**The application of Hidden Markov Chain on market statement research and stock price prediction**

**Xiao LI**

**Introduction**

According to some behavioral Finance research, during some crisis, not just the volatility of asset price will be change a lot and become much higher than before, but the huge fund managers or some other trader may think more about the current condition than its history. To be specific, when the economic are not in crisis, trader may design their trading strategies based more on history data. However, when the asset price decrease dramatically, they might change their trading strategy, and make decision more on current condition.

In the study of Markov Chain, this phenomenon can be reflect on the order of Markov Chain. Thus, this paper is prepared to explore the relationship between the economic condition and the Markov Chain order of certain asset price or index. Based on this, I will explore whether some index can be the indicator of some other’s price change. Last, make prediction according to the result.

**Data Selection**

Market data: SPY, ^DJI

Sector data: XLE, XLK, XLF

Stock data: BAC, AAPL, EOG, XOM, ENB, ATW, CRK etc.

**Model Description**

In this paper, I try to find the relationship between the crisis time interval and the Markov Chain order of certain stock. To do this, the first question is how to determine a crisis period. Since most stock are influenced a lot by market or its own sector, I choose the data of market and the sector data as the indicators. To be specific, I conduct a hidden Markov Chain model to identify whether the macroeconomic environment or the sector are in crisis, denoted by a dramatic decrease. The second problem is that in each crisis period and non-crisis period, data is not enough to do the statistical test for the order of Markov Chain, thus, I combine all the crisis period together and calculate the statistic test value.



The third question is how to determine the crisis period. I first choose three statement model, up (U), normal (I) and down (D), whose return below 33%. As for the crisis period, I first consider the 1 day before and 3days after the crisis day. In addition, in terms of the crisis day, I consider the days whose probability of down statement over 95% to be crisis day while the days whose probability of up statement over 95% to be boom days

After dealing with these problems, I can conduct the model to do the research.

First, I choose the SP500 and Dow Jones as the market indicator and the sector data as the sector indicator. I use HMM model to identify, around Jan 1, 2000 to June 29, 2017, when the market are in crisis and when not. I set three statement as mentioned before, and assume that the return are follow lognormal distribution. Based on these condition, I use the EM[[1]](#footnote-1) algorithm on 11 (2+9) groups of data. And the result is as follows. Note that the green line is the probability of down statement while the red one is the probability of up statement.

**Market statement**

First, I build the HMM model based on the SP500 and Dow Jones Data, we can see that both results result are very close,

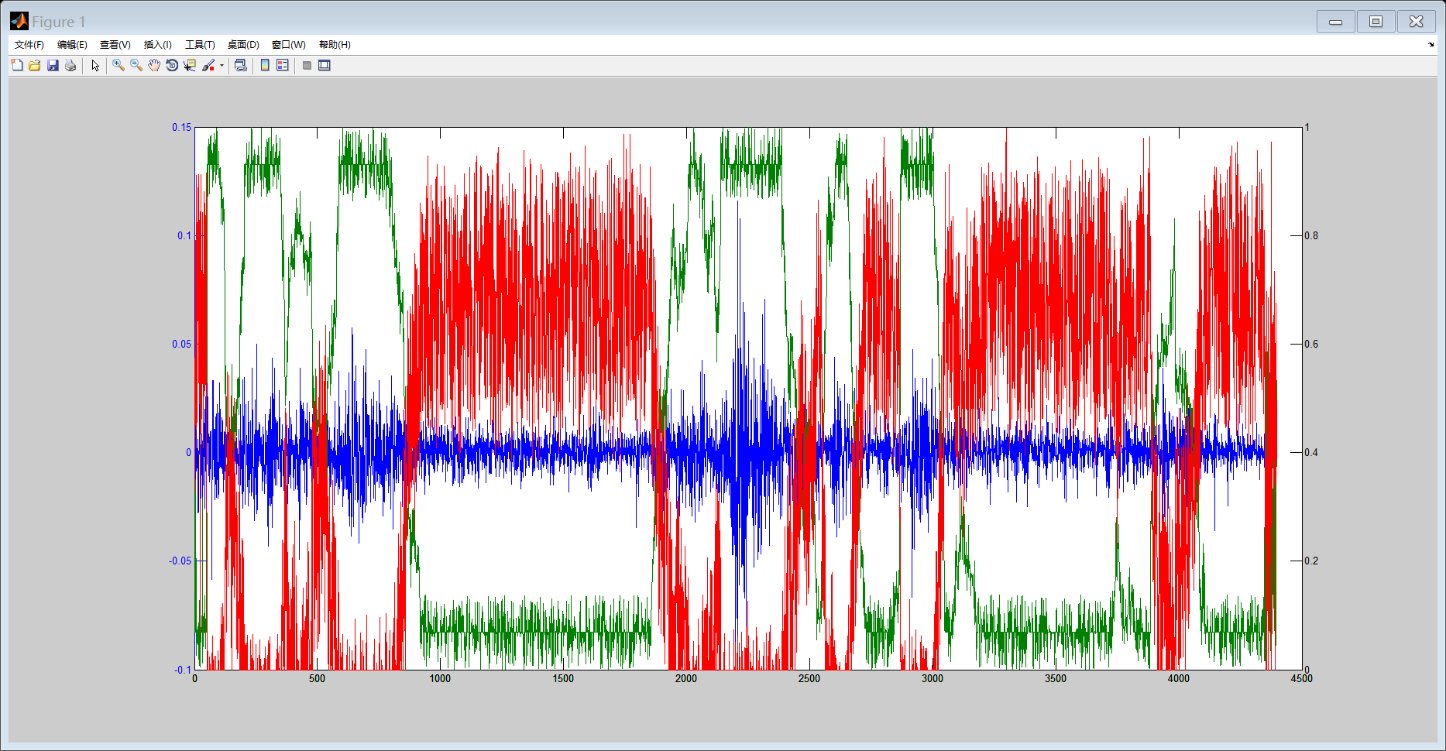


Figure1 the three statement HMM model of SP500 data

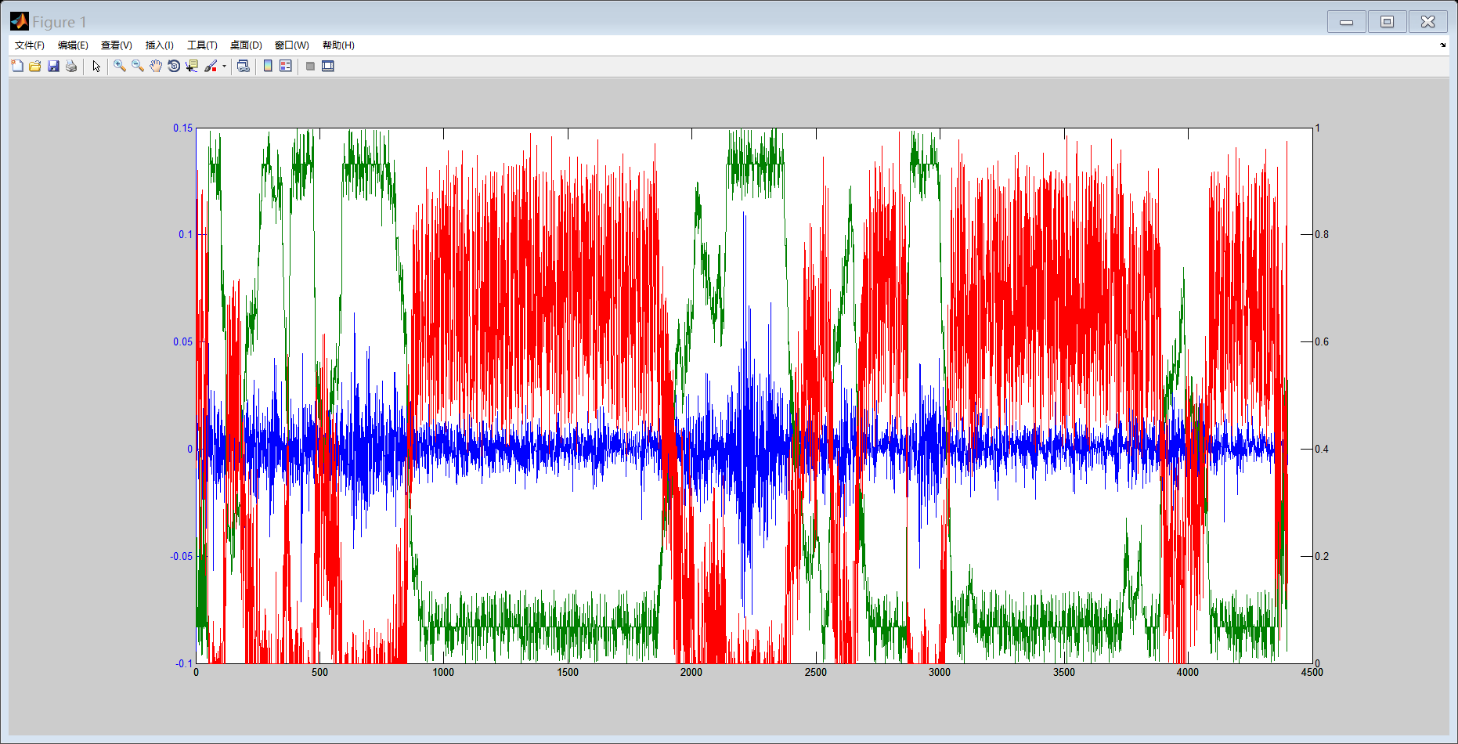


Figure2 the three statement HMM model of DJI data

**Sector Statement and Order determination**

In the **Energy sector**, we can see that the difference between up statement and down statement is much more clearly than the market data.

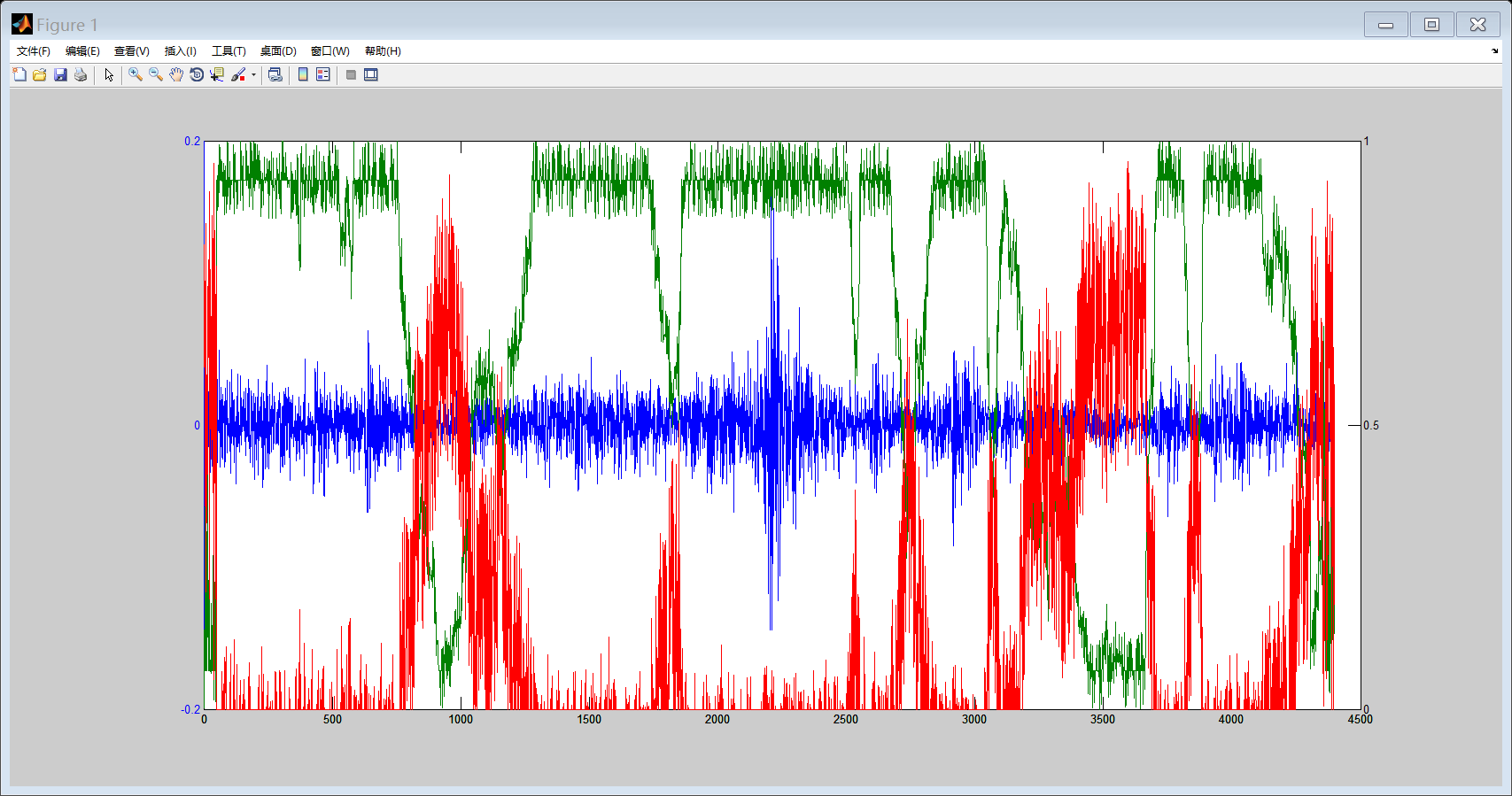


Figure3 the three statement HMM model of Energy sector data

Then I conduct two methods to select the crisis. In the first one, I choose the intersection of the down statement in each data as follows,



In the other standard, I choose the union of market statement and sector statement as follows.



After choosing the right crisis period, I select 5 most active stocks[[2]](#footnote-2) in that sector, and choose the days between 1 day before and 3 days after the crisis days as the test time interval. For example, in the crisis period mentioned above, if June 30, 2016 selected as crisis period, then in the stock data, the June 29, 2016 to July 3, 2016 will be used to do the statistical test. Then, to guarantee the number of variables, I combine all the stocks data, I use the following model to get statistic value

For the IID to first order,



For the first order to second order,



For the second order to third order,



And the result is as follows

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard[[3]](#footnote-3) | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 8.052 | 0.9103 |
| First to Second | 1 | 12 | 17.776 | 0.8773 |
| Second to Third | 1 | 36 | 36.981 | 0.5765 |
| IID to First | 2 | 4 | 7.825 | 0.9018 |
| First to Second | 2 | 12 | 13.006 | 0.6314 |
| Second to Third | 2 | 36 | 32.733 | 0.3752 |

We can see from the chart that the p-value of IID to First order test is significant on 90% confidence level but not exceed 95% confidence level, which means there is likely that the Markov Chain of ROR are follow either IID or First order Markov Chain. However, we can see from the p-value of the other two test that the Markov Chain is not second or third order.

On the other hand, I test the boom period, which has more than 95% probability to be up statement, with the same standard



And



The result is as follows

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard[[4]](#footnote-4) | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 20.210 | 0.9995 |
| First to Second | 1 | 12 | 22.661 | 0.9693 |
| Second to Third | 1 | 36 | 51.939 | 0.9584 |
| IID to First | 2 | 4 | 20.199 | 0.9995 |
| First to Second | 2 | 12 | 21.731 | 0.9594 |
| Second to Third | 2 | 36 | 51.726 | 0.9566 |

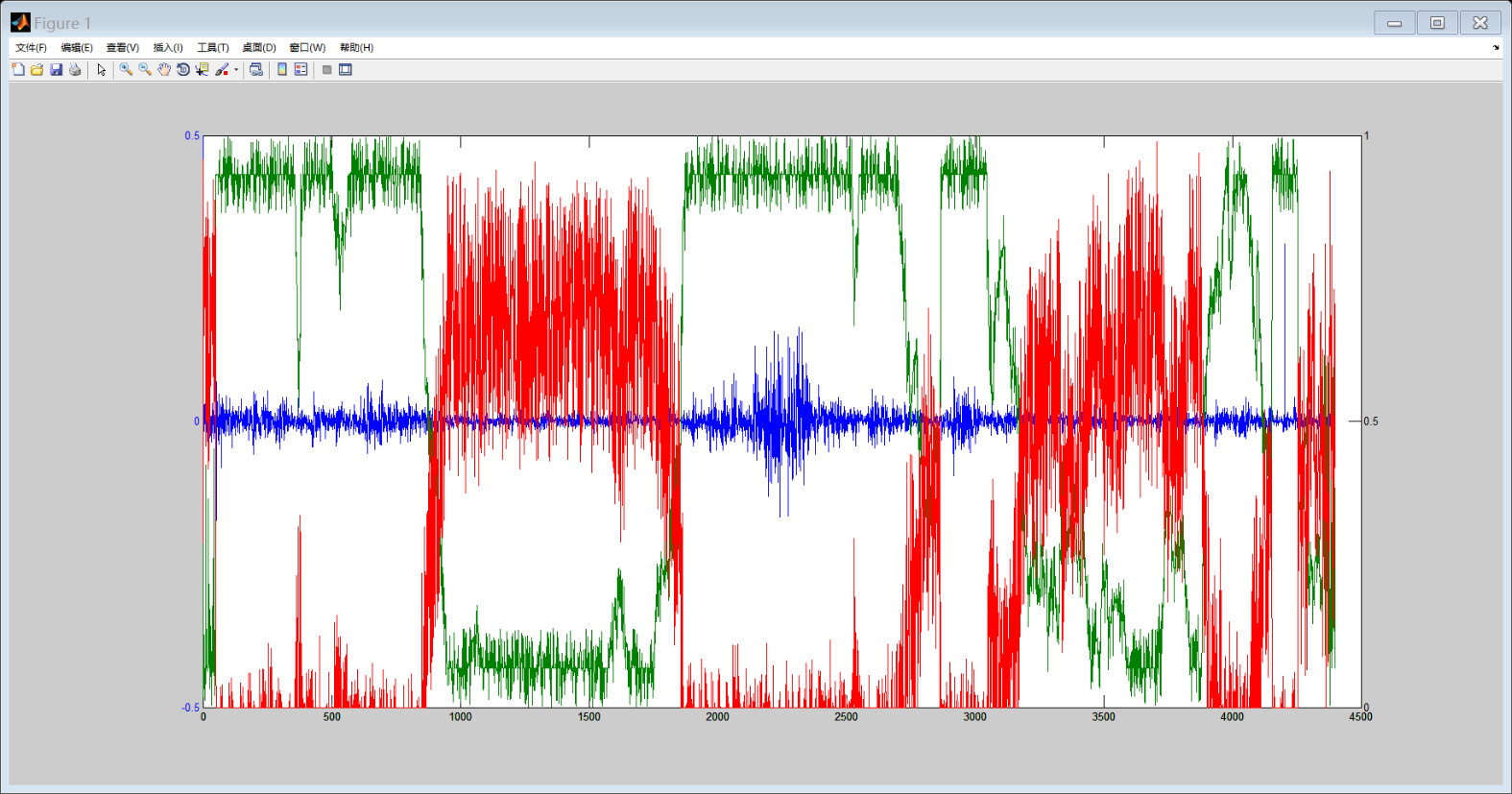
We can see from the chart that the p-values are all significant on 95% confidence level, which means that it is likely that the order of the Markov Chain of the boom period is third or more.

Therefore we can make conclusion that in the energy sector, the order will be lower when the market and the industry is in crisis while the order become higher when the economy recover. And the difference is much clear.

And the result of normal statement is as follows

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard[[5]](#footnote-5) | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 20.721 | 0.9995 |
| First to Second | 1 | 12 | 23.785 | 0.9782 |
| Second to Third | 1 | 36 | 43.063 | 0.8054 |
| IID to First | 2 | 4 | 21.642 | 0.9998 |
| First to Second | 2 | 12 | 22.096 | 0.9635 |
| Second to Third | 2 | 36 | 42.433 | 0.7866 |

**Finance sector**



With the same select standard, I get the crisis group[[6]](#footnote-6) and boom group[[7]](#footnote-7), choosing 5 stocks in this sector[[8]](#footnote-8) and the result of orders test are as follows

The crisis period

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard[[9]](#footnote-9) | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 6.836 | 0.8552 |
| First to Second | 1 | 12 | 16.522 | 0.8315 |
| Second to Third | 1 | 36 | 32.505 | 0.3644 |
| IID to First | 2 | 4 | 6.817 | 0.8541 |
| First to Second | 2 | 12 | 15.939 | 0.8060 |
| Second to Third | 2 | 36 | 31.977 | 0.3396 |

The boom period

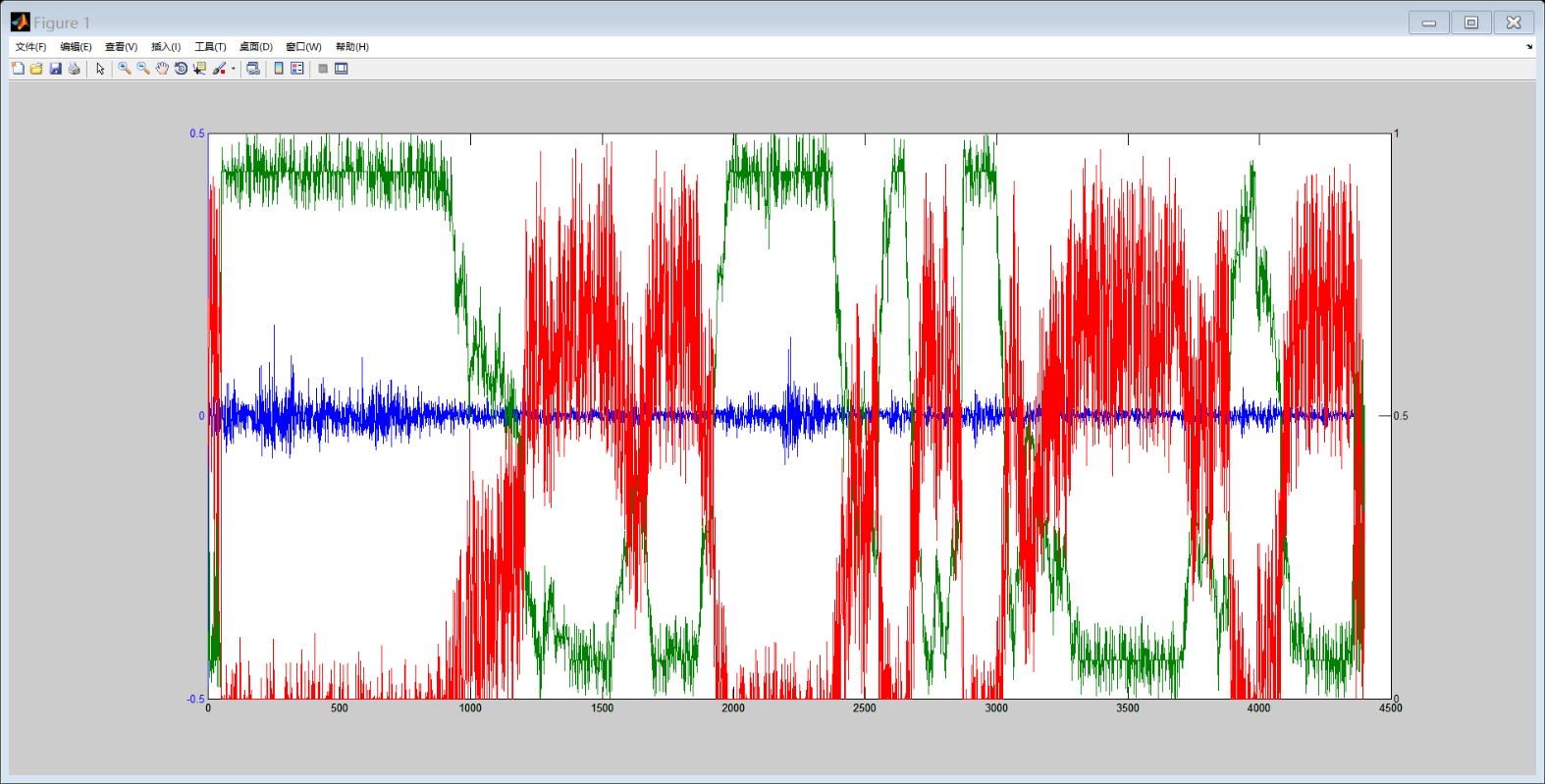
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard[[10]](#footnote-10) | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 8.921 | 0.9369 |
| First to Second | 1 | 12 | 18.359 | 0.8948 |
| Second to Third | 1 | 36 | 43.707 | 0.8233 |
| IID to First | 2 | 4 | 8.639 | 0.9292 |
| First to Second | 2 | 12 | 17.997 | 0.8842 |
| Second to Third | 2 | 36 | 43.211 | 0.8096 |

The normal period

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard[[11]](#footnote-11) | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 9.676 | 0.9538 |
| First to Second | 1 | 12 | 18.634 | 0.9023 |
| Second to Third | 1 | 36 | 44.273 | 0.8380 |
| IID to First | 2 | 4 | 9.134 | 0.9422 |
| First to Second | 2 | 12 | 18.374 | 0.8952 |
| Second to Third | 2 | 36 | 43.865 | 0.8275 |

We can see from the tables that during the crisis period, the ROR data turn to be an IID series while in the boom period, we can reject the IID hypothesis on 90% confidence interval but cannot reject on 95% confidence level. The difference between two intervals is not as clear as that in energy sector, one reason may be that the traders trait both condition in financial sector not so distinct. Another reason may be the data in this sector, especially the boom period, is not enough (only approximately 100).

**Technology Select Sector**

****

With the same select standard, I get the crisis group[[12]](#footnote-12) and boom group[[13]](#footnote-13), choosing 5 stocks in this sector[[14]](#footnote-14) and the result of orders test are as follows

The crisis period

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 8.117 | 0.9126 |
| First to Second | 1 | 12 | 15.697 | 0.7945 |
| Second to Third | 1 | 36 | 33.828 | 0.4277 |
| IID to First | 2 | 4 | 7.899 | 0.9047 |
| First to Second | 2 | 12 | 15.336 | 0.7764 |
| Second to Third | 2 | 36 | 33.533 | 0.4135 |

The boom period

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 13.986 | 0.9927 |
| First to Second | 1 | 12 | 26.797 | 0.9917 |
| Second to Third | 1 | 36 | 51.533 | 0.8233 |
| IID to First | 2 | 4 | 13.752 | 0.9919 |
| First to Second | 2 | 12 | 26.003 | 0.9893 |
| Second to Third | 2 | 36 | 51.112 | 0.9511 |

The normal period

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 14.406 | 0.9939 |
| First to Second | 1 | 12 | 27.317 | 0.9930 |
| Second to Third | 1 | 36 | 52.295 | 0.9612 |
| IID to First | 2 | 4 | 13.780 | 0.9920 |
| First to Second | 2 | 12 | 26.463 | 0.9908 |
| Second to Third | 2 | 36 | 51.807 | 0.9573 |

It is clear that in crisis period, we can reject IID hypothesis on 90% confidence interval but cannot on 95%. However, in boom period, we can consider the order up to three or more.

**Conclusion**

Among the three sector, we can find that the difference of traders’ decision making between three time intervals are distinct in Energy and Technology sector (both up to third order in boom condition but lower than one order in crisis) while unclear in Finance sector.

In addition, from the graphs of market index and each sector, we can see that the predict of crisis is much clear, which means the probability of down statement are most around 90% or 10% while as for up statement, the forecast is not so clear, which lead to fewer data for boom period to conduct order test. And it also can be seen from the tables that the orders in boom period and the normal period are much similar but both different from the crisis’.

**Stock price tendency Prediction**

In this part, I use the idea mention above to make forecast. First, I employ the transition matrix gotten above to predict the statement in the future 1 week, and then predict the certain ROR of each stocks. I will select one stock in each sector[[15]](#footnote-15).

According to the results above, I have gotten the transition probability matrix for each sector from EM algorithm, as follows,

For Energy sector



For Finance sector



For Technology sector



With the parameters gained from EM algorithm, we can get the the initial statement (the probability of each statement on June 30,2017) from the following formula



For Energy sector



For Finance sector



For Technology sector



Then on each sector, I calculate the probability in future one week.

Probability of Energy sector

|  |  |  |
| --- | --- | --- |
| Down | Normal | UP |
| 0.431534 | 0.297844 | 0.271622 |
| 0.397418 | 0.301218 | 0.302364 |
| 0.385267 | 0.30141 | 0.314323 |
| 0.380869 | 0.301329 | 0.318802 |
| 0.379267 | 0.301278 | 0.320455 |
| 0.378682 | 0.301255 | 0.321062 |
| 0.378468 | 0.301247 | 0.321285 |

Probability of Finance sector

|  |  |  |
| --- | --- | --- |
| Down | Normal | UP |
| 0.247052 | 0.451281 | 0.301667 |
| 0.252192 | 0.441553 | 0.306255 |
| 0.253988 | 0.438887 | 0.307126 |
| 0.25468 | 0.438147 | 0.307174 |
| 0.254972 | 0.437937 | 0.307092 |
| 0.255104 | 0.437875 | 0.307021 |
| 0.255167 | 0.437856 | 0.306978 |

Probability of Technology sector

|  |  |  |
| --- | --- | --- |
| Down | Normal | UP |
| 0.292076 | 0.414009 | 0.293915 |
| 0.255849 | 0.414872 | 0.329279 |
| 0.24114 | 0.413371 | 0.345489 |
| 0.235092 | 0.412252 | 0.352656 |
| 0.232585 | 0.411654 | 0.355761 |
| 0.23154 | 0.411369 | 0.357091 |
| 0.231103 | 0.41124 | 0.357657 |

Then, I conduct the forecast on stock under certain statement. For example, in Energy sector, the crisis statement has been consider to be first order Markov Chain[[16]](#footnote-16), and I have get the first order transition matrix as follows.



However, in boom statement, it is considered as third order Markov chain. We can get the third order Markov Chain probability with the following formula



Since andthe formula can be written as



Thus, I make the prediction based on the third order transition probability. And the result is as follows.

For EOC

|  |  |  |  |
| --- | --- | --- | --- |
|  | Stock down | Stock invariant | Stock up |
| Crisis | 0.489 | 0.327 | 0.184 |
| Normal | 0.279 | 0.433 | 0.288 |
| Boom | 0.238 | 0.296 | 0.466 |

For BAC

|  |  |  |  |
| --- | --- | --- | --- |
|  | Stock down | Stock invariant | Stock up |
| Crisis | 0.428 | 0.393 | 0.178 |
| Normal | 0.306 | 0.487 | 0.207 |
| Boom | 0.179 | 0.261 | 0.560 |

For AAPL

|  |  |  |  |
| --- | --- | --- | --- |
|  | Stock down | Stock invariant | Stock up |
| Crisis | 0.453 | 0.415 | 0.132 |
| Normal | 0.283 | 0.520 | 0.197 |
| Boom | 0.206 | 0.346 | 0.447 |

Thus, the probability of the statement of certain stock can be calculate by the following formula,



And the result is as follows

|  |  |  |  |
| --- | --- | --- | --- |
|  | Down | Invariant | Up |
| EOG | 0.346 | 0.349 | 0.306 |
| BAC | 0.298 | 0.394 | 0.308 |
| Boom | 0.295 | 0.434 | 0.271 |

**Extension: Viterbi algorithm**

Viterbi algorithm is a method based on the optimal path. In this algorithm, it put forward the maximum possible step by step, and ultimately get a path for start to end. Then it determines the end point according to the probability, from which we can put forward the optimal path.



And the process is as follows,

1. Initialization





1. Put forward for t =2,3,4,5,6,7…T





1. Make selection





1. We can get the path



1. The path is 

I conduct the algorithm on the model building part to get the new predicted probability and statement period. Take the Energy sector as an example,

In crisis period

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard[[17]](#footnote-17) | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 7.662 | 0. 8952 |
| First to Second | 1 | 12 | 16.882 | 0. 8459 |
| Second to Third | 1 | 36 | 35.092 | 0. 4884 |
| IID to First | 2 | 4 | 7.447 | 0. 8859 |
| First to Second | 2 | 12 | 12.359 | 0. 5827 |
| Second to Third | 2 | 36 | 31.064 | 0. 2977 |

In boom period

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard[[18]](#footnote-18) | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 17.141 | 0. 9982 |
| First to Second | 1 | 12 | 19.224 | 0. 9167 |
| Second to Third | 1 | 36 | 44.111 | 0. 8339 |
| IID to First | 2 | 4 | 17.131 | 0. 9982 |
| First to Second | 2 | 12 | 18.433 | 0. 8969 |
| Second to Third | 2 | 36 | 43.930 | 0. 8292 |

In normal period

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Standard[[19]](#footnote-19) | Degree of Freedom | Chi-Square Value | P-Value |
| IID to First | 1 | 4 | 16.820 | 0. 9979 |
| First to Second | 1 | 12 | 18.865 | 0. 9082 |
| Second to Third | 1 | 36 | 43.286 | 0. 8117 |
| IID to First | 2 | 4 | 16.811 | 0. 9979 |
| First to Second | 2 | 12 | 18.089 | 0. 8870 |
| Second to Third | 2 | 36 | 43.109 | 0. 8067 |

We can see that the result is different, which happens may because the EM algorithm gives probability of statement while the Viterbi gives certain value. Thus due to different standards to choose each statement in EM, period selected may be different. The other explanation is that the Viterbi algorithm does not work well in long time series, since it depends on the probability at the end of path, which may be not accurate in long run.

Reference

Kevin, Atteson. "Theory and Applications of Markov Chains to Finance" *http://atteson.com/Markov/Markov.html*

Krogh, Anders, et al. "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." *Journal of molecular biology* 305.3 (2001): 567-580.

Bahl, Lalit, et al. "Maximum mutual information estimation of hidden Markov model parameters for speech recognition." *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*. Vol. 11. IEEE, 1986.

Rabiner, Lawrence, and B. Juang. "An introduction to hidden Markov models." *ieee assp magazine* 3.1 (1986): 4-16.

Billingsley, Patrick. Statistical Inference for Markov Processes. The University of Chicago Press. 1961.

Mamon, Rogemar S. and Elliott, Robert J. Hidden Markov Models in Finance, Volume II. Springer, 2014.

1. With 300 times setting initial values [↑](#footnote-ref-1)
2. Include EOG, XOM, ENB, ATW, CRK [↑](#footnote-ref-2)
3. 1 means the first methods to select the crisis period which has about 500 values and method 2 has about 700 [↑](#footnote-ref-3)
4. Standard 1 has about 300 values and standard 2 has about 300 either [↑](#footnote-ref-4)
5. Standard 1 has about 300 values and standard 2 has about 300 either [↑](#footnote-ref-5)
6. Standard 1 has about 800 values and standard 2 has around 900 [↑](#footnote-ref-6)
7. Standard 1 has about 100 values and standard 2 has around 100 either [↑](#footnote-ref-7)
8. Including BOC, JPM, WFC, C, BOFI [↑](#footnote-ref-8)
9. 1 means the first methods to select the crisis period which has about 500 values and method 2 has about 700 [↑](#footnote-ref-9)
10. 1 means the first methods to select the crisis period which has about 500 values and method 2 has about 700 [↑](#footnote-ref-10)
11. 1 means the first methods to select the crisis period which has about 500 values and method 2 has about 700 [↑](#footnote-ref-11)
12. Standard 1 has about 800 values and standard 2 has around 800 either [↑](#footnote-ref-12)
13. Standard 1 has about 200 values and standard 2 has around 200 either [↑](#footnote-ref-13)
14. Including AAPL, MSFT, FB, GOOG, T [↑](#footnote-ref-14)
15. EOG in Energy sector, BOC in Finance sector and AAPL in Technology sector [↑](#footnote-ref-15)
16. Based on the table of Chi2 above [↑](#footnote-ref-16)
17. 1 means the first methods to select the crisis period which has about 500 values and method 2 has about 700 [↑](#footnote-ref-17)
18. Standard 1 has about 300 values and standard 2 has about 300 either [↑](#footnote-ref-18)
19. Standard 1 has about 300 values and standard 2 has about 300 either [↑](#footnote-ref-19)