Research on Stock Price Multi-Factor Prediction Model Based on Bayesian Model Averaging

Ruiqi Zhu *

School of Finance, Nankai University, Tianjin, China, 300071

* Corresponding Author Email: shaloss@163.com

Abstract. Factor mining is a crucial component in constructing stock price prediction and quantitative models, where factor mining methods based on feature dimensionality reduction, such as Principal Component Analysis (PCA) and sufficient dimensionality reduction, are widely utilized. However, the connection structure between the quantitative factors extracted using these methods and stock prices is unknown, leading to potential issues like overfitting or underfitting in prediction models. In light of this, this paper proposes a multi-factor prediction model based on Bayesian model averaging. On one hand, the proposed method employs the concept of model averaging instead of model selection, effectively balancing the variance and bias of prediction models. On the other hand, it can adaptively choose sub-models that play a crucial role in predicting stock prices, thereby enhancing overall prediction accuracy. Empirical data analysis indicates that the proposed method, compared to PCA-based Lasso and Ridge regression, exhibits smaller mean squared error and possesses a certain level of robustness. Lastly, by incorporating other model averaging techniques such as the Bagging algorithm, the generalization ability of the proposed method can be further improved.

Keywords: Model Averaging, Stock Price Prediction, Sufficient Dimensionality Reduction.

1. Introduction

Quantitative investment is a type of investment strategy that involves analyzing and modeling financial markets using mathematical and statistical methods. The origins of quantitative investment can be traced back to the 1950s when Markowitz (1952) [1] introduced the Mean-Variance model, incorporating mathematical statistics into the study of asset investment. Sharpe, Lintner, and Mossin (1964) [2] proposed the Capital Asset Pricing Model (CAPM), suggesting that investing in high-risk stocks could yield higher returns for investors. Building upon CAPM, Fama and French (1992) [3] introduced the "Fama-French Three-Factor Model," further refining and improving previous models.

Stock price prediction models are generally divided into two types: time series-based prediction models and multi-factor-based prediction models. The former typically use price volatility information as input variables to predict future price trends. For example, Jarrett et al. (2011) [4] forecasted Chinese stock data using an improved ARIMA model, yielding accurate results. In recent years, Long Short-Term Memory (LSTM) neural networks have been applied to stock data prediction. Di Persio and Honchar (2017)[5] utilized RNN, LSTM, and GRU neural networks for Google stock price trend prediction, with LSTM showing more precise results in financial sequence forecasting. Ma (2020)[6] conducted a comparative study on ARIMA, ANN, and LSTM in time series prediction, demonstrating the significant advantage of LSTM in stock price prediction, albeit influenced by data processing methods. However, common time series models not only require stationarity of stock price sequences but also may lack sufficient volatility information to predict future trends accurately. Therefore, this paper focuses on using multi-factor models for stock price prediction, originating from the Fama-French Three-Factor Model mentioned earlier. The literature on multi-factor prediction models is extensive in academia. For example, Sirohi et al. (2015) [7] constructed a multi-kernel learning model based on technical factors such as opening and closing prices to predict stock price movements. Regarding the selection of value factor indicators, scholars argue that the earnings-to-price ratio is more suitable for the Chinese stock market (Liu et al., 2019) [8], with the earnings-to-price ratio factor showing a better value factor effect in China than the book-to-market ratio effect. Nevertheless, traditional multi-factor models face issues such as high collinearity among independent

variables, violating model assumptions, known as the dimensionality catastrophe. To reduce inter-factor correlations and decrease data noise, techniques like feature selection and dimensionality reduction can be employed. This paper focuses on dimensionality reduction methods due to their lower computational complexity. For instance, Chong et al. used principal component analysis to process financial time series data for better stock price prediction. The experiments showed that the dimensionality reduction applied to prediction models effectively alleviated the dimensionality catastrophe problem and improved stock price prediction accuracy [9]. However, unsupervised dimensionality reduction methods do not consider the impact of independent variables on the response variable. Supervised dimensionality reduction methods theoretically collect more information favorable for prediction than unsupervised methods, and Sufficient Dimensions Reduction is a representative method that does not rely on specific model assumptions [10]. This method theoretically reduces dimensionality without losing effective information about independent variables concerning the response variable. Nonetheless, whether using principal component analysis or sufficient dimension reduction to handle the dimensionality catastrophe, constructing prediction models often leads to issues of overfitting and underfitting. Scholars have proposed model selection criteria to address this problem, including stepwise regression [11], AIC, BIC [12], generalized cross-validation [13], Lasso [14], among others. However, model selection methods cannot avoid problems such as model lack of robustness and information loss [15].

Based on this, the paper develops a multi-factor stock price prediction model based on Bayesian model averaging. Firstly, the proposed method effectively balances the bias and variance of the prediction model through model averaging and addresses the dimension selection problem in sufficient dimension reduction and principal component analysis. Secondly, when using sufficient dimension reduction as the factor mining method, the model's flexibility is further enhanced by applying the Bagging algorithm to extend the proposed model, given its model freedom. In practical data analysis, the paper employs principal component analysis (PCA), sliced inverse regression (SIR), and sliced average variance estimation (SAVE) for dimensionality reduction of stock factors. Bayesian model averaging (BMA) is utilized as the prediction method. Compared with linear models based on principal component analysis and sufficient dimensionality reduction, such as ridge regression and LASSO regression, the proposed method's superior performance in stock price prediction is numerically demonstrated.

2. Theory and Methods

2.1. Principal Component Analysis

The fundamental principle of PCA is to transform a set of correlated random vectors into a set of new, uncorrelated random vectors through an orthogonal transformation. This is achieved by aligning these new vectors along the p orthogonal directions where the sample points are most dispersed. Subsequently, a dimensionality reduction process is applied to the multidimensional variable system, allowing it to be transformed into a lower-dimensional variable system with higher accuracy.

In practical terms, the first step involves standardizing and normalizing the sample matrix (dividing the difference between matrix values and column means by column standard deviations). Then, the correlation matrix is computed:

$$R = [r_{ij}]_{p*p} = \frac{Z^T Z}{n-1} \tag{1}$$

Where Z is the sample matrix after standardization and normalization. Then, the eigenvalues of the correlation matrix are calculated:

$$R - \lambda I_p \mid = 0 \tag{2}$$

Determine the retention of the first m eigenvalues after establishing the information utilization ratio i:

$$\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{j=1}^{p}\lambda_j} \geq i \tag{3}$$

Following the formula, calculate the eigenvectors corresponding to each eigenvalue. Standardize the eigenvectors according to $b_j^0 = \frac{b_j}{||b_j||}$, then multiply them by the standardized correlation matrix to obtain the components of the decision matrix after dimensionality reduction.

$$u_{ij} = z_i^T b_j^0 \tag{4}$$

2.2. Sufficient Dimension Reduction

Sliced Inverse Regression (SIR) and Sliced Average Variance Estimation (SAVE) are classical methods in sufficient dimension reduction. Unlike the usual focus on $E(Y|X)$ in PCA, in sufficient dimension reduction, $X = (X_1, \ldots, X_p)^T$ can be replaced by $\beta^T X$ without losing predictive information for Y. This means that the conditional distributions $Y|X$ and $Y|\beta^T X$ remain the same. The dimensionality reduction process fully considers the impact of the dependent variable on the distribution of the independent variables, which is different from the typical emphasis on $E(Y|X)$ in PCA.

A brief introduction to the Sliced Inverse Regression (SIR) method is as follows. First, standardize the observed data (y_i, x_i) to obtain the matrix (y_i, z_i). Then, divide the dependent variable Y into H slices $I_1 \ldots I_H$, with p_h representing the proportion of y_i falling into the corresponding slice and n_h being the sample size of that slice. Calculate the sample mean of the independent variable within each slice:

$$m_h = \frac{1}{np_h}\sum_{y_i \in I_h} z_i = \frac{1}{n_h}\sum_{y_i \in I_h} z_i, h = 1,2,\ldots,H \tag{5}$$

Then, establish the weighted covariance matrix:

$$M_{SIR} = \sum_{h=1}^{H} p_h m_h m_h^T \tag{6}$$

This matrix is the kernel matrix estimated by the SIR method. Calculate the eigenvalues and eigenvectors of this matrix, and based on the contribution rate, select the top d eigenvectors $\eta = \{\eta_1 \ldots \eta_d\}$. Restore the data to its original scale and output $\beta = \sum_{xx}^{-1/2}\eta$.

On the other hand, Sliced Average Variance Estimation (SAVE) is a second-order improvement to Sliced Inverse Regression (SIR). Within each slice I_H, it estimates the covariance $\Sigma_h = I_p - \mathrm{cov}(Z|y \in I_h)$. The estimation method for $\mathrm{cov}(Z|y \in I_h)$ is as follows:

$$\begin{aligned} m_h &= \frac{1}{np_h}\sum_{y_i \in I_h} z_i = \frac{1}{n_h}\sum_{y_i \in I_h} z_i,\ h = 1,2,\ldots,H. \\ cov(Z|y \in I_h) &= \frac{1}{np_h}\sum_{y_i \in I_h} z_i^T z_i - m_i^T m_i; \end{aligned} \tag{7}$$

Construct the kernel matrix based on variance information.

$$M_{SAVE} = \sum_{h=1}^{H} cov(Z|y \in I_h)cov(Z|y \in I_h)^T \tag{8}$$

2.3. Bayesian Model Averaging

Bayesian Model Averaging (BMA) utilizes prior information and prior probabilities of explanatory variables to obtain an averaged model. On this basis, it fully leverages the information contained in the dataset, combines all possible models to explain the dependent variable, calculates the posterior inclusion probabilities of potential explanatory variables, and uses this to assess the

relative importance of each explanatory variable. The fundamental idea of the BMA method is to assign prior probabilities to the model Mi and its coefficients β_i. Based on this, the posterior probability of M_i is calculated. Using this as a weight, the posterior probabilities, posterior means, and posterior standard deviations of all potential explanatory variable coefficients are obtained. Applying Bayesian principles, the calculation of the posterior probability of Mi is as follows:

$$p(M_i|y,X) = \frac{p(y|M_i,X)p(M_i)}{p(y|X)} = \frac{p(y|M_i,X)p(M_i)}{\sum_{s=1}^{2^K} p(y|Ms,X)p(Ms)} \tag{9}$$

The fundamental principle of Bayesian Model Averaging (BMA) is to select the optimal model M_i, specifically by choosing the optimal posterior inclusion probability. In the formula, $\mathrm{P}(M_i)$ represents the prior probability of the model, which needs to be set beforehand. This paper adopts the approach of Ley and Steel [16], assuming that the numerical values of all prior probability distributions are the same. The model M_i is set as a multivariate linear regression model based on the dimensionality-reduced factors, and estimation is performed using the least squares method.

2.4. Modeling Process

In Sections 2.1, 2.2, and 2.3, the basic principles of Principal Component Analysis (PCA), Sufficient Dimension Reduction, and Bayesian Model Averaging (BMA) were introduced. Next, this paper will construct a multi-factor stock price prediction model based on Bayesian Model Averaging, with the specific principles outlined below. Note that steps three and four are executed during the estimation of the number of slices in step two.

Step 1: Divide the dataset into training, validation, and test sets based on the specific application scenario.

Step 2: Train the training set using Principal Component Analysis or Sufficient Dimension Reduction (Sliced Inverse Regression and Sliced Average Variance Estimation). Use the validation set to determine the number of slices for Sliced Inverse Regression and Sliced Average Variance Estimation, and obtain the estimated dimensionality reduction direction vectors (selecting a set of vectors with a larger number of dimensions).

Step 3: Train the Bayesian Model Averaging model using the dimensionality-reduced data, estimating the parameters and weights of each sub-model.

Step 4: Use the dimensionality reduction vectors to reduce the test set and predict the test data using the trained Bayesian Model Averaging model, obtaining the final predictions.

3. Data Analysis

3.1. Data Source and Experimental Setup

The experimental data for this paper is sourced from the CSMAR database. In terms of data processing, a comprehensive selection of five major indicators and 31 sub-indicators, including trading information, risk information, market information, and basic company information, were chosen as the independent variables for dimensionality reduction. ST and financial stocks were excluded, and samples with missing values were deleted. The next trading day's price change of the selected stocks was chosen as the dependent variable for prediction. Data analysis and modeling were conducted using RStudio.

In the first step, the prcomp function was used for PCA dimensionality reduction of stock factors, sorting eigenvalues in descending order, and setting the cumulative contribution rate to 99%. The dr function was used for SIR and SAVE dimensionality reduction of stock factors, sorting eigenvalues in descending order, and retaining the top ten strongest explanatory factors.

In the second step, the extracted feature vectors computed through PCA, SIR, and SAVE were multiplied with the sample data to complete data dimensionality reduction.

In the third step, the dataset was split into a training set and a test set in a 7:3 ratio. The glmnet function was used for ridge regression and lasso regression, the bicreg function was used for BMA. After training the test set with these three methods, predictions were made using the test set data, and the difference between the predicted values and the true values was calculated.

In the fourth step, the above experiments were repeated 100 times for each dataset, recording the results of each experiment. After the experiments, the average deviation and standard deviation of each method for each dataset were calculated.

To address the "stock selection" and "timing" demands in the investment process, the paper conducted two sets of experiments. In the first part, to meet the market's "stock selection" demands, the paper selected intervals of one trading month as samples. Six sample dates were selected from July to December 2018, each containing all stocks traded on the Chinese Shanghai and Shenzhen stock markets on that date, with a data volume of around 3300. In the second part, to meet the single stock's "timing" demand, the paper randomly selected five stocks from different industries on the Shanghai and Shenzhen stock markets as the prediction dataset. The stock data selected ranged from January 2018 to June 2019, covering 18 trading months and 361 trading days. Each dataset was trained using pca-BMA, pca-ridge, pca-lasso, save-BMA, and sir-BMA models. The performance of these models in predicting large and small samples was tested and compared.

3.2. Data Source and Experimental Setup

The mean bias and standard deviation of single-period, multi-stock predictions are shown in Table 1 below. From the comparison in the table, facing a large sample size of 3000, PCA dimensionality reduction generally exhibits more stability compared to the two sufficient dimensionality reduction methods, SAVE and SIR. In contrast to LASSO and ridge models, the model averaging characteristic of BMA can significantly reduce bias and increase the robustness of the estimation results. The violin plots below provide a more intuitive representation of the differences in estimation results. As observed from Figure 1, the mean bias of the PCA-BMA method is lower, with a more concentrated bias value. On the other hand, PCA-LASSO and PCA-ridge methods have slightly higher means, with more outliers. SAVE and SIR, the two dimensionality reduction estimation methods, exhibit stronger non-robustness.

Table 1. Single-period, multi-stock prediction - Mean bias and standard deviation.

		PCA-BMA	PCA-LASSO	PCA-ridge	SAVE-BMA	SIR-BMA
1	Mean	1.6456	2.1246	2.0352	2.4354	3.1167
	Sd.	0.0586	0.7830	0.6086	1.9905	2.5462
2	Mean	1.6191	2.2385	2.0221	2.5341	3.2999
	Sd.	0.2160	2.4460	1.2932	1.5236	2.4899
3	Mean	2.0403	2.4205	2.3662	3.9290	30.9611
	Sd.	0.0617	0.5384	0.4439	7.4310	62.1672
4	Mean	1.8608	1.9166	1.9023	1.9895	3.5119
	Sd.	0.0651	0.2177	0.1361	0.4405	2.7689
5	Mean	2.3510	2.6728	2.6143	2.4163	6.0918
	Sd.	0.1023	0.5627	0.3909	0.2339	5.9877
6	Mean	2.0808	2.1376	2.1166	4.0289	4.7505
	Sd.	0.1017	0.3948	0.2846	3.5470	4.8381

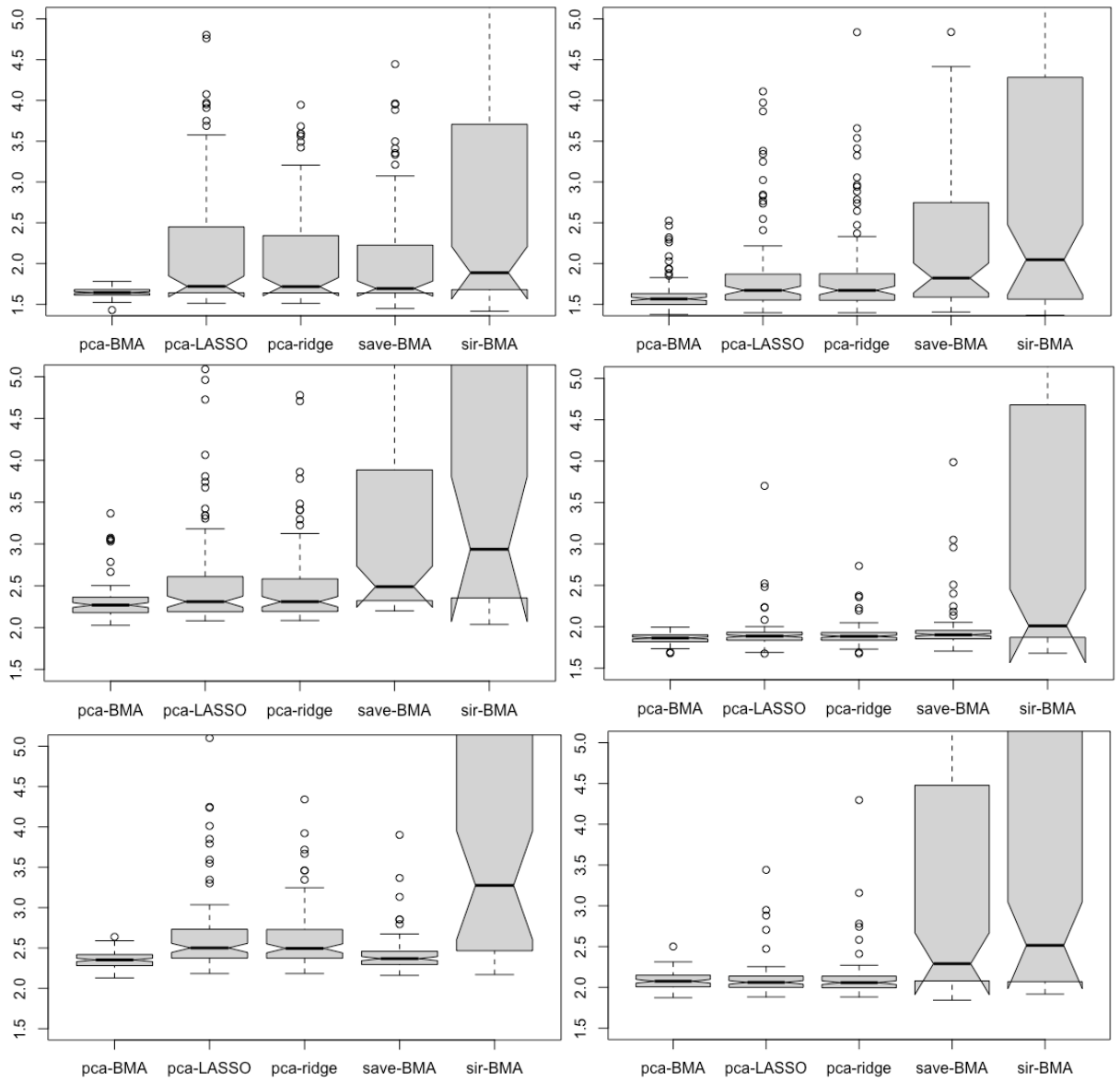

Figure 1. Single-period, multi-stock prediction -Violin plots comparing the distribution of errors.

The bias values and bias standard deviations for single-stock, multi-period predictions are shown in Table 2. From the comparison in Table 2, facing a small sample size of 300, SAVE and SIR, the two sufficient dimension reduction methods, demonstrate higher stability. SAVE and SIR methods are specifically designed as solutions for analyzing sparse high-dimensional data, exhibiting higher prediction accuracy and more robust forecasting results for small sample dimensionality reduction. PCA's dimensionality reduction also performs well in small sample reduction, making it a more versatile dimensionality reduction method. Similar to the results of single-period, multi-stock analysis, relative to the LASSO and ridge models, the model averaging characteristic of BMA can significantly reduce bias and increase the robustness of the estimation results.

The violin plot 2 below provides a more intuitive representation of the differences in estimation results. From Figure 2, it is evident that SAVE and SIR, the two sufficient dimensional reduction methods, perform well in handling sparse high-dimensional data, resulting in lower mean bias and more concentrated prediction bias values. However, the difference from PCA results is not significant. Using the BMA method after the dimensionality reduction processes of SAVE, SIR, and PCA ensures higher predictive accuracy for the models. In contrast, the LASSO and ridge modeling methods yield relatively larger bias results, with more outliers. Since the BMA method weights the different predictive models based on accuracy, considering the estimation results of multiple predictive models comprehensively, the predicted values are more accurate and robust.

Table 2. Multi-period, single-stock prediction - Mean bias and standard deviation.

		pca-BMA	pca-LASSO	pca-ridge	save-BMA	sir-BMA
1	Mean	0.0205	0.0222	0.0222	0.0202	0.0202
	Sd.	0.0028	0.0050	0.0049	0.0027	0.0027
2	Mean	0.0247	0.0276	0.0271	0.0245	0.0245
	Sd.	0.0030	0.0218	0.0160	0.0029	0.0029
3	Mean	0.0217	0.0230	0.0229	0.0216	0.0217
	Sd.	0.0032	0.0046	0.0045	0.0032	0.0031
4	Mean	0.0272	0.0276	0.0276	0.0269	0.0269
	Sd.	0.0037	0.0036	0.0036	0.0037	0.0037
5	Mean	0.0228	0.0228	0.0228	0.0225	0.0225
	Sd.	0.0045	0.0045	0.0045	0.0046	0.0045

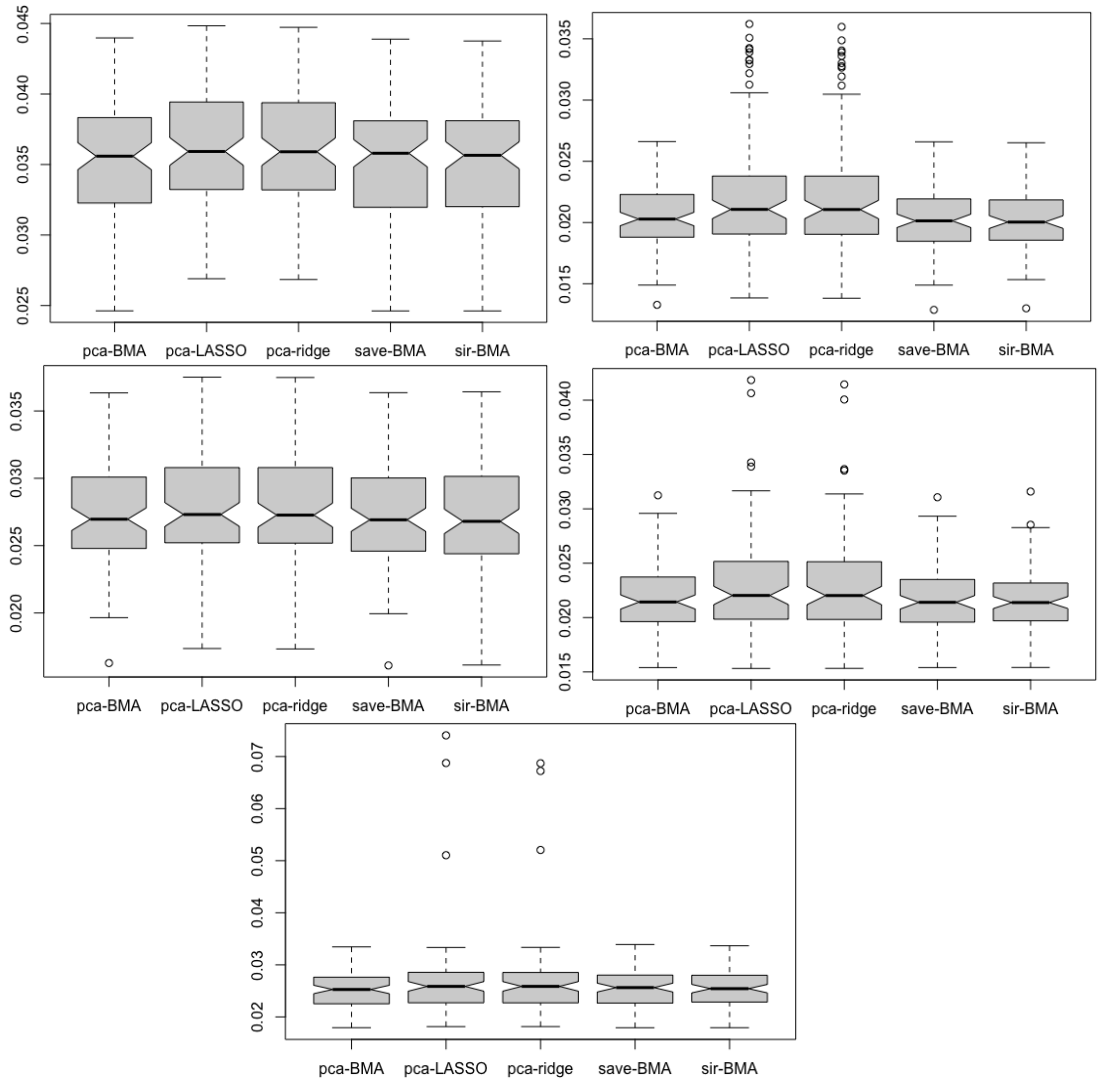

Figure 2. Multi-period, single-stock prediction -Violin plots comparing the distribution of errors.

4. Conclusion and Generalization

Model averaging methods determine the weight of each model in the averaging process based on the prediction accuracy of different models. This creative approach effectively addresses the issues of overfitting and underfitting caused by the estimation of a single model. Additionally, when dealing with high-dimensional data, this paper comprehensively selected and compared three high-dimensional data dimensionality reduction methods: PCA, SAVE, and SIR, for different sample

sizes.The study found that SAVE and SIR, as sufficient dimensionality reduction methods, perform well in predicting small sample reductions but are not suitable for large sample reductions. As for future research directions, Bayesian models have limitations on the underlying models. In contrast, Bagging, another model averaging method, can select and average multiple weak learners, making it a model averaging algorithm with stronger generalization capabilities. Furthermore, Boosting, another model averaging algorithm that continuously improves the model through iteration, offers an alternative solution to optimizing model accuracy.

References

[1] Harry Markowitz. Portfolio selection Journal of Finance [J]. March 1952, 7(1):77-91.

[2] Sharpe, William F. Capital Asset Preices: A Theory of Market Equilibrium under Conditions of Risk [J]. Journal of Finance, 1964, 19:425-442.

[3] Fama, Eugene F; French, Kenneth R. The Cross-Section of Expected Stock Returnst [J]. Journal of Finance, 1992, 47(2):427-465.

[4] Jarrett J E, Kyper E. ARIMA modeling with intervention to forecast and analyze Chinese stock prices [J]. International Journal of Engineering Business Management, 2011, 3: 17.

[5] Di Persio L, Honchar O. Recurrent Neural Networks Approach to the Financial Forecast of Google Assets [J]. International Journal of Mathematics and Computers in Simulation, 2017(11):7-13.

[6] Ma Q. Comparison of ARIMA, ANN and LSTM for stock price prediction[C]//E3S Web of Conferences. EDP Sciences, 2020, 218: 01026.

[7] Sirohi A K, Mahato P, Attar V. Multiple kernel learning for stock price direction prediction[A]. International Conference on Advances in Engineering &Technology Research [C]. IEEE,2014.

[8] Liu J, Stambaugh R F, Yuan Y. Size and value in China [J]. Journal of financial economics, 2019, 134(1): 48-69.

[9] CHONG E, HAN C, PARK F C.Deep learning networks for stock market analysis and predicti on:methodology, data representations, and case studies[J].Expert Systems with Applications,2017 (83):187-205.

[10] Li B. Sufficient dimension reduction: Methods and applications with R [M]. CRC Press, 2018.

[11] Yu T, Yu G, Li P Y, et al. Citation impact prediction for scientific papers using stepwise regression analysis [J]. Scientometrics, 2014, 101: 1233-1252.

[12] Vrieze S I. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) [J]. Psychological methods, 2012, 17(2): 228.

[13] Jansen M. Generalized cross validation in variable selection with and without shrinkage [J]. Journal of statistical planning and inference, 2015, 159: 90-104.

[14] Rasmussen M A, Bro R. A tutorial on the Lasso approach to sparse modeling [J]. Chemometrics and Intelligent Laboratory Systems, 2012, 119: 21-31.

[15] Zhang Xinyu, Zou Guohua. Model Averaging Methods and Their Applications in Prediction [J]. Statistical Research, 2011, 28(06): 97-102.

[16] Ley E, Steel M F J. Mixtures of g-priors for Bayesian model averaging with economic applications [J]. Journal of econometrics, 2012, 171(2): 251-266.