr>
<td>X312</td>
<td>X404</td>
<td>X310</td>
<td>X405</td>
<td>X102</td>
<td>X104</td>
<td>X302</td>
<td>X205</td>
<td>X304</td>
</tr>
<tr>
<td>-0.0388</td>
<td>0.0344</td>
<td>0.0343</td>
<td>0.0343</td>
<td>0.0326</td>
<td>-0.0322</td>
<td>-0.0306</td>
<td>-0.03</td>
<td>-0.0267</td>
</tr>
<tr>
<td>X602</td>
<td>X204</td>
<td>X206</td>
<td>X203</td>
<td>X604</td>
<td>X307</td>
<td>X201</td>
<td>X402</td>
<td>X207</td>
</tr>
<tr>
<td>-0.0218</td>
<td>0.0212</td>
<td>-0.0199</td>
<td>0.0194</td>
<td>-0.0187</td>
<td>0.0160</td>
<td>-0.0159</td>
<td>-0.0151</td>
<td>-0.0127</td>
</tr>
<tr>
<td>X601</td>
<td>X303</td>
<td>X401</td>
<td>X407</td>
<td>X202</td>
<td>X501</td>
<td>X403</td>
<td>X408</td>
<td>X106</td>
</tr>
<tr>
<td>-0.0106</td>
<td>0.0095</td>
<td>-0.0080</td>
<td>0.0067</td>
<td>-0.0052</td>
<td>-0.0052</td>
<td>0.00417</td>
<td>-0.0041</td>
<td>-0.0016</td>
</tr>
<tr>
<td>X502</td>
<td>X504</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.0003</td>
<td>-0.0003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

According to the regression coefficients in Table 3, we can write the following regression equation of returns corresponding 42 indicators.

\[
Ret = 0.12X101 - 0.10X103 - 0.06X311 + 0.06X105 - 0.05X409 - 0.05X406 - 0.05X308 + \cdots - 0.0016X106 - 0.0003X502 - 0.0003X504
\]  

(8)
Based on the construction of the PCA-LASSO model in the previous section, the MSE of each test model is calculated using the training data set. When the model contains the first 6 indicators in Table 3, the MSE is minimized, resulting in the following PCA-LASSO reduced model (the coefficients are only two decimal places, see Table 3 for the specific coefficients). The following model provides the 6 indicators that ultimately affect the returns.

\[
\text{Ret} = 0.12X_{101} - 0.10X_{103} - 0.06X_{311} + 0.06X_{105} - 0.05X_{409} - 0.05X_{406}
\]  

(9)

Prediction and evaluation out of sample

This part applies the multiple regression model, LASSO model, PCA-LASSO model and its simplified model to the test the out of sample of test data, and compares the predicted results with the actual results to evaluate the effectiveness of the model. We use the RMSE (Root Mean Squared Error) to measure the accuracy of the prediction model.

Based on the training data, the test data sets are predicted respectively, using the four regression equations (6)-(9). The predicted root mean square errors are denoted RMSE1 to RMSE4 respectively. The results are shown in Table 4. It can be seen from Table 4 that the prediction effect of PCA-LASSO model is superior to LASSO model and general linear regression model. Using PCA-LASSO model can not only complete the selection of important indicators that affect stock returns, but also can obtain better prediction effect.

Table 4. RMSE of the four models.

<table>
<thead>
<tr>
<th></th>
<th>RMSE1</th>
<th>RMSE2</th>
<th>RMSE3</th>
<th>RMSE4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.403</td>
<td>0.366</td>
<td>0.290</td>
<td>0.200</td>
</tr>
</tbody>
</table>

In order to further verify the effectiveness of the PCA-LASSO model, this paper also selected 501 listed companies in manufacturing, 1097 listed companies in Shenzhen and 2009 listed companies all. The same method was carried out within the sample Model estimation and sample extrapolation and the comparison results of the four models are shown in Table 5.

Table 5. Comparison of four models based on different data sets.

<table>
<thead>
<tr>
<th>Industry</th>
<th>Multiple linear regression model</th>
<th>LASSO model</th>
<th>PCA-LASSO model</th>
<th>Reduced PCA-LASSO model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturing</td>
<td>0.255</td>
<td>0.250</td>
<td>0.189</td>
<td>0.182</td>
</tr>
<tr>
<td>Shenzhen listed</td>
<td>0.200</td>
<td>0.194</td>
<td>0.199</td>
<td>0.184</td>
</tr>
<tr>
<td>All listed</td>
<td>0.194</td>
<td>0.192</td>
<td>0.188</td>
<td>0.174</td>
</tr>
</tbody>
</table>

Based on the results of Table 4 and Table 5, we can see that in terms of influencing factors of the stock returns, when we study the industry data with smaller sample size, we need to select the variables. The PCA-LASSO model method is superior to other model methods. However, when the whole market data of the sample size is large enough, the prediction effect of the four models is very close. At this time, it is generally not necessary to deepen the variables from the point of view of prediction accuracy select.

Conclusions

In this paper, the PCA-LASSO model and its reduced model are proposed. Based on the cross-sectional data of the industry, the influencing factors of the stock returns are predicted. The model proposed in this paper has universal applicability for the model prediction problem with a small number of variables. And it can effectively solve some problems and shortcomings in the traditional linear regression and LASSO regression. It completely eliminates the multicollinearity
problem in the parameter estimation process. Another feature of the model approach is the ease of use, based on the existing R package can easily achieve the relevant model calculation. Therefore, an important contribution of this paper is to propose a new simple and feasible model method.

The important factors influencing the returns of listed companies are obtained, which has a good reference for investment decision-making. For listed companies in specific industries, it is possible to make accurate estimates of stock returns based on the latest cross-sectional data. It is a good investment reference value. The method mentioned in this paper can provide reference for investors. This paper creatively combines data mining methods to achieve cross-sectional regression of returns. Due to the limited time and effort, this paper only uses the principal component analysis and LASSO method to analyze the influencing factors of the returns. In order to provide investors with more valuable suggestions, we will use the most suitable data mining method in the future study.

Acknowledgement
Lu’s research was supported by the project of the National Social Science Foundation of China (no. 15BTJ016).

References