ISSN 2616-5775 Vol. 5, Issue 10: 21-27, DOI: 10.25236/AJCIS.2022.051004

An Improved ARIMA Method Based on Functional Principal Component Analysis and Bidirectional Bootstrap and Its Application to Stock Price Forecasting

Chuyu Feng[1], Minsong Gao[2]

[1]*School of International Business Administration, South China Normal University, Guangzhou, Guangdong, 510000, China*
[2]*Department of Mathematical Sciences, Anhui University, Hefei, 230601, China*

Abstract: *This paper proposes an improved ARIMA method based on functional principal component analysis and bi-directional bootstrap. The proposed method does not require a smoothness assumption, uses intraday prices as auxiliary information and considers their functional characteristics, and effectively performs a bias-variance trade-off in the forecasting model by using a bi-directional bootstrap method. This is achieved by forming a paired sample of ARIMA forecast residuals and functional characteristics, and then fitting the forecast residuals to the regression model using the two-way bootstrap method, thereby improving the forecast accuracy.In addition, the choice of regression model is free. The empirical results show that the proposed method has better predictive performance and is more robust than the ARIMA model. Finally, the proposed method can be extended to environmental science, social science and other fields to help deal with various prediction problems.*

Keywords: *Stock Price Forecast, ARIMA, Functional Principal Component Analysis, Two-way Bootstrap*

1. Introduction

With the continuous development of our economy and the world economy, big data finance has become one of the most dynamic components of modern economic activity. The study of the intrinsic patterns and fluctuations of stock prices has become an important part of the theoretical and empirical analysis of finance and a focus of research.In order to help investors better understand the patterns and trends of stock data, and to help the financial market continue to develop steadily, it is important for the development of China's financial market to effectively build a forecasting system that can sharply grasp the patterns and trends of stock data.

The development of stock forecasting models has been the subject of much research in China. One of the most classic models is the differential integrated moving average autoregressive model (ARIMA), which is very simple and requires only endogenous variables and no other exogenous variables [1]. The main applications of ARIMA models are in medicine, environmental science, biology, physics and economics. ARIMA models are less commonly used in finance than in other fields, but their significant predictive power and flexibility have attracted the attention of some scholars in China.For example, Lin Zhengyang, Ni Xiaojie and Wang Bo (2022)[2] used 27 Building Integrated Photovoltaic (BIPV) related stocks in the Shanghai and Shenzhen stock markets as the basis for predicting the general trend of BIPV related stocks in the short term future by building a sliding window based ARIMA model analysis to predict the future short term sector index.An Xiaodan and Li Xiaoxia (2020)[3] smoothed the 1996-2018 GDP data of Yuncheng City, established an ARIMA model, used the relative errors between the real data and the fitted data in 2019 and 2020 to judge the fitting effect of the model, and used the relative errors to compare with the results of the established Holt-Winters model, finally determined the ARIMA(0,2,1)model as the forecasting model, forecasted the GDP of Yuncheng City for three years from 2023 to 2025 and put forward four suggestions.Cha Hua and Shi Sampan (2022)[4] selected the total GDP data of Jiangsu Province from 1975 to 2022 and used the functions of spss and python software data analysis to carry out comprehensive analysis of GDP data such as smoothing, determination and testing of model parameters, and finally established an ARIMA(0,1,1)model to make short-term forecasts of GDP data of Jiangsu Province in the next two years, which provided important

ISSN 2616-5775 Vol. 5, Issue 10: 21-27, DOI: 10.25236/AJCIS.2022.051004

references and bases for the Jiangsu Provincial Government and the economic development strategy and planning of Jiangsu Province.

But ARIMA also requires that the time series data be stable, or stable after differencing, which in essence only captures linear relationships and not non-linear relationships. Also, since stock price forecasting is a complex stochastic system, ARIMA models sometimes do not capture enough information to make predictions due to the varying complexity.Based on this,regarding the improvement of the ARIMA model, Yabo Zhao (2019)[5] proposed a stock price range prediction model based on information granulation and BP-Bagging, ARIMA-IGS-SVM to address the complexity and uncertainty of stock price point prediction, the limitations of traditional statistical analysis methods and the vulnerability of machine learning algorithms to overfitting, which significantly improved the accuracy of the prediction model and the effectiveness of the model in predicting R-values.Yuping Song and Yankun Sun (2021) [6] optimized the ARIMA model in accordance with the adaptive filtering method before the previous p periods of historical data and the previous q periods of prediction error before the washout, and used the optimized model to predict the 5-minute high-frequency time series represented by the log return of the CSI 300 index and the closing price of the individual stock Sany Heavy Industry, which significantly improved the prediction accuracy of the ARIMA model for high-frequency financial time series.

The novelty of this paper is an improved ARIMA model based on functional principal component analysis and bivariate Bootstrap. Unlike some ARIMA improvement models, the proposed approach does not require the assumption of smoothness and uses intraday prices as auxiliary information and considers their functional characteristics.Since the regression model between ARIMA prediction residuals and functional characteristics can be free, the proposed method can capture both linear and non-linear information. In addition, the bias and variance of the regression model predictions were effectively traded off between the ARIMA prediction residuals and the functional characteristics using a two-way bootstrap approach. Specifically, the ARIMA model was first used to obtain the predicted opening prices of the three stocks and calculate their corresponding residuals.Secondly, internal price features were extracted by means of functional principal component analysis.The function features and residuals were then modeled using a regression tree model by using a two-way Bootstrap approach.Finally, the sum of the residual predictions and the ARIMA predictions was used as the corrected prediction.

2. Research Methodology

2.1 ARIMA model

ARIMA models use historical data as the basis for estimation and also allow for rapid adjustment when stock prices are subjected to unpredictable shocks, thus predicting the data over a time period of the time series. Any non-random white noise 'non-seasonal' time series can be modelled using an ARIMA model.An ARIMA model is a combination of an autoregressive model (AR), an easily understood mean model (MA) and differencing and can be expressed as ARIMA(p,d,q), where p and q are the lagged orders of AR and MA respectively and d is the order at which the time series needs to be differenced to ensure smoothness.AR can help one to observe the similarity between past and present values. The general formula for ARIMA (p,d,q) used in this paper can be expressed as

$$\triangle P_t = \beta_0 + \beta_1 \triangle P_{t-1} + \beta_2 \triangle P_{t-p} + \varepsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} \quad (1)$$

Where P_t is the share price at time t, ε_t is the random error at time t, $\triangle P_t$ is a difference that represents the data set.

2.2 Functional Principal Component Analysis

Principal component analysis focuses on extracting information from multiple variables through the idea of dimensionality reduction and replacing this information with fewer variables, in essence transforming statistics from a high-dimensional space into a low-dimensional space, thus making the problem easier to deal with.In studying the data analysis of online education stocks, Lili Dai (2019)[7] applied a functional principal component analysis method based on B-sample basis function expansion to it, and obtained the cumulative contribution of several principal components and what they reflect with the help of R software programming to better enable investors to understand the patterns and trends of stock data and make accurate trading strategies.Functional principal component analysis is

ISSN 2616-5775 Vol. 5, Issue 10: 21-27, DOI: 10.25236/AJCIS.2022.051004

also called Karhunen-Loève expansion.

$$P_j(t) = \mu_p(t) + \sum_{i=1}^{\infty} \xi_i^p \left(P_j(t)\right) f_i(t) \tag{2}$$

$$Q_j(t) = \mu_p(t) + \sum_{i=1}^{\infty} \xi_i^q \left(Q_j(t)\right) f_i(t) \tag{3}$$

Where $\mu_p(t)$ is the mean function, $\xi_i^p\left(P_j(t)\right)$ is a centralised function. $\left(P_j(t) - \mu_p(t)\right)$ is the projection on the principal component function, which is $\xi_i^p\left(P_j(t)\right) = \int_T \left(P_j(t) - \mu_p(t)\right) f_i(t)dt$, Further,

$$P_j(t) = \mu_p(t) + \sum_{i=1}^{\infty} \int_T \left(P(t) - \mu_p(t) f_i(t)dt f_i(t) P_j(t) - \mu_p(t)\right) \tag{4}$$

$$= \sum_{i=1}^{\infty} \int_T \left(P_j(t) - \mu_p(t)\right) f_i(t)dt f_i(t)dt f_i(t)$$

Let $X_j(t) = P_j(t) - \mu_p(t)$, there be:

$$X_j(t) = \sum_{i=1}^{\infty} \int_T X_j(t) f_i(t)dt f_i(t) \tag{5}$$

Noting that $\xi_i^X = \int_T X_j(t) f_i(t)dt$, then we have:

$$X_j(t) = \sum_{i=1}^{\infty} \xi_i^X f_i(t) \tag{6}$$

Therefore there is $P_j(t)$ 和 $Q_j(t)$: $P_j(t) = \sum_{i=1}^{\infty} \xi_i^p f_i(t)$ $Q_j(t) = \sum_{i=1}^{\infty} \xi_i^q f_i(t)$, ξ_i^p and ξ_i^q is $P_j(t)$ and $Q(t)$s' projection on the principal component function:take $S_j = \frac{1}{\varphi L} \int_{t \epsilon T} P_j(t) Q_j(t)dt - \frac{1}{L} \int_{t \epsilon T} Q_j(t)dt$:

$$S_j = \frac{1}{\varphi L} \sum_{i=1}^{\infty} \xi_i^p \xi_i^q - \frac{1}{L} \sum_{i=1}^{\infty} \xi_i^q \int_{t \epsilon T} f_i(t)dt \tag{7}$$

When taking the first K principal components:

$$S_j = \frac{1}{\varphi L} \sum_{i=1}^{K} \xi_i^p \xi_i^q - \frac{1}{L} \sum_{i=1}^{K} \xi_i^q \int_{t \epsilon T} f_i(t)dt \tag{8}$$

The choice of K can be made by means of a cumulative contribution margin.

2.3 Two-way Bootstrap

Two-way Bootstrap is an integrated learning method in machine learning that can be combined with other classification and regression algorithms to improve their accuracy and stability while avoiding overfitting by reducing the variance of the results.The construction of the two-way Bootstrap consists of three steps. First, for a given training sample S, M training samples are drawn from the training sample S using Bootstrap in each round, and a total of n rounds are conducted to obtain a set of n samples. It should be noted that the n training sets here are all independent of each other.Secondly, after obtaining the set of samples, one prediction model is obtained using one sample set at a time. Thus, for a sample set of n, we can obtain a total of n prediction models.Finally, if the problem we need to solve is a classification problem, then we can use voting for the n models obtained earlier to obtain the classification results.As for the regression problem, we can use the method of calculating the mean of the model as the final prediction. The specific process is shown in Figure 1.

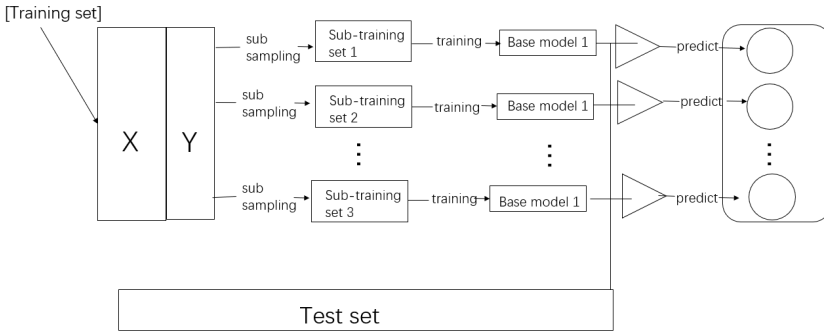


Figure 1: Two-way Bootstrap flowchart

ISSN 2616-5775 Vol. 5, Issue 10: 21-27, DOI: 10.25236/AJCIS.2022.051004

2.4 FPCA-RF

The improved ARIMA model based on functional principal component analysis and bidirectional bootstrap method proposed in this paper is called the functional principal component stochastic feature and sample bootstrap prediction method (FPCA-RF). The basic flow of the algorithm is as follows.

The input is the sample data: Y is the time series to be predicted, X is the intraday price, β is the set of model parameters and for each combination of the model parameters the following steps are performed:

1). The original time series was predicted using ARIMA to obtain the predicted values $\hat{Y}$ and calculate the residual serieset;

2). Extraction of functional features Z of the intraday price X using functional principal component analysis;

3). Pair e_t and Z from steps 1 and 2 to form the new sample data (et,Z);

4). Use the regression tree model f (Zsub) to predict et and obtain the residual prediction $\hat{e}$ and the final prediction $Y_{new} = \hat{Y}+\hat{e}$.

5). The mean squared error, mean relative error and a posteriori error are calculated using the following formulae:

Let set $N_t = \{N_1, N_2, \ldots, N_n\}$ be the prediction residuals of the one-stage ARIMA model, and set $\hat{N}_t = \{\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_n\}$ be the residuals after fitting by the machine learning algorithm. Denote the prediction residual as $\varepsilon_i = N_i - \hat{N}_i$. The indicators for calculating the prediction error of the stock price prediction.Mean squared error measures the difference between the estimator and the estimator as the sum of squared errors.

$$MSE = \frac{1}{N}\sum_{i=1}^{n} \varepsilon_i^2 \quad i = 1,2,\ldots,n \tag{9}$$

The mean relative error measures the difference in the residual value of the predicted value relative to the original data.

$$MRE = \frac{1}{n}\sum_{i=1}^{n} \frac{|\varepsilon_i|}{N_i} \times 100\% \quad i = 1,2,\ldots,n \tag{10}$$

The posterior difference test is based on the residuals of each period, and examines the probability of the occurrence of points with small residuals. The posterior error PE and the small error probability P are calculated as follows.

$$PE = \frac{S_2}{S_1} \tag{11}$$

Where, S_2 is the standard deviation of the residual sequence, and S_1 is the standard deviation of the original sequence. In the subsequent analysis, we only use the posterior error PE as the result of the posterior difference test.

6). The output Y_{new} is the predicted time series with relatively small error in 5 steps.

3. Data analysis

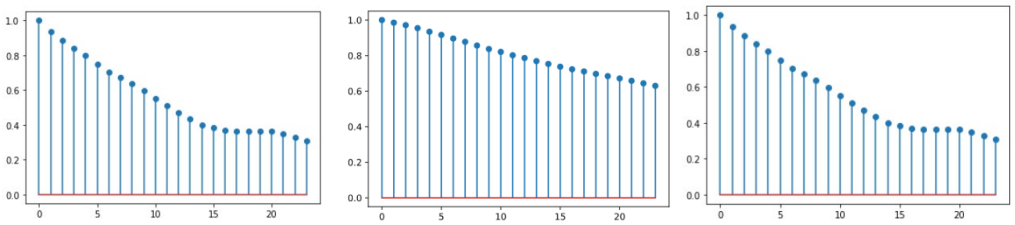
3.1 Research results

In this paper, three stocks were randomly selected from the Wind database and their trading data were collated. The three stocks selected were sh600777 (Xinchao Energy), sh600811 (Dongfang Group) and sz300741 (Huabao), which belong to the oil and gas extraction industry, investment holding enterprise group and food and beverage industry respectively. The use of three randomly selected stocks from different industries as the research sample can fully illustrate the wide availability and

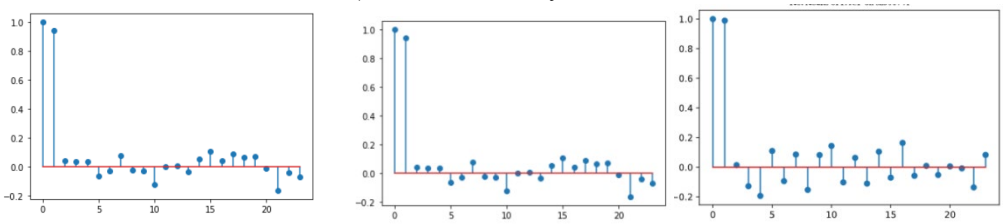
ISSN 2616-5775 Vol. 5, Issue 10: 21-27, DOI: 10.25236/AJCIS.2022.051004

accuracy of this paper's findings. The collection period was from January 2, 2020 to December 31, 2020, and 242 sets of intra-day prices and opening prices were retained for each stock as experimental research samples, for a total of 726 research samples. 60%, 70%, 80% and 90% of each stock were selected for experimental investigation in this paper.The hyperparameters of the model used in this paper were obtained by cross-validation.

To ensure the smooth conduct of the experiment, the sample data of each of the three stocks were first tested for stationarity.The volatility is illustrated by using the autocorrelation coefficient, partial autocorrelation coefficient and ADF test, provided that the data in each sample are randomly selected.The results show that from the perspective of the autocorrelation coefficient, sh600777, sh600811 and sz300741 all still take large values of the autocorrelation function when the number of lags is large and can be considered non-stationary.Meanwhile, the p-value corresponding to the ADF test statistic of sh600777 is 0.07, the p-value corresponding to the ADF test statistic of sh600811 is 0.65, and the p-value corresponding to the ADF test statistic of sz300741 is 0.07, all of which indicate that the original hypothesis is accepted, indicating that the sample data series of the three stocks are non-stationary.Specific information is shown in Figure 2.



(a) Autocorrelation functions



(b) Partial autocorrelation function

Figure 2: Autocorrelation and partial autocorrelation functions for sh600777, sh600811 and sz300741

Firstly, 80% of the data from each of the three stock samples are used to illustrate the superiority of the FPCA-RF method over the ARIMA model in predicting stock prices, and three measures of mean squared error, posterior error and a posteriori error are used to illustrate the results of the experimental study. The results show that the FPCA-RF outperforms the ARIMA model on a total of three metrics: mean square error, posterior error and posterior error, and is more appropriate to use in this context. The specific information is shown in Table 1.

Table 1: Comparison results of FPCA-RF and ARIMA

Evaluation index	sh600777		sh600811		sz300741	
	ARIMA	FPCA-RF	ARIMA	FPCA-RF	ARIMA	FPCA-RF
MSE	0.0098	0.0080	0.0346	0.0314	5.8118	1.2486
MRE	3.0165	3.0508	2.9767	2.7880	3.6262	1.5259
BE	0.6105	0.5188	0.3004	0.1912	0.1904	0.0892

The above results show that FPCA-RF has significant accuracy and superiority in stock price prediction, and compared with ARIMA model, its results are significantly smaller than ARIMA model in mean square error, posterior error and posterior error, but this relationship may be caused by other related data information and other reasons.In order to prove the stable role of FPCA-RF in stock price prediction, this paper takes 60%, 70%, 80% and 90% of the three stock sample data sets as the research object for robustness test, and finally obtains a total of 12 sample sets. Fpca-rf and ARIMA were used to predict the opening price of 12 groups of sample sets respectively. The results obtained by the two methods were compared with the actual values to obtain their respective mean square error, posterior error and posterior error, and then the same sample set was compared with the error values obtained by the two methods. Thus it shows the stability of FPCA-RF in stock price forecasting.The result is shown in Figure 3 below.

ISSN 2616-5775 Vol. 5, Issue 10: 21-27, DOI: 10.25236/AJCIS.2022.051004

The results show that the errors between the results of FPCA-RF and the real value are significantly smaller than the errors between the results of ARIMA model and the real value in terms of mean square error, mean relative error and posterior error.The above conclusions are all valid in 12 sub-sample sets, which is sufficient to show that the FPCA-RF method used in this paper has certain robustness.

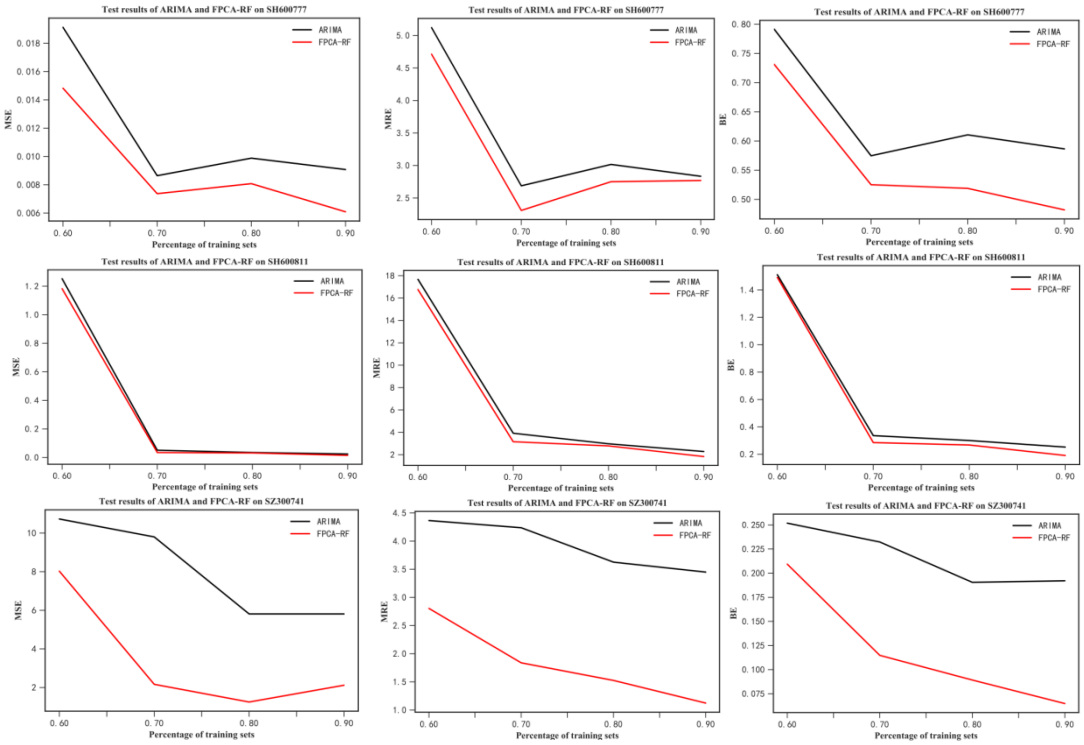


Figure 3: Robustness analysis of FPCA-RF and ARIMA

According to the above results, no matter in terms of mean square error, mean relative error or posterior error, the error between the results of FPCA-RF and the real value is significantly smaller than the error between the results of ARIMA model and the real value.The above conclusions are all valid in 12 sub-sample sets, which is sufficient to show that the FPCA-RF method used in this paper has certain robustness.

4. Conclusions and suggestions

The proposed method does not require the assumption of stationarity. In addition, in the research process of this paper, the important data of intraday price is used as auxiliary information for prediction, and its function characteristics are taken into account, so it can be used to capture nonlinear information of time series, which greatly improves the application scope of ARIMA model and the accuracy and efficiency of prediction. Finally, the method can be further improved. On the one hand, other auxiliary information other than daily price may also affect the volatility and change degree of stock price. On the other hand, the robustness of the model may be affected to varying degrees if other basis function expansion is adopted. At the same time, if different basis functions are used to extract the functional features of intraday prices, the robustness of forecast results may be significantly affected. In addition to the financial field, this method can also be applied to the prediction of other fields, such as traffic flow prediction, daily mean temperature prediction, ferry cargo load prediction and so on.

References

[1] Chen D J, Du F X & Xia H. (2022). Stock forecasting based on combination of ARIMA and SVR rolling residual models. Computer Times (05), 76-81. doi:10.16644/j.cnki.cn33-1094/ TP.2022.05.019.
[2] Lin Zhengyang, Ni Xiaojie & Wang Bo. (2022). Short-term Prediction of Photovoltaic Building Integration Stock Based on improved ARIMA Model. Computer Times (07), 40-43. doi:10. 16644/j.cnki.cn33-1094/tp. 2022.07.010.
[3] An xiaodan & Li Xiaoxia. (2022). GDP Forecasting analysis of Yuncheng City based on ARIMA

ISSN 2616-5775 Vol. 5, Issue 10: 21-27, DOI: 10.25236/AJCIS.2022.051004

Model. Journal of Yuncheng University (03), 69-73+90. doi:10.15967/ j.cnki.cn 14-1316/ g4. 2022.03.018.

[4] Zha, H. & Shi, D. (2022). GDP Prediction of Jiangsu Province based on ARIMA Model. Journal of Lanzhou University of Arts and Sciences (Natural Science edition) (03),33-36+54. doi:10.13804/j.cnki. 2095-6991.2022.03.010.

[5] Zhao Yabo. (2019).Research on Quantitative Investment Based on Data Analysis (Master's thesis, Hebei University of Technology). https://kns.cnki.net/KCMS/detail/ detail.aspx?dbname= CMFD 202201&filename=1021873458.nh

[6] Song Yuping & Sun Yankun. (2021). High frequency Financial time Series Prediction: An improved ARIMA model based on adaptive filtering method. Journal of Jilin University of Business and Technology (02), 82-86. doi:10.19520/j.cnki.issn1674-3288.2021.02.012.

[7] Sun Lili, Fang Hongbin, Zhu Xingxing, Hu Leiming & Qi Longwu. (2021). Journal of Fuyang Normal University (Natural Science edition) (02),97-101. doi:10.14096/j.cnki.cn 34-1069/n/ 2096-9341(2021)02-0097-05.