ILLINOIS INSTITUTE OF TECHNOLOGY

MATH 561 Algebraic and Geometric Methods in Statistics

Conditional Independency of Stocks Given US Dollar Index

Team 3

Mohammad Nabeel Khan
Amirreza Eshraghi
Ranjan Mishra

Prof. Sonja Petrovic

INTRODUCTION

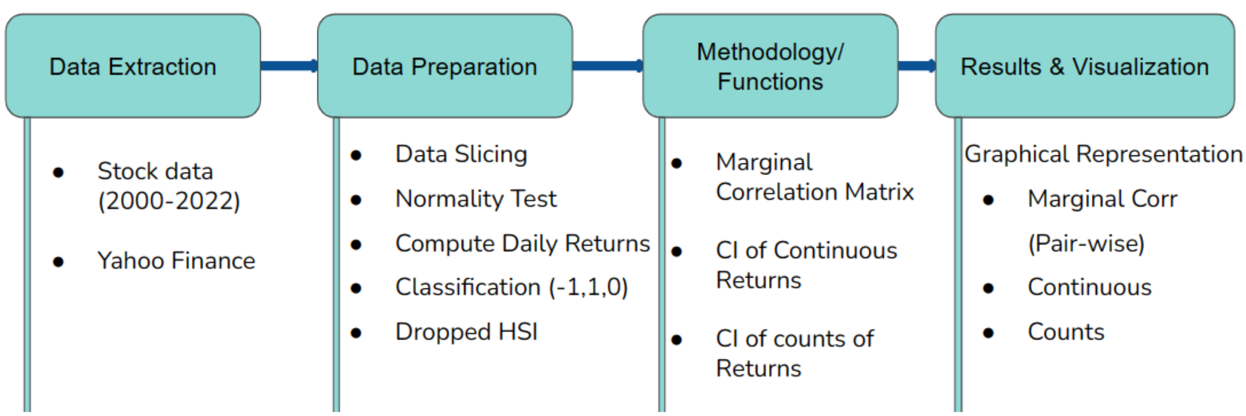
The main objective is to find out how federal reserve interest rates affect the behavior of the stock market. That is Conditional Independency of different stocks given US dollar and pairwise correlation of each stocks together.

We Sliced the data on the federal reserve based on the increase, decrease and stability of interest rates and plotted the graphs based on conditional independence results. How do the stocks differ?

We are focusing on different sectors of stocks such as energy, technology, consumer staple, finance and healthcare. These have high correlation with US dollars, If we condition on US dollar we expect the dependency to diminish.

Different sectors of stock historically showed to be highly correlated to US dollar in specific times and slices. Representation of each slice would be in a form of graph that would show how the Conditional Independency differ with continuous daily returns, Counts of daily returns and inverse correlation of each sectors of daily returns.

WORKFLOW OVERVIEW



First we perform data extraction of Stocks for the year 2000 to 2022 using Yahoo Finance (Yfinance) library.

Then we perform data preparation, where we first slice the data based on time period, then check if data is normally distributed by doing normality test.

Further we compute the daily returns, and classify the data into 3 categories -1,1 and 0 based on specific conditions. During preprocessing we dropped HSI is has a anonymous behavior

The third step is implementing the methodology, where 3 functions were built to perform a marginal pairwise correlation matrix, conditional independence of continuous returns and conditional independence of counts returns.

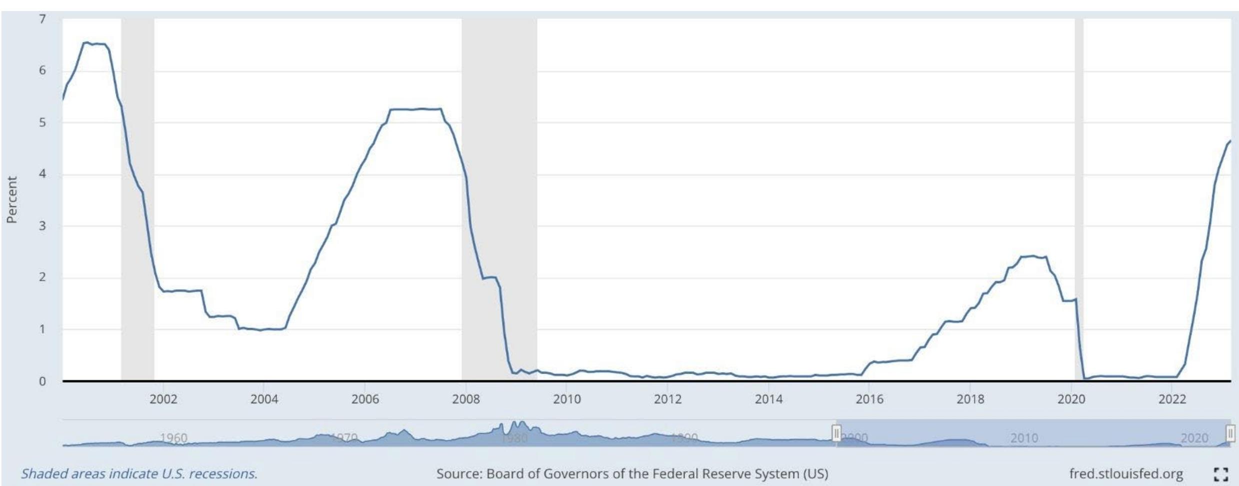
The fourth step was to build graph representation of marginal pair-wise correlation, continuous returns and counts. Finally concluded with the interpretation based on the results.

DATA EXTRACTION

All the financial data can be obtained from yahoo finance. To visualize the difference of the performance between each stock/ funds, we choose them from different area: apple stock; Nasdaq- 100 index; healthcare SPDR funds; consumer staples SPDR funds; technology SPDR funds; utilities SPDR funds; energy SPDR funds; financial SPDR funds; UK market index; Hong Kong market index; and US- dollar. We set the time range starting from 2000 to 2022.

DATA PREPARATION

Since the whole dataset has more than 20 years length, using the entire one to analyze may not be a good idea as the public free-marketing can be affected by all kinds of factors including economic status, company status, seasonal cyclic behavior or interest rate. Other unexpected reasons may also impact the price, such as coronavirus and financial crisis. Under such circumstances, we decide to slice the dataset according to the federal reserve interest rate and keep each time frame not shorter than 3 months. As seen in fig[], the US dollars index are highly correlated with the federal reserve interest rate



Normality test

Normality test is a common test used to check if the data follows normal distribution. With p-value less than the critical value, we can reject the null hypothesis and conclude that the dataset is not following the normal distribution.

In our project, we use the function "*scipy.stats.normaltest*" for normality checks. Based on D'Agostino and Pearson [6], this test combines skew and kurtosis to produce an omnibus test of normality**.** Initially we set the null hypothesis as: the datasets do follow a normal distribution with critical value $\alpha = 0.01$. By performing the test, all the p-values for different tickers are way below 0.01 (for example, p-value for US-dollar is 4.19e-26), which indicates that we can confidently reject the null hypothesis and conclude that none of them follows the normal distribution.

Dataframe preparation

Based on the experiments, we need both continuous datasets and classification datasets. We chose the "adjusted closing price" and removed all the other information, then combined all the "Adj price" together. So the whole data frame including 12 sliced lists, in each list, contains the adjusted closing price for all the tickers ordered by time. Then we computed the returns based on the adjusted closing price. So the return dataframe will be our first dataframe that can be used for continuous conditional independence tests.

To prepare for the classification test, we prepared the second dataframe based on the return table as: if the return value is positive, we just set it as 1; if the return value is negative, we set it as -1; if the return value is zero, we keep it as zero. In such a way, we have our classification table that contains only 0, -1, and 1.

Correlation matrix can also be used for independence tests, so we prepare our correlation matrix based on the returns by using the “cor()” function in r.

Now the datasets are well prepared and ready for use.

METHODOLOGY/ FUNCTIONS

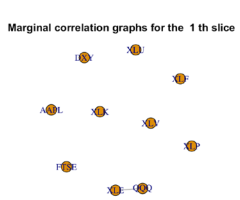
Function	Input	Output
Marginal Correlation Graphs	Correlation matrix and a threshold value	Dataframe with pairs of variables and an edge (1 or 0) indicating significant relationships
CI Function	Dataframe of returns, type of test, and fixed variable (DXY)	Dataframe with p-values and correlations for every pair of stocks
Graph Plotting Function	Dataframe with pairs of variables and an edge value, and a boolean value for graph complementation	A plotted graph using the ‘igraph’ library

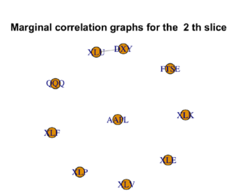
VISUALIZATION AND RESULTS

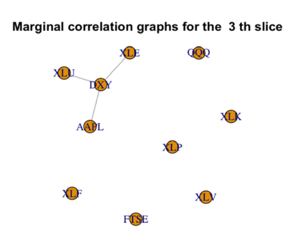
Looking at the marginal correlation graphs as the main source of independence, we can see based on results, in half of the time-slices, the US dollar index i.e “DXY” has many dependencies. This test simply computes the pearson correlation coefficient and based on a threshold which in our case we chose 0.08. The formula has been given by the following:

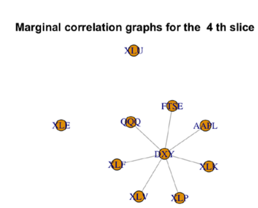
$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

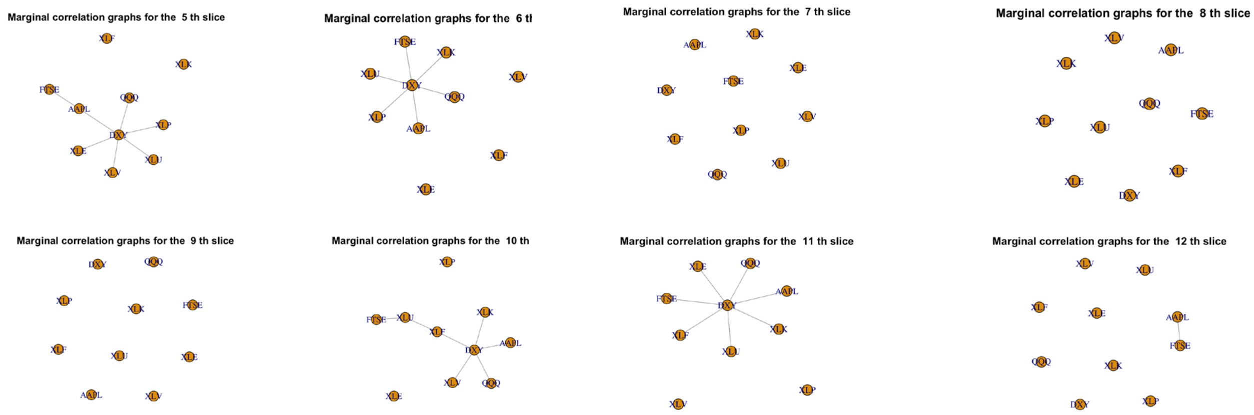
Although we should keep this in mind that in the other half of the time-slices there is no dependency on any sectors of the stocks that we studied with DXY. It’s important to note that the following graphs are the complement of the independence graph which means, edges in the graphs represent **dependency** and missing edges correspond to independent variables.





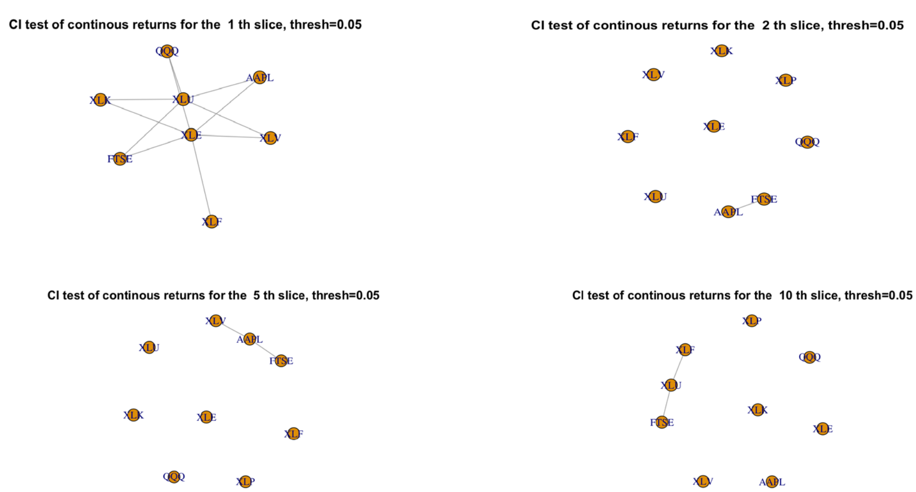


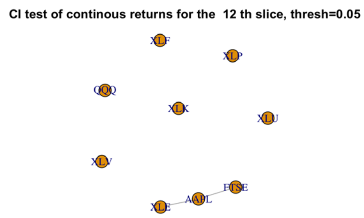




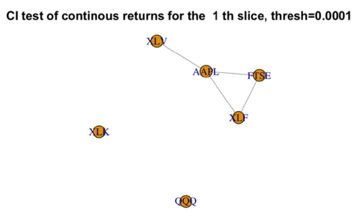
This shows us that conditioning on DXY is a good idea to test whether different sectors given DXY are still independent of each other or not. Since the graphs above confirm the idea of testing the CI of different pairs, to be able to differentiate the time-slices, we performed conditional independency test for every pair and for every slice for both the exact returns and the classified(count) returns. Since we were interested in using the p-value from the test to have a graphical representation (NOT A MODEL), we need a threshold for constructing the graphs.

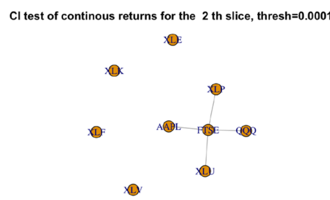
After choosing a bunch of different values for the thresholds, we found out that if we set the threshold to be 0.8, then all the pairs for every slice are independent of each other given DXY.The choice of the threshold is a parameter in our function that the experimenter can choose based on his preferences and sensitiveness, the lower the threshold goes, the more connected graphs will be obtained. For example, in our project we did set the threshold to be 0.05 and 0.0001 (we are only including the graphs which are not empty i.e fully independent) :

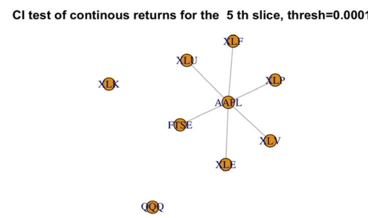


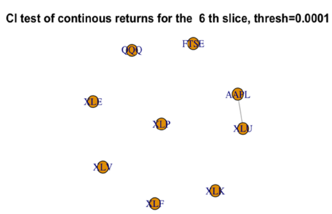


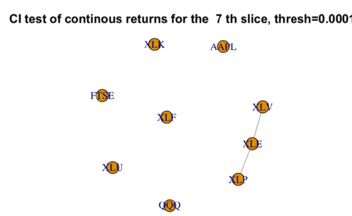
For threshold=0.0001:

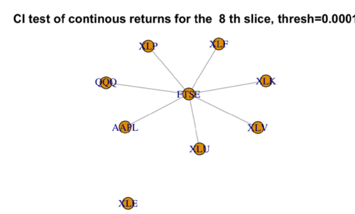


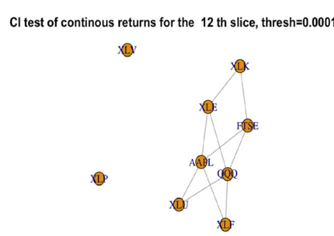




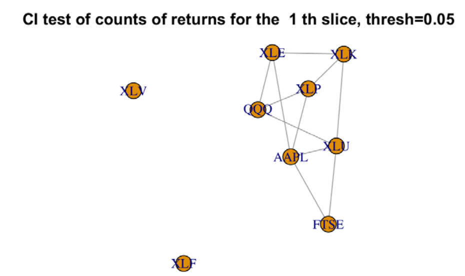


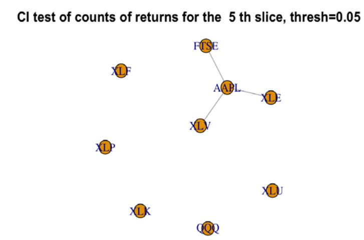


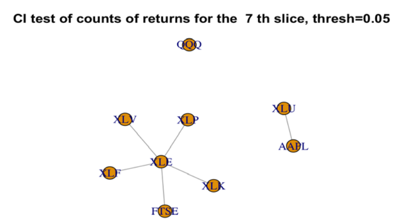


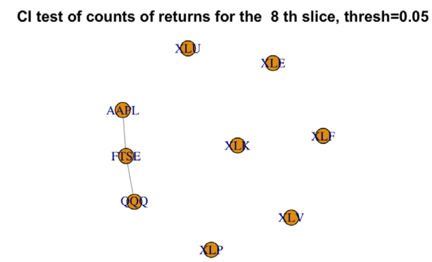


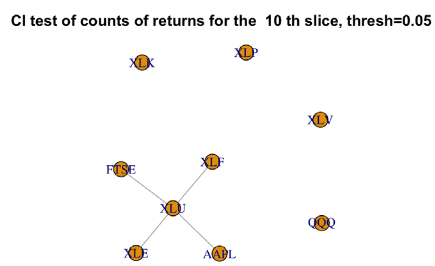
Repeating the same thing for classified data counts, threshold=0.05:

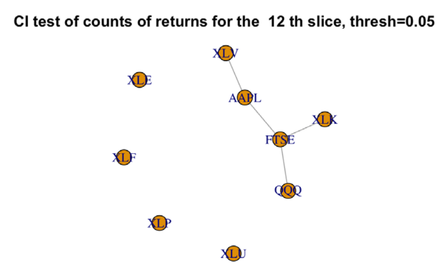




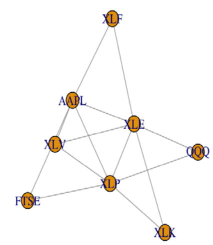




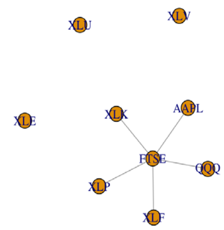




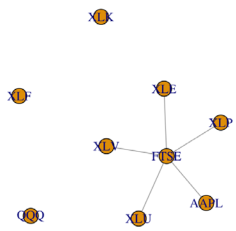
Threshold=0.0001 for counts:



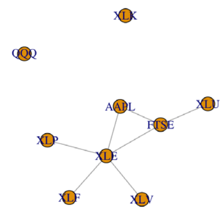
CI test of counts of returns for the 1 th slice, thresh=0.0001



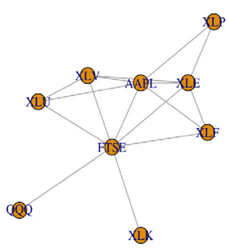
CI test of counts of returns for the 2 th slice, thresh=0.0001



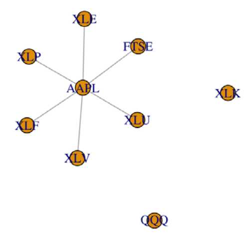
CI test of counts of returns for the 3 th slice, thresh=0.0001



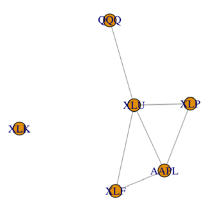
CI test of counts of returns for the 4 th slice, thresh=0.0001



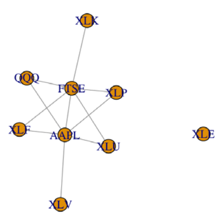
CI test of counts of returns for the 5 th slice, thresh=0.0001



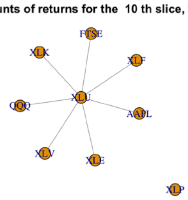
CI test of counts of returns for the 6 th slice, thresh=0.0001



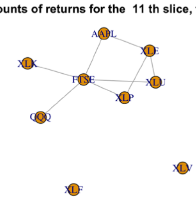
CI test of counts of returns for the 7 th slice, thresh=0.0001



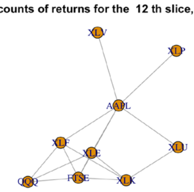
CI test of counts of returns for the 8 th slice, thresh=0.0001



CI test of counts of returns for the 10 th slice, thresh=0.0001



CI test of counts of returns for the 11 th slice, thresh=0.0001



CI test of counts of returns for the 12 th slice, thresh=0.05

CONCLUSION

Federal reserve interest rates in some periods of time can definitely impact the dependency of the different sectors which the effect can be translated to the behavior of the US dollar index which on those eras, it can directly impact the stock market specially QQQ and the technology sector of S&P 500 i.e XLK.

What is more important than hiking/cutting/pausing interest rates that affect the market, was the rate or rates i.e the change in the increments of each those. For example during summer 2022 rates were being hiked by 75 basis points and the market was in a downtrend. In the beginning of 2023 the rate was increased by 50 basis points and going forward it did increase by 25 basis points which the market was clearly in an uptrend.

Oftentimes this dependency and the characteristics of the period of time can be used as an indicator of the performance of different aspects of the stock market. However, during the rest of the periods, the dependency on DXY will vanish and even conditioning on DXY and using the characteristics of the graphs that we presented can be misleading in understanding the behavior of the market. An investor should be careful regarding using these tools in the cases that the reason behind hiking/cutting/pausing are not the typical reason.

The tools that we provided can be used to gain insights on current market conditions, depending on the choice of the threshold and analyzing the behavior of DXY. In recession themes and inflation themes, a pop in DXY will be a warning for the market if the federal reserve is in the hiking cycle of the interest rates.

FUTURE WORK

The same methods can be applied to different time-slices, not necessarily based on the federal reserve interest rates but other variables of interest. For example one might seek the performance of different sectors and conditional independence of pairs given DXY by studying the data quarter by quarter.

Using the p-value and correlation coefficient as features of a time-slice i.e period, can be leveraged to do clustering analysis. In our project, since we chose 9 different sectors and we computed pairs, the set of features contains 46 columns, and since we have only 12 time slices, doing K-Means or other clustering can be dangerous. But using more time slices and reducing the period will lead to a bigger dataset and the clustering can be performed.

Definitely the choice of DXY to be conditioned was something that we had in mind in the beginning but we can repeat the same procedure and condition on different indexes or even stocks to perform the analysis.

REFERENCES

1. https://fred.stlouisfed.org/series/FEDFUNDS
2. https://www.stat.cmu.edu/~larry/=sml/GraphicalModels.pdf
3. https://kateto.net/netscix2016.html
4. https://www.indmoney.com/articles/us-stocks/how-does-rising-affect-the-stock-market
5. https://finance.yahoo.com/quote/%5EGSPC/
6. Bnlearn-Conditional independence in R https://www.bnlearn.com/examples/ci.test
7. D'Agostino, R. and Pearson, E. S. (1973), "Tests for departure from normality", Biometrika, 60, 613-622