Volatility Prediction Based on the Multifactorial Linear Model

Tingyu Dai[1,a *, †] Jiawen Song[1,b *, †] Shenghui Wang[2,c *, †]

*dait@miamioh.edu[a], *songj21@miamioh.edu[b], *shenghui.wang@outlook.com[c]

Department of Mathematics, Miami University, Ohio, United States[1]
Department of Economics & Finance, Durham University, Durham, United Kingdom[2]
[†] These authors contributed equally.

Abstract—Volatility prediction serves as an important role in financial analysis, which could assist the investors in their financial decisions with the appropriate techniques of prediction on volatility. Meanwhile, volatility models play key roles in the academic literature for testing the fundamental trade-off between risk and return of financial assets and investigating the causes and consequences of the volatility dynamics in the economy. The paper investigates the feasibility of predicting volatility based on the multifactorial linear model. Specifically, 10 highly related factors, selected from systematically scanning, are utilized to predict 10 minutes, 30 minutes, 60 minutes, and 240 minutes of return volatility from the CSI 300 index futures. Three linear regression models including OLS, Lasso, and Ridge are constructed to predict volatility. According to the results, it is verified the prediction method is valid, and the overall goodness of fit is approximately 20%. In addition, each frequency has reported similar effects as expected. Based on the analysis, it is noted that the lower the frequency of the volatility, the better the result will be. These results shed light on volatility forecasting based on multifactorial models.

Keywords-Volatility prediction; multifactorial model; linear model; OLS; Ridge; Lasso.

1. INTRODUCTION

Since Markowitz proposed the Portfolio Theory in 1952, more and more investors could analyse and calculate the market on a large scale. However, the failure of the CAPM model has put scholars into an impasse. In 1970, Eugene Fama from the University of Chicago pointed out the potential contradiction in CAPM. He stressed that the available information assumptions reflected by prices must be tested in the context of expected return models (e.g., CAPM) [1]. With the development of the stock market, stock index futures have become a derivative in today's era. Market volatility has become an important factor in determining the price. Nevertheless, there is no stable specific pattern of volatility. Thus, the prediction of volatility

ICEMME 2022, November 18-20, Nanjing, People's Republic of China

DOI 10.4108/eai.18-11-2022.2327125

has become a significant step. In financial derivatives, futures are a very important tool, and the volatility of assets significantly impacts the futures price.

The main purpose of this paper is to find out the effective way to predict volatility and ensure the accuracy of prediction. In order to make the simulated trading in the most realistic situation, different from Ref. [2], the linear regression models are implemented into the past data of the CSI 300 index futures to compare the obtained volatility with the present volatility. If the higher the volatility and the higher the accuracy with the present volatility, then it has greater risk. Other strategies will be used to hedge the risk.

Investigation on volatility is one of the important topics in the contemporary financial field [3, 4]. It reflects the volatility of the underlying asset and is an index to measure the asset risk. Volatility is an important factor determining the price, i.e., the greater the volatility, the higher the price will be. This research is meaningful to the participants of the financial market. The effective prediction of volatility will be a great help to improve asset risk management level and investment decision-making. It also has a significant impact on financial risk prevention and control and the financial derivatives market. In addition, this paper also enriches the relevant literature in the field of volatility prediction. Most scholars' research is based on conventional volatility models, e.g., ARCH. However, the corresponding prediction ability is relatively weak, which cannot provide a good reference because the data frequency is relatively low. This paper selects the historical intraday high-frequency data of the CSI 300 index futures to predict future volatility. Relevant factors are selected and the paper holds the assumption that there is a linear relationship between the factors and the future window volatility. Therefore, three linear regression models including Ordinary Least Squares (OLS), Lasso, and Ridge are primarily used for prediction. It can be seen from the conclusion that this method can be used for reference to predict the volatility, and the overall interpretation rate is about 20%.

In the following sections, the paper will first introduce the data samples and the prediction models. Then, the regression results will be presented, analysed, and discussed. Finally, the main conclusions are summarized, and the limitations of this paper are explained to encourage further research on this topic.

2. DATA AND METHOD

This paper uses the relevant high-frequency data in minutes from CSI 300 index futures obtained from the Wind financial database [5], including information of volume, open interest, open, close, high, and low as the main research data. There are totally 326970 collected samples, ranging from 9:15 a.m. on January 5th 2015 to 2:59 p.m. on June 30th 2020. Figure 1 gives an overview of the samples by demonstrating their close price trends.

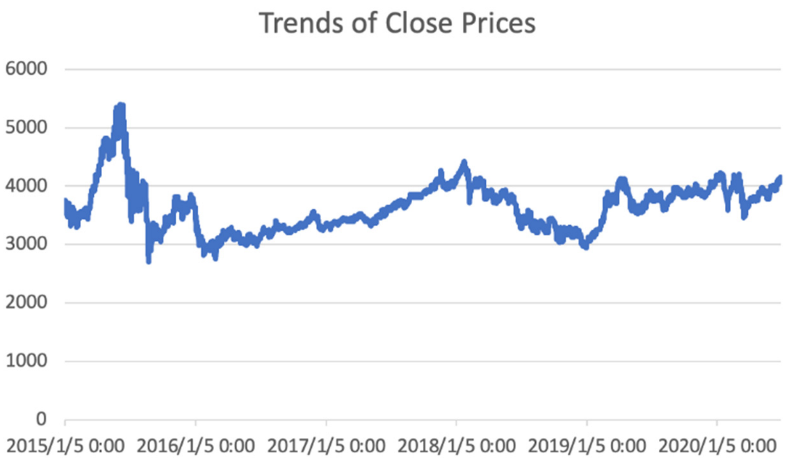


Figure 1. Close Price Trends of CSI300 future.

TABLE I. CALCULATION OF VOLATILITY AND FACTORS

Volatility (d_{close}, t, n) =	$\sqrt{\sum_{i=1}^{n}[\ln\left(\frac{d_{close}(t+i)}{d_{close}(t+i-1)}\right) - \sum_{i=1}^{n}\ln\left(\frac{d_{close}(t+i)}{d_{close}(t+i-1)}\right)/n]^{2} /n}$
Factor 1 ($d_{close},\ t,\ n_1 = 30, n_2 =$ 60) =	$\lvert\frac{1}{n_2}\sum_{k=0}^{n_2-1}\left(d_{close}(t-k) - d_{close}(t-n_1-k)\right)\rvert$
Factor 2 ($d_{high}, d_{low}, t, n_1 = 600, n_2 =$ 300) =	$\frac{1}{n_2}\sum_{k=0}^{n_2}(\max_{t\le j<t+n_1} d_{high}(j-k) - \min_{t\le j<t+n_1} d_{low}(j-k))$
Factor 3 ($d_{open}, d_{volume}, t, n_1 =$ 100, $n_2 = 1$) =	$\frac{1}{n_2}\sum_{k=0}^{n_2-1}(d_{open}(t-k) + d_{volume}(t-n_1-k))$
Factor 4 ($d_{high}, d_{low}, t, n_1 = 600, n_2 =$ 240) =	$\frac{1}{n_2}\sum_{k=0}^{n_2-1}(\max_{t\le j<t+n_1} d_{high}(j-k) \min_{t\le j<t+n_1} d_{low}(j-k)$ $- d_{close}(t-k)d_{close}(t-k))$
Factor 5 ($d_{close}, t, n_1 = 300, n_2 =$ 2) =	$\sqrt{\frac{1}{n_2}\sum_{k=0}^{n_2}\left\{d_{close}(t-k) - d_{close}(t-n_1-k) - \frac{1}{n_2}\sum_{j=t}^{t-n_2-1}(d_{close}(\right.}$
Factor 6 ($d_{open\ interest}, d_{volume}, t, n_1 =$ 10, $n_2 = 600$) =	$\frac{1}{n_2}\sum_{k=0}^{n_2}(\frac{\sum_{j=t}^{t-n_1} d_{volume}(j-k)}{\max_{t\le j<t+n_1} d_{open\ interest}(j-k) - \min_{t\le j<t+n_1} d_{open\ interest}(j-}$
Factor 7 ($d_{close}, t, n_1 = 5, n_2 = 10$) =	$\left\lvert\frac{1}{n_1}\sum_{j=t}^{t-n_2-1}(d_{close}(j) - \frac{1}{n_2}\sum_{j=t}^{t-n_2-1} d_{close}(j)\right\rvert$

Factor 8 $(d_{volume}, d_{close}, t, n_1 = 1, n_2 = 10) =$	$\left\|\frac{1}{n_1}\sum_{j=t}^{t-n_2-1}(d_{close}(j) - \frac{1}{n_2}\sum_{j=t}^{t-n_2-1} d_{close}(j)\right\| d_{volume}(t)$
Factor 9 $(d_{open\ interest}, d_{close}, t, n_1 = 10, n_2 = 30) =$	$\left\|\frac{1}{n_1}\sum_{j=t}^{t-n_2-1}(d_{close}(j) - \frac{1}{n_2}\sum_{j=t}^{t-n_2-1} d_{close}(j)\right\| d_{open\ interest}(t)$
Factor 10 $(d_{open\ interest}, t, n_1 = 10, n_2 = 240) =$	$\frac{1}{n_2}\sum_{k=0}^{n_2-1}(d_{open\ interest}\ \ (t-k) + d_{open\ interest}(t-n_1-k))$

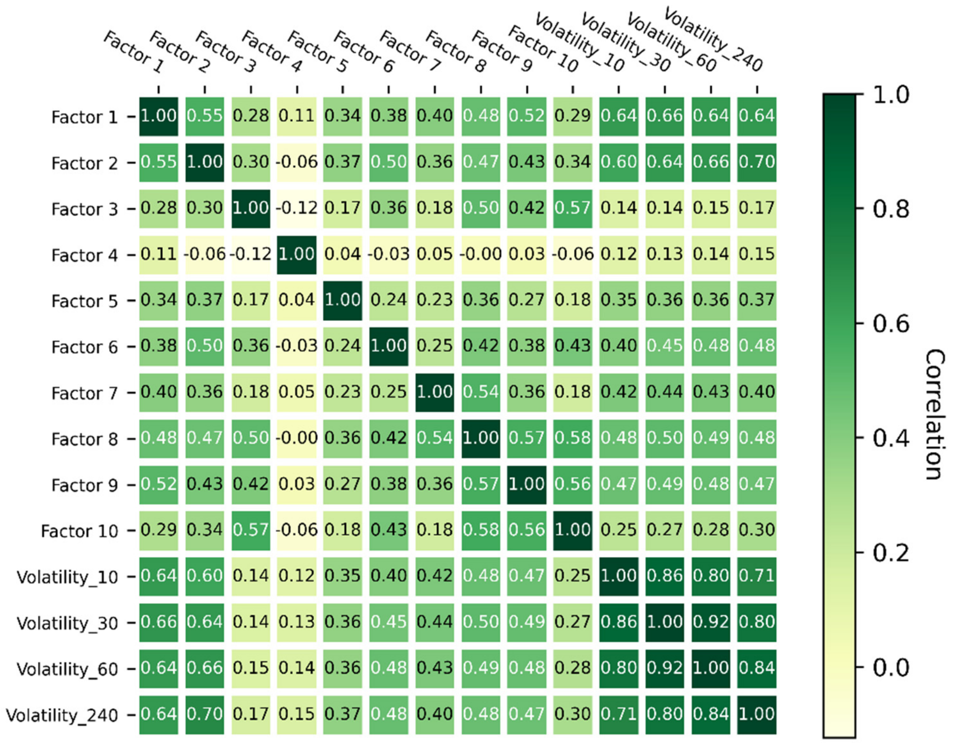
This paper investigates the volatility of four frequencies, which are 10 minutes, 30 minutes, 60 minutes, and 240 minutes respectively. The calculation formula of volatility is given in Table I., where d_{close} is the close price of samples every minute. In order to find the appropriate factor, Spearman's Rank Correlation Coefficient is used to find the factors with large volatility correlation as the relevant factor in this paper. Table I. lists all the construction methods for the 10 factors, where d_{high} is the highest price, d_{low} is the lowest price, d_{open} is the open price, d_{volume} is the trading volume, $d_{open\ interest}$ is the open interest of collected samples.

After factor selection, three linear regression methods, OLS, Lasso, and Ridge, are used to predict volatility. Data processing and chart making are implemented in Python with the support of Scikit-Learn [6], Pandas [7] and Matplotlib [8]. In statistics, Ordinary Least Squares (OLS) or linear least squares estimate the unknown parameters in a linear regression model. This method minimizes the sum of squared vertical distances between the observed responses in the dataset and the responses are predicted by the linear approximation. The resulting estimator can be expressed by a simple formula, especially in the case of a single regressor on the right-hand side [9]. In order to improve the fitting accuracy, OLS will also fit the noise, resulting in over-fitting. Therefore, it turns to Lasso and Ridge models to avoid overfitting through these two models. Lasso represents the minimum absolute contraction and selection operator. In order to avoid overfitting, it penalizes the absolute value of its coefficient on the basis of OLS. Ridge Regression is a technique for analysing multiple regression data that suffer from multicollinearity [10]. The limitation of the Ridge model is that Ridge reduces the complexity of the model, but does not reduce the number of variables, because it will never lead to zero coefficients, but only minimize them. Therefore, the model is not conducive to feature reduction. The difference between Ridge and Lasso regression is that it tends to make coefficients to absolute zero compared to Ridge, which never sets the coefficient to absolute zero [11].

3. RESULTS & DISCUSSION

As shown in Figure 2, the correlation coefficient of the selected 10 factors and volatility y is relatively large (maintained at greater than 0.1), indicating that the factors are highly correlated with volatility. Besides, the autocorrelation coefficient between each factor is relatively small (less than 0.6), indicating that each factor is independent and there is no similar factor. These two points can prove that the 10 factors selected in this paper are effective and can predict volatility.

In the model prediction, changing the alpha value in the OLS regression model will not affect its prediction effect. In this paper, 0.5 is randomly selected as the alpha value of the OLS regression model. However, changing the alpha value will have an impact on the results in the other two regressions. In the lasso regression model, the smaller the alpha value is, the greater the R^2 value, but the goodness of fit will become weaker when reduced to a certain extent. 1e-05 is the alpha value of the lasso regression model selected in this paper. The Ridge regression model is contrary to the Lasso regression model. The alpha value is directly proportional to the R^2 value, but the fitting effect will be weakened to a certain extent. In this paper, 10000 is selected as the alpha value in the Ridge regression model.



	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Volatility_10	Volatility_30	Volatility_60	Volatility_240
Factor 1	1.00	0.55	0.28	0.11	0.34	0.38	0.40	0.48	0.52	0.29	0.64	0.66	0.64	0.64
Factor 2	0.55	1.00	0.30	-0.06	0.37	0.50	0.36	0.47	0.43	0.34	0.60	0.64	0.66	0.70
Factor 3	0.28	0.30	1.00	-0.12	0.17	0.36	0.18	0.50	0.42	0.57	0.14	0.14	0.15	0.17
Factor 4	0.11	-0.06	-0.12	1.00	0.04	-0.03	0.05	-0.00	0.03	-0.06	0.12	0.13	0.14	0.15
Factor 5	0.34	0.37	0.17	0.04	1.00	0.24	0.23	0.36	0.27	0.18	0.35	0.36	0.36	0.37
Factor 6	0.38	0.50	0.36	-0.03	0.24	1.00	0.25	0.42	0.38	0.43	0.40	0.45	0.48	0.48
Factor 7	0.40	0.36	0.18	0.05	0.23	0.25	1.00	0.54	0.36	0.18	0.42	0.44	0.43	0.40
Factor 8	0.48	0.47	0.50	-0.00	0.36	0.42	0.54	1.00	0.57	0.58	0.48	0.50	0.49	0.48
Factor 9	0.52	0.43	0.42	0.03	0.27	0.38	0.36	0.57	1.00	0.56	0.47	0.49	0.48	0.47
Factor 10	0.29	0.34	0.57	-0.06	0.18	0.43	0.18	0.58	0.56	1.00	0.25	0.27	0.28	0.30
Volatility_10	0.64	0.60	0.14	0.12	0.35	0.40	0.42	0.48	0.47	0.25	1.00	0.86	0.80	0.71
Volatility_30	0.66	0.64	0.14	0.13	0.36	0.45	0.44	0.50	0.49	0.27	0.86	1.00	0.92	0.80
Volatility_60	0.64	0.66	0.15	0.14	0.36	0.48	0.43	0.49	0.48	0.28	0.80	0.92	1.00	0.84
Volatility_240	0.64	0.70	0.17	0.15	0.37	0.48	0.40	0.48	0.47	0.30	0.71	0.80	0.84	1.00

Figure 2. Correlation of Volatility and Factors.

The selected 10 factors have different prediction abilities on each volatility frequency. The three regression models selected 60% data as the train set and the remaining 40% data as the test set. As shown in Table II., under the frequency of 10 minutes volatility, the prediction effects of the three models are similar. About 0.49 of the train set can be predicted accurately, but only 0.14 of the test set can be predicted. The correlation of a train set is 0.76, and that of the test set is 0.62. As listed in Table III., under the 30 minutes volatility frequency, the prediction effects of the three models are also similar, which can explain 0.58 of the train set and only 0.17 of the test set. The correlation with the train set is about 0.80, but the effect on the test set is greatly reduced, only 0.64. As displayed in Table IV., when the volatility frequency is reduced to 60 minutes, the three models can explain about 0.62 of the train set and 0.19 of the test set. The correlation with the train set is 0.81, but only 0.63 with the test set. According to Table V., when the volatility frequency is 240 minutes, the three models can explain about 0.70 of the train set, and the test set is also improved, about 0.27. Compared with the correlation of train set and test set, it has increased at the previous frequency, which is 0.83 and 0.64, respectively. The prediction effects of the three models under each volatility frequency are not too poor, and the prediction effects of the train set are better than the test set. According to the prediction results of different frequencies, it can be concluded that as the frequency of volatility becomes lower, the accuracy and correlation of regression prediction will become higher, i.e., the performances of the model are better.

However, the paper is limited by factor selection and the prediction approach using linear models. The usage of machine learning models, e.g., Gradient boosting decision trees (GBDT) and convolutional neural networks (CNN), give solutions for the limitation because they are through a large amount of data evaluation and practice so as to be able to process and analyze forecast data effectively.

TABLE II. PREDICTION RESULTS OF 10 MINUTES VOLATILITY

Model	Alpha	Train R^2 score	Test R^2 score	Correlation train	Correlation test
OLS	0.5	0.492	0.143	0.761	0.625
Lasso	1e-05	0.488	0.148	0.766	0.612
Ridge	10000	0.491	0.149	0.766	0.629

TABLE III. PREDICTION RESULTS OF 30 MINUTES VOLATILITY

Model	Alpha	Train R^2 score	Test R^2 score	Correlation train	Correlation test
OLS	0.5	0.581	0.166	0.804	0.648
Lasso	1e-05	0.577	0.176	0.809	0.642
Ridge	10000	0.580	0.176	0.809	0.651

TABLE IV. PREDICTION RESULTS OF 60 MINUTES VOLATILITY

Model	Alpha	Train R^2 score	Test R^2 score	Correlation train	Correlation test
OLS	0.5	0.627	0.181	0.811	0.630
Lasso	1e-05	0.623	0.196	0.814	0.620
Ridge	10000	0.625	0.195	0.815	0.631

TABLE V. PREDICTION RESULTS OF 240 MINUTES VOLATILITY

Model	Alpha	Train R^2 score	Test R^2 score	Correlation train	Correlation test
OLS	0.5	0.700	0.241	0.827	0.640
Lasso	1e-05	0.695	0.275	0.829	0.630
Ridge	10000	0.697	0.270	0.830	0.641

4. CONCLUSION

In summary, 10 effective factors are selected to forecast the return volatility of CSI 300 index futures based on OLS, Lasso, and Ridge for 10 minutes, 30 minutes, 60 minutes, and 240 minutes. According to the results, it can be concluded that it is feasible to use multi factors to predict volatility. The accuracy of the test set of the three models increases with the decrease of volatility frequency. To be specific, the interpretation ability metricated by R^2 scores increases from 0.14 to 0.27 (i.e., indicating better performances of the models). However, this paper has some limitations. One is that the selected factors may be insufficient for prediction. The other is that the paper assumes a linear relationship between factors and volatility, which may not be accurate, resulting in an unsatisfactory prediction effect. Further study ought to be carried out from these two perspectives. These results offer a guideline for the prediction of volatility that allows us to price financial derivatives (e.g., options and futures) and effectively predict the risk factor of the market.

REFERENCES

[1]Fama, Eugene, F., and Kenneth R. French, “The Capital Asset Pricing Model: Theory and Evidence. Journal of Economic Perspectives,” 18 (3): 25-46, 2004.

[2]Wei Yu, “Volatility forecasting models for CSI300 index futures. Journal Of Management Sciences In China,” 13 (2): 67-75, 2010. (In Chinese)

[3]Brook, C. and Persand, G, “Volatility Forecasting for Risk Management,” University Reading, 22(1): 1-22, 2003.

[4]Timothy J. Brailford, Robert W. Faff, “An Evaluation of Volatility Forecasting Techniques,” Journal of Banking & Finance, 20 (3): 419-438, 1996.

[5]Wind official website. Available at: https://www.wind.com.cn

[6]Scikit-Learn official website. Available at: https://scikit-learn.org/stable/index.html

[7]Pandas official website. Available at: https://pandas.pydata.org/

[8]Matplotlib official website. Available at: https://matplotlib.org/

[9]Hayashi, Fumio, Econometrics. Princeton University Press. ISBN 0-691-01018-8, 2020.

[10] Hoerl, A.E. and Kennard, R.W, “Ridge Regression: Biased estimation for nonorthogonal problems,” Technometrics 12, 55-82, 1970.

[11] Gareth James et al, “An Introduction to Statistical Learning: With Applications in r,” Springer, 219–220, ISBN 978-1-4614-7137-0, 2017.