Cloud and Service-Oriented Computing (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/csoc.2022.020103
ISSN 2516-399X Vol. 2 Num. 1

An Optimization Framework for Stock Price Prediction Based on Statistical Information and Recursive Model Average -- Taking ARIMA Model as an Example

Jialan Xing[1], Yawen Li[2]

[1]*College of Arts and Sciences, Northeast Agricultural University, Harbin, Heilongjiang, 150006, China*
[2]*School of Statistics, Capital University of Economics and Business, Beijing, 100070, China*

Keywords: stock price forecasting, statistical information quantity, recursive model average method, functional characteristics, ARIMA

Abstract: Based on statistical information and recursive model average method, this paper proposes an optimization framework for stock price forecasting models. The proposed method uses intraday prices as auxiliary information and considers their functional and statistical characteristics. This framework continuously fits the residuals obtained from the original model prediction by a recursive model average method, weights the bias and variance of the prediction model, captures the functional statistical characteristics of intraday prices and the model structure among response variables, and finally optimizes the prediction accuracy of the original model. This framework is model free in theory. Taking the optimized ARIMA model as an example, the data analysis results show that the proposed method has better fitting performance and is robust compared to the ARIMA model. In addition, the proposed method can be extended in application scenarios such as average temperature prediction, traffic flow monitoring, and port cargo capacity prediction.

1. Introduction

As the most important component of financial markets, stocks have become the subject of long and continuous research by scholars. Investors' decisions rely heavily on financial time series analysis and forecasting. The core of the time series forecasting problem is to mine the patterns of random variables over time from the data, use statistical models to build mathematical models of historical time series data, and use them to make predictions about future data. Time series data are often treated as a unique kind of data with the following main characteristics [1]: first, because the values of the series are somewhat random, the prediction of our historical data on present or future data often has errors; second, Time series data are characterized by high noise and high dimension and the information of the original series may not be sufficient for future forecasting.

Many scholars have applied financial time series forecasting to different subjects such as economics and mathematics, and the forecasting models have undergone a transformation from linear to nonlinear models. The classical time series analysis methods are Moving Average (MA), Auto Regressive (AR), and Auto Regressive Moving Average (ARMA) models. Wu Rongliang [2] used monthly data of Shanghai A-share index returns as time series empirical analysis data, and found that

the AR model is feasible as a short-term prediction model for Shanghai A-share index returns; Wu Yuxia and Wen Xin [3] used ARIMA model to empirically analyze the closing price of "Huatai Securities" for 250 periods. The ARIMA model has good results for linear smooth systems, but the prediction accuracy is not high enough for nonlinear or non-smooth problems. With the rapid development of computer technology and the advent of the era of big data, the amount of data that can be recorded and stored in the financial market shows exponential growth. Due to the increased sampling frequency, the information that high-frequency data can carry is substantially higher, allowing for a more detailed reflection of market changes, but also subject to more market noise. Because of this noise effect, the previous research methods for time series analysis of low frequency data used for high frequency data are not very effective. Machine learning methods can largely compensate for the shortcomings and deficiencies of traditional methods. Zhao Xuan used a combined prediction model of ARIMA and LSTM to analyze the actual passenger reservation problem of an airline, and the empirical results showed that the accuracy of the combined model was better than each single model [4]. Meng Yi selected three methods of ARIMA time series, BP neural network and BP-ARIMA combined model to study and forecast CPI in China. The results show that the combined BP-ARIMA model has the best forecasting effect [5]. Xiaomeng Tu and Qiangguo Chen used the ARIMA-LSSVM hybrid model crime data for forecasting. The results showed that the model has higher prediction accuracy and validity for crime time series with small samples compared to the ARIMA-BP hybrid model [6]. However, whether the predictor variables as well as the original series can provide effective and sufficient information to support the high accuracy of the prediction model is still an issue that needs to be studied in depth. In summary, the existing methods mainly have the following problems: first, the information provided by the original time series may not be sufficient for prediction, second, they do not deal with the problem of high dimensionality and noise of the time series, and they do not take into account the functional and statistical characteristics of the time series, and third, the degree of influence of the extracted information on the response variable is undetermined. How to accurately and adequately use the extracted information for prediction of the response variable.

In summary, the main contribution of this study is to propose an optimization framework for stock price prediction models, which provides a new idea for stock research. On the one hand, it takes into account the functional and statistical characteristics of intraday stock prices, which can effectively solve the challenges posed by high-dimensional data. On the other hand, the use of recursive model average makes the proposed method have a good ability to weigh the deviation variance of the forecasting model, and can solve the problem of unknown model structure between the functional statistical characteristics and the response variables. The empirical results show that the model can make full use of the information carried by high frequency data, weaken the effect of time series noise, and improve the prediction accuracy of the ARIMA model.

2. Theory and Methodology

2.1. ARIMA model

The basic idea of the ARIMA model: the sequence of predictor variables over time is treated as a random series, after which a specific mathematical model is used to describe the random series based on the autocorrelation of the time series. the structure of the ARIMA (p,d,q) model is as follows.

$$\begin{cases} \varphi(B)\nabla^d X_t = \theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s) = 0, t \neq s \\ E(X_s\varepsilon_t) = 0, \forall s < t \end{cases} \tag{1}$$

2.2. Feature extraction of Intraday price

For function feature extraction, we use Functional Principal Components Analysis (FPCA) [7-9]. This method can transform infinite dimensional function-based data into finite dimensional score vectors, thus achieving dimensionality reduction. By expanding and truncating the principal component bases of $X_i(t)$, it is represented as follows:

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) \tag{2}$$

Where $X_i(t)$ is the fitted function curve, $\Phi(t) = \{\phi_1(t), \cdots, \phi_K(t)\}$ is the principal component basis function, the coefficient vector $c_i = (c_{i1}, \cdots, c_{ik})'$, and K is the number of basis functions. For the statistical information of the original time series, we can study the statistical characteristics of the time series data by several common data statistics, and the statistics selected in this paper are mean, standard deviation, polar deviation, quartiles, kurtosis, skewness, and coefficient of variation. Define the characteristic information matrix of its composition as Z, and define the principal component coefficient score matrix as C. Construct the information matrix of the combination of the two defined as $M = (C, Z)$, where C is an n*K matrix.

2.3. Recursive model average method

The main principle of the recursive model average method is to generate multiple base learners without dependencies in parallel and use the average method to deal with the regression problem. The recursive uses the the Boostrap method to perform random sampling with put-back, so that there is a difference between the training sets of the training learners, and let different samples train a base learner separately, which in turn makes the trained base learners have certain differences, and finally combine the base learners to get the final strong learner. The steps are mainly as follows.

(1) Initialize the weight distribution of the training samples.

$$D_1 = (\omega_{11}, \cdots, \omega_{1i}, \cdots \omega_{1N}),\ \omega_{1i} = \frac{1}{N}, i = 1,2, \cdots N \tag{3}$$

(2) Repeating the following steps for the iteration times $t = 1,2, \cdots, T$:

a. The training data set with the current weight distribution D_t is used for training and learning to obtain a basic regressor $G_t(x)$.

b. Calculate the relative error of the ith sample on the training data set. Here we take the absolute error as an example:

$$e_{ti}\frac{|y_i - D_i(x_i)|}{max|y_i - D_i(x_i)|}, i = 1,2, \cdots, N \tag{4}$$

c. Calculate the weight coefficient α_t for the base learner $G_t(x)$:

$$\alpha_t = \frac{\sum_{i=1}^{N}\omega_{ti}e_{1i}}{1-\sum_{i=1}^{N}\omega_{ti}e_{1i}} \tag{5}$$

d. Update the weight distribution of the training set:

$$D_{t+1} = (\omega_{t+1,1}, \cdots, \omega_{t+1,i}, \cdots \omega_{t+1,N}) \tag{6}$$

$$\omega_{t+1,i} = \frac{\omega_{t,i}}{\sum_{i=1}^{t}\omega_{ti}\alpha_t^{1-e_{ti}}}\alpha_t^{1-e_{ti}} \tag{7}$$

Integration of linear combinations of the base learners to obtain the final strong regressor:

$$G_x = \sum_{t=1}^{T}\ln\frac{1}{\alpha_t}D_t(x) \tag{8}$$

2.4. Recursive model average based on FPCA and statistical information matrix improves AIRMA

This paper captures the functional characteristics and statistical information of intraday prices based on functional principal component analysis and statistical information matrix, and deals with the problem of unknown importance of variables (model unknown) using recursive model average, and captures the linear and nonlinear relationships between predictor variables and response variables because the base model is free. We call the proposed method in this paper FSIMA. The specific algorithmic procedure is as follows:

1). The intraday prices of the original stock data are expanded and truncated by principal component basis for dimensionality reduction, and the estimated values of the expanded coefficient matrix c are obtained by approximation using a linear combination of K known basis functions $\varphi_k(k = 1,2,\cdots,K)$.

2). Extract the statistical information matrix Z of the original intra-day stock price data, such as mean function, variance function, skewness function, kurtosis function, etc. Use the function information matrix C and the statistical information matrix Z to form the feature matrix M.

3). Use the ARIMA model to predict the opening price of the stock, and use the recursive model average method to predict the prediction residuals of the ARIMA model.

2.5. Evaluation Metrics

(1) Mean squared error (MSE): The mean squared error is the average of the sum of squares of the errors, calculated as the average of the sum of squares of the deviations of the predicted values from the true values for each sample data. This indicator is the loss function of linear regression, and in linear regression, our goal is to minimize the loss function. Its calculation formula is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{9}$$

(2) Mean relative error (MRE): The mean relative deviation is the average value of the relative error, which is generally expressed using the absolute value, that is, the absolute value of the mean relative error. The formula for calculating the average relative error is.

$$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \tag{10}$$

(3) Posterior error (BE): the error that arises from the prediction of the true distribution $p(x, y)$, which is known in advance. In statistics, it is the lowest possible error for the random output of any classifier:

$$\mathrm{C} = \frac{S_2}{S_1} \tag{11}$$

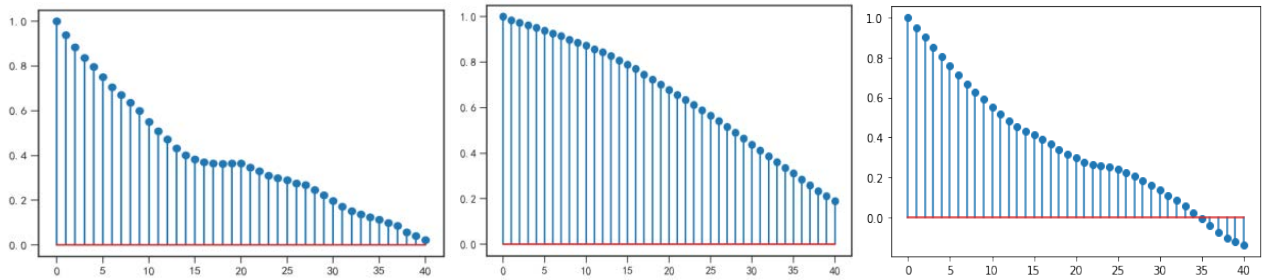
Where S_2 is the standard deviation of the relative value series and S_1 is the standard deviation of the original series.

3. Data analysis

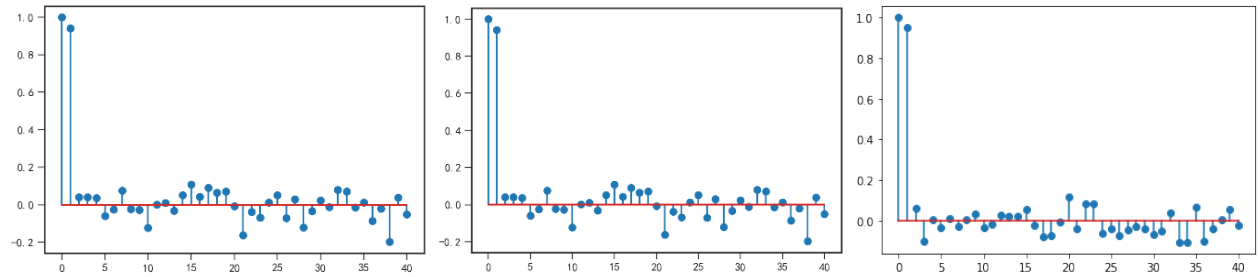
3.1. Data sources and pre-analysis

The data used in this method comes from the WIND database, and the opening price of the trading day is used as the research object. Four types of stock codes are selected, among which SH600777 represents the new energy of the mining industry, SH600811 represents the Dongfang Group of the agricultural and sideline food processing industry, and SH603833 represents the furniture The

industry's OPPEIN home furnishing. All kinds of industries are randomly selected, and all kinds of different stocks can be regarded as stochastic systems with different degrees of complexity, and what we want to explore is whether our method is generally better than ARIMA under stochastic systems with different degrees of complexity.



(a) ACF of SH600777, SH600811 and SH603833



(b) PACF plot of SH600777, SH600811 and SH603833

Figure 1: ACF and PACF of SH600777, SH600811 and SH603833

The horizontal axis is the number of days, the vertical axis is the number of autocorrelation and partial autocorrelation, the top three graphs are ACF, and the bottom three graphs are PACF. Through Python visualization and processing the raw data of the three stocks, the ACF and PACF graphs of the time series of closing prices can be obtained. The details are shown in Figure 1. The ADF test is performed on the three time series data. The ADF test P values of the three stocks are 0.07, 0.651 and 0.143, all of which are greater than 0.05. At the 0.05 significance level, the null hypothesis cannot be rejected, so the opening prices of the three stocks can be considered as non-stationary time series.

3.2. Comparison between FSIMA and ARIMA

Table 1: Comparison between the revised model and the original model (SH600777)

	FSIMA			ARIMA		
	MSE	MRE	BE	MSE	MRE	BE
170	0.0067	3.4944	0.5112	0.0354	7.6231	1.1654
175	0.0134	4.3870	0.6680	0.0354	7.8982	1.1447
180	0.0031	2.5378	0.3467	0.0319	7.6796	1.1089
185	0.0120	4.4745	0.6506	0.0417	8.9659	1.2637
190	0.0049	3.0720	0.4270	0.0295	6.6700	1.0605
195	0.0031	2.2855	0.3354	0.0340	7.4826	1.1398
200	0.0054	3.2968	0.4579	0.0488	9.8434	1.3705
210	0.0073	3.8645	0.5271	0.0466	8.6724	1.3398
215	0.0024	2.0008	0.3034	0.0350	8.3575	1.1613
220	0.0076	3.7423	0.5313	0.0391	8.2294	1.2249

Table 2: Comparison between the revised model and the original model (SH600811)

	FSIMA			ARIMA		
	MSE	MRE	BE	MSE	MRE	BE
170	0.1295	6.6661	0.5728	0.8170	17.7555	1.4704
175	0.1636	7.8054	0.6034	0.7674	18.3771	1.4100
180	0.1432	7.1888	0.5828	0.7767	18.1185	1.4131
185	0.1030	6.1316	0.5083	0.6686	15.8987	1.3295
190	0.2308	8.6431	0.7432	0.8798	19.2597	1.4988
195	0.0713	5.3441	0.4216	0.6018	15.1219	1.2623
200	0.0794	5.8734	0.4491	0.6505	15.5384	1.3117
210	0.1299	6.4133	0.5685	0.7694	17.3863	1.4241
215	0.0558	4.6632	0.3793	0.5568	14.9500	1.2116
220	0.0700	5.3036	0.4250	0.7155	17.4459	1.3752

Table 3: Comparison between the revised model and the original model (SH603833)

	FSIMA			ARIMA		
	MSE	MRE	BE	MSE	MRE	BE
170	142.2144	6.3052	0.9604	291.5232	12.3081	1.3015
175	299.6854	8.5180	1.3717	560.9919	15.4302	2.0294
180	83.1649	5.5454	0.7573	251.9428	11.6211	1.3913
185	88.7660	4.6332	0.7846	225.0459	10.6730	1.3205
190	118.7108	5.9530	0.8928	215.3620	10.7936	1.2920
195	134.9543	5.8342	0.9421	300.0839	12.0488	1.5177
200	53.3632	4.7588	0.6242	199.8444	9.9757	1.2299
210	54.1092	4.6965	0.6363	254.1465	11.1573	1.3957
215	88.5019	5.4823	0.7984	356.5986	12.8897	1.6537
220	39.5327	4.2197	0.5524	292.7670	12.2810	1.5054

We then enabled the Auto-ARIMA model to automatically stabilize a dataset consisting of stock closing prices. The training set is set to 170, 175, 180, 185, 190, 195, 200, 210, 215 and 220 respectively, and the optimal parameter performance is adjusted through automatic autoregressive training combined with AIC information criterion, and then linear prediction is performed to obtain the predicted value of the closing price, and then generate the residual sequence. Then use the method proposed in this paper to model it and get the prediction residuals of the proposed method. The comparison results of the two methods are shown in Table 1-Table 3.

The table reflects the comparison results of the three stocks on the evaluation indicators such as BE, MRE and MSE. The purpose of setting the number of samples in different training sets is to explore the robustness of FSIMA. The three error comparison charts show that: For these three stocks, the error of the revised model is smaller than that of the original model, indicating that the revision of the model is very successful. To sum up: There are different differences in the stationarity of the three stocks, and the revised models have different degrees of reduction compared with the original models, indicating that this model revision is very effective for different stocks.

4. Conclusions

At present, people mostly use time series data to study the problem of financial stock price prediction. The idea of applying functional type data to predict stock prices in this paper is relatively novel, on the one hand, using functional type principal component analysis can better mine the information carried by the data and use the statistical information matrix to supplement the statistical information of the original series. The problem of high dimensionality of the data is effectively

solved. On the other hand, it is based on the recursive model average method to capture the model structure between the statistical characteristics of the function and the ARIMA prediction residuals. It can effectively weigh the bias and variance of the prediction model and weaken the noise effect. The prediction accuracy of the traditional ARIMA model is improved. For the future work, the weight of the importance of functional and statistical features, the optimization of which type of time series forecasting method is more obvious, and whether the functional principal component analysis method can be replaced by other base expansion methods. The measures proposed in this paper have not yet been tested in more detailed and concrete practice, and the proposed measures need further refinement.

References

[1] Yang Haimin, Pan Zhisong, Bai Wei. A review of time series forecasting methods. Computer Science, 2019, 46 (01): 21-28.

[2] Wu Rongliang. An empirical analysis of the return rate of Shanghai A-share index based on AR model. Quotient, 2016 (03): 173-174.

[3] Wu Yuxia, Wen Xin. Short-term stock price prediction based on ARIMA model. Statistics and Decision, 2016, 13 (23): 83-86.

[4] Zhao Yi. An ARIMA and LSTM based combined forecasting model for civil aviation passenger reservation [J]. Computers and Modernization, 2020 (11): 65-69 + 76.

[5] Meng Yi. Application of time series ARIMA and BP neural network combination model in CPI prediction. Journal of Shandong Agricultural University (Natural Science Edition), 2018, 49 (06): 1079-1083.

[6] Tu Xiaomeng, Chen Qiangguo. Crime time series prediction based on ARIMA-LSSVM hybrid model [J]. Electronic Technology Application, 2015, 41 (02): 160-162 + 166.

[7] Wang Qingrong. Research and application of functional principal component analysis and functional linear regression model. Chongqing Technology and Business University, 2020.

[8] Tang Yi, Feng Changhuan. Research on urban population in China based on functional principal component analysis. Journal of Yili Normal University (Natural Science Edition), 2019, 13 (03): 9-16.

[9] Wu Fei and Chen Dirong. Functional data classification based on functional principal components [J]. Journal of Wuhan Textile University, 2019, 32 (02): 48-56.