Research on stock trend prediction methods - Taking CSI 300 and CSI 500 as examples

Yi Qiao[1, *], Shihao Feng[2, †], Ruizhi Wu[3, †]

[1]School of Finance and Taxation, Central University of Finance and Economics, Beijing, China

[2]School of Business, Central University of Finance and Economics, Beijing, China

[3]School of Finance, Central University of Finance and Economics, Beijing, China

* Corresponding Author Email: qyszsy2016@163.com

† These authors contributed equally.

Abstract. As an important part of the capital market, the stock market can not only reflect the quality of the macro economy in time, but also reflect the policy changes in the change of the stock price.In addition, the stock market not only provides a public financing platform with strong liquidity for enterprises, but also plays a crucial role in the redistribution of social resources.This paper takes China's CSI 300 index CSI 500 index from 2007 to 2022 as the research object, and uses ARIMA, ARCH family and other models to implement it, which provides a better solution for the prediction of stock trends.

Key words: ARIMA, ARCH Stock market, trend prediction.

1. Introduction

At present, China's stock market has a considerable scale.With the development of the market becoming more mature, the securities market has mastered a bigger and bigger voice in China's macroeconomic regulation.Although the supervision and regulation of the securities market are increasingly perfect, the maturity of the securities market in China is far lower than the effectiveness of foreign capital markets. Many problems have been exposed under the high-speed development, such as the insufficient supervision. [1] There are a lot of irrational investment. [2] The suspension supervision is loose and so on, they are the factors that cause the abnormal fluctuations of Chinese stock market, the normal fluctuations of stock price can not only bring profits to investors, but also play a key role in the development of the capital market and the stability of our internal economic environment. [3] Therefore, research on stock prices can not only provide a reference for the implementation of macro policies, but also promote the effectiveness of the market and reduce unnecessary losses due to irrational investment. [4] However, there are a large number of investment targets and financial assets in the securities market, and it is often impossible to study the securities market thoroughly. Therefore, it is of great significance to study the comprehensive stock index that can reflect the market.

2. Data preprocessing

Firstly, the two groups of temporal data to be analyzed are preprocessed, and the programming platform is introduced. [5] Secondly, the traditional models are used to analyze the two groups of data. In the empirical analysis of the traditional model, we use Stata16 as the programming platform.For the stationarity test, ADF test was used in this section to determine whether the data were stationary. The ADF statistic of the original series of China Certificate 500 was -2.45, and the critical value was -3.43 at the 1% significance level, so the null hypothesis of the existence of unit root could not be rejected. Therefore, the null hypothesis of unit root could not be rejected.The ADF statistic of the original sequence of CSI 300 is -2.117, and the critical value is -3.43 at the 1% significance level. Therefore, the hypothesis of unit root cannot be rejected, indicating that the original sequence is not stationary. [6] Therefore, the two original sequences were processed with

first difference respectively, and the ADF statistic was used to test the differenced sequences again. Both of them rejected the null hypothesis of the existence of unit root at the significance level of 1%, indicating that the differenced series were stationary. Figure 1 shows the combination of the two groups of stock index series and their first-order difference series. It can be seen from Figure 1 that the original series are not stationary, but after the first-order difference, the series are stationary.

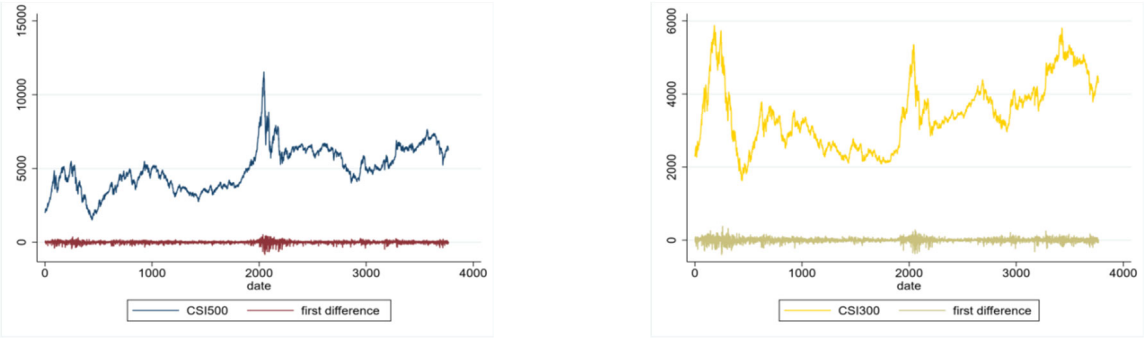

Figure 1. Combination diagram of first-order difference series

3. Analysis of experimental results

3.1 ARMA model

After data preprocessing, the data series have been stabilized, so when constructing the ARMA model, only the autoregressive order (AR) and the moving average order (MA) need to be determined. [7] Firstly, for the CSI 500 index, the PACF graph and ACF graph are drawn to judge the AR order and MA order. Figure 2 and Figure 3 show the PACF and ACF plots of the first-order difference sequence of the CSI 500 index.

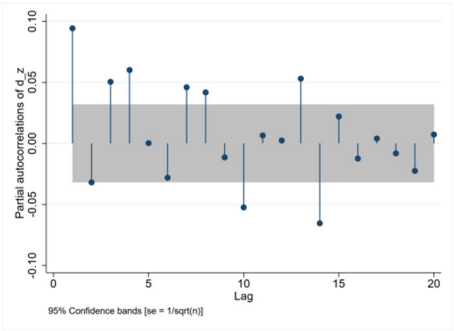

Figure 2. The PACF diagram of Syndrome 500

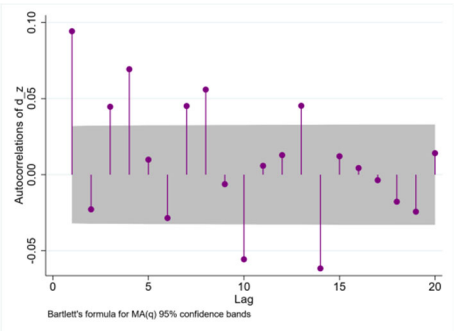

Figure 3. ACF diagram of Syndrome 500

It can be seen that the series of syndrome 500 after the first difference has autoregressive significance and moving average significance in multiple orders. After tests, we find that the ARMA model has the best fitting effect when AR(4) and MA(4) are used to construct the ARMA model. [8]Table 1 shows the specific situation of ARMA model fitting.

Table 1. ARMA of Evidence 500

ARMA	*Coef.*	*Std.Err.*	*z*	*P > |z|*
AR				
L1.	0.49217	0.099775	4.93	0
L2.	0.59973	0.057024	10.52	0
L3.	0.67147	0.086093	7.8	0
L4.	0.225806	0.068857	3.28	0.001
MA				
L1.	0.589287	0.101424	5.81	0
L2.	0.636881	0.061924	10.28	0
L3.	0.778068	0.083867	9.28	0
L4.	0.12135	0.075546	1.61	0.108

Similarly, we performed the same process for CSI 300 index. First, PACF and ACF plots were used to judge the order, and then different orders were tested to obtain the ARMA model with the optimal order. [9] Figure 4 and Figure 5 show the PACF and ACF.

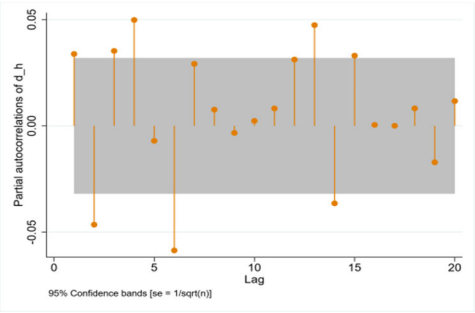

Figure 4. PACF diagram of CSI 300

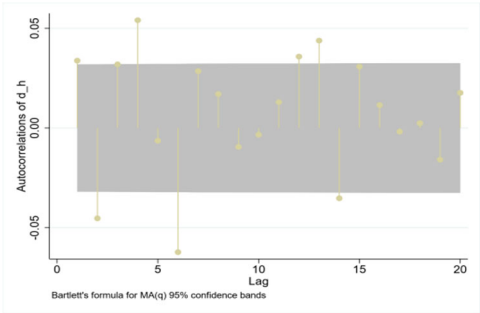

Figure 5. ACF diagram of CSI 300

Similar to CSI 500 index, CSI 300 index also has significant autoregressive order and moving average order in different orders.Through experiments, we find that the ARMA model with two orders of AR(6) and MA(6) has the best fitting effect.Table 2 shows the fitting of the ARMA model

Table 2. ARMA of csi 300

ARMA of CSI300				
ARMA	*Coef.*	*Std.Err.*	*z*	$P>\|z\|$
AR				
L1.	0.63296	0.15484	4.09	0
L2.	0.58183	0.059689	9.75	0
L3.	0.344654	0.088747	3.88	0
L4.	0.411707	0.082043	5.02	0
L5.	0.99372	0.046686	21.29	0
L6.	0.320035	0.140019	2.29	0.022
MA				
L1.	0.672764	0.151817	4.43	0
L2.	0.576165	0.059857	9.63	0
L3.	0.31122	0.086787	3.59	0
L4.	0.39392	0.082566	4.77	0
L5.	0.97028	0.04448	21.81	0
L6.	0.37587	0.131928	2.85	0.004

3.2 ARCH model

Before constructing ARCH model, we need to judge whether two sets of first-order difference series have volatility aggregation.From the data preprocessing stage, we can see that both groups of first-order difference series have volatility aggregation. [10]Therefore, it is necessary to further verify whether the two groups of sequences are suitable for ARCH model and GARCH family model.Firstly, VAR series models were used to determine the order of the best autoregressive model by comparing the values of the information criteria.Table 3 and Table 4 show the reliability criteria of the two sets of sequences respectively.

Table 3. CSI500 Information Criteria Table

VAR of CSI 500								
lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	22556.7				9548.19	12.002	12.0026	12.0036
1	- 22540.	33.447	1	0	9468.65	11.9936	11.9948	11.9969
2	22538.1	3.8944	1	0.048	9463.88	11.9931	11.9949	11.9981
3	22533.2	9.6333	1	0.002	9444.68	11.9911	11.9934	11.9977
4	22526.4	13.593	1	0	9415.6	11.988	11.9909	**11.9963 ***
5	22526.4	0.00048	1	0.983	9420.61	11.9885	11.9921	11.9985
6	22524.9	3.0035	1	0.083	9418.09	11.9883	11.9924	11.9999
7	22520.9	8.0116	1	0.005	9403.04	11.9867	11.9914	11.9999
8	22517.7	6.5587 *	1	0.01	**9391.64 ***	**11.9855 ***	**11.9908 ***	12.0004

Table 4. CSI300 Information Criteria Table

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	20721.1				3595.55	11.0253	11.0259	**11.027 ***
1	- 20719.	4.2296	1	0.040	3593.42	11.0247	11.0259	11.0281
2	20714.9	8.1518	1	0.004	3587.54	11.0231	11.0249	11.0281
3	20712.5	4.7674	1	0.029	3584.9	11.0224	11.0247	11.029
4	20707.8	9.3822	1	0.002	3577.87	11.0204	11.0233	11.0287
5	20707.7	0.18455	1	0.667	3579.6	11.0209	11.0244	11.0308
6	20701.2	13.154 *	1	0.000	3568.99	11.0179	**11.022 ***	11.0295
7	20699.5	3.2996	1	0.069	**3567.76 ***	**11.0176 ***	11.0223	11.0308
8	20699.4	0.21761	1	0.641	3569.45	11.018	11.0233	11.033

Therefore, according to the two groups of information criteria tables, the best autoregressive order of the CSI 500 series is 8, and the best autoregressive order of the CSI 300 series is 7.

Next, we constructed the autoregressive model to obtain the residuals of the two groups of models, and performed LM test on the residuals to determine whether the two groups of sequences had ARCH effect.Table 5 and Table 6 show the LM test tables of the two groups of data, respectively

Table 5. shows the 500 LM test

LM test of CSI500		
LM test for autoregressive conditional heteroskedasticity (ARCH) chi2	df	Prob>Chi2
362.626	1	0.000
635.459	2	0.000
755.223	3	0.000
850.882	4	0.000
877.693	5	0.000
877.812	6	0.000
878.734	7	0.000
889.016	8	0.000

H0: no ARCH effects vs. H1: ARCH(p) disturbance

Table 6. CSI 300 LM test

LM test of CSI300		
LM test for autoregressive conditional heteroskedasticity (ARCH) chi2	df	Prob>Chi2
159.417	1	0.000
253.803	2	0.000
348.362	3	0.000
424.367	4	0.000
488.729	5	0.000
493.399	6	0.000
497.816	7	0.000

H0: no ARCH effects vs. H1: ARCH(p) disturbance

It can be seen from Table 5 and Table 6 that both sets of sequences have strong ARCH effects, so we can use ARCH model and GARCH family model to analyze the sequences.

First we need to determine the order of the ARCH model, so we need to determine the best autoregressive order of the square of the residual term.In a similar way, VAR series models were used to determine the optimal autoregressive order of the residual values of the two sets of sequences. Finally, we found that both of them were optimal at the fourth order, so the ARCH (4) model was used to fit the sequences. Table 7 and Table 8 show the specific performance of ARCH model in two sets of different sequences.

Table 7. ARCH of CSI500

ARCH of CSI500				
	Coef.	Std.Err.	z	$P > \|z\|$
L1.	0.140448	0.016355	8.59	0
L2.	0.188895	0.017626	10.72	0
L3.	0.214808	0.018478	11.63	0
L4.	0.137144	0.014226	9.64	0
_cons	2933.939	108.4334	27.06	0

Table 8. ARCH of CSI300

ARCH of CSI300				
	Coef.	Std.Err.	z	$P > \|z\|$
L1.	0.150266	0.015898	9.45	0
L2.	0.227499	0.017089	13.31	0
L3.	0.216355	0.017489	12.37	0
L4.	0.209658	0.018318	11.45	0
_cons	1000.796	38.98599	25.67	0

Table 7 and Table 8 show that the ARCH model has a good estimation ability for the price difference series, and the coefficients of each order are significant.

4. Conclusion

The prediction of financial time series has always been a hot research field. In this paper, we take two representative stock indexes in China -- CSI 500 and CSI 300 as the research objects, and establish ARMA and ARCH models respectively. They play an irreplaceable role in analyzing the characteristics of financial series.When constructing the ARCH model, we find that the volatility of the price series of the two indexes is clustered.

In future research, traditional models can be used to analyze the characteristics of financial time series, so as to design suitable machine learning models for financial time series, so as to achieve the purpose of promoting the development of traditional models and machine learning models.

References

[1] Jiang, Z., Xu, D., & Liang, J. (2017). A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem. ArXiv, Abs / 1706.10059.

[2] Wang, J., Zhang, Y., Tang, K., Wu, J., & Xiong, Z. (2019). AlphaStock: A Buying-Winners-and-Selling-Losers Investment Strategy using Interpretable Deep Reinforcement Attention Networks. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

[3] Wang, Z., Huang, B., Tu, S., Zhang, K., & Xu, L. (2021). DeepTrader: A Deep Reinforcement Learning Approach for Risk-Return Balanced Portfolio Management with Market Conditions Embedding. AAAI.

[4] Cong, L.W., Tang, K., Wang, J., & Zhang, Y. (2020). AlphaPortfolio: Direct Construction Through Deep Reinforcement Learning and Interpretable AI. Capital Markets: Asset Pricing & Valuation eJournal.

[5] Liu, X., Yang, H., Chen, Q., Zhang, R., Yang, L., Xiao, B., & Wang, C. (2020). FinRL: A Deep Reinforcement Learning Library for Automated Stock Trading in Quantitative Finance. Capital Markets: Market Microstructure eJournal.

[6] Yang, H., Liu, X., Zhong, S., & Walid, A. (2020). Deep reinforcement learning for automated stock trading: an ensemble strategy. Proceedings of the First ACM International Conference on AI in Finance.

[7] Yang, M., Zheng, X., Liang, Q., Han, B., & Zhu, M. (2022). A Smart Trader for Portfolio Management based on Normalizing Flows. IJCAI.

[8] Wang, H., Wang, T., Li, S., Zheng, J., Guan, S., & Chen, W. (2022). Adaptive Long-Short Pattern Transformer for Stock Investment Selection. IJCAI.

[9] Mguni, D.H., Sootla, A., Ziomek, J., Slumbers, O., Dai, Z., Shao, K., & Wang, J. (2022). Timing is Everything: Learning to Act Selectively with Costly Actions and, between, regional Constraints. ArXiv, ABS /2205.15953.

[10] Pretorius, R., & van Zyl, T.L. (2022). Deep Reinforcement Learning and Convex Mean-Variance Optimisation for Portfolio Management. ArXiv, Abs / 2203.11318.