Big Data Business Actual Analysis: Stock Price Prediction Based on Time Series Model

Aiwen Rui
Tianjin University of Finance and Economics, Tianjin 300222, China
Email: 3294679254@qq.com

Abstract: This paper selects the daily closing price data of the Shanghai Composite Index from January 1, 2016 to December 31, 2017, excluding holidays, and preprocesses the data. After taking the logarithm and converting it into the rate of return data, the first-order difference is performed to make it into a stable time series, and then the ARMA(p,q) model is constructed. Through parameter significance test, residual test and characteristic root test, according to the minimum principle of AIC, the optimal model is finally determined to be ARMA(2,5) of sparse coefficient, and the expression of the model is obtained. The GARCH(1,1) model is established for the residual of ARMA(2,5), and the model expression is obtained. In order to directly predict the return rate of the Shanghai Composite Index, the ARIMA(2,1,5) model of the sparse coefficient is constructed for the return rate of the Shanghai Composite Index, and the model expression is obtained. By predicting the Shanghai Composite Index return data on January 2, 2018, it is found that the prediction error of the model is small, and it can be used for subsequent predictions.

***Keywords*:** ARMA(2,5) of sparse coefficient, ARIMA(2,1,5), GARCH(1,1)

1. Construction of ARMA(p,q) model and residual GARCH(p, q) model

1.1 Data selection and preprocessing

Due to the large volatility and differences of individual stocks in the stock market, using the Shanghai Stock Exchange Index as a reference standard to reflect the stock market situation can more systematically predict the stock market trend. Select the daily closing price data of the Shanghai Stock Exchange index from January 1, 2016 to December 31, 2017, excluding holidays, draw a relevant time series graph for the closing price data y, and conduct an autocorrelation test, and find that the data is non-stationary.

Figure 1. Time series chart of Shanghai stock index return rate from 2016 to 2017

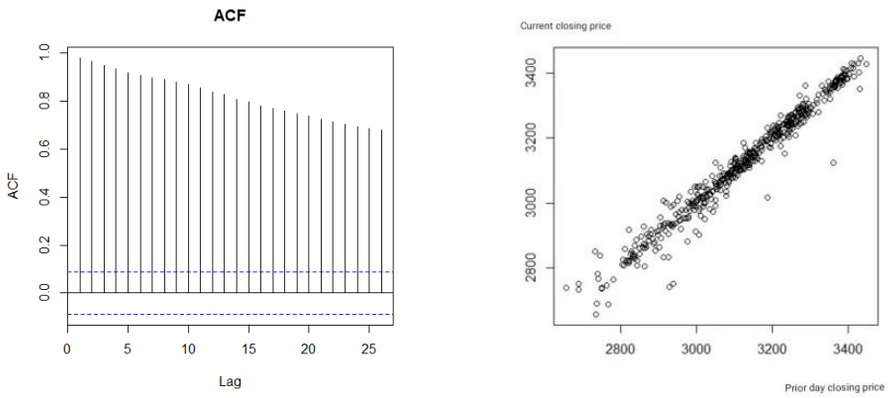

Figure 2. Autocorrelation graph of return rate and scatter plot of adjacent data

The general trend of the Shanghai Composite Index in the past two years has no obvious cycle and seasonality, and the fluctuation range is large. The ACF exhibits a tailing phenomenon, but the attenuation is slower and does not fall within the range. The adjacent data scatter chart shows the closing price of the previous day There is a clear positive correlation with the current closing price. Therefore, it can be judged that the sequence is non-stationary and does not satisfy randomness.

Next, the sequence is smoothed. Take the logarithm of the closing price sequence and transform it into a return rate sequence, and then perform a first-order difference to obtain the sequence ∇lny, namely:

$$\nabla lny = lny_t - lny_{t-1}, t = 1,2,3,\cdots,n$$

1.2 Stationarity test

It can be seen from the time series diagram that the time series after the first-order difference fluctuates at the zero mean value, and the ACF quickly decays to 0, and basically all fall within the interval, so the series is stable.

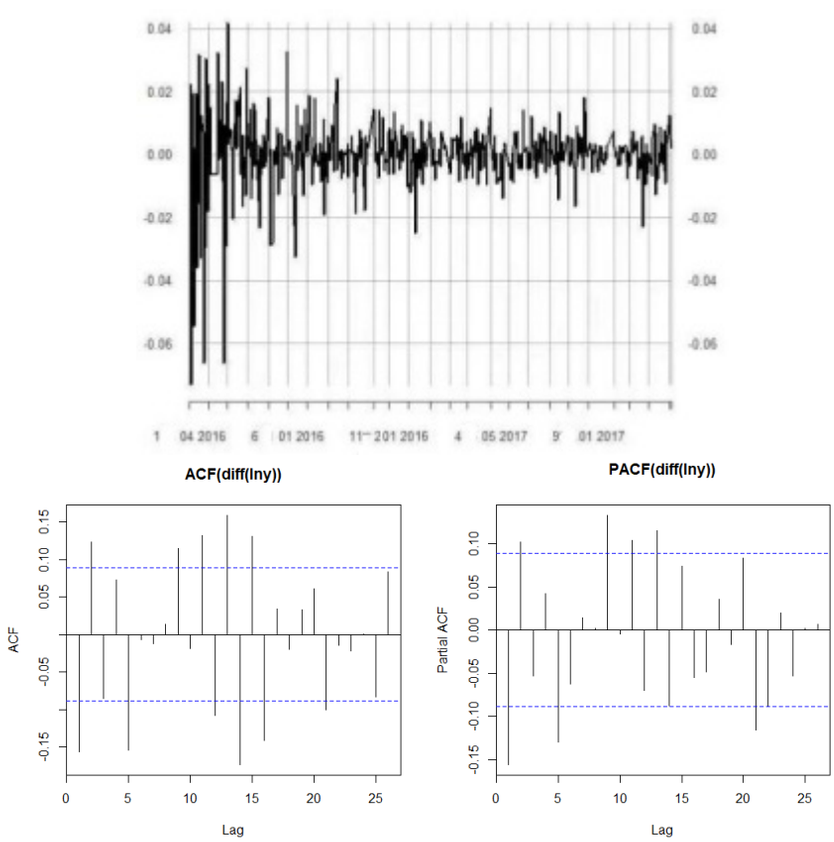

Figure 3. Time sequence diagram of return rate after first-order difference and ACF and PACF diagrams

1.3 White noise inspection

Table 1. White noise test of yield series after first-order difference

	Box – pierce test	
diff (*lnx*) – *squared*	23.519	40.441
df	6	12
p – *value*	0.00064	$6.07 \times e^{-5}$

At the 5% significance level, the null hypothesis is rejected, that is, the yield data after the first-order difference is not a white noise sequence.

1.4 Construction of ARMA(p,q) model

Since the yield data after the first-order difference is stable and non-white noise, the ARMA(p,q) model is selected to fit the data.

1.4.1 The order of the model

After the above-mentioned smoothing treatment, the order is further determined by eacf.

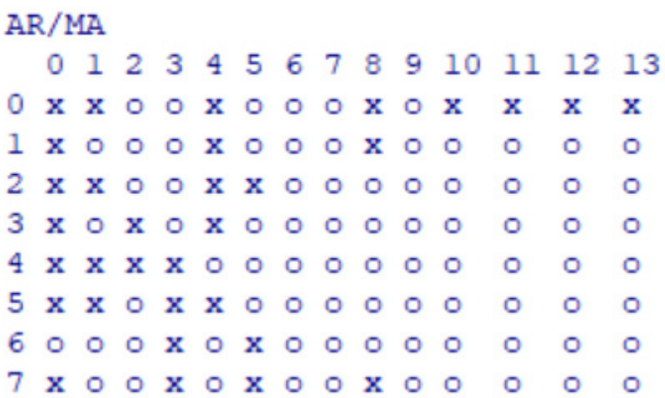

Figure 4. ARMA(p,q) model ordering diagram

It is preliminarily determined that the model is ARMA(3,5) or ARMA(1,5) model.

1.4.2 Model parameter estimation and testing

(1) Test and estimate the parameters of the model

Fit the data based on the ARMA(3,5) and ARMA(1,5) models, and adjust the parameters to make the model pass the significance test. The model ARMA(3,5) is reduced to the ARMA(2, 5), the model ARMA(1,5) is reduced to MA(5).

Table 2. Parameter estimation results of ARMA(p,q) model

Model	Parameter	Estimated value	Standard error	T test	P value	Significance
ARMA(2,5)	AR(2)	0.48324	0.09590	5.039	$4.68 \times e^{-7}$	***
	MA(1)	-0.08508	0.04008	-2.123	0.0338	*
	MA(2)	-0.47600	0.10182	-4.675	$2.94 \times e^{-6}$	***
	MA(5)	-0.11839	0.05146	-2.301	0.0214	*
MA(5)	MA(1)	-0.10452	0.04493	-2.326	0.02000	*
	MA(5)	-0.14120	0.04492	-3.144	0.00167	**

(2) Model checking and optimization

The white noise test on the residual sequence can verify the relevance of the model and the randomness of the variables, thereby verifying the rationality of the model. The residual sequence fluctuates around 0, and the autocorrelation coefficient and partial autocorrelation coefficient coefficients quickly decay to 0, and basically fall within the interval, so it can be judged that the residual is a stationary sequence.

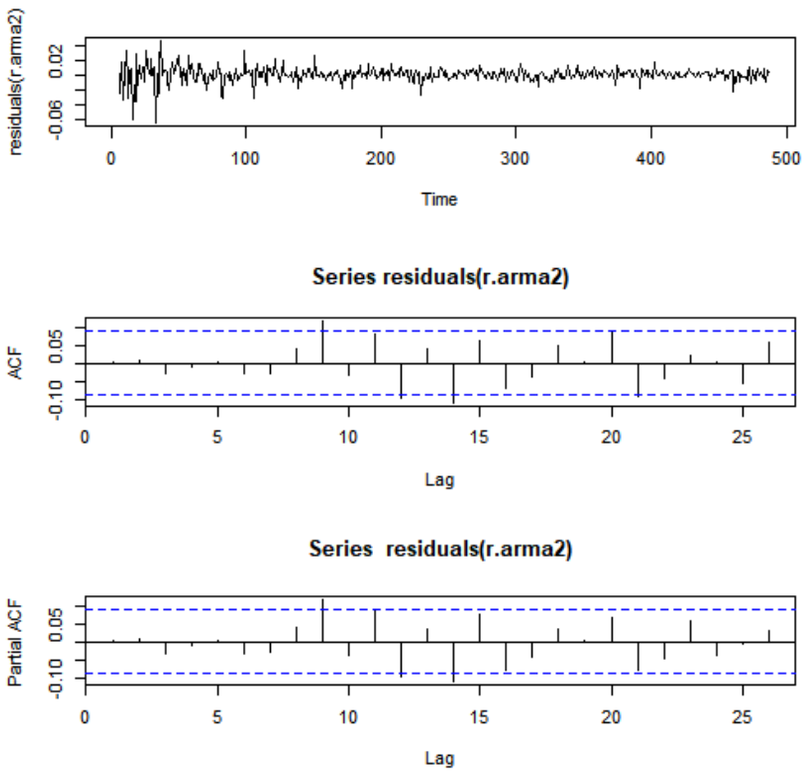

Figure 5. ARMA(2,5) residual sequence test chart

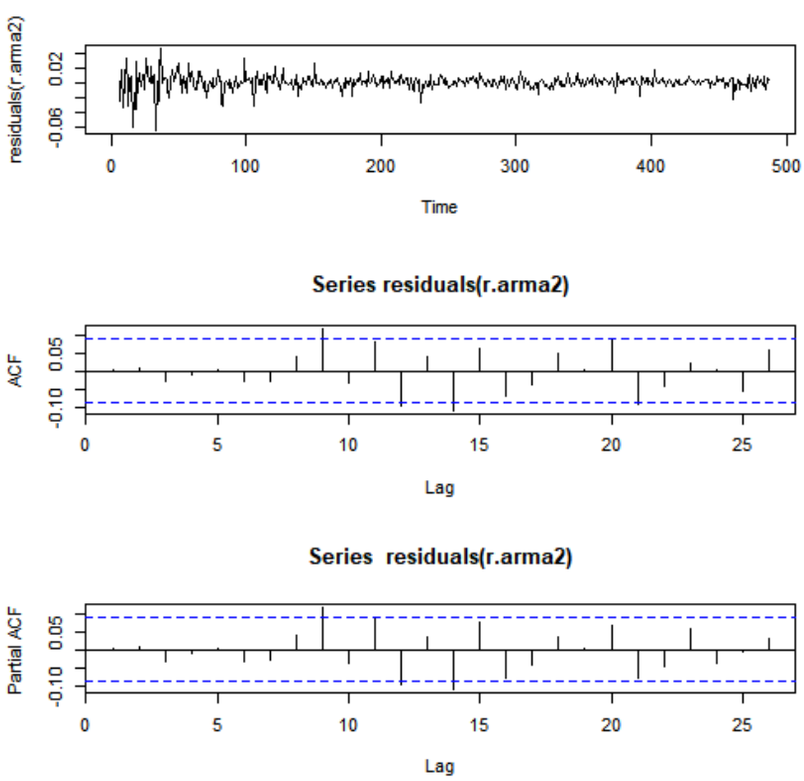

Figure 6. MA(5) residual sequence test chart

Draw a scatter plot of ARMA(2,5) model and MA(5) residual adjacent data, and find that there is no correlation between the residual data and it is random.

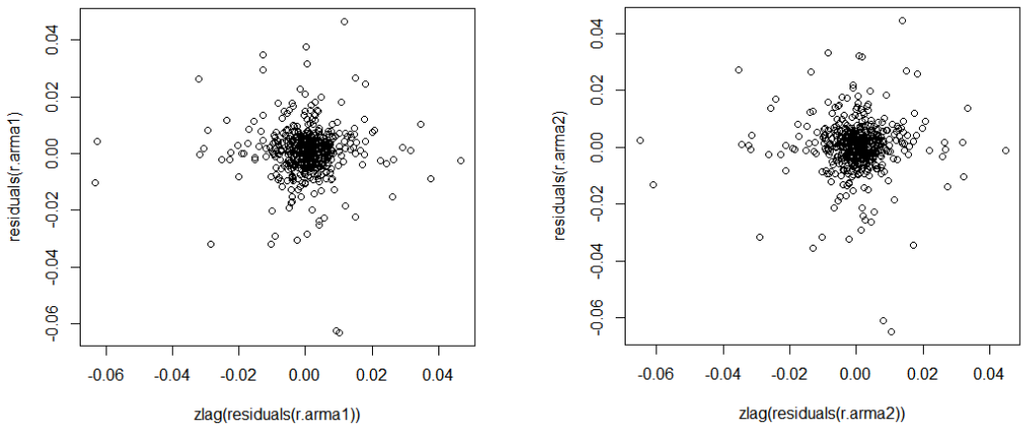

Figure 7. Scatter plot of adjacent data of residual sequence (left: ARMA(2,5); right: MA(5))

Through Box test, the original hypothesis that the residual sequence is a white noise sequence is accepted at a significance level of 5%, so the residual sequence is a white noise sequence.

Table 3. ARMA(2,5) residual sequence model parameter estimation results

	Box – pierce test		
ARMA(2,5)	$residuals(r.arma1)-squared$	1.5369	12.837
	df	6	12
	$p-value$	0.975	0.381
MA(5)	$residuals(r.arma2)-squared$	0.99372	17.488
	df	6	12
	$p-value$	0.9858	0.1321

In summary, the residuals of the ARMA(2,5) and MA(5) sequences are stationary white noise sequences, which proves that the conclusions of the model are valid, and the model already contains all the information of the data; in order to further test the rationality of the model , Discriminate the characteristic roots of the model, and find that the value of the characteristic roots of the two models is greater than 1, that is, all the characteristic roots are outside the unit circle. The model is determined to be the ARMA(2,5) of the sparse coefficient. The expression is:

$$\begin{cases} \widehat{d\ln y} = 0.483239X_{t-2} + \varepsilon_t - 0.0850833\varepsilon_{t-1} - 0.4759967\varepsilon_{t-2} - 0.1183926\varepsilon_{t-5} \\ \sigma_\varepsilon^2 = 9.009 \times e^{-5} \end{cases}$$

And the MA(5) of the sum sparse coefficient, the expression is:

$$\begin{cases} \widehat{d\ln y} = \varepsilon_t - 0.1045183\varepsilon_{t-1} - 0.1412021\varepsilon_{t-5} \\ \sigma_\varepsilon^2 = 9.243 \times e^{-5} \end{cases}$$

Choose the best model ARMA(2,5) through the AIC minimum principle.

Model	*ARMA*(2,5)	*MA*(5)
AIC	-3146.22	-3146.22

1.5 ARMA(2,5) model prediction

Next, use the model to simulate and predict the three working days from January 2 to 4, 2018.

Table 4. ARMA(2,5) model prediction values

Time	Observed value	ARMA(2,5) model prediction value
2018.01.02	0.01236702	0.001051217
2018.01.03	0.06187650	-0.00353813
2018.01.04	0.04915553	0.0008742533

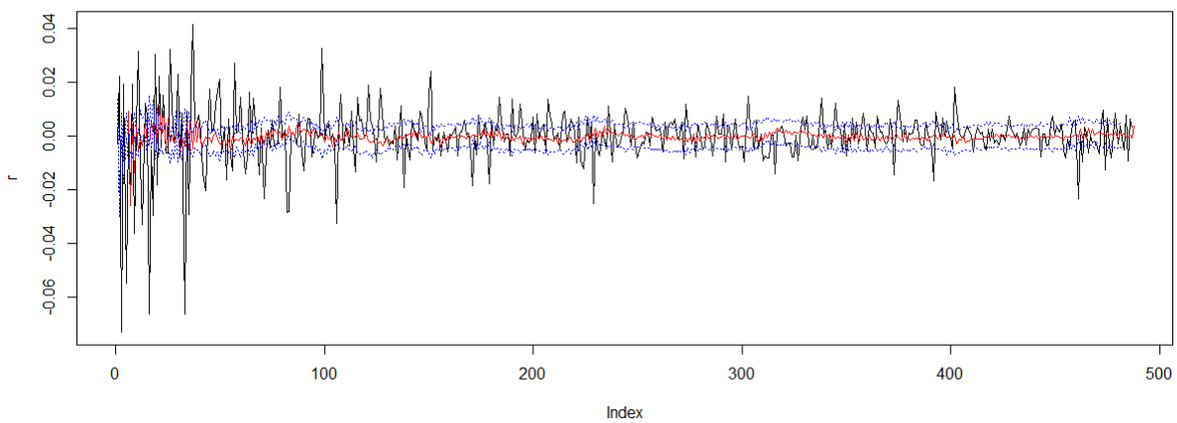

Figure 8. ARMA(2,5) model prediction graph

In Figure 8, the black curve represents the fitted sequence curve, the red curve represents the predicted value, and the blue dotted line represents the 95% confidence upper and lower limits of the fitted sequence, and the predicted value is basically within the confidence interval.

1.6 Conditional heteroscedasticity modeling

After obtaining the mean equation of the ARMA(2,5) model with sparse coefficients, the correlation test and conditional heteroscedasticity test are performed on the residual sequence of the model, and the GARCH(q,p) model is established for the sparse coefficient ARMA(2,5). The residual term of the model is fitted.

1.6.1 Heteroskedasticity autocorrelation test

Conditional heteroscedasticity is the ARCH effect. By drawing a time series diagram of the residuals, which is the sum of squares of the residuals, it is found that the series residuals have a fluctuating clustering effect. Afterwards, the residuals are tested and the test results are significant, most of which are less than 0.05, indicating the model The residual sequence of has ARCH effect, you can try to build a GARCH model for the residuals.

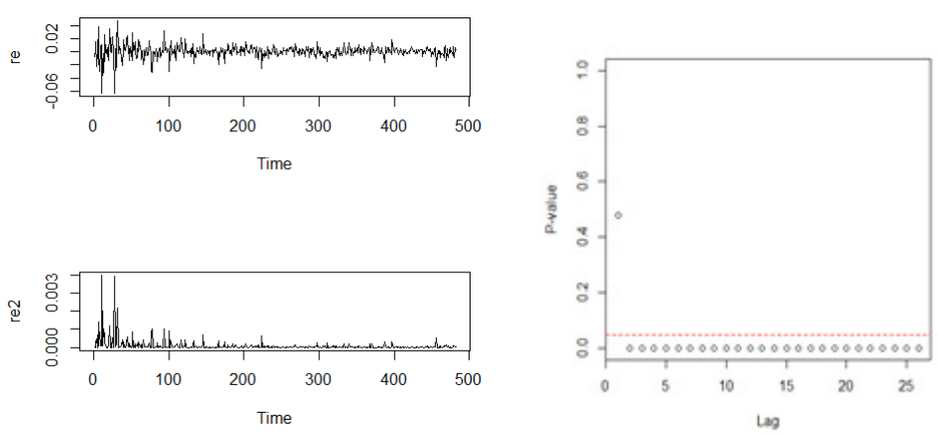

Figure 9. Residual sequence diagram and test diagram

1.6.2 GARCH model order determination and parameter estimation

First, establish the ARMA(p,q) model to determine the order of the square of the residual.

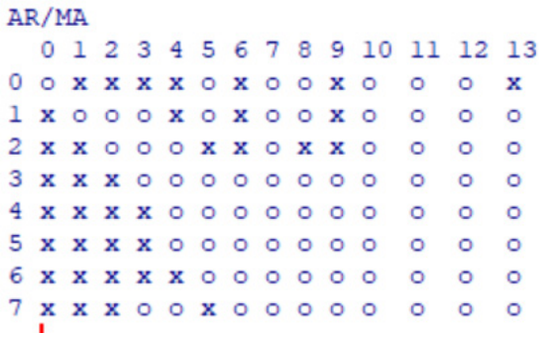

Figure 10. Model ordering diagram

Finally, the model is determined to be ARMA(1,1), and then the GARCH(1,1) model is fitted to the residual sequence. Adjust the parameters to make the model pass the significance test, and obtain the GARCH(1,1) model structure as:

$$\varepsilon_{t|t-1}^2 = 6.751\times e^{-7} + 0.9391\varepsilon_{t-1|t-2}^2 + 4.814\times e^{-2}\times\gamma_{t-1}^2$$

Table 5. GARCH(1,1) model parameter estimation results

Parameter	Estimated value	Standard error	T test	P value	Significance
a_0	$6.751\times e^{-7}$	$2.061\times e^{-7}$	3.176	0.00105	**
a_1	$4.814\times e^{-2}$	$7.798\times e^{-3}$	6.174	$6.68\times e^{-10}$	***
b_1	0.9391	$9.030\times e^{-3}$	103.998	$<2\times e^{-3}$	***

1.6.3 Normality test of residuals

To test the residual of GARCH(1,1), the p value is less than 0.05, and the null hypothesis of normality test cannot be accepted.

Table 6. *Jarque Bera* Checklist

Jarque Bera Test	
$X-squared$	253.74
df	2
$p-value$	$2.2\times e^{-16}$

Through Box test, it is found that the p-values are all greater than 0.05, which shows that the residuals of the GARCH(1,1) model are white noise sequences.

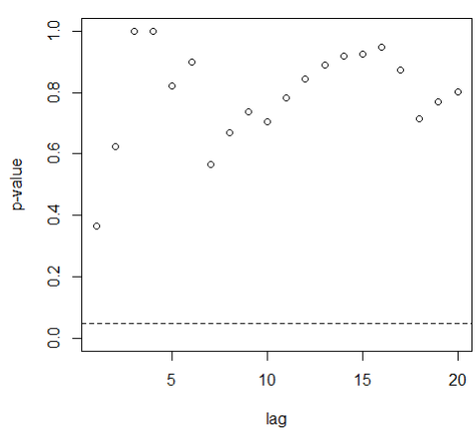

Figure 11. The residual Box test

In summary, the ARMA(2,5)-GARCH(1,1) model is fitted to the return data after the first-order difference. The expression to get its mean and variance is:

$$\widehat{d\ln y} = 0.483239X_{t-2} + \varepsilon_t - 0.0850833\varepsilon_{t-1} - 0.4759967\varepsilon_{t-2} - 0.1183926\varepsilon_{t-5}$$

$$\varepsilon^2_{t|t-1} = 6.751\times e^{-7} + 0.9391\varepsilon^2_{t-1|t-2} + 4.814\times e^{-2}\times\gamma^2_{t-1}$$

2. ARIMA (p, d, q) model construction and prediction

Above we constructed the ARMA(2,5) model to predict the return sequence after the first-order difference, and constructed the GARCH(1,1) model for the residual items. In order to more intuitively predict the return rate of the Shanghai Composite Index in the next few days, we build an ARIMA(2,1,5) model for the non-stationary return rate series based on the above analysis.

2.1 Unit root test

In order to further test the stationarity of the return rate series, an ADF test is performed on the data. At a significance level of 1%, the return rate series is not stable.

Table 7. ADF inspection

Augmented Dickey-Fuller Test	
Dickey–Fuller	−3.6401
Lag order	7
p–value	0.02889

2.2 ARIMA (p, d, q) model construction

Based on the time series ARMA(2,5), by adjusting the parameters, the model passes the significance test, and the model is determined to be the ARIMA(2,1,5) of the sparse coefficient. The expression is:

$$\begin{cases} \widehat{\nabla\ln y} = 0.3457X_{t-2} + \varepsilon_t - 0.1184\varepsilon_{t-1} - 0.2272\varepsilon_{t-2} - 0.1586\varepsilon_{t-5} \\ \sigma^2_\varepsilon = 0.0001061 \end{cases}$$

Table 8. ARIMA(2,1,5) parameter estimation results

Parameter	Estimated value	Standard Error
AR(2)	0.3457	0.1843
MA(1)	-0.1184	0.0455
MA(2)	-0.2272	0.1748
MA(5)	-0.1586	0.0475
$\widehat{\sigma^2_\varepsilon} = 0.0001061$; Log likelihood value=1537.26; *AIC*=-3066.51		

2.3 ARIMA(2,1,5) model prediction

Use the ARIMA(2,1,5) model to simulate and predict the return rate data for one working day on January 2, 2018, and draw the forecast map.

Table 9. ARIMA(2,1,5) predicted values

Time	Observation value	ARIMA(2,1,5) model prediction value	Error ratio
2018.01.02	8.10754	8.116216	0.107%

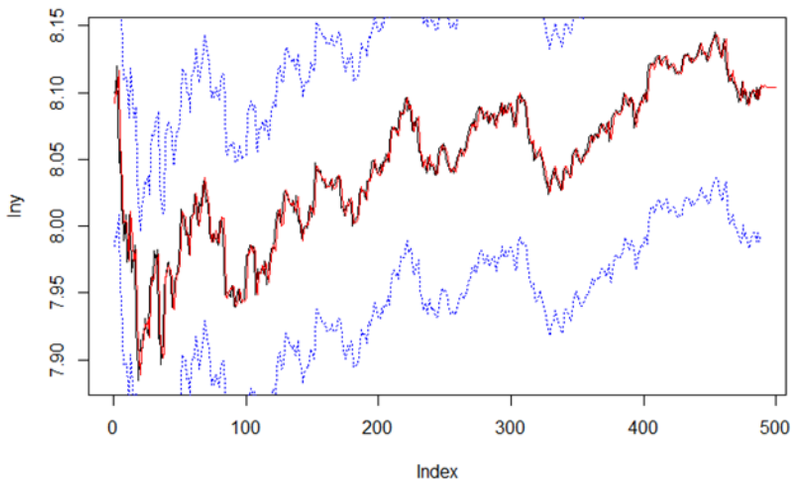

Figure 12. ARIMA(2,1,5) forecast chart

As can be seen from Figure 12, the black curve represents the fitted sequence curve, the red curve represents the predicted value, and the blue dashed line represents the 95% confidence upper and lower limits of the fitted sequence. The predicted value is basically within the range of the confidence interval, and the prediction is relatively accurate.

3. Conclusion

Fit the sparse coefficient ARMA(2,5) to the return data after the first-order difference, the expression of the model is: $\widehat{d\ln y} = 0.483239X_{t-2} + \varepsilon_t - 0.0850833\varepsilon_{t-1} - 0.4759967\varepsilon_{t-2} - 0.1183926\varepsilon_{t-5}$ And build the GARCH(1,1) model for the residuals of ARMA(2,5), and get the model expression as: $\varepsilon^2_{t|t-1} = 6.751\times e^{-7} + 0.9391\varepsilon^2_{t-1|t-2} + 4.814\times e^{-2}\times\gamma^2_{t-1}$

In order to directly predict the rate of return of the Shanghai Composite Index, the ARIMA (2,1,5) model of the sparse coefficient is constructed for the rate of return of the Shanghai Composite Index, and the model expression is obtained as:

$$\nabla\widehat{\ln y} = 0.3457X_{t-2} + \varepsilon_t - 0.1184\varepsilon_{t-1} - 0.2272\varepsilon_{t-2} - 0.1586\varepsilon_{t-5}$$

By predicting the Shanghai Composite Index return data on January 2, 2018, it is found that the prediction error of the model is small, and it can be used for subsequent predictions.

References

[1] Xu Jun. Empirical analysis of gold futures prices based on the ARMA model-523 sets of data from the New York Stock Exchange from 2006 to 2016. *Industrial Economic Forum*. 2017; 04(04): 16-22.

[2] Guo Xue, Wang Yanbo. Forecast of Shanghai stock index based on ARMA model. *Times Economics and Trade*. 2006; (S3): 58-59.

[3] Deng Jun, Yang Xuan, Wang Wei, Jiang Zhehui. Empirical research on stock price prediction using ARMA model. *Enterprise Herald*. 2010; (06): 266-267.

[4] Huang Lixia. Analysis and forecast of stock price based on ARIMA model — Taking Ping An of China as an example. *Science and Technology Economic Market*. 2020; (10): 62-63.