A stock multi-factor model based on slice inverse regression and the Bootstrap method

Jing Wu[1, #], Wenqi Zhou[2, #], Jian Liu[3, *, #]

[1]School of Statistics, University of International Business and Economics, Beijing, China, 100020

[2]School of Statistics and Data Science, Qufu Normal University, Shandong, Jining, 273165

[3]School of Mathematics, Jilin University, Changchun, Jilin, 130000

*Corresponding author:15840993768@163.com

#These authors contributed equally.

Abstract. With the increasingly prominent position of machine learning algorithms in the field of financial quantification, the research on factor quantification and prediction models for high-frequency stock price time series has emerged as a prominent area of study. This paper presents a novel factor quantization prediction model based on slice inverse regression and Bootstrap. The proposed method effectively addresses the challenge of data dimensionality disaster and enables factor quantification while preserving the information of the original predictor variables. Additionally, the method incorporates Bootstrap technology and adopts the concept of model averaging instead of model selection, which enables the adaptive capturing of the unknown connection structure between factors and stock prices. Moreover, it effectively balances the bias and variance of individual prediction models. The empirical analysis using actual data demonstrates the superiority of the proposed method over the comparison method in terms of evaluation indicators such as mean square error, thus highlighting its robustness.

Keywords: stock price prediction, Factor quantization, Full dimensionality reduction, Model average.

1. Introduction

Over the long term, the stock market holds great significance as a vital investment market. The accurate prediction of stock prices has consistently been a pivotal aspect within the realm of financial quantification [1]. Precisely forecasting stock prices can significantly enhance the profitability and safety of investments, while also providing valuable reference for making informed investment decisions.

Stock price forecasting commonly employs time series forecasting methods. For instance, Kong Yuping et al.[2] utilized a combination of time series prediction models and machine algorithms to predict and analyze ICBC stock. Similarly, Chen Dengjian et al.[3] employed a combination of ARIMA and SVR rolling residual models for stock prediction. It is widely recognized that models such as ARIMA, GARCH, and SVR can enable linear predictions of stock prices. However, classical time series models are typically more suitable for low-frequency and short-term time series data modeling tasks. In the case of high-frequency data requiring long-term adaptive updating, it is often necessary to transform it into a multiple regression problem. This involves identifying factors highly correlated with stock price time series and constructing a multi-factor model for stock price prediction. For example, Wang Peidong[4] used some variables such as transaction amount, lowest price, trading volume, next day opening, opening price, closing price, and highest price as explanatory variables. The next day's closing price of the stock was taken as the dependent variable to establish a multiple linear regression model. However, the presence of numerous predictors often leads to the challenge of dimensionality disaster. To address this, there are three common approaches.

The first approach is the parameterized model based on regularization techniques, such as ridge regression and LASSO. For instance, Xiong He [5] employed LASSO and maximum likelihood estimation to estimate parameters of simulated data, demonstrating that LASSO performs well. Gu Zhiting et al.[6] focused on the CSI 300 index and utilized the LASSO variable selection method and

a multi-factor model to construct an enhanced index fund. Yang Guangyu[7] compared and analyzed the prediction results of five methods, including least square method, ridge regression method, LASSO method, Bayesian ridge regression method, and polynomial regression method. The findings indicated that regularization-based prediction methods yield better results. However, it is important to note that regression models based on regularization techniques are single prediction models, which may not effectively balance the trade-off between variance and bias of the prediction model [8]. The assumption of equal importance for each model in random forest is clearly unreasonable. The second approach involves feature selection techniques, such as chi-square test, Pearson coefficient, and mutual information criterion. However, these methods suffer from high information loss and computational costs [9].The third approach is based on dimensionality reduction techniques, such as principal component analysis, kernel principal component analysis, independent component analysis, and factor analysis. For instance, Li Jun [10] optimized the traditional multiple linear regression model using factor analysis to address multicollinearity and achieve a better fitting effect. Unsupervised dimensionality reduction methods, however, may also lead to information loss as the dimensionality reduction process is unrelated to the prediction task. With the advancement of deep learning, neural network technologies have emerged. Techniques like backpropagation (BP) neural networks and Long Short-Term Memory (LSTM) neural networks can adaptively capture relevant predictor information for the prediction task and are not affected by predictor dimensionality. For example, Zhang Jigang and Liang Na [11] divided samples with dispersed characteristics using self-organizing map (SOM) neural networks. However, deep learning technologies currently lack explainability and a solid theoretical foundation, and some factors cannot be effectively quantified and explained.

Therefore, this paper proposes a factor quantization and prediction model based on sliced inverse regression and the Bootstrap method. The aim is to address the challenge of data dimensionality disaster and achieve interpretability in factor quantification. The proposed method ensures that the effective information from the original predictors is theoretically preserved during the quantization process. Additionally, instead of model selection, the method adopts a model averaging approach that adaptively captures the unknown connection structure between factors and stock prices, effectively balancing the bias and variance of the model. The results of data analysis demonstrate that the proposed method outperforms the pairwise method in terms of mean square error and exhibits a certain level of robustness.

2. theories and methods

2.1. Slice inverse regression

Full dimensionality reduction [12] is to find several linear combinations of the original variables without assuming the specific form of the model, and the variables generated by these linear combinations can replace all the original variables without loss of effective information. For a one-dimensional response variable Y and a p-dimensional independent variable $X=(X_1, X_2, ..., X_p)$, dimensionality is reduced sufficiently to find a matrix B of $p\times d$ such that Y and X are independent of each other in the given case of $B^T X : Y \perp\!\!\!\perp X \mid B^T X$, where b represents conditional independence. Sliced Inverse Regression (SIR) is an adequate dimensionality reduction method proposed by Li [13]. Slice inverse regression can effectively avoid the problem of dimension disaster. The basic principle is to first obtain the sample covariance matrix and sample mean of the derived variables, then divide the response variables into several non-covariance slices in order of size, and calculate the average value of the independent variables in each slice, and then carry out weighted principal component analysis on the sample mean to obtain the weighted covariance matrix, and calculate its eigenvalues and eigenvectors. Finally, the largest eigenvector is used. Output the original scale. The specific algorithm is as follows: for $(y_i, \mathbf{x}_i),\ i=1,\ ...,\ n$, standardize $\mathbf{x}$, get

$$\tilde{\mathbf{x}}_i = \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}}(\mathbf{x}_i - \bar{\mathbf{x}}),\ (i = 1,\ ...,\ n) \tag{1}$$

Where $\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}$ and $\bar{\mathbf{x}}$ are sample covariance matrix and sample mean respectively; Slice the value range of y into segment H, I_1, ..., I_H, then the probability of y_i falling in I_h is

$$\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n}\delta_h(y_i) \tag{2}$$

Where $\delta_h(y_i)$ is a binary variable, 1 is taken when y_i falls in I_h, and 0 is taken otherwise.Within each section, calculate the sample mean $\tilde{\mathbf{x}}_i$ and $\hat{m}_h$, $h = 1,\ ...,\ H$, namely

$$\hat{m}_h = \frac{1}{n\hat{p}_h}\sum\nolimits_{y_i \in I_h}\tilde{x}_i \tag{3}$$

A (weighted) principal component analysis was performed for $\hat{m}_h$, $h = 1,\ ...,\ H$ using the following method: A weighted covariance matrix was formed

$$\hat{V} = \sum_{h=1}^{H}\hat{p}_h\hat{m}_h\hat{m}_h' \tag{4}$$

The eigenvalues and eigenvectors are found. Let the K largest eigenvectors be $\hat{\eta}_k$, $k = 1,\ ...,\ K$, all row vectors, we get:

$$\hat{\beta}_k = \hat{\eta}_k\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-\frac{1}{2}},\ k = 1,\ ...,\ K \tag{5}$$

2.2. Stock price prediction model based on slice inverse regression and Bootstrap

The main idea of Bootstrap method [14] is to obtain multiple Bootstrap samples by placing repeated samples back into the original data samples, and carry out statistical inference on these samples respectively. In general, it is difficult to obtain multiple training sets, and the two-self-help method can obtain multiple different training sets through repeated sampling, and finally average all the fitting results. The self-help method can improve the prediction effect of various regression models by reducing the variance. Specifically, building Bootstrap consists of three steps. First of all, for the given training sample S, M training samples are extracted from the training sample S by Bootstrap method in each round, and n sample sets are obtained through a total of n rounds. It should be noted that the n training sets here are all independent of each other. Second, after obtaining the sample set, the prediction model is obtained one sample set at a time. Therefore, for the set of n samples, we can get a total of n prediction models. Finally, for regression problems, we can use the method of calculating the mean of the model as the final prediction result. The specific process is shown in Figure 1:

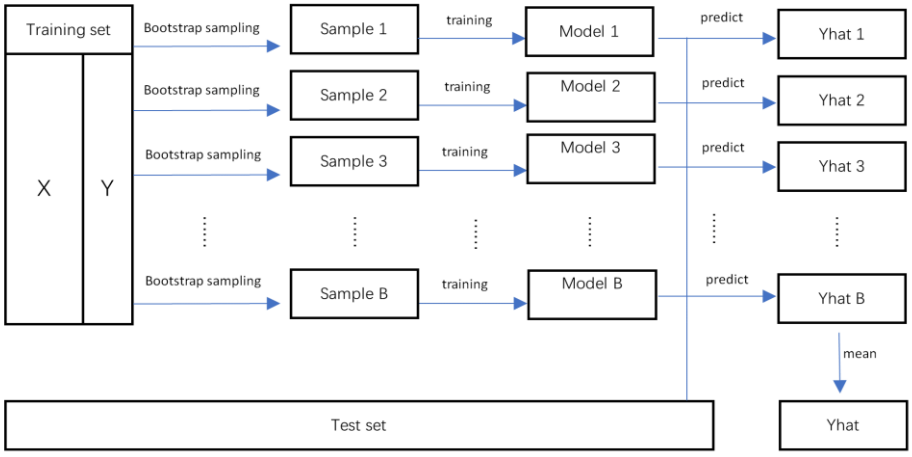


Figure 1 bootstrap principle

In this paper, the Bootstrap method is applied to each first-order factor after dimensionality reduction by slice inverse regression to fit the connection function between the quantized factor and stock price. Specifically, the prediction factors of stocks have many aspects such as profit, valuation, scale, risk and market, and each aspect contains multiple factors. This paper proposes a SIR-Bootstrap method for stock price prediction. The specific steps are as follows:

(1) Class level I impact factors were determined according to relevant literature references, and each class level I predictor had a secondary predictor;

(2) Using the method of sliding window, the time series data of stocks are divided into subsets, and the training set and the test set are divided

(3) The original data of the secondary influence factors were fully dimensionally reduced by slice inverse regression to obtain the primary predictor;

(4) Set each specific sub-model in Bootstrap, take the first-level predictor obtained by quantization after dimensionality reduction as input, and stock price as output, and train the parameters to be estimated of the model, in which the hyperparameters are obtained by generalized cross-validation method.

(5) Use the trained model to process the test set and get the predicted value on the test set.

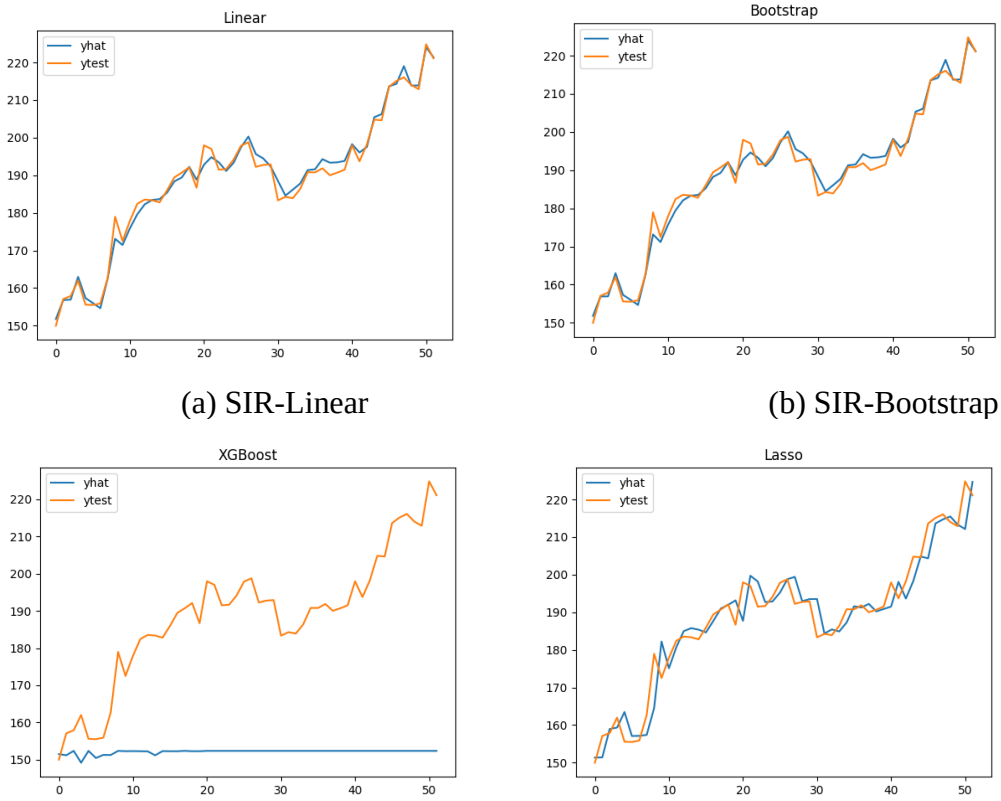
3. Data analysis

3.1. Data sources and experimental Settings

To verify the effectiveness of the proposed method, five leading stocks in the liquor sector, including Luzhou Laojiao (000568), Laobaigan Wine (600559), Shanxi Fenjiu (600809), Shide Wine (600702) and Kweichow Moutai (600519), were selected as empirical data. Shares of its 2016-2020 sequence forecasting, data are derived from the RESSET stock database (https://db.resset.com/common/main.jsp). Through literature research, this paper divides the factors affecting stock prices into five categories: profit factor, valuation factor, scale factor, risk factor and market factor, as shown in Table 1. The purpose of the experiment is to quantify these five factors and complete the task of stock price prediction.Specifically, this paper takes the stock factor of the first day as the independent variable and the closing price of the second day as the dependent variable (y) to conduct prediction research.

Table 1 Stock price factor structure framework

First-order factor	Second-order factor	Symbolic representation
Profit factor	daily return rate	Dret
	net assets income rate	ROE
	Operating earnings per share	OpPrfPS
	Earnings per share	EPS
Valuation factor	Price-earnings ratio	PE
	price-to-sales ratio	PB
	Market rate	PCF
	Price/book ratio	PS
Size factor	Daily total market value	Dmc
	trading volume	Trdvol
	Transaction amount	Trdsum
Risk factor	Diurnal amplitude	Dampltd
	volatility	Garch
	Exponential weighted moving average	Ewma
	Risk factor	Beta_tmv
	Adjusted R square	adj_tmv
Quotation factor	opening price	Oppr
	The highest price	Hipr
	The lowest price	Lopr



(a) SIR-Linear (b) SIR-Bootstrap

(c) SIR-XGBoost (d) Lasso

Figure.2 The predicted result of Luzhou Laojiao (000568)

As for comparison methods, Lasso regression is used as a comparison method in terms of dimensional disaster processing methods. In fitting prediction function methods, multiple linear regression, LSTM neural network and XGBoost model are selected in this paper. The data is divided into training set and test set by sliding window processing method. Specifically, this paper divides the data into 7 samples, and divides each sample into a training set and a test set according to a ratio of 7:3, so as to further study the comparative results of the five methods. After fitting the model, this paper judges the fit of the model according to the mean square error MSE, absolute error AE and the mean and variance of relative error RE of the test set. The calculation formula is as follows:

$$MSE = \sum_{i=1}^{n} \frac{1}{n} \left(\hat{y}_i - y_i \right)^2 \tag{6}$$

$$AE = \sum_{i=1}^{n} \frac{1}{n} \left| \hat{y}_i - y_i \right| \tag{7}$$

$$RE = \sum_{i=1}^{n} \frac{1}{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{8}$$

3.2. Comparison result

Empirical analysis was conducted on five methods, namely SIR-Bootstrap self-help method, Lasso regression, multiple linear regression, LSTM and XGBoost, and the results were shown in Table 3-Table 7 and Fig.2. Specifically, Table 3-Table 7 shows the error comparison between the SIR-Bootstrap method used in this paper and other commonly used machine learning methods for each stock. The evaluation indexes were the mean absolute error (MAE_mean), mean square error (MSE_mean), standard deviation of absolute error (MAE_std) and standard deviation of mean square error (MSE_std) of 7 experiments, respectively. Figure 2 show the comparison of the predicted value of the forecasting method with the real stock price series value. As can be seen from the comparison

results, the MAE_mean and MSE_mean of the SIR-Bootstrap algorithm used in this paper are the smallest in all stocks, and in most cases, the MAE_std and MSE_std are also the smallest. This indicates that the proposed SIR-Bootstrap method has smaller absolute error and mean square error compared with other methods, and the results are robust to a certain extent.

Table 2 Error table of Luzhou Laojiao fitting results

	MSE_Mean	MSE_Std	AE_Mean	AE_Std	RE_Mean	RE_Std
LSTM	7651.2932	11190.950	71.9670	49.1447	0.9391	0.0291
Linear	2.031300	1.777700	1.0663	0.5861	0.0151	0.0059
SIR-Bootstrap	0.144600	0.712700	0.0753	0.3169	0.0011	0.0042
XGBoost	242.9788	563.4993	8.7610	11.4874	0.0885	0.0474
lasso	13.719500	18.61140	6.3675	5.4573	0.0760	0.0263

In addition, LSTM and XGBoost methods cannot effectively fit and predict stock prices, Lasso regression has a large prediction difference, multiple linear regression results are better SIR-Bootstrap self-help regression fitting results are the best the relative error of prediction is below 0.01and the absolute error is within 1.Mean square error is less than 10. Compared with other examples the proposed method is more effective for the following reasons: First it solves the problem of dimensionality disaster by using sufficient dimensionality reduction technology identifies the low-dimensional representation in the dimensionality feature space, and does not lose the effective prediction information of stock price. Second the month Bootstrap technology can generate multiple subsets from the initial data set and then train multiple prediction submodels, which can effectively weigh the deviation and variance of the prediction model so as to alleviate the overfitting or underfitting problems of a single prediction model.

Table 3 Error table of Laobaiganjiu fitting results

	MSE_Mean	MSE_Std	AE_Mean	AE_Std	RE_Mean	RE_Std
LSTM	371.7624	357.0010	16.4741	8.2072	0.7901	0.0746
Linear	5.2443	12.3727	0.9551	1.7205	0.0391	0.0652
SIR-Bootstrap	0.3431	3.2958	0.0651	0.4990	0.0027	0.0192
XGBoost	50.1042	85.6761	4.6778	5.4579	0.1889	0.1941
lasso	40.5058	91.8970	4.8786	5.3088	0.2175	0.1859

Table 4 Error table of Shanxi Fenjiu fitting results

	MSE_Mean	MSE_Std	AE_Mean	AE_Std	RE_Mean	RE_Std
LSTM	12475.2488	23941.6205	77.9994	78.3186	0.9310	0.0375
Linear	60.2484	76.9502	5.1056	4.1542	0.0849	0.0916
SIR-Bootstrap	4.7203	28.8700	0.3765	1.8195	0.0062	0.0345
XGBoost	656.8277	1457.1610	12.1163	16.5684	0.1194	0.0742
lasso	52.2297	61.9277	10.3006	13.5094	0.1029	0.0419

Table 5 Error table of fitting results of Shede Wine industry

	MSE_Mean	MSE_Std	AE_Mean	AE_Std	RE_Mean	RE_Std
LSTM	1060.2238	1037.8265	29.3155	12.7965	0.8885	0.0271
Linear	0.5112	0.5604	0.5091	0.2813	0.0164	0.0091
SIR-Bootstrap	0.0376	0.2062	0.0366	0.1544	0.0012	0.0050
XGBoost	76.5511	180.4545	4.4956	5.7009	0.1059	0.0698
lasso	5.3544	10.4974	4.3230	5.0897	0.1093	0.0732

Table 6 Error table of fitting results of Kweichow Moutai

	MSE_Mean	MSE_Std	AE_Mean	AE_Std	RE_Mean	RE_Std
LSTM	899417.02	945557.96	834.6129	448.3848	0.9956	0.0014
Linear	102.0281	85.6945	7.5106	3.8457	0.0094	0.0033
SIR-Bootstrap	7.2715	35.2508	0.5307	2.1963	0.0007	0.0026
XGBoost	811.6653	462.3177	41.5455	11.9513	0.0561	0.0133
lasso	1266.6851	1339.3667	45.2096	21.7527	0.0548	0.0174

4. Conclusions

Aiming at a series of problems such as factor quantization dimension disaster andconnection function fitting this paper proposes a multi-factor prediction model based on full enhancement and self-help method. On the one hand the proposed method can quantify relevant factors while solving dimension disaster and theoretically ensure that there is no loss of effective information on stock price. On the other hand, the idea of modular average is used to improve the generalization ability of prediction model. In addition, when the interpretable machine learning method is selected for the base model the proposed method can further give the importance degree of each prefactor to the stock price, so that the natural person can make an effective decision on the portfolio strategy more comprehensively. The future research directions are as follows: First the sliced inverse regression used in this paper is a linear dimensionality reduction method. Since the relationship between real data may be nonlinear it can be considered to use the sliced inverse regression nonlinear dimensionality reduction.

References

[1] Cheng Mengfei, Gao Shuping. Multi-scale stock prediction based on Deep transfer Learning [J].Computer Engineering and Applications, 2022,58(12):249-259.

[2] Kong Yuping. Stock predictionanalysis of ICBC based on time series and machine algorithm [J].China Management Information Chem, 2023,26(06):146-148.

[3] Chen Dengjian, Du Feixia, Xiahuan, Stock prediction based on ARIMA and SVR Rolling Residualmodel combination [J].Computer time generation,2022,359(05):76-81.

[4] Wang Peidong. Stock price analysis and forcast based on multiple linear regression [J].Science and Technology Economic Market .2020(01):84-85.

[5] Xiong He. LASSO analysis method and its application in stock price prediction[D].Jinan University,2017.

[6] GuZhiting, SongZefang, Li Yuan. Research on the construction of Enhancedindex fund based on LASSO variable selection and multi-factor model [J]. Mathematical Statistics and Management,2020,39(03):417-428.

[7] Yang Guangyu. Comparative analysis of five kinds of stock prediction by linear regression methods based on stock correlation [J]. Modern Business,2022,No.654(29):42-45.

[8] Fang Quannan, Wu Jianbin,Zhu Jianping et al. A review of random forestmethods [J]. Statistics and Information Forum,2011,26(03):32-38.

[9] Li Zhi-Qin,Du Jian-Qiang,Nie Bin,etal.review of feature selection methods[J]. Computer Engineering and Applications,2019,55(24):10-19.

[10] Li Jun, Multiple Linear Regression Method based on factor Analysis and its application in Stock price prediction [D]. Nanjing University,2014.

[11] Zhang Jigang,Liang NaStock Price Prediction based on som Network-Principal Component-BP Network [J]. Statistics and Decision,2008, 258(06):158-160.

[12] LiXiangjie,WuYanyan,Zhang Jingxiao.Statistical Research,2018,35(07):115-124.

[13] Li K C. Sliced Inverse Regression for Dimension Reduction (With Discussion) [J]. Journal of the American Statistical Association, 1991, 86(414): 316-342.

[14] Mao Ping. Bootstrap method and application [D]. Xiangtan University, 2013.