A hybrid model of stock price prediction based on the PCA-ARIMA-BP

Hua Luo[1] Shuang Wang[2]

(Received 13 June 2016; revised 17 February 2017)

Abstract

A PCA-ARIMA-BP hybrid model is proposed to study the Shanghai Composite Index. The model is based on principle component analysis (PCA), autoregressive integrated moving average model (ARIMA), and backward propagation (BP) neural network. We use data mining methods to select data. BP neural network, PCA-BP model and PCA-ARIMA-BP hybrid model prediction results are compared. The results show that the PCA-ARIMA-BP hybrid model can effectively improve the prediction precision. This can guide investors to avoid risks and improve benefit.

Keywords: PCA-ARIMA-BP hybrid model; the Shanghai Composite Index; risk-averse

DOI:10.21914/anziamj.v58i0.10991, Published July 20, 2017, as part of the Proceedings of the 2016 Joint Conference of ANZIAM and Zhejiang Provincial Applied Mathematics Association. ISSN 1445-8810. (Print two pages per sheet of paper.)

Contents

1 Introduction

Stock price prediction is a very important topic for financial markets. The stock market has the characteristics of high yield and high risk. In order to obtain higher benefits, many scholars seek effective ways to explore stock market internal rules.

When people want to analyse financial markets, seeking an adaptation model is crucial. Autoregressive moving average model (ARMA) and Autoregressive Integrated Moving Average model (ARIMA) have been widely used in time series forecasting. For example, Corbier et al. [1] found that addressed the existence of autoregressive moving average (ARMA) model with reduced

order to get neurodegenerative disorder signals by using a Huberian approach. Since gait rhythm dynamics between Parkinson's disease (PD) or Huntington's disease (HD) and healthy control (CO) differ, and since the stride interval presents great variability, they propose an ARMA modelling approach based on a Huberian function to assess parameters. Rounaghi et al. [5] compared the volatility dynamics of the S&P 500 and the London Stock Exchange using an ARMA model, and showed that an ARMA model for the S&P 500 outperforms the London stock exchange and it is capable to predict medium or long horizons using real known values. The statistical analysis in London Stock Exchange shows that an ARMA model for monthly stock returns outperforms the yearly stock. A comparison between the S&P 500 and London Stock Exchange shows that both markets are efficient and are financially stable during periods of boom and bust. In 2015, Liu et al. [3] used an autoregressive-moving-average model (ARMA) to quantitatively analyse the influence of the air temperature and the precipitation on the streamflow that originated from mountain glaciers. The ARMA model was subsequently applied to analyse the transformed time series. Results of this analysis indicated that the runoff was related to the temperature and the precipitation at any given time and that the precipitation was more important than the temperature in controlling the streamflow. The run off in the upstream of the Urumqi River increased approximately $1\,m^3/s$ every 10 years due to the climate change. Ramos et al. [4] compared the forecasting performance of state space models and ARIMA models. The results show that the overall out-of-sample forecasting performance of state space and ARIMA models evaluated via RMSE, MAE and MAPE is quite similar on both one-step and multi-step forecasts. Khashei et al. [2] found that both theoretical and empirical findings indicate that integration of different models can be an effective way of improving their predictive performance, especially when the models in the ensemble are quite different. In their paper, a new hybrid model of the autoregressive integrated moving average (ARIMA) and probabilistic neural network (PNN) yielded more accurate results than traditional ARIMA models.

Based on the traditional model we propose here a new hybrid model (PCA-

ARIMA-BP hybrid model). The results of BP neural network model, the principal component analysis and BP neural network model, and PCA-ARIMA-BP mixture model are compared. We find the PCA-ARIMA-BP hybrid model outperforms the other two models. Selecting data also increases the data mining method, greatly improving the prediction accuracy of the model.

2 Principal component analysis

Principal component analysis linearly combines the variables of the original structure. Each linear combination is as unrelated to each other as much as possible to reflect the original variable information. Assume that $X_1, X_2, \ldots, X_n$ is involved in the problem of n random variables, the model is defined as

$$\begin{aligned} Y_1 &= l_1^T X = l_{11}X_1 + l_{12}X_2 + \cdots + l_{1n}X_n \,, \\ Y_2 &= l_2^T X = l_{21}X_1 + l_{22}X_2 + \cdots + l_{2n}X_n \,, \\ &\vdots \\ Y_m &= l_m^T X = l_{m1}X_1 + l_{m2}X_2 + \cdots + l_{mn}X_n \,, \end{aligned}$$

where $l_i = (l_{i1}, l_{i2}, \ldots, l_{in})$ for $i = 1, 2, \ldots, n$ represents n constant vectors. We use the new variable Y_i instead of the original variables $X_1, X_2, \ldots, X_n$, and require that Y_i reflect the information of original variables as much as possible. The variance of Y_i, $\mathrm{Var}(Y_i)$, is maximised under orthogonality constraints. Let Σ be the covariance matrix of $X = (X_1, X_2, \ldots, X_n)$, with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant 0$ and $e_1, e_2, \ldots, e_n$ respectively representing the characteristic eigenvalues of Σ and the characteristic eigenvectors, orthonormal, then the ith principal component is

$$Y_i = e_i^T X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{in}X_n \,, \quad i = 1, 2, \ldots, n,$$

where eigenvector $e_i = (e_{i1}, e_{i2}, \ldots, e_{in})$. Some algebra shows that the variance of each main component equals the corresponding eigenvalues. Therefore

$\lambda_k / \sum_{i=1}^{n} \lambda_i$ represents the share of total information provided by the kth principal component, known as the first k principal component contribution rate of Y_k. The sum of the first m components $\sum_{j=1}^{m} \lambda_j / \sum_{i=1}^{n} \lambda_i$ is called the cumulative contribution rate of $Y_1, Y_2, \ldots, Y_m$. Generally, the number of selection variables m is less than the number of original variables n. If the m principal component rate reaches more than 85% of the total information, then we use the m principal components instead of the original variables. This not only reduces the dimensions of the input data, but also retains the most information from the original data.

3 ARIMA model

Time series modelling method is primarily based on historical data. For non-stationary time series, an ARIMA model is mainly used. The autoregressive moving average model ARIMA(p, d, q) is

$$\begin{aligned} &\Phi(B)\nabla^d y_t = \Theta(B)\varepsilon_t, \\ &E(\varepsilon_t) = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2, \quad E(\varepsilon_t \varepsilon_s) = 0, \quad s \neq t, \\ &E(y_s \varepsilon_t) = 0, \quad \text{for all } s \leqslant t, \end{aligned}$$

where $\nabla^d = (1-B)^d$; $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$; $\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ is a smooth move smoothly reversible coefficient ARMA(p, q) model polynomial; p is the autoregressive order; q is moving average order; d is differential order; $\phi_1, \phi_2, \ldots, \phi_p$ are the regression coefficients; $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average coefficients; B is difference operator; and ε_t is a white noise sequence.

The modelling process of an ARIMA model is divided into three parts. Firstly, the raw data is preprocessed. The actual transaction data are usually non-stationary time series, and so before the modelling the non-stationary time series needs smooth processing into a stationary time series. An ARIMA model in d-order difference sequence can be made into stationary time series. Second

is the model recognition and model parameter estimation. We judge the order of an ARIMA model using the minimum information AIC criterion and BIC criterion, where for n samples

$$\begin{aligned} \text{AIC} &= n \log(\hat{\sigma}_\varepsilon^2) + 2\,(p + q + 1)\,, \\ \text{BIC} &= n \log(\hat{\sigma}_\varepsilon^2) + \log(n)\,(p + q + 1)\,. \end{aligned}$$

Finally, the last step is an adaptive test model. After the model identification and parameter estimation, if the model error is tested by white noise, then the model is established. If it cannot pass the test, then you need to reset the model to determine the order and parameter estimation.

4 BP neural network

Neural network is a kind of intelligent information processing method. Based on the error of a back-propagation neural network (BP), which is a function of nonlinear differential weights training multi-layer network. The main idea of a BP neural network is to use the negative gradient descent learning algorithm to modify weights and thresholds in a least square method. It enables the error to be minimised.

A BP neural network mainly consists of an input layer, a hidden layer, and an output layer. The learning process consists of two parts: forward and backward. In the forward propagation, information from input layer passes through intermediate layers is processed and transmitted to the output layer. Comparing the output to the desired output values is the start of back propagation. The reverse propagation process modifies the weights of the neural connections. Through iteration, the error is is reduced to reach the allowable range.

In a basic BP neural network we denote that the input signal is x_i ($i = 1, 2, \ldots, m$), the hidden layer output signal is y_j ($j = 1, 2, \ldots, n$), and the output signal for the output node z_k ($k = 1, 2, \ldots, l$), The weights from x_i

to y_j are set to ω_{ij} with the connection threshold value θ_i, the weights from y_j to z_k is set to ω_{jk} with connection threshold θ_j, and the desired output of the output node is $\hat{z}_k$. The output neurone model is

$$y_j = f\left(\sum_{i=1}^{m} \omega_{ij} x_i - \theta_i\right), \quad z_k = f\left(\sum_{j=1}^{n} \omega_{jk} y_j - \theta_j\right),$$

where the activation function is the Sigmoid function $f(x) = 1/(1+e^{-x})$. The error between the output value and the expected value of the model is defined as $\Delta = \frac{1}{2}\sum_{k=1}^{l}(z_k - \hat{z}_k)^2$. Therefore the error is

$$\Delta = \frac{1}{2}\sum_{k=1}^{l}\left\{f\left[\sum_{j=1}^{n}\omega_{jk} f\left(\sum_{i=1}^{m}\omega_{ij}x_i - \theta_i\right) - \theta_j\right] - \hat{z}_k\right\}^2.$$

The network output error is a function of the weights and threshold values. To modify the weights and the thresholds and change the error, the back propagation modifies the connections layer by layer according to

$$\begin{aligned}
\delta_{jk}^{(l)} &= \frac{\partial E_{jk}^{(l)}}{\partial I_{jk}^{l}} f'\left(I_{jk}^{(l)}\right), \\
\delta_{jk}^{(l)} &= f'\left(I_{jk}^{(l)}\right)\sum_{p=1}^{n^{(l+1)}}\delta_{pk}^{(l+1)}\omega_{jp}^{(l)}, \\
\Delta\omega_{ij}^{(l-1)}(t) &= \eta\delta_{jk}^{(l)} O_{ik}^{(l-1)}, \\
\omega_{ij}^{(l-1)}(t+1) &= \omega_{ij}^{(l-1)}(t) + \Delta\omega_{ij}^{(l-1)}(t),
\end{aligned}$$

where $l = L-1, L-2, \ldots, 1$; $j = 1, 2, \ldots, n^{(l)}$; $i = 1, 2, \ldots, n^{(l-1)}$; $I_{jk}^{(l)}$ denote examples of k input vector x_k output, through layer l node j output; $O_{ik}^{(l-1)}$ denotes layer l node j output; $\omega_{ij}^{(l)}$ for layer l node i connect the layer $l+1$ node j weights; $n^{(l)}$ denotes the layer l node number; η denotes the learning efficiency, $0 < \eta < 1$; $E_{jk}^{(l)}$ denotes the expected value of the network output error.

5 The proposed hybrid model

In financial time series, most are nonlinear sequences. That is, the sequences contain linear and nonlinear components. The principal component analysis method extracts the main part of numerous factors. An ARIMA model can fit a linear time series very well. A BP neural network better fits nonlinear sequences. However, a single model often fail to achieve good forecast performance, so here the three models are combined. Such combination can give the model the advantages of linear and nonlinear methods and improve the prediction accuracy of the model.

Using the method of principal component analysis to extract the factors, one of the main components is recorded as $\{X_t\}$. Then $\{X_t\}$ is decomposed into two parts, linear and nonlinear, $X_t = M_t + N_t$, where M_t and N_t respectively denote linear and nonlinear components in the sequence. First, an ARIMA model for time series is used to get the linear composition forecast $\hat{M}_t$. Then the error between the actual value and the predicted value is the residual $\varepsilon_t = X_t - \hat{M}_t$.

The residual series $\{\varepsilon_t\}$ contains the nonlinear part of the original sequence. A BP neural network to fit residuals gives the predicted value $\hat{N}_t$. Finally, combining the predictive value of the two parts together, we get the time t original sequence predicted value $\hat{X}_t = \hat{M}_t + \hat{N}_t$.

The hybrid model consists of three parts. First of all, using principal component analysis will extract data on the principal components. Second, the linear part of the ARIMA model analyses the data. The residual of the ARIMA model is modelled using a BP neural network model. As a result, the advantages of the hybrid model is to be able to find the main factors from the influence of many factors, and use different models to simulate the data. The linear and nonlinear fitting is combined with the data to thus improve the prediction performance of the model.

Table 1: root test

	t-statistic	1% test	5% test	10% test	P value
value	−3.87	−3.65	−2.95	−2.61	0.71

6 Results analysis and discussion

We analyse the Shanghai index monthly data including the closing price, open price, high price, low price, volume and 20-day moving average from January 2013 to December 2015. In order to improve the forecast results and increase the impact of the macro economy on the stock market, the CPI and the exchange rate data are also included.

6.1 Principal component analysis

In the stock market, there are many associated stock price indexes. However, too many will increase the complexity of the scalar analysis and calculation, which will be difficult for the analysis of stock prices. So principal component analysis (PCA) extracts from the original variables those that contain most of the information.

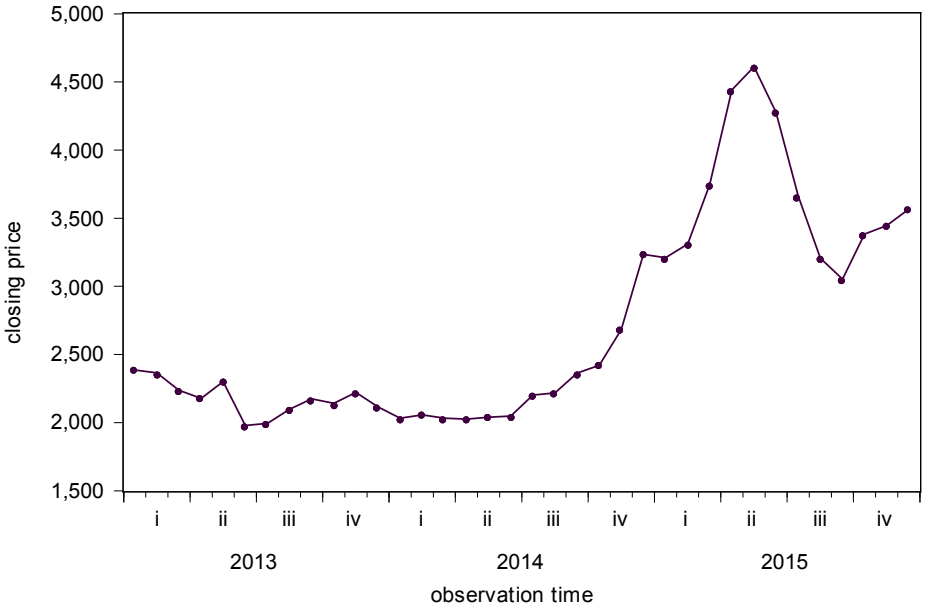
The contribution rate of the two main components is greater than 85%. Thus the numbers of principal components are two.

6.2 Establishing ARIMA model

First we do stability test. The Shanghai index is shown in Figure 1. The Shanghai Composite Index sequence needs a stationarity test using the unit root test. The sequence is first-order cointegration.

According to the results of unit root test, Table 1, the value of the t statistics is less than the one percent level, so we accept the null hypothesis. That

Figure 1: Shanghai Composite Index closing price timing chart



is, the sequence is non-stationary. Then Table 2 compares the SBC and AIC values of those models.

Here, we establish the ARIMA$(2,1,2)$ model.

6.3 BP neural network

We reconsider the data of the Shanghai Composite Index from January 2013 to November 2015. The training set, from January 2013 to May 2015, is used to set up the neural network. The test set, from June 2015 to November 2015, is used to test the results of network prediction. The reason is to improve

Table 2: Model order determination

AIC/SBC	ma(0)	ma(1)	ma(2)	ma(3)
ar(0)		13.66	13.67	13.71
		13.75	13.81	13.89
ar(1)	13.73	13.73	13.90	13.65
	13.82	13.86	13.91	13.89
ar(2)	13.71	13.76	13.55	13.68
	13.85	13.94	13.78	13.95

Table 3: List of variables

	Variable		Variable
x1	Closing price	x2	20 day moving average
x3	Volume	x4	Opening price
x5	Highest Price	x6	Lowest price
x7	CPI	x8	exchange rate

the prediction accuracy of the neural network, and increase the index of the relevant variables, as shown in Table 3.

The closing price, opening price, trading volume, the highest price, the lowest price, the 20 day moving average, CPI and exchange rate of Shanghai Composite Index are the input variables. The closing price in second month is the output variable. So the input number is eight, and there is one output layer. The number of hidden layers is generally set from $l = \sqrt{m+n} + \alpha$ where $\alpha \in [1, 10]$, m is the input layer number, n is the number of output layer. Using the method of trial and error gives Table 4. As shown in Table 4, we choose ten hidden layers. Under the condition of determining the hidden layer, the impulse and the learning efficiency can be changed. Choosing eight input layers, ten hidden layers, one output layer, the training function of `traingdx`, impulse items is 0.9, learning efficiency is 0.1, training times is 5000, and the mean square error (MSE) is 0.65×10^3, aBP neural network is established.

Table 4: Select the hidden layer

hidden layer	4	5	6	7	8
Regression coefficients	0.9972	0.996	0.997	0.9963	0.9973
Hidden layer	9	10	11	12	
Regression coefficients	0.9974	0.9977	0.9971	0.9975	

Table 5: The regression coefficient table of adjusting impulse and learning efficiency

Learning efficiency/impulse	0.95	0.9	0.85	0.8	0.7
0.05	0.9977	0.9966	0.9956	0.9962	0.9961
0.1	0.9984	0.9976	0.9972	0.9961	0.9958
0.2	0.9971	0.9961	0.9964	0.9969	0.9959
0.3	0.9968	0.9977	0.9964	0.9977	0.9972

6.4 Principal component analysis and BP neural network

The Shanghai composite index monthly closing price data is the input variable, while the next monthly closing price is the output variable. Then we establish the BP neural network. By trial and error the method determines the nine hidden layers, the training function is `traingdx` and momentum is 0.95, learning efficiency is 0.2, training times is 5000, and the mean square error is 0.65×10^{-3}. Figure 3 illustrates the results.

6.5 Model predictions comparison

Figure 4 shows the results of the three models and the prediction of the true value. Empirical studies show that the combination of PCA-ARIMA-BP hybrid model has the best prediction. The method of combining Component Analysis and BP neural network is better than the BP network which directly

Figure 2: Within the sample forecast figure compared with the actual figure.

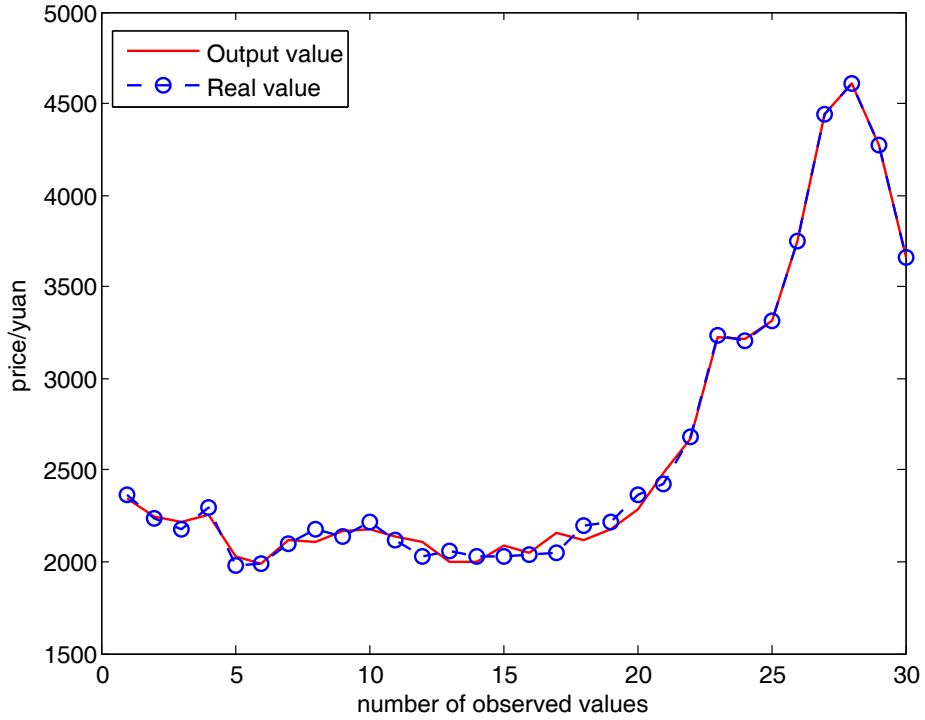
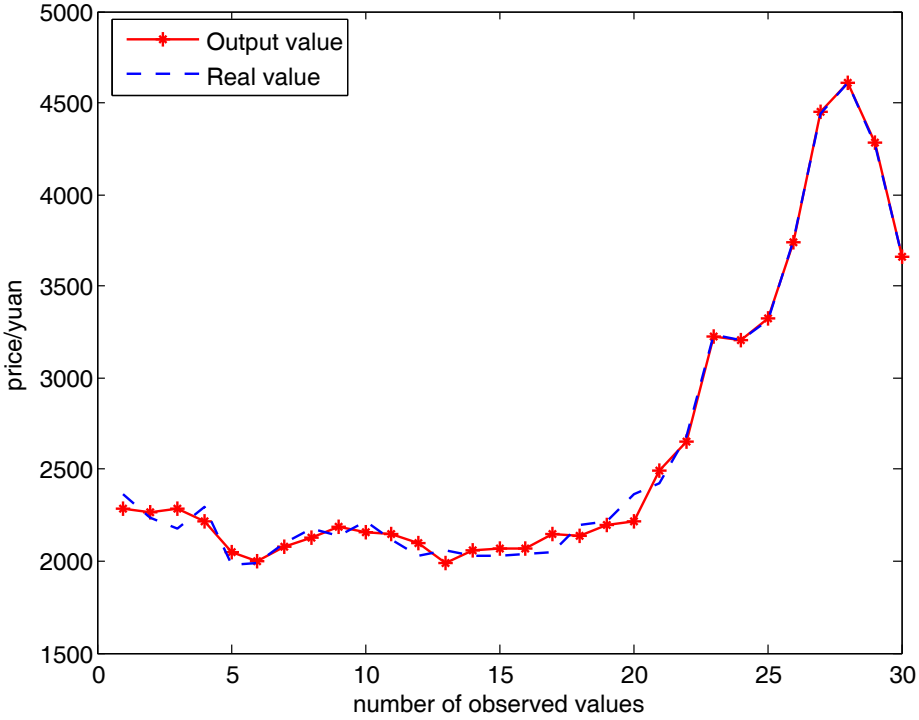


Table 6: Model predictions comparison chart

real value	4214	3614	3157	3156	3337
BP neural network	4149	3743	3935	4968	4639
BP-PCA model	4217	5790	4480	5203	5409
PCA-ARIMA-BP hybrid model	4198	3664	3297	3079	3413

Figure 3: Principal component analysis and BP neural network to predict the effect of fitting Figure

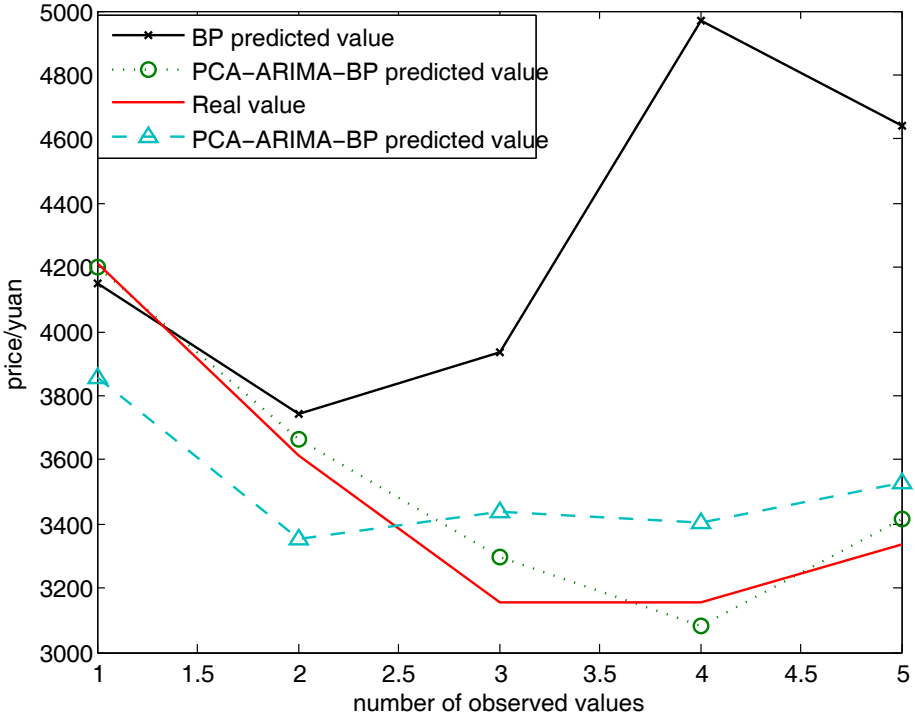


input the original variable. It effectively prevents excessive input variables generated by over-fitting.

7 Conclusions

This paper focuses on the historical data of price, and makes use of data mining methods to study the stock price model based onPCA-ARIMA-BP.

Figure 4: Model predictions comparison figure



Stock market technical analysis indicators, GDP and the exchange rate index, are introduced into prediction model to improve the prediction accuracy of the model. The research results show that the PCA-ARIMA-BP forecasting model is effective and feasible.

References

[1] Christophe Corbier, Mohamed El Badaoui, and Hector Manuel Romero Ugalde. "Huberian approach for reduced order ARMA modeling of neurodegenerative disorder signal". In: *Signal Processing* 113 (2015), pp. 273–284. DOI: 10.1016/j.sigpro.2015.02.010 (cit. on p. E163).

[2] Mehdi Khashei, Mehdi Bijari, and Gholam Ali Raissi Ardali. "Hybridization of autoregressive integrated moving average (ARIMA) with probabilistic neural networks (PNNs)". In: *Computers & Industrial Engineering* 63.1 (2012), pp. 37–45. DOI: 10.1016/j.cie.2012.01.017 (cit. on p. E164).

[3] Youcun Liu et al. "Analyzing effects of climate change on streamflow in a glacier mountain catchment using an ARMA model". In: *Quaternary International* 358 (2015), pp. 137–145. DOI: 10.1016/j.quaint.2014.10.001 (cit. on p. E164).

[4] PatrÍlcia Ramos, Nicolau Santos, and Rebelo Rui. "Performance of state space and ARIMA models for consumer retail sales forecasting". In: *Robotics and Computer-Integrated Manufacturing* 34 (2015), pp. 151–163. DOI: 10.1016/j.rcim.2014.12.015 (cit. on p. E164).

[5] Mohammad Mahdi Rounaghi and Farzaneh Nassir Zadeh. "Investigation of market efficiency and Financial Stability between S&P 500 and London Stock Exchange: Monthly and yearly Forecasting of Time Series Stock Returns using ARMA model". In: *Physica A Statistical Mechanics & Its Applications* 456 (2016), pp. 10–21. DOI: 10.1016/j.physa.2016.03.006 (cit. on p. E164).

Author addresses

1. **Hua Luo**, Department of Mathematics, College of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
mailto:luohuahill@163.com
orcid:0000-0002-3272-3831

2. **Shuang Wang**, Department of Mathematics, College of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
mailto:619458360@qq.com
orcid:0000-0003-2343-5819