Research on stock index forecasting based on ARIMA-GARCH and SVM mixed model

Xin Cao*, Ziqing Zhao

College of Science, Chongqing University of Technology, Chongqing, 400050

*Corresponding author. Email: caoxin20220322@163.com

Abstract. With the development of information technology, financial forecasting is the research direction of traditional mathematical statistics and combines emerging data mining technologies such as machine learning and deep learning to make forecasts. This paper starts from the traditional time series model $ARIMA-GARCH$, which can handle the conditional heteroscedasticity of time series data and combines the advantages of a support vector machine (SVM) that can handle the nonlinear part of financial time data $ARIMA-GARCH-SVM$. Stock market forecasts. In order to test the prediction effect of the model and compare traditional prediction methods with modern machine learning prediction methods, this paper establishes ARIMA-GARCH, SVM, and the combination ARIMA-GARCH- SVM model of these two types, respectively.

Keywords: ARIMA-GARCH-SVM model, support vector machine, tushare, stock index forecast.

1. Introduction

After the reform and opening up, the country's financial market has been continuously developed and improved since its establishment, the income level has been continuously improved, and more and more individual investors have invested in the stock market. Yield, but with its high yield comes high risk. The erratic changes in stock prices make investors feel the complexity of this market. In order to obtain better investment returns and avoid investment risks, people urgently need a scientific forecasting method to guide investment, and they need to find a theory to explain the reasons for price changes. However, making predictions on financial data is very challenging due to the noisy, nonlinear, and non-stationary nature of financial data and being affected by sample size. Although financial market price forecasting is a global problem, it remains a very active area of research. Since the birth of financial markets, people have sought ways to predict changes in asset prices. Forecasting is an intriguing and challenging research topic, whether past, present, or future.

2. Empirical Analysis of Stock Index Forecast Based on Mixed Model

2.1. Source and description of data

This article is based on the R language and obtains data through the API port of tushare-pro. The selected data is the Shanghai Stock Exchange Index (000001), the time span is from January 1, 2018, to March 31, 2020, with a total of 543 pieces of data, and the selected variable indicators are open (opening price), pre_open (yesterday's opening price)), close (closing price), pre _close (yesterday's closing price), high (the highest price), pre_high (yesterday's highest price), low (the lowest price), pre_low (yesterday's lowest price). At the same time, 486 pieces of data from January 2018 to the end of December 2019 were selected as train_set (training set), and 58 pieces of data from January to March 2020 were selected as test_set (test set).

2.2. Stock index prediction based on the portfolio model

2.2.1. (1) Based on ARIMA- GARCH model

(1) Stationarity test

The stationarity test is carried out on the closing price data of the training set. First, the training set's time series, ACF, and PACF are visualized, as shown in Figure 1. We found that in the training set, the market index showed the characteristics of first falling and then rising, and at the same time, there was a considerable correlation before and after the sequence, and the autocorrelation function graph showed the characteristics of tailing, indicating that the data was unstable.

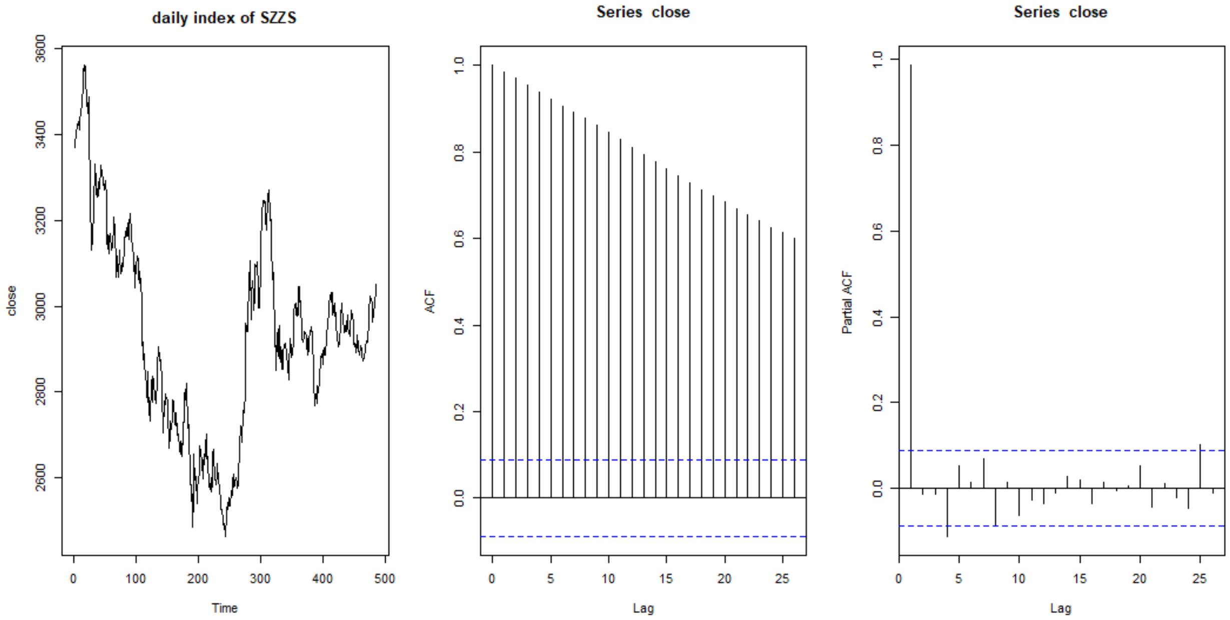

Figure 1. Correlation visualization of the training set

We perform the first-order difference on the original data and perform the ADF test on the horizontal and first-order difference series. The results are shown in Table 1. It can be seen from the test results that the horizontal sequence has a unit root in the three modes, but the three modes of the first-order difference sequence all pass the ADF test. We visualize the ACF and PACF of the differenced sequence, as shown in Figure 2. We can see that the acf graph presents the first-order truncation. That is, there is almost no before-and-after correlation after the first-order lag. $\{diff\_close_t\}$ Therefore, we select the closing price series after the first difference to establish the ARIMA-GARCH model.

Table 1. ADF Test Results

Model settings	original sequence		first-order difference sequence	
	ADF statistic	P_Value	ADF statistic	P_Value
No intercept term, no trend	-0.6103	0.4218	-15.3607	0.01
Intercept term with no trend	-2.1588	0.2537	-15.3542	0.01
There is an intercept term with a trend	-1.8547	0.6396	-15.4223	0.01

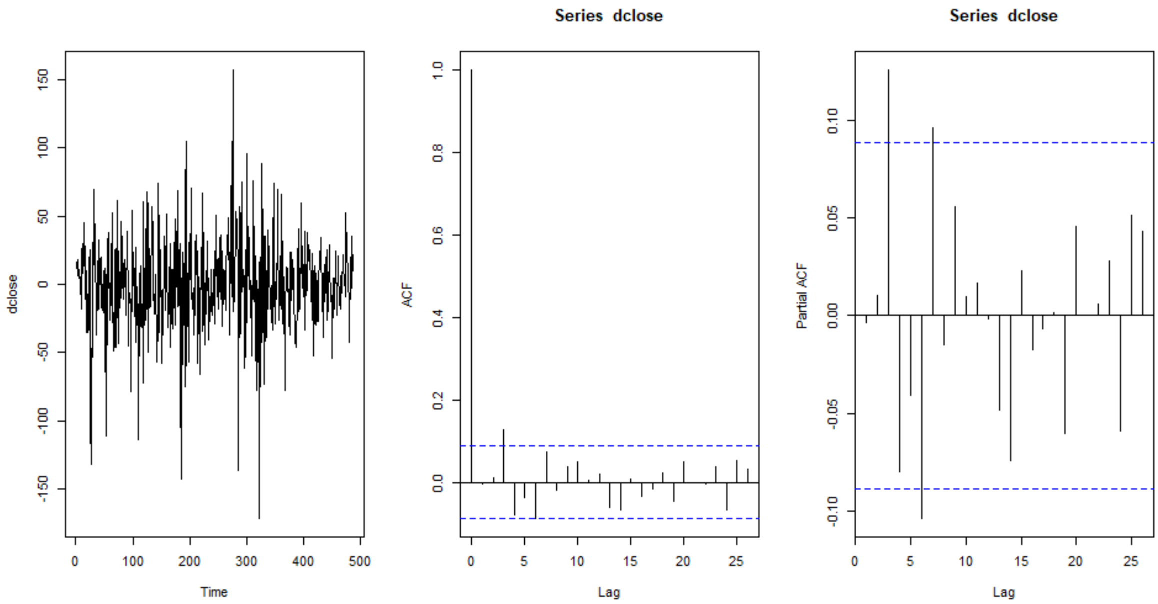

Figure 2. Correlation visualization of the training set after first differencing

(2) Determination of Order - Based on AIC Information Criterion

AIC is a standard for measuring the goodness of fitting of statistical models. Japanese statistician Hiroji Akaike proposed it in 1974. It is based on the concept of entropy, which provides a standard for weighing the complexity of the estimated model and the goodness of fitting data. Typically, it is a weighted function of the fitting accuracy and the number of unknown parameters, and AIC can be defined as follows:

$$AIC = -\ln(L) + 2K$$

Where K is the number of unknown parameters in the L model and is the maximum likelihood function in the model. When choosing the best model from a set of alternative models, the model with the smallest AIC is usually chosen.

By writing a function, traversing $P \in (0,5)$, $q \in (0,5)$, and calculating the AIC value of the model under different orders, we select a set of orders with the smallest AIC. The traversal results are shown in Table 2. We choose $p = 1, q = 3$ that the AIC at this time is the smallest, that is, to build the model $ARIMA(1,1,3)$.

(3) Fitting and testing of ARIMA(1,1,3) model

The model is fitted to the training set, and the parameter estimation results of the model are shown in Table $ARIMA(1,1,3)$3. Meanwhile, we tested the model's residuals to detect the model's fitting effect. As shown in Figure 3, we found that the residuals of the model had a weak correlation before and after, which indicates that the residuals were close to the white noise sequence, which could indicate the fitting of the model. Accuracy is very high.

Table 2. Parameter Estimation Results

	ar1	**ma1**	**ma2**	**ma3**
estimated value	-0.55	0.56	0.015	0.167
se	0.1766	0.1766	0.0513	0.0594

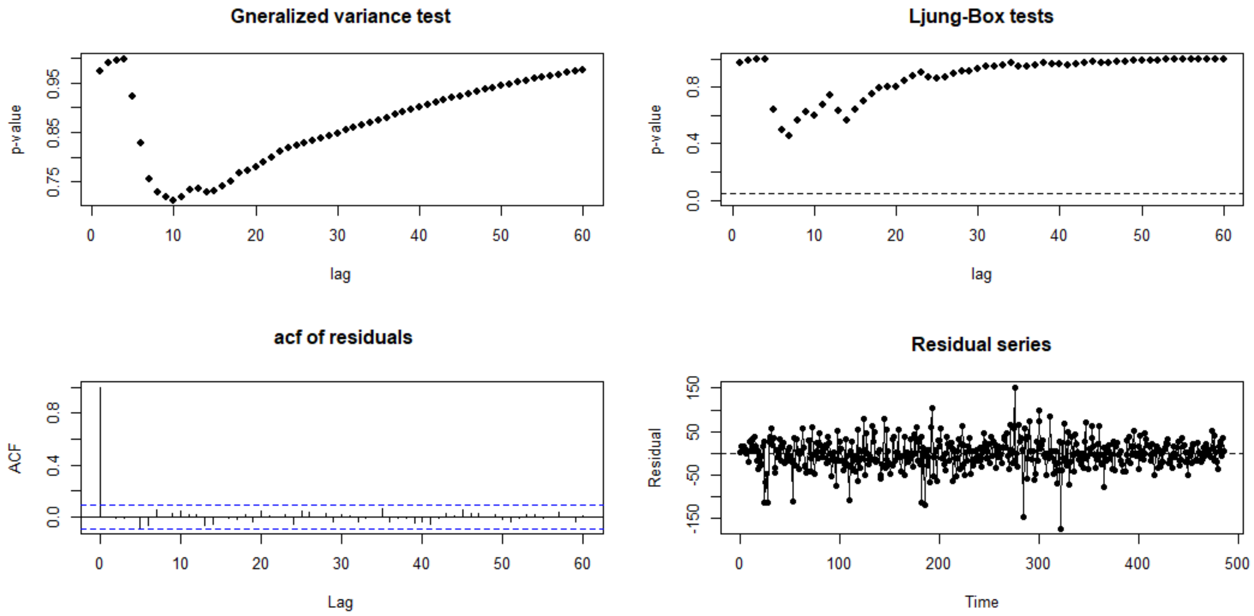

Figure 3. Residual test results

(4) LM test

The established above is $ARIMA(1,1,3)$ carried out. That is, the LM test is carried out. The test results are shown in Figure 4. It can be found that with the increase of the lag order, the more the ARCH effect is, which means that the training set has a cluster effect. Therefore, the GARCH model needs to be introduced into the original model. That is, the ARIMA-GARCH model must be established. At this time, we do not verify the m and s parameters of the GARCH model and adopt $m=1, s=1$ the parameters, that is, the most commonly used GARCH(1,1) model.

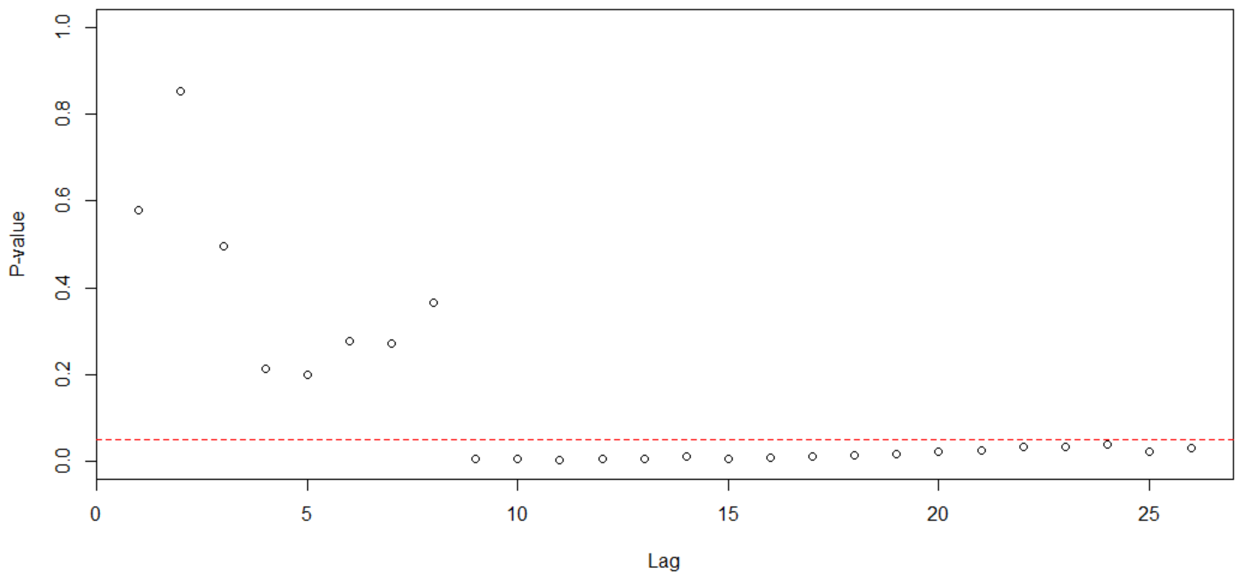

Figure 4. LM test results

(5) Establishment and fitting of ARIMA(1,1,3)-GARCH (1,1) model

For the training set, we choose a model with a lag order of (1,1), and the specific expression of the $ARIMA(1,1,3)$-$GARCH\ (1,1)$ model is:

$$\begin{cases} diff\_close_t = \mu_t + a_t \\ \mu_t = \phi_0 + \phi_1\mu_{t-1} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \theta_3 a_{t-3} \\ a_t = \sigma_t \cdot \varepsilon_t \\ \sigma_t^{\ 2} = \alpha_t + \alpha_1 a^2_{\ t-i} + \beta_1 \sigma^2_{\ t-1} \end{cases}$$

We use the latest rugarch package to perform all the following GARCH model fitting and prediction based on R software. The parameter estimation results of the ARIMA-GARCH model are shown below. We found that the parameter estimates passed the significance test based on the test statistic and P-value. Meanwhile, we found $\alpha_1 + \beta_1 \approx 1$ that the volatility is persistent. In other words, the model can well explain the clustering phenomenon of the Shanghai Composite Index, and it can also reflect the continuous fluctuation of the series after the shock.

Table 3. Parameter Estimation Results

	mu	**ar1**	**ma1**	**ma2**	**ma3**	**omega**	**alpha1**	**beta1**
estimated value	3367.7	0.998	-0.002	0.037	0.052	36.25	0.053	0.92
se	24.51	0.0028	0.04	0.043	0.047	28.28	0.023	0.033
t	137.38	355.11	-13.2	16.85	10.099	3.28	2.28	27.88
P value	0	0	0.03	0.039	0.002	0.02	0.022	0

(6) Prediction based on ARIMA(1,1,3)-GARCH(1,1) model

Use the fitted model $ARIMA(1,1,3)$-$GARCH(1,1)$ to predict the test set determined in Section 4.1. For better prediction, we adopt rolling prediction, the step size is 1, and the sliding window is the length of the training set. That is, a window is used for one-step prediction to ensure the accuracy of the prediction.

2.2.2. Based on the SVM model

(1) Feature selection

We artificially selected the characteristics of the fundamental indicators of the Shanghai Composite Index, namely yesterday's closing price $close_{t-1}$, yesterday's opening price $open_{t-1}$, yesterday's highest price $high_{t-1}$, yesterday's lowest price low_{t-1}, today's opening price $open_t$, today's highest price $high_t$, and today's lowest price low_t. The above features are used as the feature input of the model, and the output is today's closing price $close_t$. The specific formula can be written using the black-box model. After that, the SVM model is written as the SVR model.

(2) Establishment and solution of SVR model

The SVR black-box model can be written as:

$$close_t = SVR\left(open_t, high_t, low_t, open_{t-1}, close_{t-1}, high_{t-1}, low_{t-1}\right) + e_t$$

We cannot determine which kernel function to use here, so the Gaussian radial basis kernel function is now used, and the formula is defined as follows,

$$K\left(x_i, x\right) = \exp(-\gamma \|x_i - x\|^2)$$

After many selections of parameters and references from the paper, we finally determined the penalty coefficient $C = 0.03$, kernel function parameters $\gamma = 0.1$, and $\varepsilon = 0.009$.

The fitting of the SVM model is realized based on the e1071 package of R software.

(3) Prediction based on SVR model

Since the SVR model is a black-box model, the output value can be obtained only by given parameters and feature vectors, which is here $close_t$. We use the e1071 package to make predictions. In order to better display the prediction results, we also use the ggplot2 package to visualize the results.

2.2.3. Based on the ARIMA-GARCH-SV M model

ARIMA-GARCH-SVM model has explained above, which treats the SVR model as a semi-parametric nonlinear tool, combined with the GARCH model fitted to the training set σ_t^2 and predicted to the test set σ_t^2, as input The features are added to the SVR black-box model, and the specific model is as follows:

$$\begin{cases} diff\_close_t = \mu_t + a_t \\ \mu_t = \phi_0 + \phi_1\mu_{t-1} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \theta_3 a_{t-3} \\ a_t = \sigma_t \cdot \varepsilon_t \\ \sigma_t^2 = \alpha_t + \alpha_1 a^2_{t-i} + \beta_1 \sigma^2_{t-1} \\ close_t = SVR\left(open_t, high_t, low_t, open_{t-1}, close_{t-1}, high_{t-1}, low_{t-1}, \sigma_t^2\right) + e_t \end{cases}$$

We found that the SVR regression combined with the GARCH model has a better prediction effect than the pure SVR regression. This improvement lies in the large volatility, with a better trend effect, such as 40 days and close to 60 days. Compared with the SVR single model, it has a better prediction effect on volatility. The combination method combines the ability of traditional models to deal with heteroscedasticity, and the support vector machine can better capture nonlinear information.

3. Conclusion

We draw the following conclusions through the above research: (1) Financial time series data is characterized by nonlinearity, instability, and high noise. We compare traditional time series analysis with emerging machine learning methods and find that machine learning is applied in financial The advantage of prediction is that the traditional ARIMA-GARCH model can only handle linear information but cannot handle nonlinear information. However, the SVM model can handle nonlinear data better. After the empirical test, the traditional time series model combined with the machine learning method can show the excellent characteristics of its prediction. (2) The traditional model ARIMA-GARCH can handle time series with conditional heteroscedasticity and eliminate the arch effect by constructing a model, eliminating the model's heteroscedasticity. However, it is not good at capturing nonlinear information. (3) The parameters in the SVR model we use are artificially set. In order to predict the accuracy, the verification process based on grid search can be used to select the parameters C, ε, γ . At the same time, different kernel functions can be compared, and the continuation of this paper can be further studied. (4) We are thinking about the advantages of machine learning in processing financial data. Next, we can study the combination of machine learning, deep learning, and financial markets other than SVM support vector machines. For example, BP neural network combined with ARIMA- GARCH model for prediction, There are many works of literature and practices in this area, but it is still very challenging to predict financial markets from a purely mathematical perspective. Therefore, we can consider applying other theories when building the model, such as technical indicators and irrational investor sentiment indicators for forecasting and other indicators that can reflect the trend of the index, such as price-earnings ratios, etc.

References

[1] Gao Tong, Zhu Hailong. Volatility Prediction Based on GARCH-like Model and BP Neural Network Model——Based on Daily Data of Shanghai Composite Index [J]. Financial Economics, 2020(09):37-45.

[2] Cai Zedong. Stock index futures price prediction based on underlying stock index and machine learning [J]. China Business Review, 2020(17): 90-92.

[3] Cao Wei, Ren Siru. Research on RMB exchange rate fluctuation prediction based on LSTM and GARCH family mixed model [J]. Computer Application Research, 2020, 37(S1): 79-82.

[4] Zhu Wenxiu. Prediction and Analysis of CSI 300 Index Closing Price Based on Time Series Analysis [D]. Shandong University, 2020.

[5] Zhang Lu. Research on stock index prediction based on neural network [D]. Chongqing Technology and Business University, 2020.

[6] Sun Hao. fTGARCH model and GARCH-SVR model for predicting volatility of financial time series [D]. Dalian University of Technology, 2020.