An improved ARIMA model based on regularized Gaussian basis function and its application to stock price forecasting

Xinyu Xu[1, #, *], Yuqing Du[2, #], Jie Lie[3, #]

[1]School of mathematics, Sun Yat-sen University, Guangzhou, China, 510275

[2]School of Management, Shanghai University of International Business and Economics, Shanghai, China, 201620

[3]School of mathematics, Beijing Normal University, Zhuhai, China, 519087

*Corresponding author: xuxinyu18@outlook.com

#These authors contributed equally.

Abstract. We propose an improved ARIMA model based on regularized Gaussian basis expansion, which is a generalized linear model. This approach takes into account the auxiliary information of intra-day prices, does not require the assumption of linear smoothing and is able to capture the functional features in high-dimensional time series. Specifically, the discrete series of intra-day prices are first functionalized by Gaussian basis smoothing and fitted to the residuals obtained from the ARIMA model using the basis function coefficients, and the fitted model is chosen as a random forest. The empirical results show that the RT-ARIMA model considering auxiliary information is more accurate than the original ARIMA model. And we performed a robustness test by adjusting the size of the training set, and the results show that the method is robust. This model is helpful to improve the accuracy of stock price prediction and other complicated stochastic systems.

Keywords: Stock price prediction, High-frequency data, ARIMA model, Gaussian basis expansion

1. Introduction

In time series model forecasting, the forecasting of stock prices has been a key concern in the academic field. Price forecasts are extremely instructive for the investment market and many investors need to base their trading strategies on the forecast values. Stock price movements are influenced by a variety of factors, including monetary policy, macroeconomic trends, or changes in traders' expectations. This makes stock prices a complex system, based on which building an effective forecasting model that decomposes multiple factors is also an important topic for the field of time series forecasting (Chen, 2020) [1]. Therefore, building a model that can accurately predict stock price movements is a challenging and attractive problem.

Traditional stock price forecasting models are mainly based on time series data modeling of historical stock prices, such as the autoregressive moving average model (ARIMA) for non-stationary time series. The basic idea of the model is to consider the data series of the object to be predicted over time as a random series, and then to approximate this series by certain mathematical models (Ullah and Finch, 2013) [2]. However, the ARIMA model has shortcomings in stock price forecasting, such as the model's linearity assumption has limitations and cannot achieve accurate forecasting when there are non-linear factors in the actual data sample. In addition, since the ARIMA model requires the data sample to be a smooth series or to be able to achieve smoothness by differencing, it is generally used for modeling low-frequency data samples, while in the case of high-frequency data, the model's assumption of smoothness is generally difficult to establish (Tsutsui and Hirayama, 2004) [3].

In response to these problems, scholars are turning to more sophisticated hybrid models or incorporating auxiliary information in the original model. Pai and Lin (2005) used a mixture of support vector machines and ARIMA models, which also better handled the nonlinear part of the actual stock prices [4]. In addition, some scholars have also considered the role of other relevant information in stock market forecasting, such as P/E and P/N ratios (Fama, et al., 1988), investor sentiment (Baker and Wurgler, 2007), technical indicators (Brock et al.,1992) [5-7]. These models

incorporate auxiliary information other than prices in their forecasts, adding to the predictive power of the models for stock price movements.

In recent years, with the rapid development of financial markets, the trading frequency in the securities market has become more and more rapid, the trading volume has become larger, and the price of securities has changed more and more frequently, making it difficult for forecasting based on historical data and low-frequency data to meet the requirements of market development, and high-frequency trading information has become more and more important. There are quite a few literatures that propose processing methods for the characteristics of high-frequency data, for example, Kawano and Konishi (2007) established a nonlinear regression model using Gaussian basis function that can extract information from data with complex structure [8]. In terms of stock price forecasting, Ait-Sahalia (2017) used matrix analysis methods such as principal component analysis to extract predictor variables with strong explanations from high-frequency data, and Bollerslev (2009) obtained indicators such as market volatility risk premium and probability of extreme market events from high-frequency data in options and futures markets to assist in stock price movement forecasting [9-10].

Compared with previous models, the innovations of this paper are as follows:

(1) The generalized linear model breaks through the limitations of the linear smooth assumption of the traditional ARIMA model and can better handle the problems in actual stock price forecasting.

(2) We use intra-day minute prices as auxiliary information to obtain the functional characteristics of this high-dimensional time series by using Gaussian basis expansion. We use its basis function coefficients to fit together with the residuals obtained from the ARIMA model, which has a better fitting effect.

This paper is organized as follows: the second part is an introduction to the theoretical method and model, the third part is an empirical study and results, and finally a summary and discussion.

2. Theory and Methods

2.1 ARIMA

The ARIMA model, known as the differential autoregressive moving average model, is a time series forecasting method proposed by George Box and Gwilym Jenkins. The basic idea is to describe a sequence of numbers that changes over time approximately by a corresponding mathematical model, and then predict the future values of the sequence from the past and present values of the time series by the model. ARIMA model consists of three parts together, which are containing three p, d, and q p is the order of the autoregressive model, which indicates the number of previous time breakpoint intervals used in the autoregressive model at runtime. p=1 means that the model will use data from one previous time breakpoint. q is the order of the moving average model, and d is the number of differences needed to change a non-stationary time series into a stationary time series.

For non-seasonal time series$\{Y_t: t = 1,2,3,\cdots,T\}$, the ARIMA (p,d,q) model can be used to process the entire series by first making it smooth through the d-order difference operation, and then using the ARMA model for forecasting.

The general form of the ARMA model is

$$Y_t^d = \varphi_1 Y_{t-1}^d + \varphi_2 Y_{t-2}^d + \cdots + \varphi_p Y_{t-p}^d + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots \theta_q \varepsilon_{t-q} \tag{1}$$

where $\{\varepsilon_{t-j}\}_{j=1}^{q}$are the residual term, and φ_i、 $\theta_j(i = 1,\cdots p, j = 1,\cdots,q)$are the autoregressive coefficients and moving average coefficients to be estimated, respectively.

The modeling process of the ARIMA model can be divided into four steps as follows.

(1) smoothness test: ADF test is usually used to test the unit root of the original series, and if the series does not satisfy the smoothness condition, the d-difference transformation is performed to transform the non-smooth time series into a smooth time series

(2) Determine the model order: the autocorrelation coefficient and bias correlation coefficient are used to initially determine the possible values of p and q, and then the optimal model is selected according to the AIC information criterion of the estimated model for further judgment

(3) Model test: test whether the residuals generated by the model are white noise, if the model passes the test, the model is correct; otherwise, the model needs to be readjusted and tested until the correct model is obtained.

(4) ARIMA model is used to forecast the time series.

2.2 Regularized Gaussian basis with unit orthogonalization

Assume there is an independent variable$\{ (x_{\alpha i}, t_{\alpha i}); i = 1, \cdots, N_\alpha, t_{\alpha i} \in \mathcal{T} \subset R \}$, and consider the functionalization of its data without considering the dependence for the sign in the variable, i.e. $\{(x_I, t_I); i = 1, \cdots, N \}$Assume that the observations in the main body are derived from the regression model :

$$X_i(t) = \sum_{k \geq 1} a_{ik}\, u_k(t) + \varepsilon_I \ , \quad i = 1, \cdots, N \tag{2}$$

where the residuals $\varepsilon_i$obey an independent normal distribution with a mean of 0 and a variance of σ^2,wherea_{ik}denotes the coefficients, $\varphi_k(t)$is a set of orthogonal basis functions, each of which forms a local acceptance domain in the input control, and the specific expression for the Gaussian basis function is

$$u_k\left(t; \mu_k, \eta_k^2, v\right) = \exp\left\{ - \frac{(t-\mu_k)^2}{2v\eta_k^2} \right\} \tag{3}$$

Where μ_k is the location of the decision center, η_k^2 is the discrete parameter, and v is the hyperparameter, which can be adjusted by adjusting the amount of overlap between the basis functions in order to estimate the regression function to capture the data structure of the independent variables and to obtain information on the dependent variable.

A clustering algorithm is first used to determine the centers and discrete parameters of the Gaussian basis functions, and then a policy drawing approach is used to estimate their weights. This two-stage learning approach is used to solve the convergence problem and the identification problem. However, since this paper is based on the assumption that B in S is mutually orthogonal, this implies that the integration matrix of the basis functions should be a unitary array. For this case, a method of constructing Gaussian unitary orthogonal basis is used in this position, and the procedure is as follows.

$$V_i = \frac{u_i}{\sqrt{\tau_i}}; \tau_i = (\pi v \hat{\eta}_i^2)^{\frac{1}{2}}; i, j = 1,2,3, \cdots, K \tag{4}$$

$$\langle V_I, V_j \rangle_{\mathcal{H}_x} = \left(\frac{2\pi \hat{\eta}_i^2 \hat{\eta}_j^2}{\tau_i \tau_j \hat{\eta}_i^2 + \tau_i \tau_j \hat{\eta}_j^2} \right)^{\frac{1}{2}} \exp\left\{ - \frac{\left(\hat{\mu}_i - \hat{\mu}_j\right)^2}{2v\left(\hat{\eta}_i^2 + \hat{\eta}_j^2\right)} \right\} \tag{5}$$

$$\langle V_i, \Phi_j \rangle_{\mathcal{H}_x} = \langle V_i, V_j \rangle_{\mathcal{H}_x} - \sum_{k=1}^{i-1} \langle V_i, V_k \rangle_{\mathcal{H}_x} \langle V_j, V_k \rangle_{\mathcal{H}_x} \tag{6}$$

$$\begin{cases} Q_1 = 1 \\ Q_i = 1 - \sum_{j=1}^{i-1} \langle V_i, \Phi_j \rangle_{\mathcal{H}_x}{}^2 ; i > 1 \end{cases} \tag{7}$$

$$\varphi_1 = \frac{V_1}{Q_1}; \varphi_i = \frac{\left(V_i - \sum_{j=1}^{i-1} \langle V_i, \Phi_j \rangle_{\mathcal{H}_x}\right)}{Q_i}; i > 1 \tag{8}$$

After orthogonal basis functions, we consider the estimation of the coefficients against the probability density expressed as

$$f(x_i | (x_{\alpha i}, t_{\alpha i}); a, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[- \frac{\{x_i - a^T \varphi(x_{\alpha i}, t_{\alpha i})\}^2}{2\sigma^2} \right] \tag{9}$$

The regularization method is penalized by maximizing the log-likelihood function as

$$\ell_{\lambda_a}(a, \sigma^2) = \sum_{a=1}^{n} \log f(x_a | t_a; \theta) - \frac{n\lambda_a}{2} a^T R_a a \tag{10}$$

where $\lambda_a > 0$ is the regularization parameter controlling the smoothness of the fitted model, R_a is a known non-negative definite $N \times N$matrix, and the maximum penalized likelihood estimator (a, σ^2) is

$$\hat{a} = (\varphi^T \varphi + n\lambda_a \hat{\sigma}^2 R_a)^{-1} \varphi^T x_i \; ; \; \hat{\sigma}^2 = \frac{1}{n} (x_i - \varphi^T \hat{a})^T (x_i - \varphi^T \hat{a}) \tag{11}$$

In practice, the GIC is used for each curve to obtain the optimal number of basis functions, and then the longest number of basis functions chosen among N samples, K, is determined, and once K is fixed, we choose the optimal value of the smoothing parameter I and select the minimum value of u for each set of discrete data as a criterion, and finally convert the observed discrete data into the form of a function of.

$$x(t) \equiv \sum_{k=1}^{K} \hat{a}_k \varphi_k(t) \tag{12}$$

Here, it can be seen that for each object $x_i$it is estimated independently using its own observation vector. Also once a fixed basis function is given, this function is uniquely reflected on the coefficient vector. Finally the coefficient matrix is obtained.

$$\Lambda = (\hat{a}_1, \hat{a}_2, \hat{a}_3, \cdots, \hat{a}_n)^T \tag{13}$$

2.3 RT-ARIMA

For the prediction series$\{Y_t : t = 1,2,3, \cdots, T\}$, using our method yields a representation of the prediction results as

$$Y = f_1(Y_{\mathcal{T}}) + f_2(\hat{\Lambda}) \tag{14}$$

Where $f_1$denotes processing using the ARIMA model, $f_2$denotes processing using the random forest model, $Y_{\mathcal{T}}$denotes the antecedent $\mathcal{T}$of the prediction sequence, and Λdenotes the coefficients of the basis expansion term after the Gaussian basis expansion of the auxiliary information sequence $x(t)$.

The specific algorithms are as follows.

(1) Fitting the time series $\{Y_t: t = 1,2,3,\cdots,T\}$ using the ARIMA model to obtain the one-stage prediction results $\hat{Y}$, where $\hat{Y} = f_1(Y_t)$.

(2) Calculate the residuals on the basis of the one-stage prediction results $\hat{e}, \hat{e} = Y - \hat{Y}$.

(3) The auxiliary information series $x(t)$are processed using Gaussian basis expansion to calculate the coefficient matrix $\hat{\Lambda}$.

(4) Fitting the obtained coefficient matrix with a random forest model to obtain the prediction results of the second stage, the new residual series$\hat{e_a}, \hat{e_a} = f_2(\hat{\Lambda})$.

(5) The results of the two stages are summed to obtain the prediction results of the RT-ARIMA model, finally $\hat{Y}_{new}, \hat{Y}_{new} = \hat{Y} + \hat{e_a}$.

2.4 Model Accuracy Check

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.

1) Mean Squared Error

Mean squared error measures the difference between the estimator and the estimator as the sum of squared errors.

$$MSE = \frac{1}{N}\sum_{i=1}^{n} \varepsilon_i^2 \quad i = 1,2,\ldots,n \tag{15}$$

2) Mean Relative Error

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

$$MRE = \frac{1}{n}\sum_{i=1}^{n} \frac{|\varepsilon_i|}{N_i} \times 100\% \quad i = 1,2,\ldots,n \tag{16}$$

3) Posterior Difference Test

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.

$$PE = \frac{S_2}{S_1} \tag{17}$$

$$P = P\{|\varepsilon_i - \bar{\varepsilon_i}| < 0.674S_1\} \quad i = 1,2,\ldots,n \tag{18}$$

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.

3. Actual data analysis

In this paper, the intraday price data of SZ300583, SZ300594 and SH600811 are downloaded from the Wind terminal in minutes. The three stocks are randomly selected from wind terminal, and the three stocks are randomly selected from different industries. The random selection is to verify whether the model has good performance in random systems with different levels of complexity (due to different industry backgrounds, the internal fluctuation mechanisms of stocks are different, so they have different levels of complexity). This paper focuses on the establishment of the model. Then the validity of the model is verified by empirical study.Take 200, 220 and 180 as training set days and 200 as test set days to expand the test. The evaluation indicators we used mainly include posterior error, mean square error and mean square relative error. The comparison results are shown in Table 1, 2 and 3. Figure 1, 2, 3, respectively for the three stocks sz300583, sz300594, sh600811 training sets 200, 220 and 180 predicted residual error .

Through the comparison of the above chart data, we can see that the residual sequence of the improved ARIMA model is smaller than that of the original evaluation model. Within the range of 0-200, the curve of the improved ARIMA model remains roughly unchanged, but the ARIMA model still has a relatively large fluctuation. The mean square error of the improved model is smaller than that of the original model. The smaller the mean square error is, the better the prediction model can describe the experimental data. Because the proposed method considers intraday price information and describes its function characteristics, it can capture more potential information, and because of the smooth method, the nonlinear information part is also taken into account. So it makes sense to have improved on all the measures.

Table 1. Model evaluation Results (training sets of 200 days)

Stocks code	indicators	ARIMA	RT-ARIMA
SSZ300583	MSE	7.50557	5.46164
	MRE	5.03037	3.91921
	PE	1.12646	0.95703
SZ300594	MSE	3.70733	2.00703
	MRE	3.71484	2.61535
	PE	0.80388	0.58355
sh600811	MSE	0.07400	0.05386
	MRE	4.24201	3.09539
	PE	0.42356	0.35001

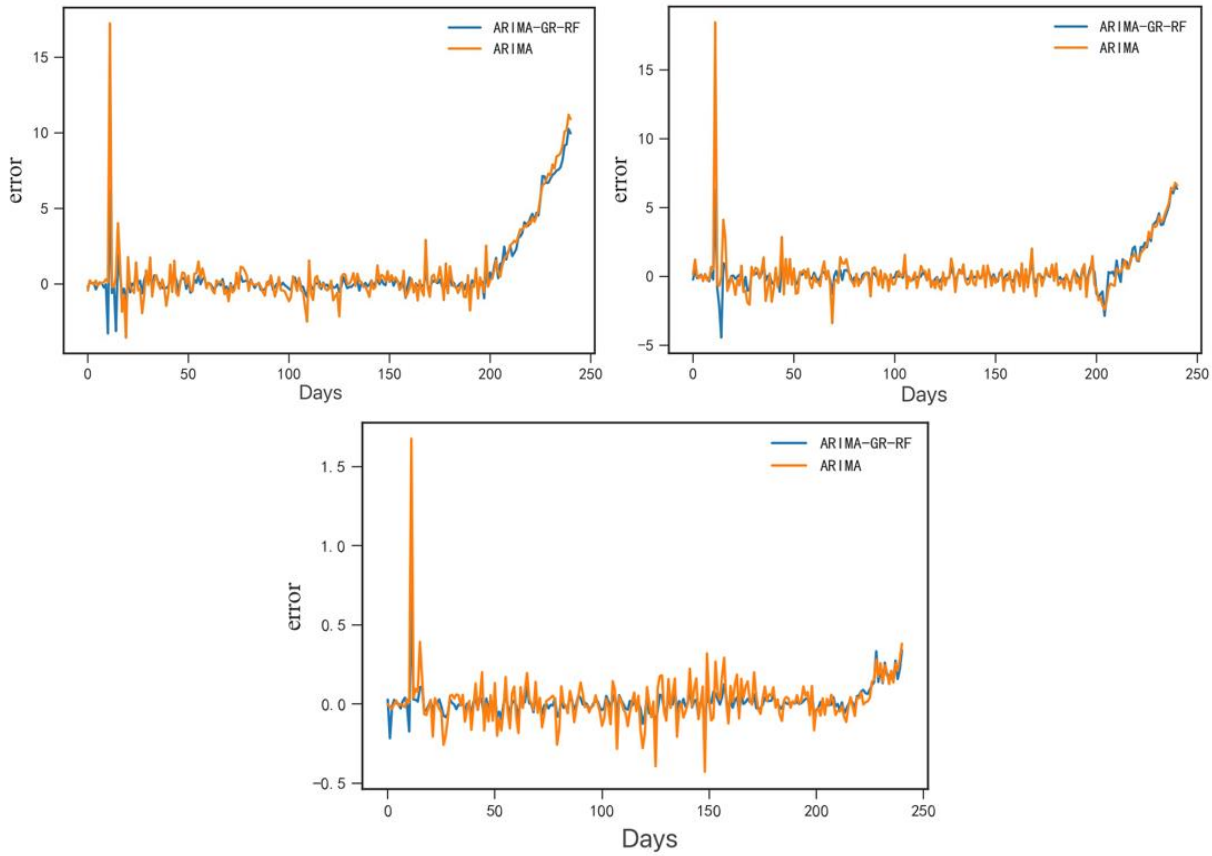

Figure 1. residual error (training sets of 200 days)

According to the above figure 1, when the training set is 180, the residual of the improved model fluctuates in the range of 0-180, and the random fluctuation can be considered to be roughly constant in the whole time series, and the mean value is close to 0. Similarly, when the training set is 220, the residual of the improved model can be considered roughly unchanged within the range of 0-220, and the mean value is close to 0, indicating that the model tends to be stable.

Table 2. Model evaluation Results (training sets of 220 days)

Stocks code	indicators	ARIMA	RT-ARIMA
SSZ300583	MSE	2.54367	0.90854
	MRE	2.82364	1.72540
	PE	0.68613	0.40772
SZ300594	MSE	2.97127	1.23303
	MRE	3.22065	2.05543
	PE	0.71945	0.46247
sh600811	MSE	0.02606	0.00700
	MRE	2.41681	1.24370
	PE	0.26029	0.13409

Table 3. Model evaluation Results (training sets of 180 days)

Stocks code	indicators	ARIMA	RT-ARIMA
SSZ300583	MSE	2.81186	1.28312
	MRE	3.20413	2.24484
	PE	0.72904	0.49181
SZ300594	MSE	3.68780	1.70323
	MRE	3.86594	2.67245
	PE	0.79738	0.53301
sh600811	MSE	0.02852	0.01093
	MRE	2.69985	1.80332
	PE	0.27477	0.16995

Firstly, the research goal of this paper is to develop a stock price prediction method for complex stochastic systems. Considering external shocks and the volatility of its own system, we test the prediction of different stocks and set different window periods to test the robustness (robustness) of the model. Under the background of stock volatility, every stock can be seen as a random system, and the random system does not know the assumptions behind the model, so the general ARIMA model can't accurately to predict, in addition, more important is price easily affected by different external shocks, such as some policy factors, thus become more complicated, The experimental results show that the proposed method makes full use of the auxiliary information of intraday price to better describe the stock price fluctuation. Empirically, we believe that the proposed model has the ability to predict more complex stochastic systems.

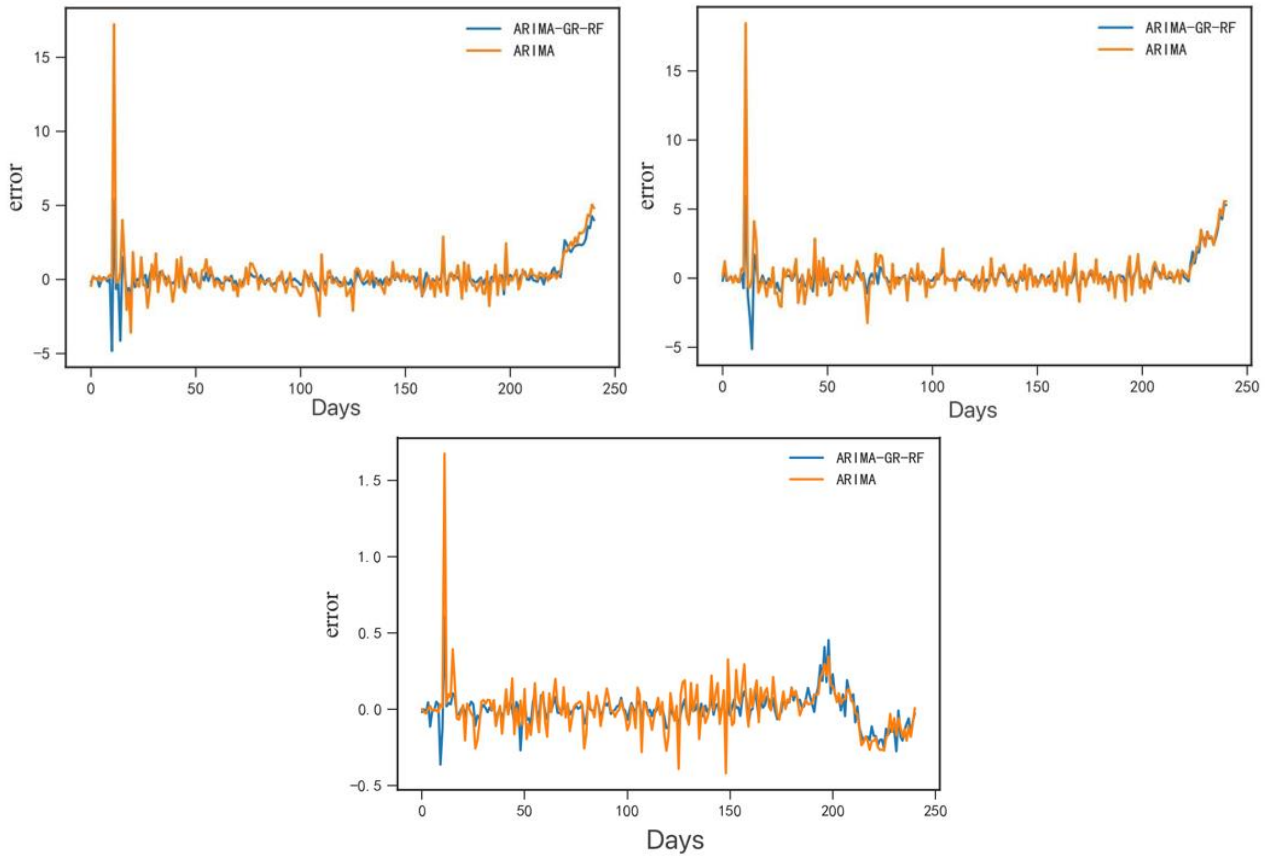

Figure 2. residual error (training sets of 220 days)

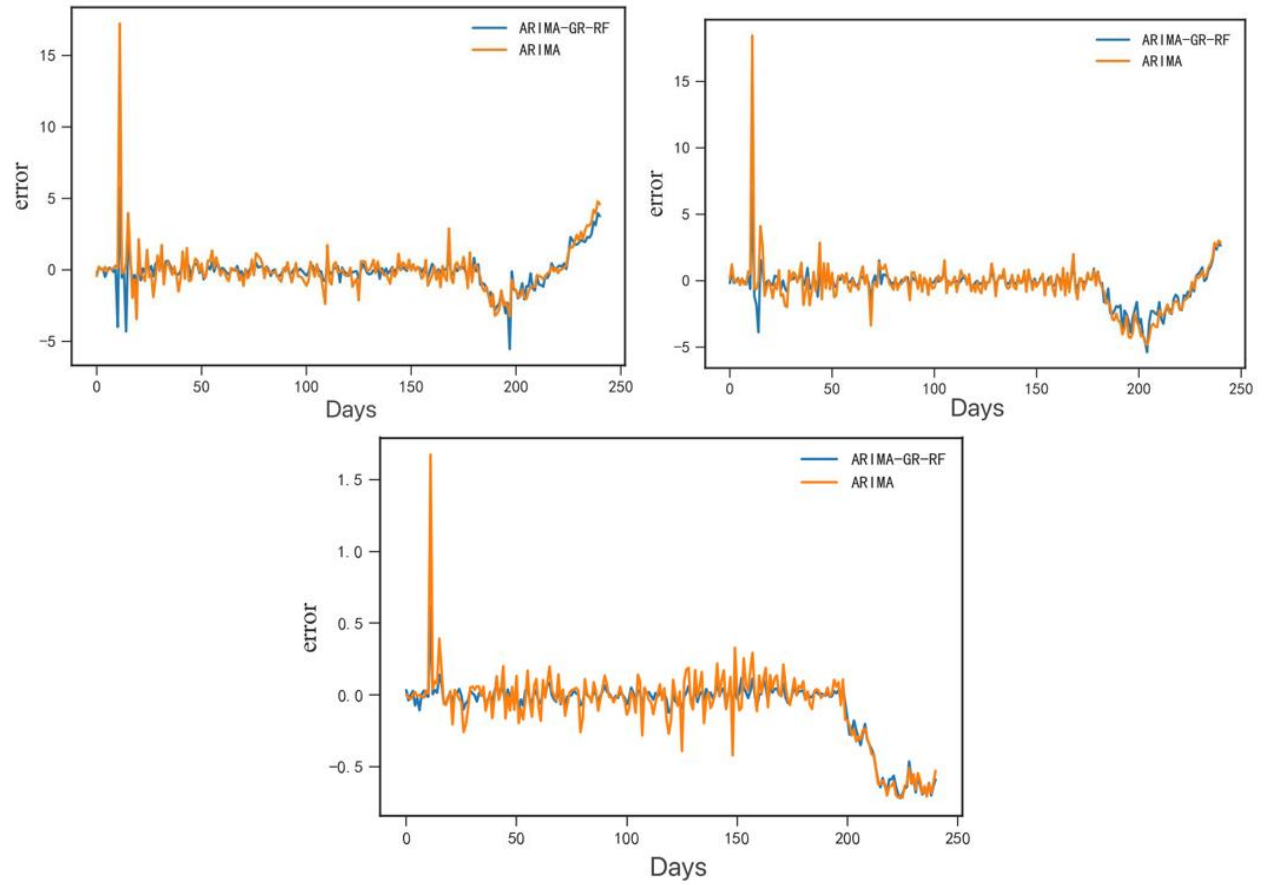

Figure 3. residual error (training sets of 180 days)

4. Conclusion and discussion

Based on ARIMA model, this paper carries out an empirical test on Shanghai and Shenzhen A-share stocks. By using gaussian basis expansion to process auxiliary information, the coefficient matrix is obtained and the random forest model is used to fit. The new residual sequence is obtained and added with the first-stage predicted value, and finally the new predicted value is obtained. The results show that the new prediction model with auxiliary information is more accurate than the original ARIMA model. Secondly, the nonlinear model of auxiliary information is constructed by considering the nonlinear information of auxiliary information when using ARIMA model. Since this model is not the optimal choice, subsequent studies try to reduce the dimension of function type, model average, sparse type and dense type auxiliary information to further improve the prediction accuracy.

References

[1] Chen, H. Can high frequency data improve stock prices forecasting?—Empirical evidence based on functional data analysis [J]. China Journal of econimetrics, 2020, 1(2).

[2] Ullah, S., Finch, C.F. Applications of functional data analysis: A systematic review [J]. BMC Med Res Methodol, 2013.

[3] Tsutsui, Y. and Hirayama, K. Market Efficiency and International Linkage of Stock Prices: An Analysis with High-Frequency Data [J]. SSRN Electronic Journal, 2004.

[4] Pai, P. and Lin, C. A hybrid ARIMA and support vector machines model in stock price forecasting [J]. Omega, 2005, 33(6), pp.497-505.

[5] Fama, E. and French, K. Dividend yields and expected stock returns [J]. Journal of Financial Economics, 1988, 22(1), pp.3-25.

[6] Baker, M. and Wurgler, J. Investor Sentiment in the Stock Market [J]. SSRN Electronic Journal, 2007.

[7] BROCK, W., LAKONISHOK, J., LeBARON, B. Simple Technical Trading Rules and the Stochastic Properties of Stock Returns [J]. The Journal of Finance, 1992, 47(5), pp.1731-1764.

[8] Kawano, S. and Konishi, S. Nonlinear Regression Modeling via Regularized Gaussian Basis Functions [J]. Bulletin of informatics and cybernetics, 2007, 39, pp.83-96.

[9] Aït-Sahalia, Y. and Xiu, D. Using principal component analysis to estimate a high dimensional factor model with high-frequency data [J]. Journal of Econometrics, 2017, 201(2), pp.384-399.

[10] Bollerslev, T., Tauchen, G. and Zhou, H. Expected Stock Returns and Variance Risk Premia. Review of Financial Studies, 2009, 22(11), pp.4463-4492.