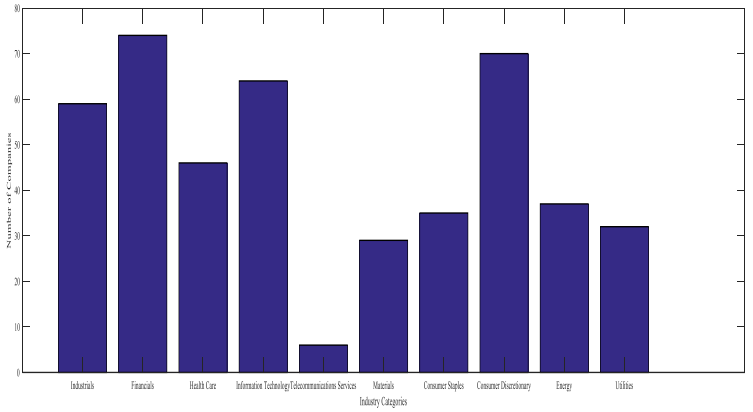
CSIC 5011 Mini-Project 1:Principle Component Analysis on Finance Data

Lui Go Nam[1] gnlui@connect.ust.hk
[1]: Department of Mechanical and Aerospace Engineering, HKUST

1. Introduction

For the first mini project. I want to discover the development trend for different industry based on the SNP500 data. This dataset includes the closed price of work days in four year for 452 different company. To discover inner information from this dataset, it is necessary to rearrange this enormous dataset. All the operations are done in MATLAB.

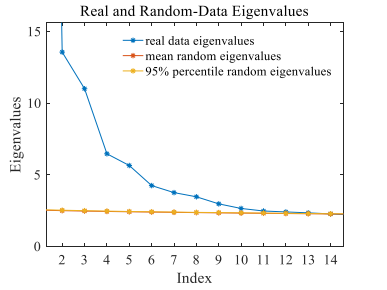
2. Data Rearrangement



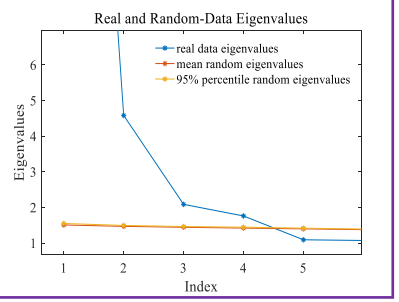
Based on the dataset, the rearrangement is shown above. The figure illustrates that this dataset includes companies from 10 different categories. Which includes Industrials, Financials, Health Care etc. 74 of these companies are from financials area, which constructs the largest group of this dataset. Telecommunication services area has the lowest number of companies which is 6. To induce the stock price trend for different industries, I tend to use principle component analysis. Next step is to discuss the capability of PCA in this case.

3. Parallel Analysis

To determine the number of principle component which cannot be ignored, I construct a parallel analysis for the whole data set and financial group. For the whole dataset 12 principle component should be consider. For the financial group, 4 principle components should be consider to better explain the dataset. Other groups has lower number for explanation.

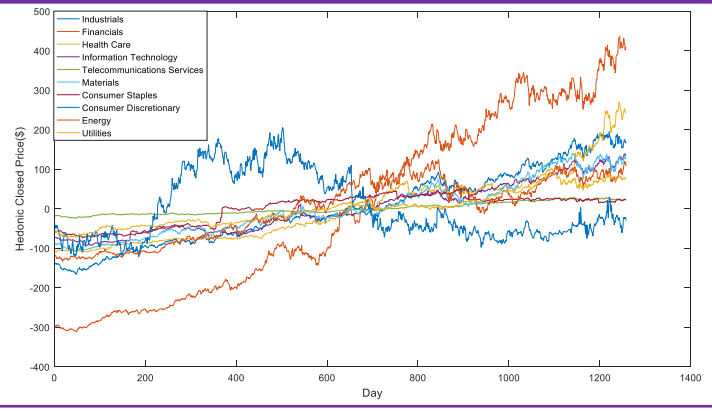


Top 12 Principal components for explained the whole dataset.



Top 4 Principal components for explained the Financial group.

4. Comparison by PCA



5. Analysis

From the parallel analysis for the dataset I know that to use PCA for different groups of data is reasonable in this case, which is shown that the first principle component can largely explain the price trend for different industries. Thus, I construct PCA from each 10 groups of data to get their first principle component, and use it for the inter-industries comparison.

From the time series plot for different industry we know that most of the industries are increasingly developing in this 4 years, and financial industry has the rapidest increasing speed. Other increasing industries includes industrials, health care and materials etc.

However, there are some industries have relatively flat trend and decreasing trend. The telecommunication services and consumer stable have relatively flat trend and the consumer discretionary stock price increases first and then decrease. Interestingly, the consumer discretionary decreases while the financial trend still increases.

6. Conclusion

Using Principle Component Analysis is acceptable for the time series stock price. By applying this technology, use one principle component to represent the whole industry group, to check the trend and investigate the relationship of different industry.

Further investigation can be involving other data analysis algorithm into the analysis process, such as Multi-dimensional scaling. Also, more variables for the companies should be considered, such as the location of the company, to construct a more thorough research.

7. References

Yao, Yuan. “A Mathematical Introduction to Data Science." (2019)