Analysis of SNP500 Prices Using PCA

Huiqing Wang
Student ID 20405646
Department of Mechanical and Aerospace Engineering

March 13, 2017

1 Introduction to the raw data

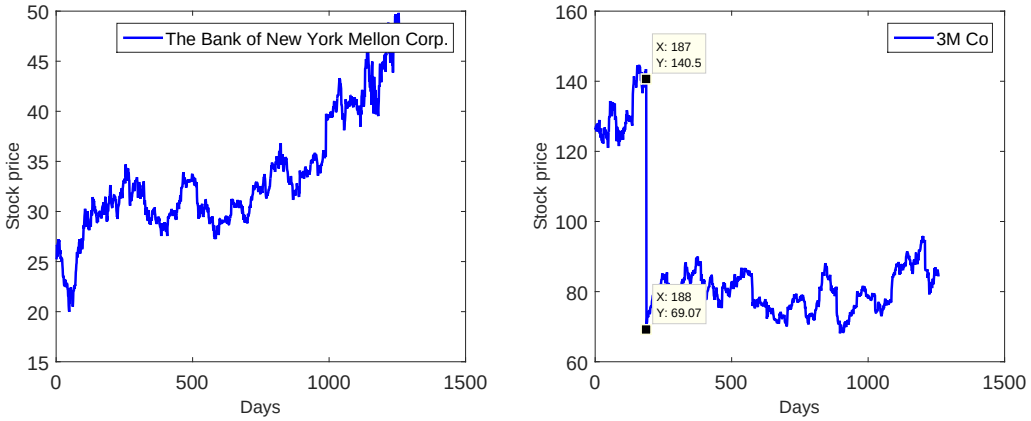
The data used in this assignment are obtained from the course 'A Mathematical Introduction to Data Analysis Spring 2017' [1]. 'Statistics and Machine Learning Toolbox' in MATLAB is used, and the main references of this project assignment are the lecture notes written by Prof Yao [7] and the book written by Friedman *et.al* [2].

The data contains 1258 times 452 matrix with closed prices of 452 stocks in SNP500 for workdays in 4 years, and a 452 cell matrix that tells code, name, and class of the 452 stocks. Based on the given information, these 452 stocks belong to ten sectors as following:

- Industrials
- Financials
- Health Care
- Consumer Discretionary
- Information Technology
- Utilities
- Materials
- Energy
- Telecommunications Services

Time plots of all 452 stocks are made to observe their evolution, and it is found that there are some stocks that has 'weird' behaviours on some days. In Figure 1 Two of these stocks are selected to demonstrate their evolutions over time horizon. Figure 1 shows the time plots of daily price of 'the bank of New York Mellon Corp' in (a) and '3M Co' in (b). It demonstrates that there is a huge gap between prices of day 187 and day 188 for the price of '3M Co', while the prices of 'The bank of New York Mellon Corp' evolves smoothly. After reading some financial

[1] The MATLAB data is obtained from http://math.stanford.edu/ yuany/course/data/snp452-data.mat



(a) The bank of New York Mellon Corp (b) 3M Co.

Figure 1: Time plots of daily stock prices over 1258 days.

literature, it is quite possible that a *corporate action* occured on day 188 for stock '3M Co' [3]. Some examples of a corporate action are *stock splits*, *dividends*, *mergers* and so on [3]. Here some knowledge about stock split and dividend is provided:

- A stock split is an action that the number of shares increases by a specific multiple, and the price is adjusted such that the market capitalization of the company remains the same [6]. One of the most common split ratios is 2-for-1, which means that the stock holder will have double shares after. The price of each share is then adjusted to half of the previous value, and the total market values of stocks held by the stockholder are the same as before split.
- A dividend is to distribute a portion of a company's earning to its shareholders. On the ex-dividend date, the stock price is adjusted downward by the amount of the dividend by the exchange on which the stock trades [5]. Similarly, dividends do not change the total market value of a company' share.

Data of price on the *ex-split date* or *ex-dividend date* should not be directly used for analysis. There is no information about the stock split or dividends in the obtained raw data, it is difficult to judge the reason for the sudden change in the market price of '3M Co' on day 188 simply from the time plot of stock price. However, it would become easier if we study log-return of stocks instead of its price, and to study the return of stocks is a common method in time series analysis [4].

2 Processing the raw data

The log-return R_t of the stock price is defined as following [4]:

$$R_t = \log\left(S_{t+1}/S_t\right), \tag{2.1}$$

where S_t represents the stock price at time t, $t = 1, 2, \ldots, 1257$.

The time plot of log-return for the two stocks are presented in Figure 2. Now it is easy to detect the unusual change in the price of '3M Co' on day 187 from Figure 2(b). Because there is a clear gap between the changes on day 187 and its previous and after days, I think we have good reason to think that an corporate action (most likely, a stock split) was implemented for the stock '3M Co' on day 187.

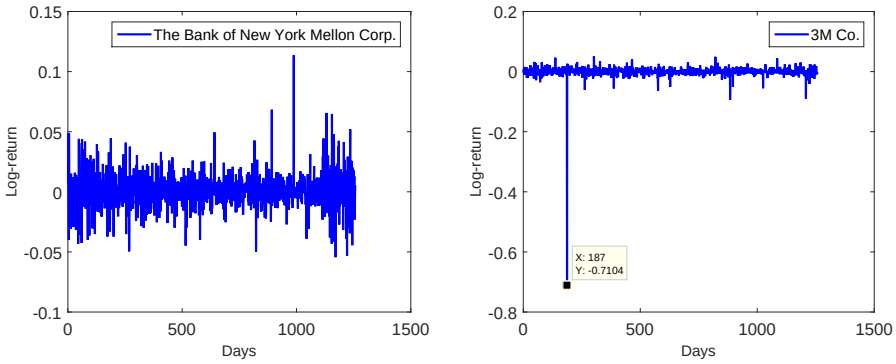


Figure 2: Time plots of daily log-returns over 1257 days: (a) the bank of New York Mellon Corp, and (b) 3M Co.

These direct impacts of corporate actions should not be considered into market movement, and these changes in the stock prices because of corporate actions can be considered as 'outliers'. I define the following criterion for identifying 'outliers', if:

- The absolute value of $(R_j - R_{j-1})$ and the absolute value of $(R_j - R_{j+1})$ are greater than 0.5;
- Or, the absolute value of R_j is greater than 0.4,

then the data R_j will be considered as an outlier.

After the raw data is 'processed', the patterns within the ten market sectors are studied. The correlation plot of log-return within each market sectors are presented in Figure 3. It shows that within sectors of Financials, Utilities, Materials and Energy, log-return of stocks are highly correlated. The highest correlations within these four sectors are respectively 0.8213, 0.7393, 0.7413 and 0.8290. The number of stocks within the four market sectors are presented in Table 2.1.

In the following section, principal component analysis (PCA) is employed to find patterns within these four sectors.

Financials	Utilities	Materials	Engergy
74	32	29	37

Table 2.1: The number of stocks in sectors of Financials, Utilities, Materials and Engery.

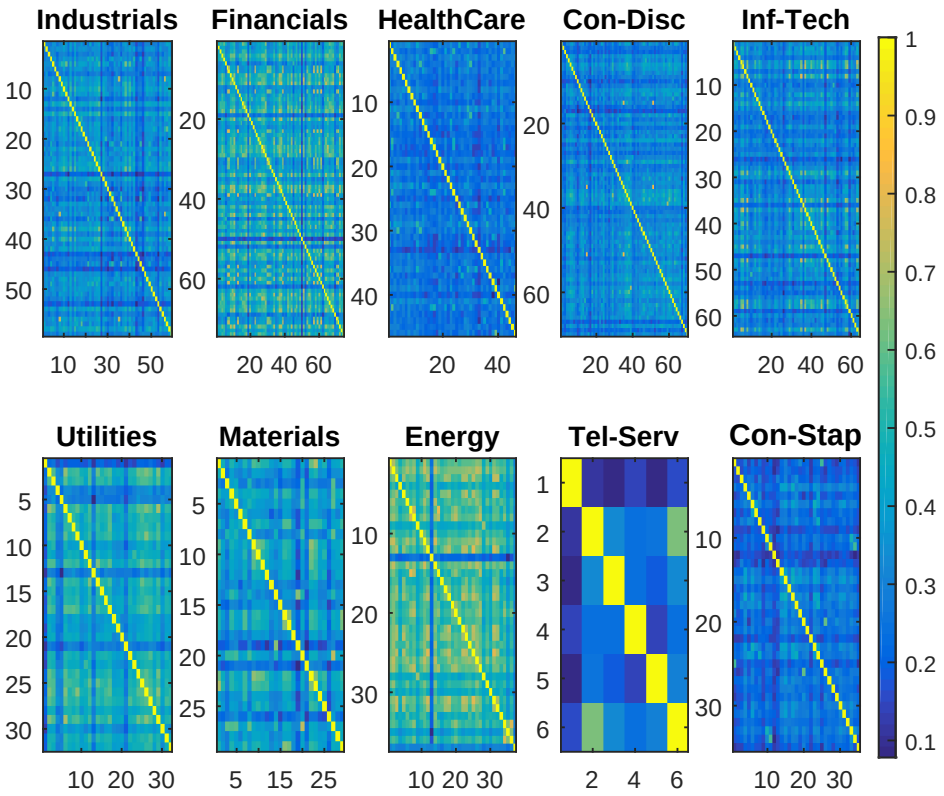


Figure 3: Correlation plots of log-returns in ten market sectors.

3 Principle component analysis

In this section, the analysis on the processed log-return data of Financials sector is presented to demonstrate the procedures of using principal component analysis (PCA).

There are 37 stocks within the Energy sector, and the pairwise correlation between these stocks has been checked in Figure 3: the correlation among some variables can be as high as 0.8290. Each stock can be regarded as an indicator of the Energy sector, and there are 1257 daily observations. Denote the log-return value of the k-th stock on the j-th day by $R_{j,k}$, $j = 1, 2, \ldots, 1257$, $k = 1, 2, \ldots, 37$. Consider a $p-$dimensional variable $X \in \mathbb{R}^p$, where $p = 37$. The j-th observation of the variable X is $X_j = [R_{j,1}, R_{j,2}, \ldots, R_{j,p}]$, $j = 1, 2, \ldots, 1257$ [2].

Firstly a box plot is drawn to check the variability of log-returns of these stocks. Figure 4 shows that some of stocks have much higher variability than others.

Figure 5 demonstrates the percent variance explained by each principal component [7]. It only presents variance percents explained by the first 10 components instead of the total 37. There is a clear break in the variance explained by the first and second component. The first component

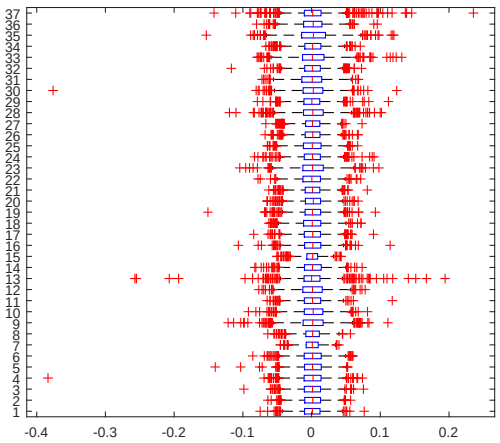


Figure 4: Box plot of the 37 stocks within the Energy sector.

can explain more than 60% percent of the total variance, and the first ten components in total can explain around 80% percent of the variance. It indicates that it is reasonable to choose the first one or two components to reduce the dimension.

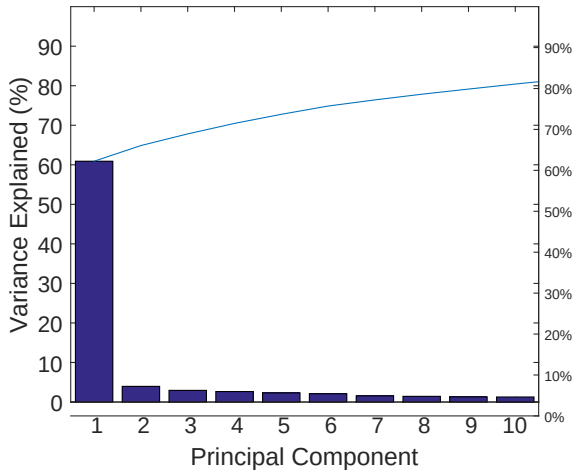


Figure 5: Scree plot of the percent variance explained by each principal component.

Make the scatter plot with the first two components, as presented in Figure 6. Because the data is centered and scaled when employing PCA, we can see that the center of the two components is the zero point. In addition, the scale of the first component is much larger than that of the second, as the first component can explain much more variance than the second one.

In addition, Figure 6 labels those days that are furthest from the center. It implies that the market movement on some days, such as day 1213, day 706 and day 51, is different from other days. when applying Multidimensional Scaling (MDS) using the distance matrix defined by one

minus correlation matrix, we will get the same results as in Figure 6.

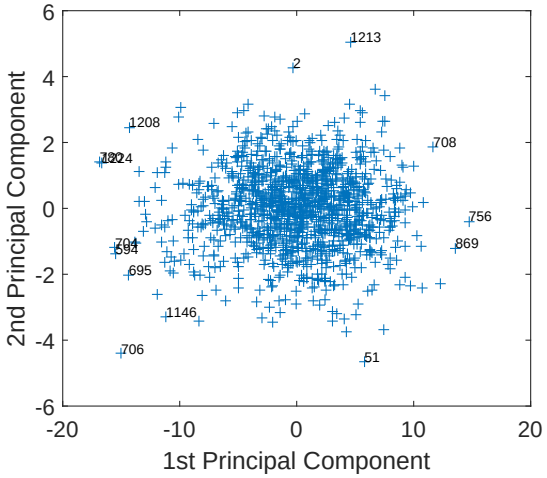


Figure 6: Scatter plots of the first two components.

In a similar way, PCA can be performed for the other three sectors: Financials, Utilities and Materials, in order to check if there is some similar patterns among these four sectors.

An output in pca function of MATLAB is *Hotelling's* T^2, which is a statistical measure of the multivariate distance of each observation from the center of the data set [1]. Using this statistics, one can find out the most extreme 15 points for each section. Figure 7 presents the most extreme 15 days against the corresponding average scaled log-return (over stocks). It shows that there are some common periods that all four sectors suffer or profit from the most extreme days, such as the period between day 157 to day 159,

4 Multidimensional Scaling

In this section, I will try to visualize the distance between stocks among the Energy sector. The pairwise distance between the k_1-th and k_2-th stocks is determined by vector $[R_{1,k_1}, R_{2,k_1}, \ldots, R_{1257,k_1}]$ and vector $[R_{1,k_2}, R_{2,k_2}, \ldots, R_{1257,k_2}]$.
The 'cmdscale' function in matlab is used, and the distance option of 'correlation' and 'Chebychev' respectively for computing distance matrix is employed. Plot of Eigenvalues is shown in Figure 8. It shows that at least 5 eigenvalues are required to restruct the origonal distance matrix in both method. It indicates that the two-dimensional visualization will not be very good.
Figure 9 presents the obtained map of these 37 stocks within the Energy sector. Figure 9 (a) shows that the stocks with code 'CNX' , 'SUN', 'TSO', 'WMB' and 'VLO' are furthest from the others, and in Figure 9 (b), the stocks with code 'RRC', 'TSO', 'WMB', 'EP' and 'COG' are different from others. Using these two way of defining distance matrix, companies with code 'TSO' and 'WMB' both show clear difference from others. From the public information of these stocks, it shows that all these companies operates business with oil or/and petroleum products. It is a bit difficult to find out the financial reason that why these companies stand out of line.

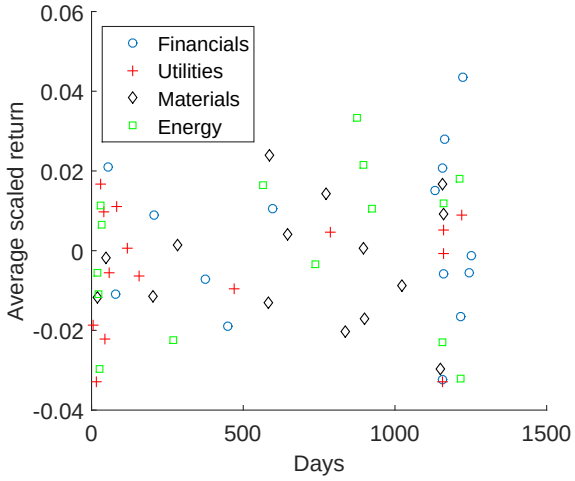
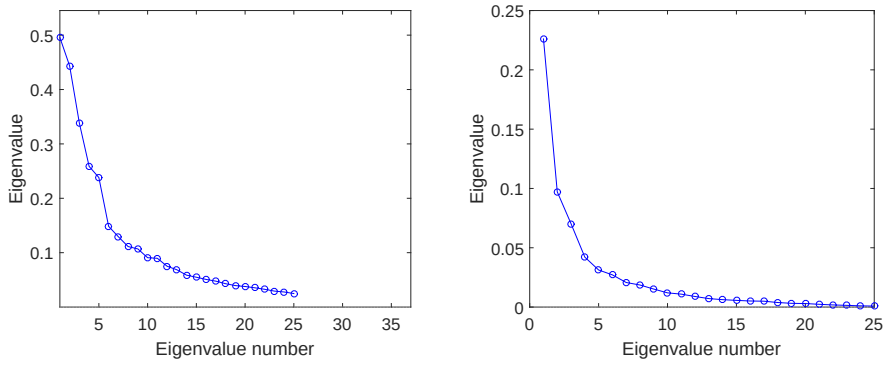


Figure 7: The most extreme 15 days within sectors of Financials, Utilities, Materials and Energy, based on Hotelling's T^2. X axis presents days and y axis presents the corresponding average scaled log-return for those day.



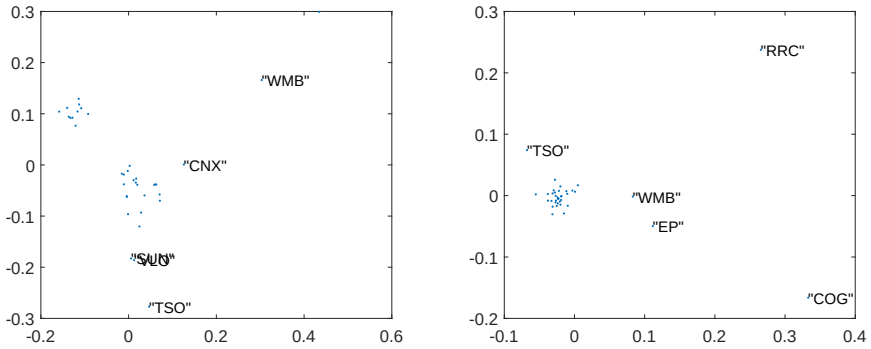
(a) correlation (b) Chebychev

Figure 8: The first 25 eigenvalues: (a) correlation, and (b) Chebychev.

However, as mentioned, the eigenvalue plot implies that using a two-dimensional visualization will not be very good.

5 Conclusion

The technique of PCA and MDS is used to analyze the SNP500 market data over four years. In order to exclude the impacts of corporate actions, the raw data is processed, and the log-returns are chosen to be studied. Details of analysis within the Energy sectors are presented. Using PCA, it shows that only use a one-dimensional principle can explain over 60% variance of the toal 37 stocks, and the common volatile periods can be identified for sectors of Financials, Utilities, Materials and Energy. Another try to give a location map for the 37 stocks within the



(a) correlation (b) Chebychev

Figure 9: Location map of the 37 stocks: (a) correlation, and (b) chebychev

Energy sector using MDS is made.

References

[1] MATLAB documentation: pca. https://nl.mathworks.com/help/stats/pca.html, 2015. [Online; accessed 12-March-2017].

[2] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[3] R. Heakal. What are corporate actions. http://www.investopedia.com/articles/03/081303.asp. [Online; accessed 12-March-2017].

[4] R. S. Tsay. *Analysis of Financial Time Series*, volume 543. John Wiley & Sons, 2005.

[5] U.S. Securities and Exchange Commission. Ex-Dividend Dates: When Are You Entitled to Stock and Cash Dividends. https://www.sec.gov/fast-answers/answers-dividenhtm.html. [Online; accessed 12-March-2017].

[6] U.S. Securities and Exchange Commission. Stock splits. https://www.sec.gov/fast-answers/answersstocksplithtm.html. [Online; accessed 12-March-2017].

[7] Y. Yao. *A Mathematical Introduction to Data Science*.