# Analyze the stocks in S&P 500 Index by PCA

### Overview

PCA (principal component analysis) is a useful data-reduction tool to analyze correlated variables and reduce data dimensions. It uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components[[1]](#footnote-1). The principal components could explain the different variation in the original data but in reduced number.

The stock market contains complicated data due to frequent changes, highly correlated prices and so on. The price changes of a stock might be caused by the market effect, the industry influence, or the effect from the opposite part, for example, a sudden increase in the price of oil, will be positive to oil companies but decrease the benefits of airlines industry.

Thus, the discussion of the project focuses on showing and explaining the similarity and variation of the stocks’ return by PCA. The stock list is from S&P 500 Index[[2]](#footnote-2), which is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE, NASDAQ, or the Cboe BZX Exchange.

### Data

I used the 504 stocks in the S&P 500 index (updated 2019/04/19). Through this list of stocks, I performed PCA on the return of 504 companies over a varying period of time from 2009/03/31 to 2019/04/15. The reason to choose the time period after 2009 is to avoid the turbulence caused by the economic crisis in 2008.

The stocks’ price data is downloaded from Yahoo Finance[[3]](#footnote-3), an open and free financial data resources online. But the number of stocks is too much to download by hand, which means opening one website for each stock and choosing the date ranges then downloading. I chose to use pdr.get\_data\_yahoo() function in the pandas\_datareader library by python .

The pandas\_datareader library is an example of an external library for remote access to an Application Programming Interface (API) and the pdr.get\_data\_yahoo() function provide directly connection with Yahoo Finance[[4]](#footnote-4). They are so efficient that I could download all data in 5 minutes and continue to the next step.

The return matrix is derived by the adjusted closing price of 504 stocks in the S&P500 Index portfolios from 2009/03/31 to 2019/04/15.

The daily return matrix is calculated by:

\*100

The is the return of the day of all dates for the stock.

The adjusted close is the adjusted close price of the day of all dates for the stock.

i is from 1 to 2528, standing for the weekdays between 2009/03/31 and 2019/04/15.

j is from 1 to 504, standing for the 504 stocks.

The return from start date is to standardize the stock price by the adjusted closing price of 2009/03/31.

\*100

You could imagine one scenario that you invest one million dollars into the stock on the date and how much return will you get in the following days.

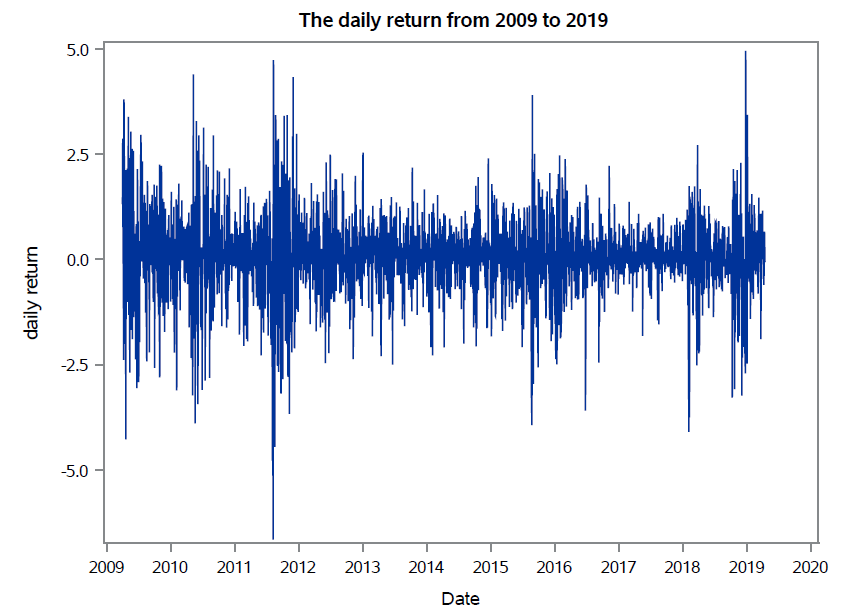
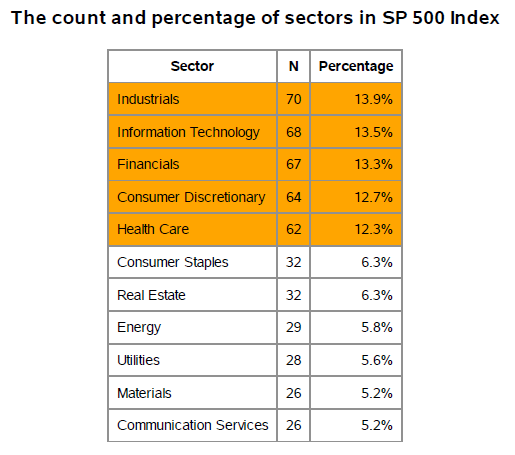
### Analysis

**PCA in Python:**

By the return matrix above, we could calculate the covariance matrix by cov(.) function. Then put the covariance matrix into la.eig() function and get the eigenvalues and eigenvectors from the result.

Later, I wrote the data frames of them into csv files, then used SAS to generate tables and graphs for further analysis based on the PCA results.

**Sectors distribution in S&P500 Index**



There are totally 504 stocks in S&P 500 Index and the most stocks come from industrials, Information Technology, Financials, Consumer Discretionary and Health Care, of which percentages over 10%. The other sectors are minority compared to them.

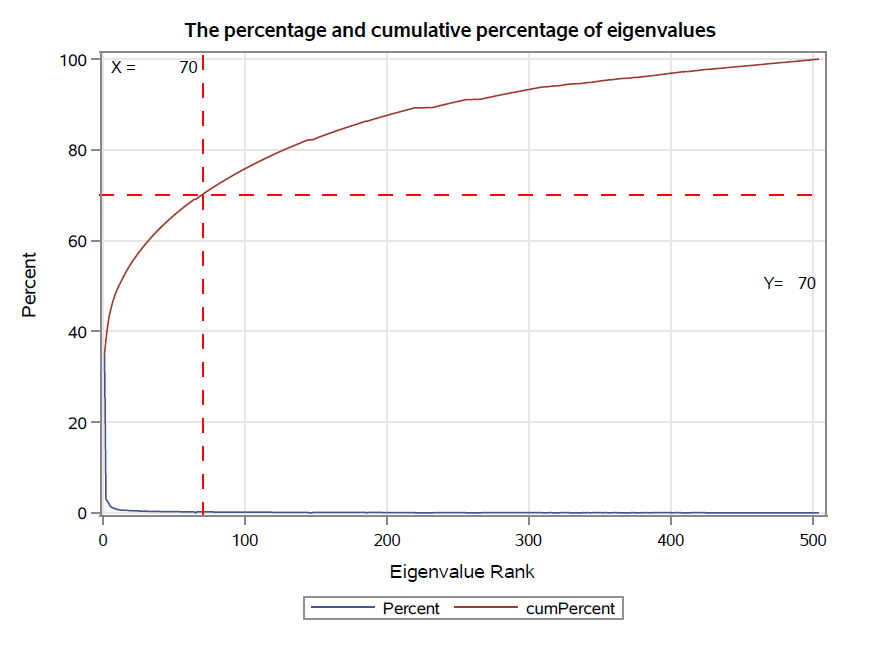
The daily return of the S&P 500 Index looks like a brain wave, with random trend with the date. Because the daily return is actually comparing the price of current day to the previous day, the return could won’t increases or decrease all the time but fluctuate around zero. When the return experience

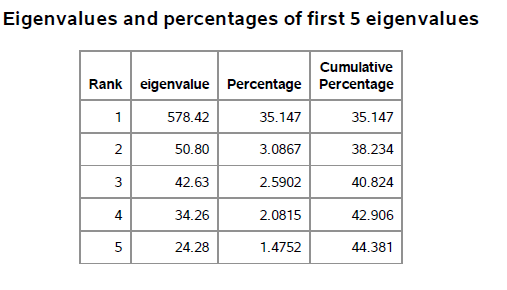
Dramatic fluctuation, the whole market might be affected by some shocking news or under unstable economic environment.

For example, the daily return changed dramatically between mid-2011 to mid-2012, ranging from -0.7 to 0.5. During the time period, the United States debt-ceiling crisis of 2011 might be the most significant factor, which sparked the most volatile week for financial markets since the 2008 crisis, with the stock market trending significantly downward.

Thus, the S&P 500 is an index to reflect the change of stock market and later I will analyze three eigenvectors from the PCA and discussion the relationship between S&P 500 and them.

**The distribution of the eigenvalues:**

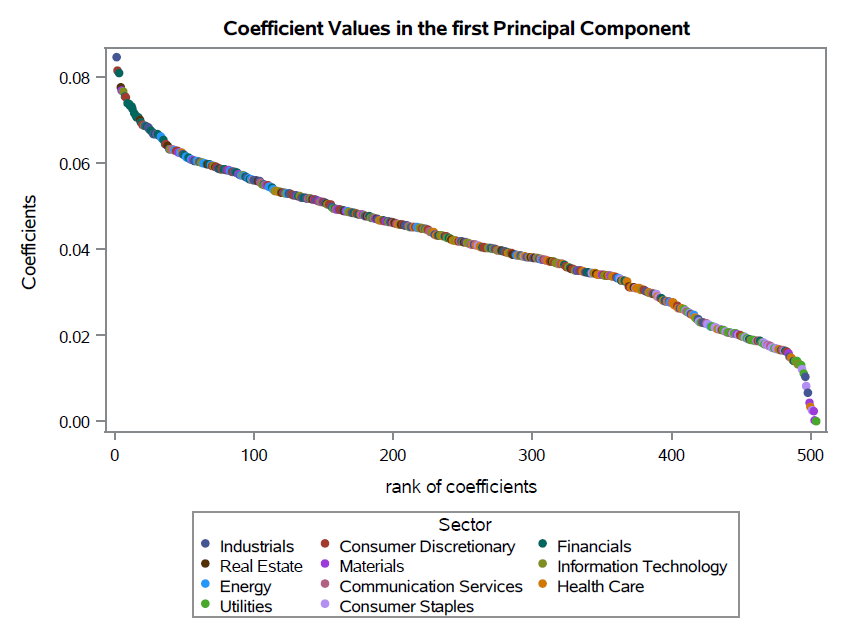


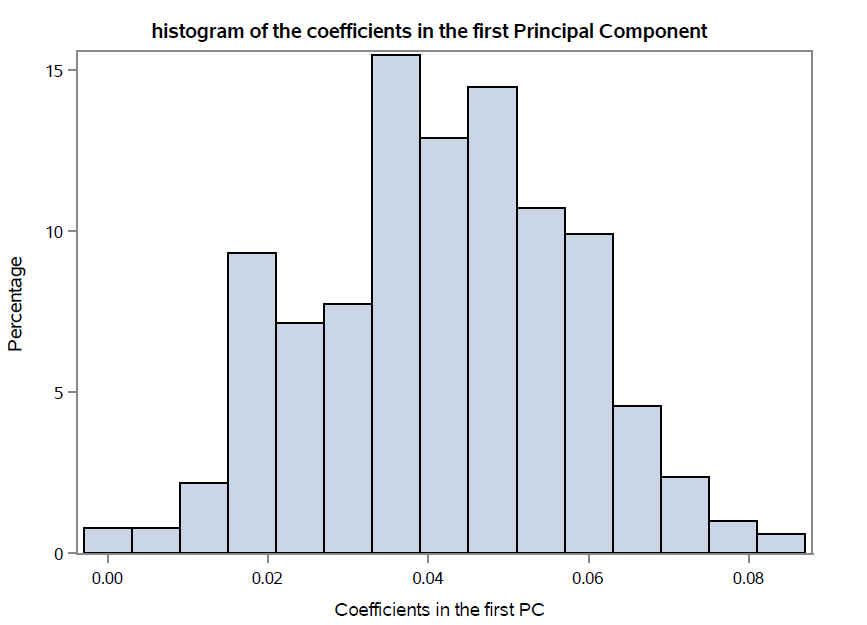


From the distribution of eigenvalue percentage of each stocks, we could notice the eigenvalue of first principal component takes account of more than 35% of summation of all eigenvalues.

Except the first few eigenvalues, the other eigenvalues have similar values close to zero. The first 70 eigenvalues stand for 70% of summation of all eigenvalues, which means 70% variance of the raw return matrix could be explained by the first 70 principal components (ranked by eigenvalues). Thus, PCA show its effective way to concentrate variation and reduce the data dimension.

**First Eigenvalues:**



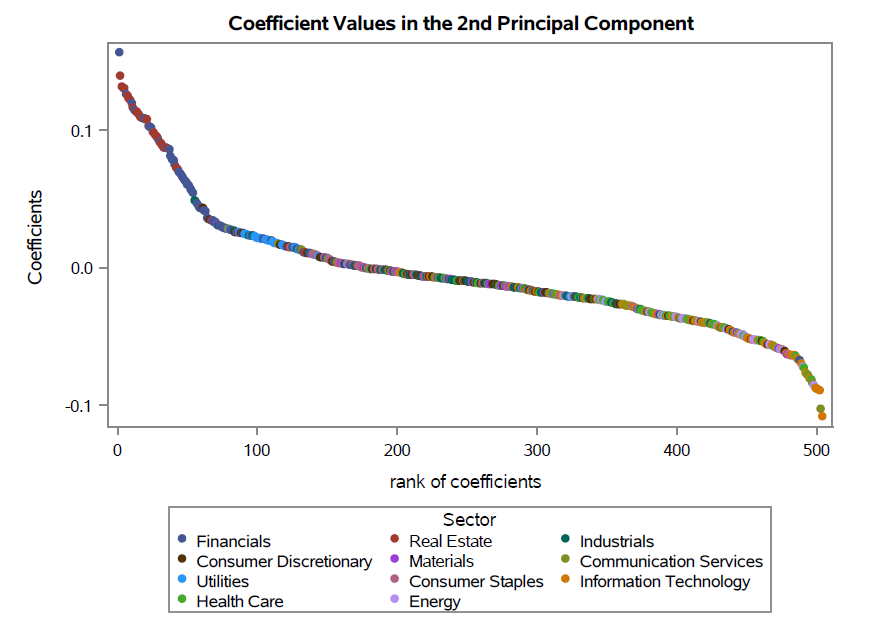


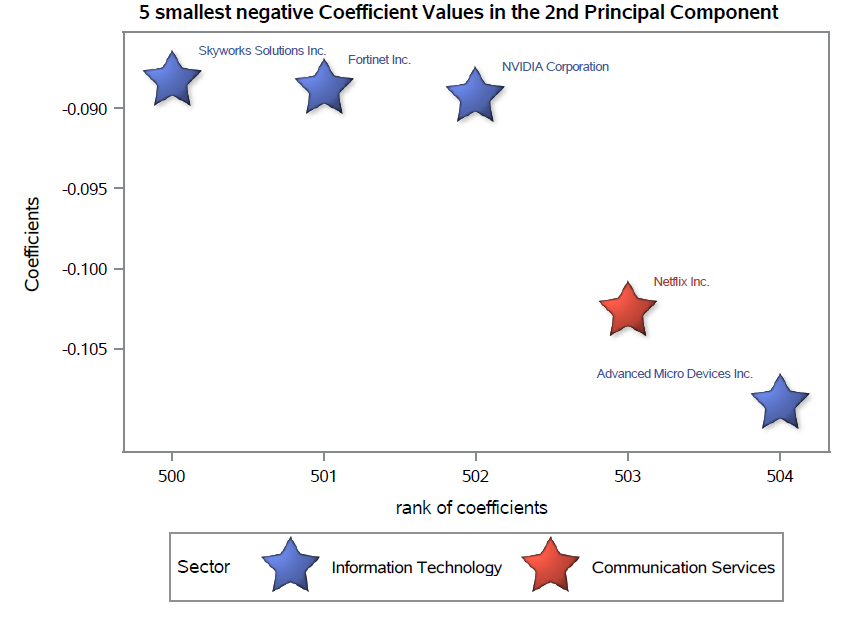
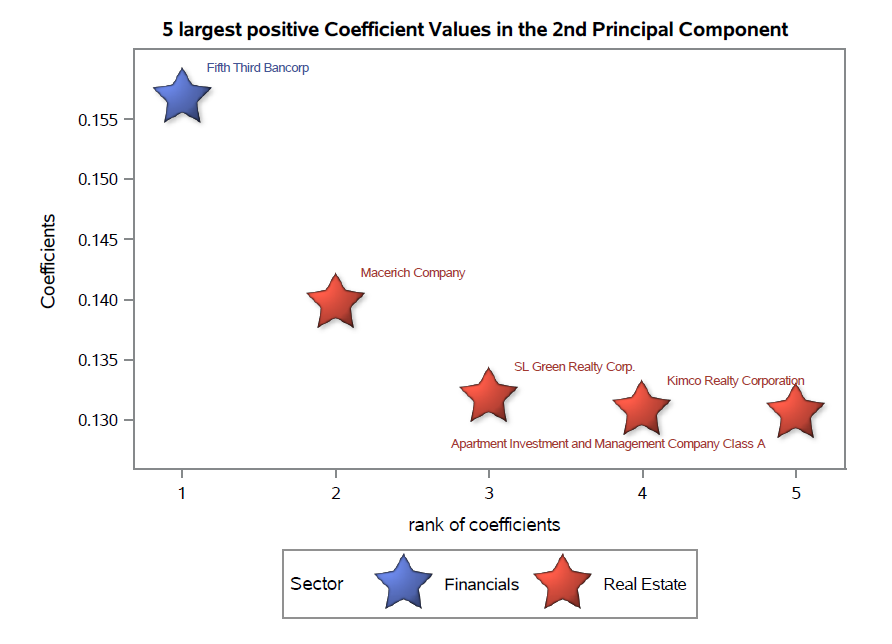
From the distribution of the coefficients in the first eigenvalues, we could notice that all coefficients are larger than zero and the distribution is quite symmetric.

Normally the first principal component explains the most variation, which is understood as the market component[[5]](#footnote-5).

The first eigenvalue is 578.42 and its percentage among summation of all eigenvalues is 35.147, much higher than the other principal components. If we take the square of coefficients as the weight of stocks and construct a stock portfolio, the correlation of the daily return between the portfolio by first eigenvector and S&P500 index is 96.6%, which reflect the conclusion in former literature.

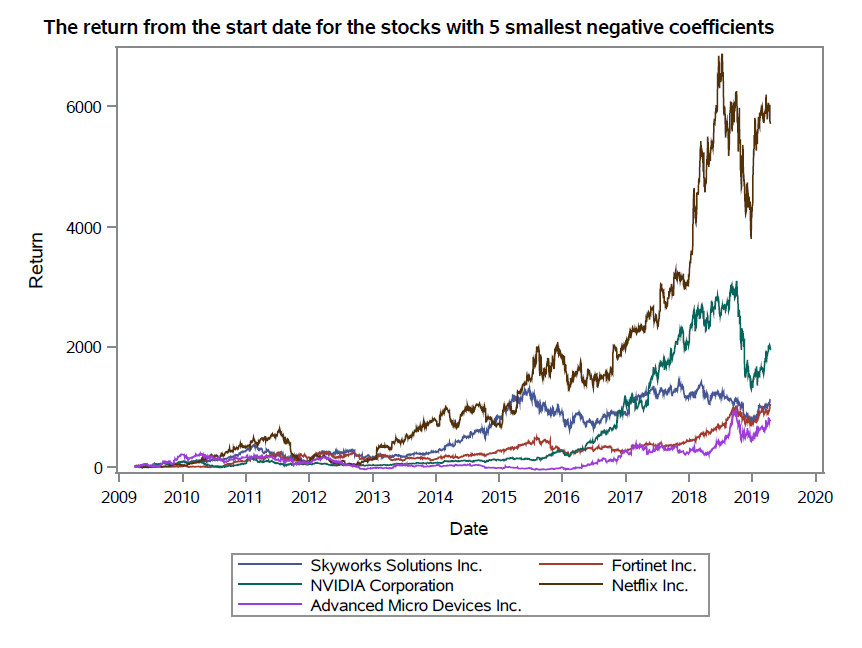
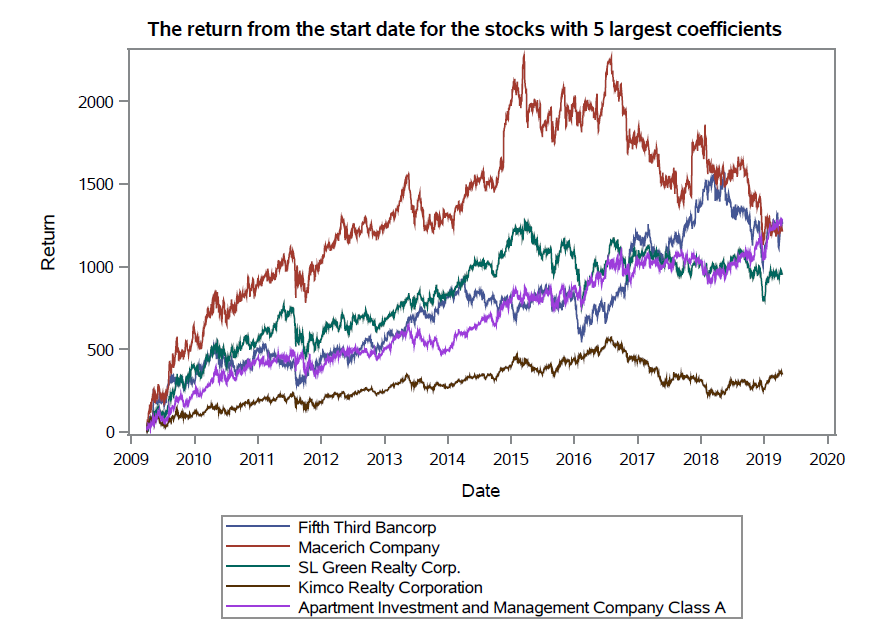
**Second Eigenvalues:**





From the distribution of second eigenvectors, we could observe the clusters of real estate and finance at the upper left corner, the positive coefficient value. On the lower right corner, there is a cluster of information technology and communication Services.

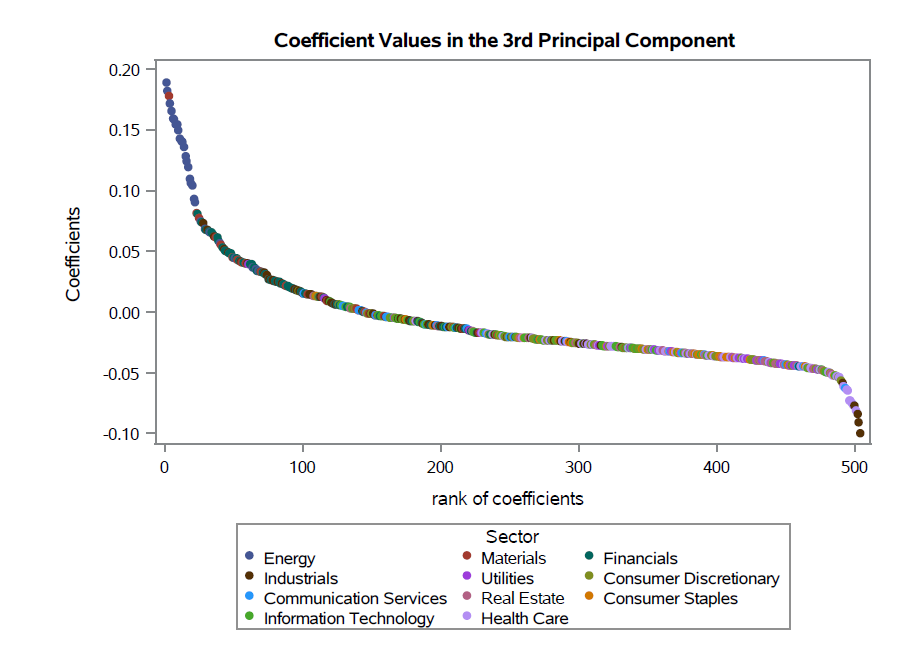
If we just look into the 10 stocks with largest 5 positive coefficients and smallest 5 negative coefficients, the 5 stocks with positive coefficients are one financial company and four real estate companies. On the other side, the 5 stocks with negative coefficients are one information technology company and four communication services companies.

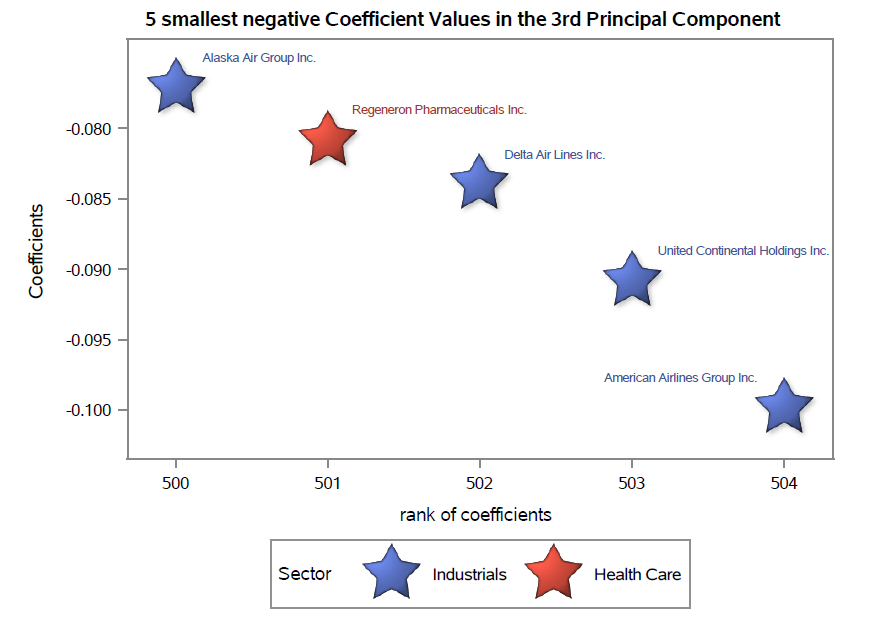
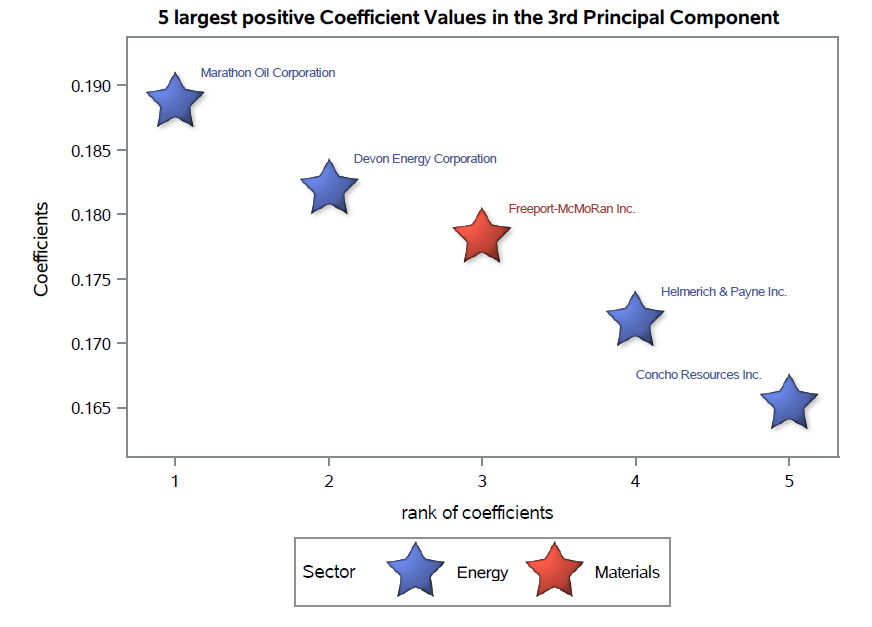


I also compared the return of the 10 stocks among all dates. The stocks with same sign of coefficients have similar patterns of return change. The financial and real estate companies increase quite flat compared to the information technology and communication services companies. The maximum of the left side plot is less than 2000%, but the return rate from the start data for Netflix could achieve 6000%. However, the return fluctuation of the right side is also larger than the left side, which means lack of stabilization and more risks

Thus, the second Principal Component explains the variation caused by high profited information technology and communication services sectors and ‘safer’ but lower profited real estate and finance sectors.

**Third Eigenvalues:**

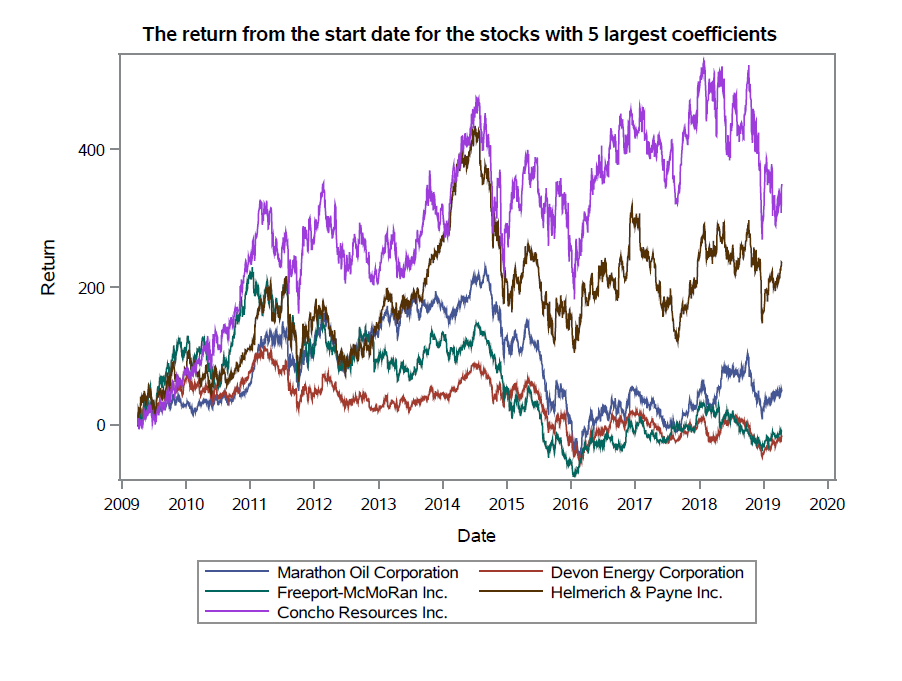


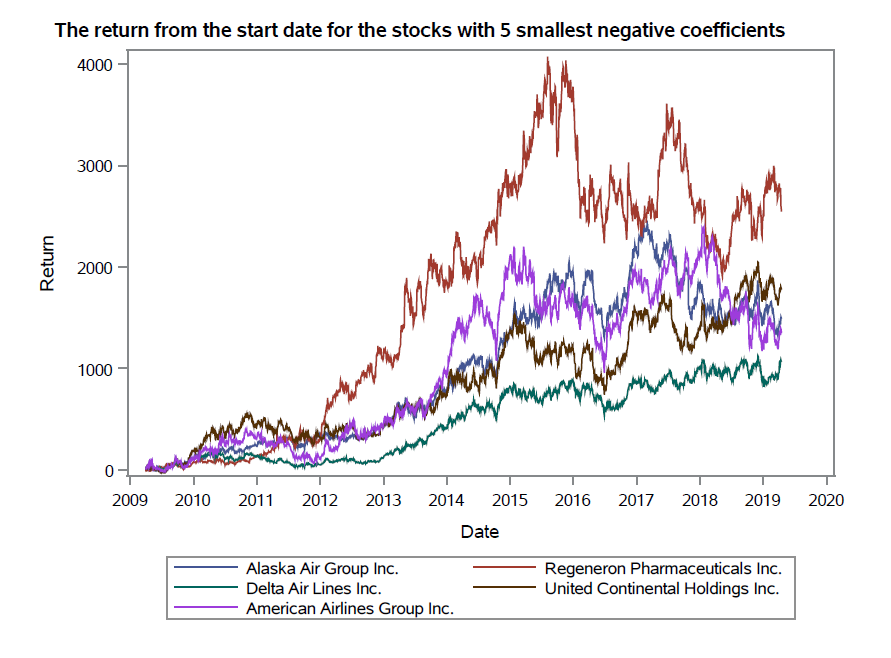


The analysis process for the third Principal Component is similar to the former analysis.

Firstly, we could notice the cluster of the largest few positive coefficients on energy and materials and the smallest few negative coefficients on industrials and health care.

The interesting part is on the right side, among the 5 stocks with smallest negative coefficients contains 4 famous airline companies: Alaska Airline, Delta Airline, United Airline and American Airline.





Then we check the return of the 10 stocks separated by the sign of coefficients. The trends of change with time is similar within plots. But the trends are different between plots, for example, the return increment of energy and materials (ranging from 0-450) is much lower than that of industrials and health care (ranging from 0-4000). The most interesting trend is the opposite changes of price between 2 plots, which means senses because the price of oil and materials plays an important role in the profit of airline companies and pharmaceutical companies. Thus, when you consider buying some stocks in industry sector, it is wise to check the price change of energy and materials ahead. Some stocks portfolios have both investments on industry and energy and materials to decrease the sharp price changes among one of the sectors to reduce the total risk.

The other principal components could be explained in a similar way, so I just contain the former 3 principal components in the project to avoid redundancy.

### Summary:

From the graphs of coefficients within an eigenvector, we found an upper cluster of similarly performing industries and a lower cluster of similarly performing industries.

In the second principal component, the upper cluster is real estate and finance, performing more similarly than to the other sectors. This makes sense, since these companies engage primarily in banking and finances and they are both influenced by the interesting rate and have quite stable profits.

In the lower cluster, information technology and communication services companies behaved similar. Especially after 2016, the return soared to a highest point at mid-2019. The quick return increasement of information technology and communication services companies benefits from the technology development in computers, mobile phones and Internet. We could not leave internet for searching information and social connection. Now the lifestyle of present people is relied on the communication services and electronic products. That’s is why the performance of many stocks in information technology and communication services could be better than average in the stock market in the recent years.

In the third principal component, the lower cluster is industrials and health care, relative to other industries. In the upper cluster is energy and materials sectors, companies perform similarly to those companies because these are mostly related to mining and processing natural resources. Due to the industrials and health care sector are greatly influenced by the price of the raw materials and resources like oil and electricity. The return of the two clusters change in opposite direction.

In conclusion, PCA does seem to facilitate investigating stock similarity and variation.

The analysis of how to cluster stocks is very intuitive. I am particularly encouraged by the fact that stocks in the same industry seem to appear nearby. Stocks that are distant from one another also seem intuitively "dissimilar". Thus, I believe that principal component analysis of stock return is useful, and one could consider using it for stock analysis in order to see through complex interrelationships among stocks.

### Limitation:

1. Due to the time and economic knowledge limitation, I only explained the first 3 Principal Components and 10 stocks among the second of third Components. There could be more rigorous and completed analysis of more principal components.
2. I have no enough time to learn PROC IML, so I do all data preparation and matrix calculation in Python, which adds to the complexity of transferring data from one software to another.

1. The PCA definition in Wikipedia: <https://en.wikipedia.org/wiki/Principal_component_analysis> [↑](#footnote-ref-1)
2. The S&P 500 Index in Wikipedia: <https://en.wikipedia.org/wiki/S%26P_500_Index> [↑](#footnote-ref-2)
3. The website of Yahoo Finance: <https://in.finance.yahoo.com/> [↑](#footnote-ref-3)
4. <https://www.red-gate.com/simple-talk/sql/bi/historical-stock-prices-volumes-python-csv-file/> [↑](#footnote-ref-4)
5. An Application of Principal Component Analysis to Stock Portfolio Management, Libin Yang [↑](#footnote-ref-5)