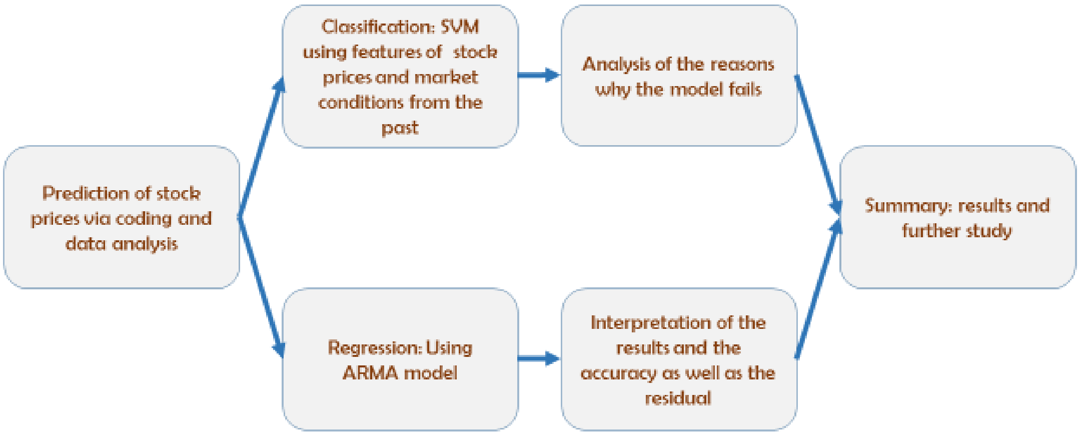
The prediction of return on stock price——Two approaches

Team10: 赖亚王, 5130719085 ；邓凡意, 515120910159 ；刘益鹏, 515120910200 ；LINJIANG LI, 713120990034

Ⅰ. Abstract

Our project mainly focuses on the research of precise and accurate ways to predict future stock prices. We first try the SVM machine learning method, using both the returns on stock price and market index of yesterday and the day before yesterday as features. After the SVM classification model failed to predict whether the return of tomorrow’ s stock will be greater than zero or not, we analyze the reason for the failure. Then, we turn to another model, Autoregressive–moving-average model. We build our regression model and tests the accuracy of the regression. Although the result is still not accurate enough, our model is more precise in certain aspect than the former one. Finally, we briefly summarize the result of our model and ways of further increase the accuracy of the prediction.


Prediction of stock prices via coding and data analysis
Classification: SVM using features of stock prices and market conditions from the past
Analysis of the reasons why the model fails
Regression: Using ARMA model
Interpretation of the results and the accuracy as well as the residual
Summary: results and further study

Ⅱ. Machine Learning: SVM Classification Method

2.1 Introduction of Support Vector Machine

In machine learning, support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.[1]

2.2 Assumptions adopted in the prediction model

We adopted sic features representing the past performance of both the stock and the market, as we assume that according to the CAPM model, the return on the stock are determined by the formula below:

$$\bar{r_a} = r_f + \beta_a * (\bar{r_m} - \bar{r_f})$$

In the formula above, r_a represents the reyurn on the asset. r_f is the riskless interest rate and r_m is the return on market. β_a represents to which extent does the stock undertake the risk of the market.

As the risk free interest rate can be seen as unchanged over a relative long time period, we can simpilfy the model and assume than the return on the stock is determined by the market condition and by the stock`s own characteristics.

Another essential assumption of the model is that the future return of the stock price can be predicted through analysis of past data.

2.3 Data, Training and Prediction Results

We choose IBM as the stcok to predict and Dow Jones index as the representation of the market. And the features we adopted are shown below:

Features	
Stock return	*Market return*
Yesterday	Yesterday
The day before yesterday	The day before yesterday
Three days ago	Three days ago
Labels	
The stock prices declines	0
Otherwise	1

We split the data into training data and testing data with a precentage of 25%. And we use the SVM function of sklearn to train the model. Then we use the function matrix.accuracy_score to attain its accuracy, which is only of 52%.

2.4 Analysis of the failure in prediction

[1] Support vector machine, From Wikipedia, the free encyclopedia

One possible explaination of the failure is that the connection between the past price data and the present price is ambiguous. Other factors such as the change of policy, investors` attitudes and the condition of the company can also influence the stock price. Besides, if many people predict that the stock price will rise tomorrow, they are likely to buy stock today, causing the price to rise immediately instead of rise tomorrow(reflexivity).

III. Regression: Using ARMA model[2]

3.1 Introduction of ARMA model

In the statistical analysis of time series, autoregressive–moving-average (ARMA) models provide a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the auto-regression and the second for the moving average. Given a time series of data Xt, the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The model is usually then referred to as the ARMA(p,q) model where p is the order of the autoregressive part and q is the order of the moving average part.[3]

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

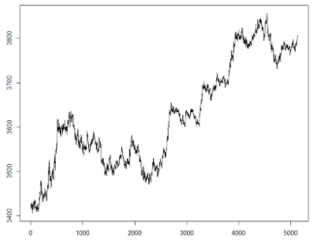
ARMA is appropriate when a system is a function of a series of unobserved shocks (the MA part) as well as its own behavior. For example, stock prices may be shocked by fundamental information as well as exhibiting technical trending and mean-reversion effects due to market participants.

3.2 Description of the data

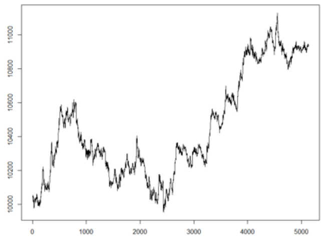
The data we adopted in the ARMA model is the high frequency prices series of HS300 Index and A50 Index between 9:15 o`clock in December 1st 2015 and 15:14 o`clock in December 25th 2015. And we then calculate the log return on both indexes.

[2] Although we build the model based on the ARMA fitting function of python, part of the data analysis and statistic test we conducted are realized through R.

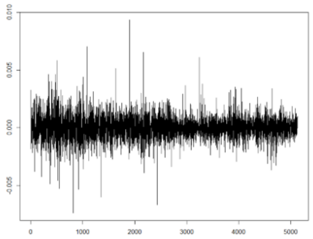
[3] Autoregressive–moving-average model,From Wikipedia, the free encyclopedia



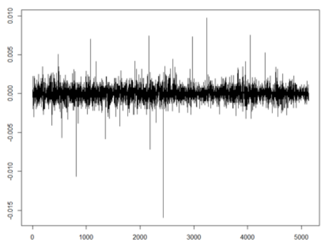
HS300 index



A50 index

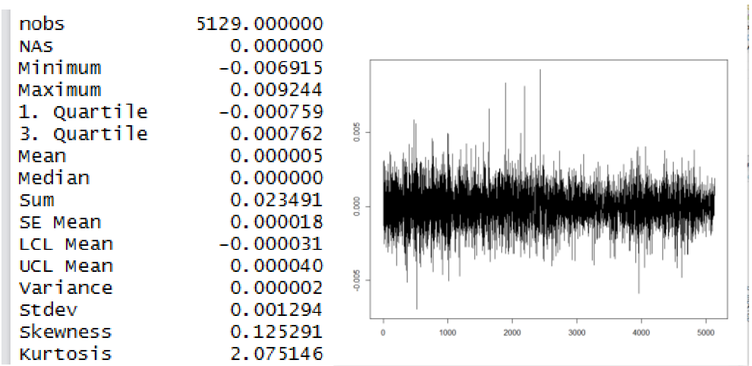


HS300 return



A50 return

As we build our model to do arbitrage between the option of two indexes, we then calculate the difference between the two log return series. Its basic statistic information is demonstrated below:



```
nobs           5129.000000
NAs               0.000000
Minimum          -0.006915
Maximum           0.009244
1. Quartile      -0.000759
3. Quartile       0.000762
Mean              0.000005
Median            0.000000
Sum               0.023491
SE Mean           0.000018
LCL Mean         -0.000031
UCL Mean          0.000040
Variance          0.000002
Stdev             0.001294
Skewness          0.125291
Kurtosis          2.075146
```

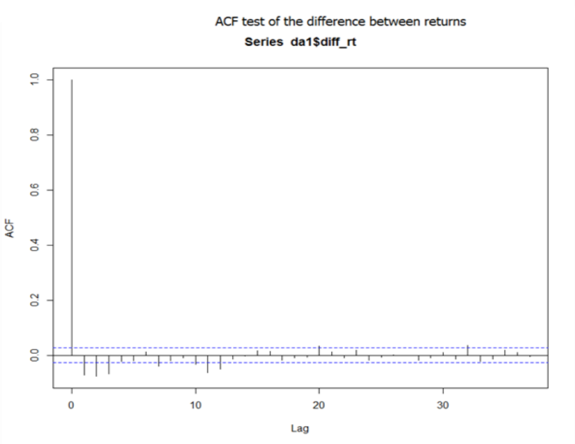
3.3 Assumptions and Test on the return series

Our model`s most important precondition is stationarity, which has 2 forms. The strict form is the data`s joint distributions of 1<t<T are time-invariant. The weak form is the first 2 moments are time-invariant. The stationarity in practice is that in the past, time plot of {rt} varies around a fixed level within a finite range; in the future, the first 2 moments of future rt are the same as those of the data so that meaningful inferences can be made.

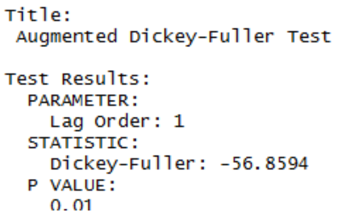
Besides, we conducted several statistic tests on the series we are going to build ARMA model on, including Ljung-Box test and Dickey-Fuller test. The first tests we conducted is to test the autocorrelation of the data.

Box-Ljung test	R_diff
X-squared	83.369
df	5
p-value	< 2.2e-16

As we can see from the table above, for the difference between the returns, the p-value is small enough to reject the null hypothesis, which is that the series don`t have autocorrelation. And the ACF graphics are shown below:



It can be seen from the images that the series has relatively nice properties when it comes to autocorrelation. And the nest test we conducted is to figure out whether the series we are modelling is Unit Root Stable.



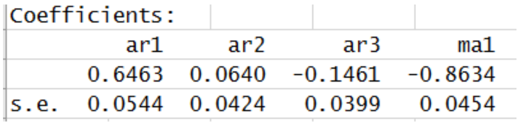
```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 1
  STATISTIC:
    Dickey-Fuller: -56.8594
  P VALUE:
    0.01
```

And from the above result we can see that there is no unit root stable in the series.

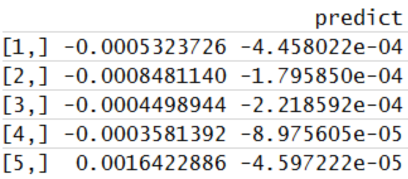
3.4 Training and Prediction

We can combine the AR(p) and MA(q) model to be ARMA(p,q) model. To identify the ARMA model, we still use AIC method. To estimate the model, we still use the conditional or exact likelihood method, which is achieved by the training based on the data series.



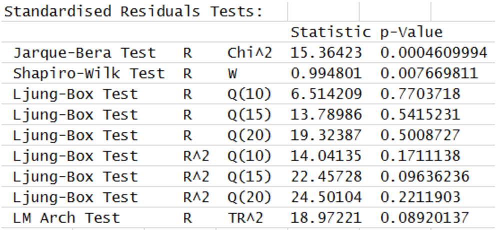
```
Coefficients:
          ar1     ar2      ar3      ma1
       0.6463  0.0640  -0.1461  -0.8634
s.e.   0.0544  0.0424   0.0399   0.0454
```

The prediction conducted by the model we trained are shown below. And it is a prediction of how the difference between the returns on indexes will be in the next 5 minutes, through which we are able to do arbitrage with a combination of longing the call option of one index and shorting the call option of another.

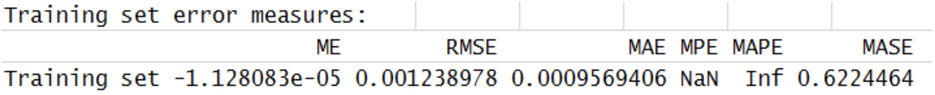


```
                             predict
[1,] -0.0005323726 -4.458022e-04
[2,] -0.0008481140 -1.795850e-04
[3,] -0.0004498944 -2.218592e-04
[4,] -0.0003581392 -8.975605e-05
[5,]  0.0016422886 -4.597222e-05
```

And the analysis of the residual shows that the residual series is more of a white noise series and tends to be relatively small.



```
Standardised Residuals Tests:
                                  Statistic p-Value
 Jarque-Bera Test   R    Chi^2  15.36423  0.0004609994
 Shapiro-Wilk Test  R    W      0.994801  0.007669811
 Ljung-Box Test     R    Q(10)  6.514209  0.7703718
 Ljung-Box Test     R    Q(15)  13.78986  0.5415231
 Ljung-Box Test     R    Q(20)  19.32387  0.5008727
 Ljung-Box Test     R^2  Q(10)  14.04135  0.1711138
 Ljung-Box Test     R^2  Q(15)  22.45728  0.09636236
 Ljung-Box Test     R^2  Q(20)  24.50104  0.2211903
 LM Arch Test       R    TR^2   18.97221  0.08920137
```



```
Training set error measures:
                        ME        RMSE          MAE MPE MAPE      MASE
Training set -1.128083e-05 0.001238978 0.0009569406 NaN  Inf 0.6224464
```

3.5 Summary of the ARMA Prediction

Although the model tends to be accurate in many circumstances, there are still enough wrong prediction for us to suffer from severe loses. We blame the errors of the prediction on the ARCH effect of the series, which can be solved through the modelling of the volatility.

Volatility is not directly observable. The basic idea behind volatility study is that the asset return series is either serially uncorrelated or with minor lower order serial correlations, but it is a dependent series, which means serial dependence in asset returns is nonlinear.

Volatility models attempt to capture such dependence in the return series. As we said before, $r_t = E(u_t|F_{t-1}) + \sigma_t \varepsilon_t$ where μ_t is estimated by ARMA. Volatility models are concerned with time-evolution of the conditional variance of the return: volatility equation—risk.

$$\sigma_t^2 = \mathrm{Var}(r_t|F_{t-1}) = \mathrm{Var}(a_t|F_{t-1})$$

We can get a better prediction with the GARCH model of the series volatility:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2$$

IV. Summary of the project and Future study

Although the issue is of great complexity and none of the model manages to solve the problem accurately, there are certain circumstances where we can use machine learning and time series analysis to get accurate predictions. For example, the second model is extremely powerful, or at least extremely powerful in some circumstances, in the prediction of return series with certain qualities.

SVM can be combined with data mining of social network(Twitter and Weibo) to get people`s attitudes towards certain stock, which can also be applied to the prediction of stock price. And our future study focuses on the combination of machine learning, time series analysis and data mining to build a trading system with stable profit and controllable risk and drawdown.