High-frequency trading data volatility prediction based on deep learning method

Anonymous CVPR submission

Paper ID

Abstract

In a quantitative trading strategy, forecasting future market conditions is an important issue, and more accurate forecasts can help build a better performing quantitative trading strategy. In our group's research, we use time series analysis methods and RNN (recurrent neural network) methods to predict the volatility of stocks in the short term. In the comparison experiment using the stock asset data contained in the ETF50, multiple stock asset data are used. We found that RNN-based prediction has higher accuracy.

1. Introduction

In this section, I will introduce some relevant background knowledge, includeing a simple set of technical indicators that can be used to build trading strategies.

1.1. Quantitative trading strategies

Quantitative trading consists of trading strategies based on quantitative analysis, which rely on mathematical computations and number crunching to identify trading opportunities. As quantitative trading is generally used by financial institutions and hedge funds, the transactions are usually large in size and may involve the purchase and sale of hundreds of thousands of shares and other securities. However, quantitative trading is becoming more commonly used by individual investors. It has the advantages of disciplinary, systematic, timeliness, accuracy, decentralization and other traditional trading methods.
The idea of quantifying trading strategies comes from related investment theories.Theoretical development history:

In 1952, Markowitz, mean-variance model
1964-1966, Sharp, Lintner, CAPM model
1965, Samuelson, Fama, Efficient Market Hypothesis
1973, Black, Scholes, option pricing model
1976, Rose, APT model
Backward Stochastic Differential Equations in the 1980s
The 1990s, VaR model
The 1990s, Behavioral Finance

1.2. A simple quantified trading strategy based on Bollinger Bands

The Bollinger Bands Index, the BOLL Indicator, has the full English name "Bollinger Bands". The Bollinger Bands (BOLL) was created by Mr. John Brin. Using statistical principles, the standard deviation of stock prices and their confidence intervals are determined. The fluctuation range of the stock price and its future trend use the band to show the high and low price level of the stock price.Defining the asset price as S_t:

$$MA_t = \frac{1}{N} \sum_{i=t-N+1}^{t} S_i \tag{1}$$

$$MD_t = \sqrt{\frac{1}{N} \sum_{i=t-N+1}^{t} (S_i - MA_t)^2} \tag{2}$$

$$MB_t = \frac{1}{N-1} \sum_{i=t-N+2}^{t} S_i \tag{3}$$

$$UP_t = MB_t + 2MD_t \tag{4}$$

$$DN_t = MB_t - 2MD_t \tag{5}$$

Among them: MB is called the Bollinger Band Middle Track, UP is called the Bollinger Band Upper Track, and DN is called the Bollinger Band Lower Rail. These technical indicators can be used to construct a very simple but effective trading strategy:

Example of trading strategy:
Buy a unit of assets:
$if: S_t > MB_t + p_1 * MD_t$ and Increased trading volume
Sell a unit of assets:

$$if: S_t < MB_t$$

Hold on:

$$otherwise$$

1.3. Strategy backtracking

Use the strategy mentioned in the previous section to trace the daily trading data of Vanke A stock (code 000002).Data from 2010/1/4 to 2017/12/29:
With the implementation of the strategy, the change in asset account amount over time is shown in Figure 1.

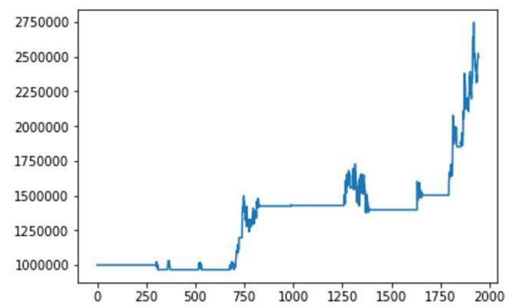


Figure 1. Strategic asset changes

It can be calculated:
The rate of return (annualized) is: 12.75136900410212%
The volatility (annualized) is: 0.21275937296589917
The Sharpe ratio (annualized) is: 0.4113270725565063
The maximum retreat is: 0.2034904811633779
The maximum number of uncreated high days is 514 days

1.4. A mistake in one experiment

In an experiment, after the strategy was implemented, the results in Figure 2 were obtained, and the corresponding calculations were:

The rate of return (annualized) is: 93.40629900018347%
The volatility (annualized) is: 0.2986558491684209
The Sharpe ratio (annualized) is: 2.9936229023850323
The maximum retreat is: 0.1837691661188275
The maximum number of uncreated high days is 153 days

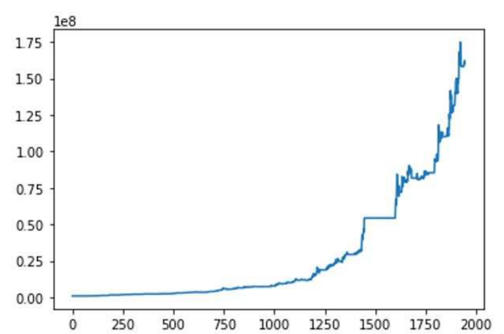


Figure 2. Strategic asset changes(A mistake)

This is a result that is not in line with reality: the annualized rate of return exceeds 90% and is extremely robust. After a review of the program, it was found that in this experiment, due to a small error in the assignment of the loop variable, the program used the market information of the next day in the judgment of the transaction. Therefore, the effectiveness of the strategy has been completely changed. It is a pity that this is the result of a mistake, but it also suggests that we began to consider and try to make predictions about future short-term market information.

1.5. Predicting prices is exceptionally difficult

The Efficient Markets Hypothesis (EMH) was proposed by Eugene Fama in 1970. The effective market hypothesis originated in the early 20th century. The founder of this hypothesis was a French mathematician named Louis Bachelier who applied statistical analysis methods to the analysis of stock returns and found that Its mathematical expectation is always zero.
The effective market hypothesis actually states that the future value of freely traded assets in the market cannot be expected, or that all expectations about the future have been fully reflected in the current price level of the market,So,predicting prices is exceptionally difficult!

1.6. Volatility forecast

Since price forecasting is almost impossible, and secondarily, we consider studying the volatility of asset returns. The true volatility cannot be directly observed. We can only approximate the volatility by some statistics. One common estimation method is to use the squared average of the logarithmic returns r_t for a period of time as a measure of volatility $\sigma_t^{history}$. This measure is called historical volatility.In accordance with the definition of historical volatility, we use σ_t as a measure of the volatility of assets at time t.

$$r_t = \log \frac{S_t}{S_{t-1}} \tag{6}$$

$$\sigma_t^{history} = \frac{1}{n} \sum_{i=t-n+1}^{t} r_i^2 \tag{7}$$

$$\sigma_t = \frac{1}{2n+1} \sum_{i=t-n}^{t+n} r_i^2 \tag{8}$$

In the study we take n=5. In the minute data, the volatility measure is about 10 minutes. Ten minutes is long enough for high frequency data.

1.7. Prediction effect evaluation

Loss function is a traditional index used to evaluate volatility prediction accuracy. Although Bollerslev (1994), Diebold, Lopez (1996) and others pointed out that loss function is not necessarily the best evaluation method, but it is still a very effective Evaluation means. We use several loss functions: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE),The smaller the loss function value, the better the prediction effect.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\sigma_t - \sigma_t^{predict})^2 \tag{9}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\sigma_t - \sigma_t^{predict}| \tag{10}$$

$$RMSE = [\frac{1}{N}\sum_{i=1}^{N}(\sigma_t - \sigma_t^{predict})^2]^{1/2} \tag{11}$$

$$QMSE = \frac{1}{N}\sum_{i=1}^{N}(\sqrt{\sigma_t} - \sqrt{\sigma_t^{predict}})^2 \tag{12}$$

$$QMAE = \frac{1}{N}\sum_{i=1}^{N}|\sqrt{\sigma_t} - \sqrt{\sigma_t^{predict}}| \tag{13}$$

2. Date Set

In this section, I will describe the data set we use in the study, including its source and necessary preprocessing.

2.1. 50ETF component stocks

The 50 ETF Index selects 50 representative stocks with large market size and good liquidity as the constituent stocks to comprehensively reflect the overall situation of a group of high-quality large-cap companies with the most market influence in the Shanghai securities market. The SSE 50 Index was compiled by the Shanghai Stock Exchange. The index is abbreviated as 50ETF, the index number is 000016, and the base day is December 31, 2003. The base point is 1000 points.
The Shanghai 50 ETF was formally released on January 2, 2004 and was listed and traded on the Shanghai Stock Exchange. The fund manager is Huaxia Fund Management Co., Ltd. The investment objective of the Shanghai 50 ETF is to closely track the Shanghai 50 Index, minimizing tracking deviations and tracking errors. The fund adopts a passive investment strategy, and the specific method of tracking index investment is mainly a full copying method, pursuing similar risk and return characteristics as the 50ETF index.

We selected the 50 stocks(in Table1) that were used as weighted stocks in the ETF50 index in 2015 as data for our research. Mainly due to: 1. ETF50 index of heavy stock market trading is relatively large, it is not easy to be artificially controlled and can reflect the general laws of the market. 2. The ETF50 index has available options for trading derivatives, providing possibilities for volatility trading applications.
Among them, China CNR (code 601299) and Oriental Pearl (code 600832) did not conduct transactions after 15 years. Excluding their data.

2.2. Data and preprocessing

The data set we use comes from a company called ”FinTech”, whose software ”QuantDesk” provides data interfaces. The trading data of each stock is recorded once every minute during the trading period, including open, high, low, close, amount, and volume. (volume). Within each trading day, 9:30 am to 11:30 am and 1:30 pm to 3:30 pm are trading hours. There are 240 data points recorded during each trading day, with approximately 220-230 transactions each year. day. Data was recorded from January 4, 2012 until May 16, 2018, with approximately 370,000 data points per stock.An example of data is given in Table 2.
We use close at each time point as the asset price at that point in time. It should be noted that in the course of stock trading, stock prices may sometimes change suddenly due to ex-rights or private placement, but the fact that the total value of the stock is invariable in fact needs to take into account the price caused by the ex-rights when calculating the logarithmic profitability. Variety. Since we have not obtained specific data on the ex-rights of each stock, we determine whether to ex-rights by limiting the stock's rise and fall. As the Chinese stock market is subject to price increases $(-10\% to + 10\%)$, if the absolute value of the logarithmic rate of return exceeds $log(1.1)$ then the sudden change in this price is due to an ex-rights or targeted increase. At this point, the logarithmic profitability of this point is recorded as 0.Finally, we calculate the volatility for each time point based on the volatility calculation formula mentioned in the previous article.

3. Traditional Method

In real life, it is obvious that many economic variables are related to each other. Building structural models in which variables are linked to each other gives insight into the interrelationships between variables. In time series approach, one is more concerned with predicting future values, including future uncertainty.

3.1. Background

The objective of studying financial time series analysis is to predict future values of some financial variable. This objective is achieved by following methodology based on four steps, which are model specification, model estimation, model validation and forecasting. The model to be used for forecasting is in fact estimate from history, i.e. data from the past is used to make probabilistic statements about future values.

3.2. ARMA Model

As we have mentioned before, historical values are used to identify the dynamic properties of the time series and these are used to forecast the future. A critical assumption behind the methodology is that such dynamics do not change over time. In a heuristic sense, we require the future to be like the past in a probabilistic sense. Usually, it is sufficient to require the mean, variance and covariances of a time series to be independent of time, rather than the entire distribution. This is referred to as stationarity.

Formally,a process $\{Y_t\}$ is said to be weakly stationary if, for all t, it holds that

$$\begin{aligned} E(Y_t) &= \mu < \infty \\ \mathrm{Var}(Y_t) &= E(Y_t - \mu)^2 \\ \mathrm{Cov}(Y_t, Y_{t-k}) &= E(Y_t - \mu)(Y_{t-k} - \mu) = \gamma_k, k = 1, 2, 3 \cdots \end{aligned} \tag{14}$$

To illustrate the relationship between a structural model and a time series model, assume that the following regression model describes the relationship between two variables Y_t and X_t

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \tag{15}$$

where ε_t is a white noise error them. If X_t can be described by some ARMA model, then Y_t is the sum of an ARMA process and a white noise process and therefore also follow an ARMA process.

For example, if Y_t is modeled by an $ARMA(p, q)$ model

$$\begin{aligned} Y_t = \beta_0 + \beta_1(X_t + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p}) \\ + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \end{aligned} \tag{16}$$

3.3. ARIMA Model

As we have mentioned before, one of the necessary conditions of ARMA model is stationary. However, it is impossible for every time series to satisfy this condition. As a matter of this kind of situation, we created a new model called ARMIA, when the d times differences of a time series Y_t (denote as $W_t = \nabla^d Y_t$) is a stationary ARMA process. If W_t is modeled by $ARMA(p, q)$ model, then we call Y_t is a $ARIMA(p, d, q)$ process. Fortunately, we value d at most 2.

We concerned $ARIMA(p, 1, q)$ as followed. Let $W_t = Y_t - Y_{t-1}$, then we have

$$\begin{aligned} W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_p W_{t-p} \\ + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \end{aligned} \tag{17}$$

4. Proposed Methodology

We usually utilize Recurrent Neural Networks(RNN) to solve time series problem, and utilize Convolution Neural Networks(CNN) to improve the feature map scales. Trading data is a kind of classic time-series data, we design our deep learning network with convolution layers and Long Short-Term Memory(LSTM) layers to estimate future volatility of trading data. We validate our results on 50 trading data and compare them with the traditional method result, which prove our framework effective. Our best result is 1.48*10e-10 on MSE, 6.03*10e-6 on MAE, 1.22*10e-5 RMSE, 2.97*10e-6 on RMAE, 1.54*10e-3 on QMAE. The function of LSTM gate are as follows:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t + b_f]) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t + b_i]) \\ \widetilde{C}_t &= tanh(W_C \cdot [h_{t-1}, x_t + b_C]) \\ C_t &= f_t * C_{t-1} + i_t * \widetilde{C}_t \end{aligned} \tag{18}$$

4.1. Data preprocessing

Trading data volatility can be easily influenced by some unstable factors such as financial policy and industry trends. This kind of noise affects on our trading volatility prediction seriously. To solve this problem, we add a preprocessing step at the beginning of our framework which we name i Input-Smooth part. We have trained our network model on data without treatment as comparison, but the results are totally underfitting.

Input-Smooth part is a processing to the input Logarithmic rate of return. The evolution standard is shown above which generally express the volatility, but the trend of volatility is submerged by the by the over-limit noise so that we cannot predict well. The frequency of the noise signal is much higher than the variation of the basic volatility, and our smooth filter works well.To get more information of different scale. We add two kinds of smooth filters to the input data with the size of 10 and 30 embedding them into the two channels of our network, the validation is shown behind in 3. Besides of controlling the noise, the larger filter size is the lower frequency of information we can get.

4.2. Network Architecture

The architecture of our network is composed of two branch networks with a convolution later at the beginning to get information of different scales, then we merge them following with LSTM layers to get the the feature maps of the time series. The network architecture is illustrated in Figure 1.

We calculate the Logarithmic rate of return from the raw trading data, then cut the result series into patches with

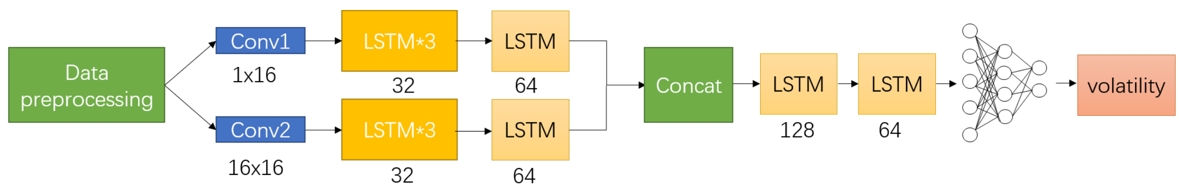


Figure 3. Network Architecture

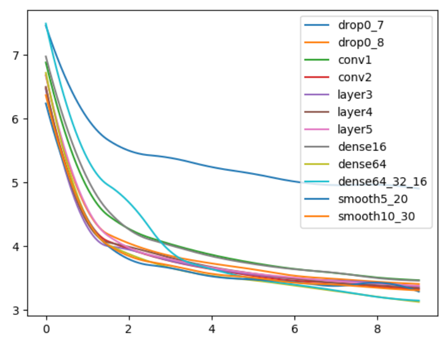


Figure 4. train loss

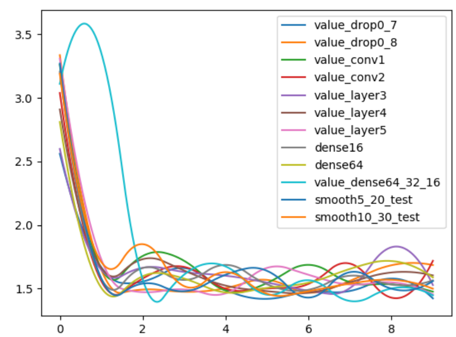


Figure 5. test loss

length of 128 following with the preprocessing step. We use the future 10 data to calculate the real volatility with function above and set then as Ground Truth. We select 14500 patches from the whole 50 trading data with date of deadline 2017.1.1 and shuffled them to form our training set. We validate our result on the data in 2017. **??** shows our validation result.

4.3. RNN neural network advantages over traditional methods

Although traditional methods can make some effective predictions, the future time points that can be predicted are relatively short. Moreover, the traditional method is too monotonous in the establishment of the model and can not provide good time and parameter complexity to predict the information in the next 5 to 10 minutes. The sensitivity of traditional networks to data is too strong to predict effectively based on known long-term information. The depth model solves this problem very well. LSTM has good time and parameter complexity. LSTM has a wealth of gating units to control long and short time. It can well memorize my short time information and long time information, and can effectively predict based on the long time in the past. At the same time, the RNN method has better robustness, can deal with different stocks, and does not need to spend a lot of time on the training of different models, which also largely saves a large part of the computational overhead.

5. Comparisons between different configurations of our network

The ablation analysis aims to investigate the effectiveness of our network architecture. 5 and **??** shows the results of the network components. We set the number of branch as 2, 3 and 4, the convlution kernel size as (1&16), (16&32) and (16&48), the dropout rate as 0.2, 0.4 and 0.8, the smooth window size as (5&10), (10&30), and (10&20), the loss function as MAE and MSE, the optimizer as Adam and AdamGrad.

5.1. Analysis of results

we first use the data about the logarithmic rate of return to obtain the data about volatility of stock prices so that We in turn model 50 stock volatility data and predicted the corresponding volatility.

Specifically, we selected the data from 2013 to 2017 to fit, predicted the volatility of stocks in 2018, and compared the fitted value with the real volatility. And the results about stock 44 are diaplayed in Fig.9

From the above two figures, we can see that no matter which stock, the blue curve has the similar shape with the red one. At the same time, the MSE and MAE of the fitted value and real value are used to measure the fitting accuracy, which is shown in Table 2. Through the above anaylsis, it is obviously seen that traditional time serise models can indeed fit and predict the data of volatility.

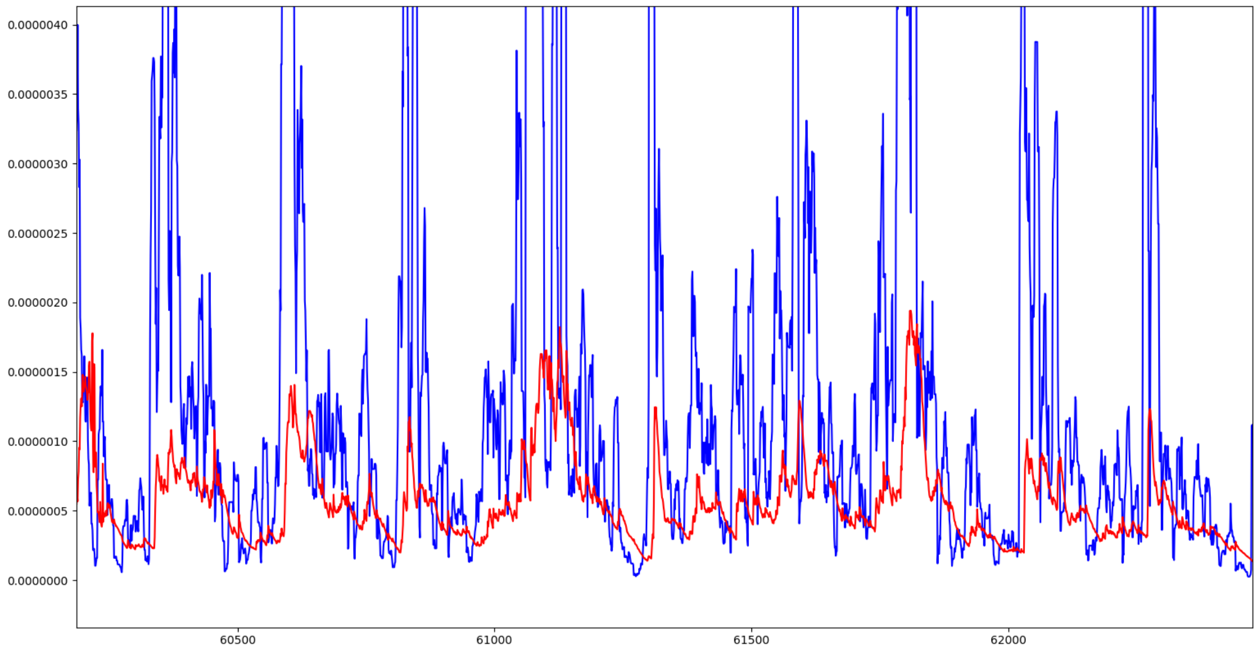


Figure 6. Network Architecture

Table 1. Comparing with the state-of-the-art method

proposed	mean	var	median	Q(0.25)	Q(0.75)	min	max
MSE	3.59E-11	1.28E-21	2.06E-11	1.21E-11	5.32E-11	2.03E-12	1.48E-10
MAE	1.42E-06	8.11E-13	1.16E-06	8.53E-07	1.61E-06	4.88E-07	6.03E-06
RMSE	5.35E-06	7.45E-12	4.54E-06	3.48E-06	7.30E-06	1.42E-06	1.22E-05
QMSE	6.45E-07	2.07E-13	5.75E-07	3.67E-07	8.34E-07	1.80E-07	2.97E-06
QMAE	0.000496	4.86E-08	0.000451	0.000364	0.000565	0.000217	0.001547
tradition	mean	var	median	Q(0.25)	Q(0.75)	min	max
MSE	3.51E-11	1.35E-21	2.05E-11	1.12E-11	4.82E-11	2.29E-12	1.69E-10
MAE	1.69E-06	8.22E-13	1.63E-06	1.16E-06	2.01E-06	6.46E-07	6.14E-06
RMSE	5.27E-06	7.51E-12	4.53E-06	3.35E-06	6.94E-06	1.51E-06	1.30E-05
QMSE	9.53E-07	2.83E-13	9.04E-07	5.67E-07	1.19E-06	2.39E-07	3.19E-06
QMAE	0.000718	5.79E-08	0.000707	0.000529	0.000881	0.000366	0.001613

Stock Number	MSE	MAE	RMSE
44	1.56×10^{-11}	1.85×10^{-6}	3.95×10^{-6}

Table 2. The values of MSE, MAE and RMSE

6. Application of Volatility Prediction in Quantifying Transactions

In this section, we discuss how the prediction of volatility can be applied to actual transactions. For general assets, volatility forecasts need to be combined with directional forecasts to play a role in transactions. Or it can be used for risk estimation and hedging operations. In particular, in option trading, it is possible to carry out arbitrage trading

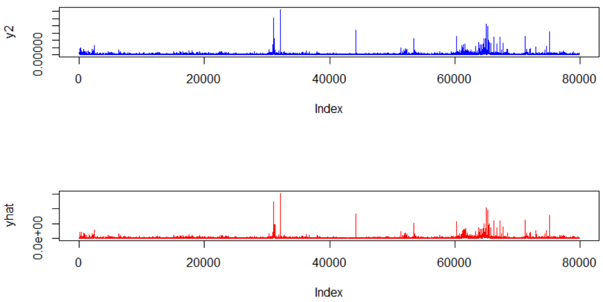


Figure 7. Comparison between fitted and true values about stock 44. The blue line is about real volatility from 2017, while the red one is about fitted value called 'yhat'.

```
> arma_y1
Series: y1
ARIMA(5,1,0)

Coefficients:
         ar1     ar2     ar3     ar4     ar5
      0.0539  0.0275  0.0156  0.0148  0.0109
s.e.  0.0021  0.0021  0.0021  0.0021  0.0021

sigma^2 estimated as 3.108e-11:  log likelihood=2488474
AIC=-4976935   AICc=-4976935   BIC=-4976873
```

Figure 8. Details of the simulation about stock 44 using arima model, including coefficients, log likelihood, AIC and BIC (which are index to evaluate the degree of fitting)

on volatility through some option combinations.

6.1. Option portfolio

An option is an option that refers to the right to buy or sell a certain amount of a particular commodity at a specific price at a specific time in the future. It is a financial instrument based on futures that gives the buyer (or holder) the right to purchase or sell the underlying asset. The holder of the option may choose to buy or not to buy, sell or sell in the time specified in the option, he may exercise the right, or he may waive the right, and the seller of the option only has the option contract. Obligations.

By making investment portfolios for options with different execution prices, some special portfolios can be constructed. Among them, wide-span options and butterfly options can be used for volatility arbitrage transactions. The relationship between the value of buying a butterfly option and the price of a calibrated asset is shown in the first chart of Figure 3. The relationship between the value of a short-span option and the price of a denominated asset is shown in the second chart in Figure 3. As shown, in the case of very low volatility in the future, you can buy butterfly options for profit. Conversely, short-term short-term straddling options will benefit in the future with large volatility.

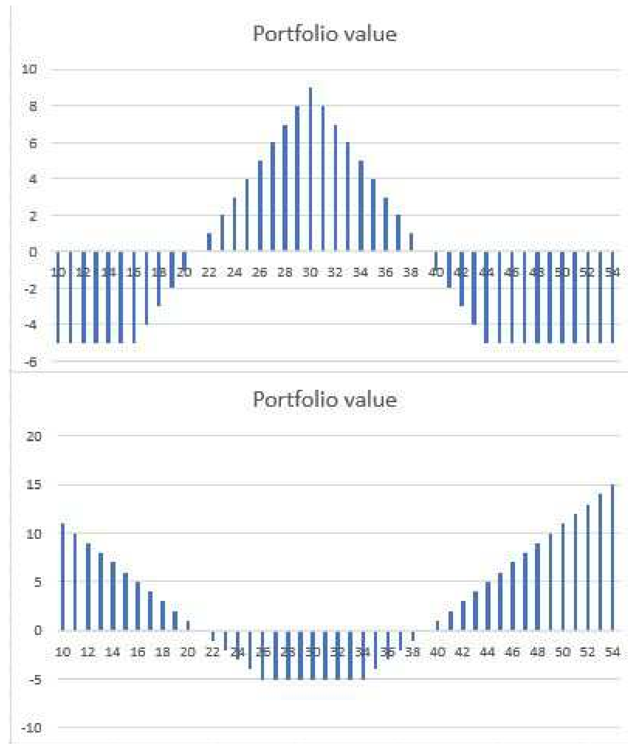
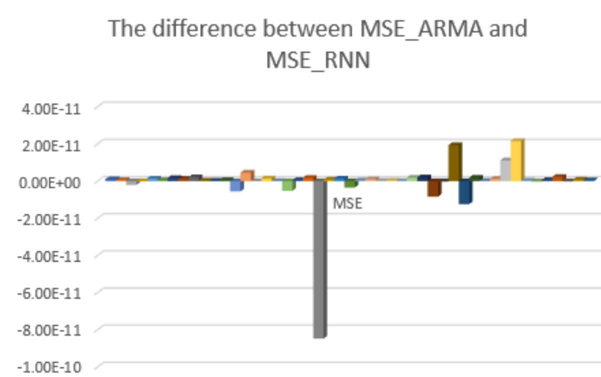


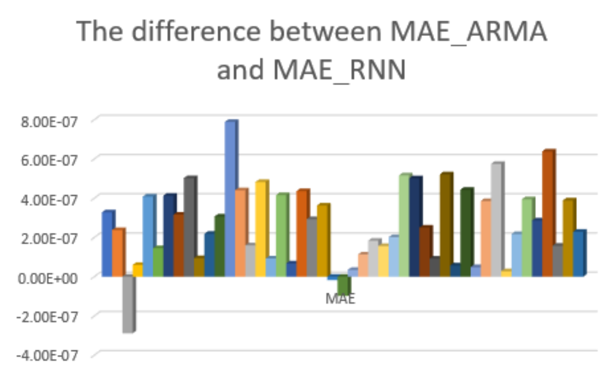
Figure 9.

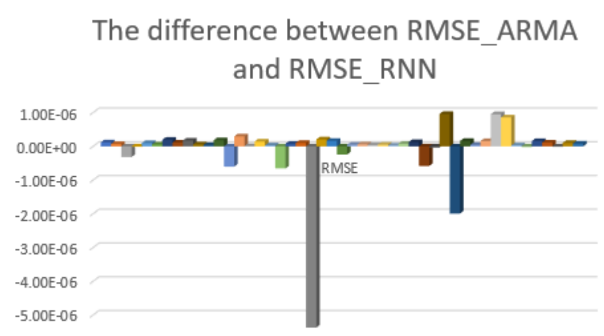
6.2. Options of 50ETF

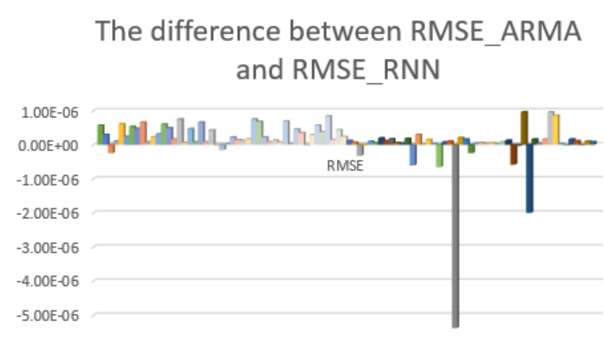


With the approval of the China Securities Regulatory Commission, Shanghai Stock Exchange decided to list and trade 50 ETF option contract varieties on February 9, 2015. The subject of the SSE 50 ETF option is the "Shanghai 50 Open-ended Index Securities Investment Fund". Since February 9, 2015, the Shanghai Stock Exchange has listed the corresponding SSE 50 ETF options contracts according to different contract types, expiration months and exercise prices.

Using the 50 ETF fund-based options traded on the market,

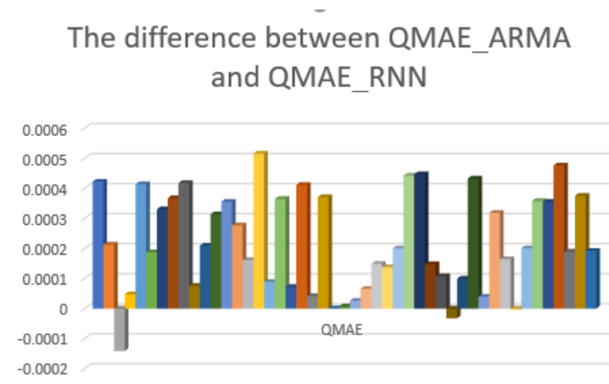






a combination of butterfly options and wide-span options can be constructed. Then the volatility can be arbitrarily quantified based on the output of the volatility prediction model. Unfortunately, we did not get the minute data of the option trading in the corresponding period, so we did not conduct further testing on the application effect of the forecast results.

To sum up, we generate a table to compare the varies indexes of two methods. In the table below,"method 2" represents time series method while "method 1" represents deep learning method. It is crystal clear that the mean of deep learning MAE is obviously smaller than the other method.



But considering another index MSE, deep learning doesn't show a great advantage.

7. Acknowledgements

It is my honor to study deep learning course with students from the 33rd group of the Deep Learning Group in the spring semester of 2018 and I am so fortunate to have the privilege of attending and benifiting frome the lessons and reports given by teachers Wang Wei and Wang Liang. Also I want to thank our teaching assistant Wang Junbo for his patience . In the study of the final project, we especially need to thank the students Yang Qili and Qiu Xinling for their help and suggestions in the project. Finally, we would like to thank the six students of our group for their mutual help in their studies and their joint efforts in the project.

References