Spectral Clustering of Stock Market Graph

Sina Shahsavari,
Sina Malekian

***Abstract*—This document is about classifying companies based on their stock's price fluctuation by using spectral clustering algorithms; especially graph laplacian. We will use correlation coefficient for constructing similarity matrix and MDS for decreasing dimension.**

***Index Terms*—spectral clustering, graph laplacian, correlation coefficient matrix, Multidimensional scaling.**

I. Introduction

THERE are a myriad of analysis which have been done on the data coming from the exciting world of finance and stock market. One of the most significant purposes in all of them is to minimize the risk. The industry to which a given stock belongs is one of the most crucial factors for determining the risk of investment on the mentioned stock. Since previous investigations show that stocks from one particular industry are highly interdependent, clustering related companies based on their industry and market zone may be an interesting factor which can be influential on determining the risk of investment.

We chose to do this project as it was an effective intersection of the ideas put forth in this class and the ideas in some of our other classes this quarter. Also it seems a highly useful investigation on the stock and finance area which is quite attractive for us.

II. Problem Formulation

Our objective is to evaluate this hypothesis: fluctuation of stocks price of companies is related to their industry and companies in same industry have a highly interdependent stocks price and in result dependent risk of investing.

for this purpose, we will use a stock price data set to classify some famous companies based on their stock price fluctuation. To clarify, we want to group a bunch of companies to check whether companies which are in a same group have highly dependent stock price and risk of investing in one will be affected by the others or not.

Daily price of the stocks of companies is the only input feature we have and we would implement our clustering based on that, also In order to evaluate the classification and decide to accept the hypothesis or reject it, we will use industries data set containing the corresponding industry for each stock as the ground truth.

III. Data description

We used "S&P 500" stock index data set which was downloaded from Yahoo Finance website [3] and includes two tables. One contains the the price of one equity of 473 Companies in 2517 days, the other presents the industries each of these 473 companies are belong to and the total number of industries is 11.

IV. Prepocessing Data

the absolute value of the daily price of equity is not an useful feature since if two or more companies are related to each other we expect to observe this in the variation of their price. Therefore the growth and day to day changes of price are more valuable for clustering; however, the difference of price in two conceding day has a big deficiency which is that for a small companies the absolute value of changes in price are always lower than changes for big companies, hence the ratio of variations will address this issue since with ratio of day to day changes we can compare companies. In the economics literature, the logarithm of this ratio will be used as a describing feature for a company and it is called *Log Return*. Log return distribution is roughly Gaussian like since previous works shew that prices have log normal distribution. From now we will use $r(t)$ as feature and it will be calculated by following equation:

$$r(t) = \log_{10}(\frac{p(t)}{p(t-1)}) \tag{1}$$

where $p(t)$ denotes the price of equity on day t.

Advisor: Prof Orlitsky, ECE Depratment, University of California, San Diego

Before implementing any clustering algorithms, the data should be standardized which lead to values that can be compared easily. this can be done by just subtracting the mean from all points and then dividing them by their std, same as what we learned in this course. However it is better to use a kernel for standardizing instead of using whole data set because with kernels local events will not have global effects.
We have used a window of size 100 for this purpose which seems reasonable.

V. Implementation

We have used Spectral Clustering algorithm which are mainly used in network clustering applications. The paper [1] has reviewed the literature very well, hence we want to implement its graph based method for our classifying.

A. Why spectral Clustering?

Spectral clustering makes no assumption on the shape of clusters, can handle any non-convex shapes like intertwined and spirals etc. It is a non iterative method, while other existing methods like EM and K-means algorithm are iterative and needs many restart to find local minima.

B. Graph Laplacian

In this method at first we need to define a weighted graph $G = (V, E)$ and Each vertex v_i in this graph represents a company. Since the whole procedure of clustering process will be done base of this graph's characteristics, we should construct it for our data at first step. For this goal we need a criterion of similarity between companies in order to identify graph's edges weight. The best candidates can be correlation coefficients between companies' log return. As what we have learned in this course it can express connection between points. Let $X \in \mathbb{R}^{T\times N}$ be the data matrix. Correlation matrix C from N variables can be estimated from T observations by

$$C_{ij} = \frac{\langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle}{\sqrt{(\langle x_i^2\rangle - \langle x_i\rangle^2)(\langle x_j^2\rangle - \langle x_j\rangle^2)}} \tag{2}$$

where x_i and x_j are i_{th} and j_{th} columns of data matrix X, respectively. Therefore, C will be a $N \times N$ symmetric square matrix and we should use absolute value of C elements to weighting our similarity graph since its eigen values will be positive and it is required in further steps. Next, with having adjacency matrix of our similarity graph we form degree matrix which is diagonal matrix with weighted degree of verteces at its diagonal. Then by using algorithms from [1] the graph laplacian will be made and we use its first k smallest eigenvectors to made a new matrix. The rows of this matrix are used to run a K-means algorithm with k clusters in k dimensional space. At last we assign the original points to the corresponding cluster.
The complete algorithms is as follows:

Algorithm 1 Spectral Clustering/Normalized Graph Laplacain

```
1: Input data← Adjacency matrix $W \in \mathbb{R}^{N\times N}$ ,
   Number of clusters k
2: Compute Vertex Degree matrix as:
   $D_{ij} = \begin{cases} \Sigma_{l=1}^{N} W_{ij}, & i = j \\ 0, & i \neq j \end{cases}$   (3)
3: Obtain Unnormalized Laplacian of adjacency matrix as:
   $L = D - W$   (4)
4: Obtain Normalized Laplacian of adjacency matrix as
   $L_{sym} = D^{-1/2} L D^{-1/2}$   (5)
5: Find k smallest eigenvectors of $L_{sym}$.
   Form a matrix $U = \{u_1, u_2, \cdots, u_k\} \in \mathbb{R}^{N\times(k)}$ by stacking
   the k smallest eigenvectors as columns.
6: Form a matrix $T$ from $U$ by renormalizing each of $U'$s
   rows to have unit length as:
   $T_{ij} = \frac{U_{ij}}{(\Sigma_{l=1}^{k} U_{lj}^2)^{1/2}}$   (6)
7: Treating each row of $T$ as a point in $\mathbb{R}^k$, cluster them
   into k cluster via k-means algorithm.
8: Assign the original points $x_i$ to cluster $j$ if row $i$ of the
   matrix $T$ was assigned to cluster $j$.
```

VI. Evaluation

In this part, we want to test the hypothesis that the stocks belonging to the same industry have almost the same risk. In order to that, we will evaluate the clusters obtained through the implementation of the previous part. The evaluation will be done using the ground truth classification which specifies the corresponding industry for each company. Since we have implemented an unsupervised clustering for our data, there will not be error percentage for

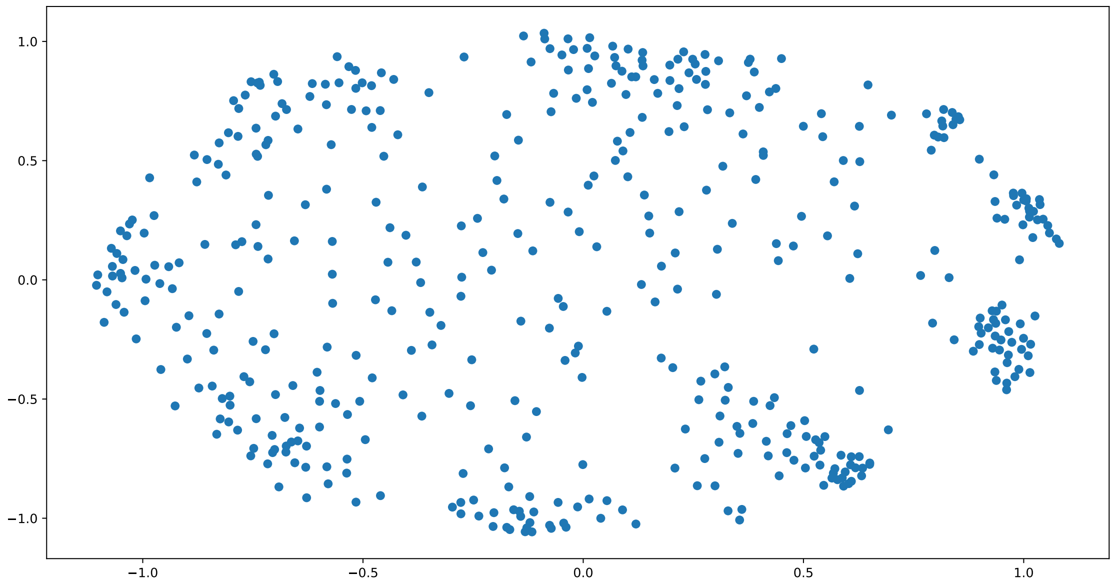

Fig. 1: Data points without market sector information.

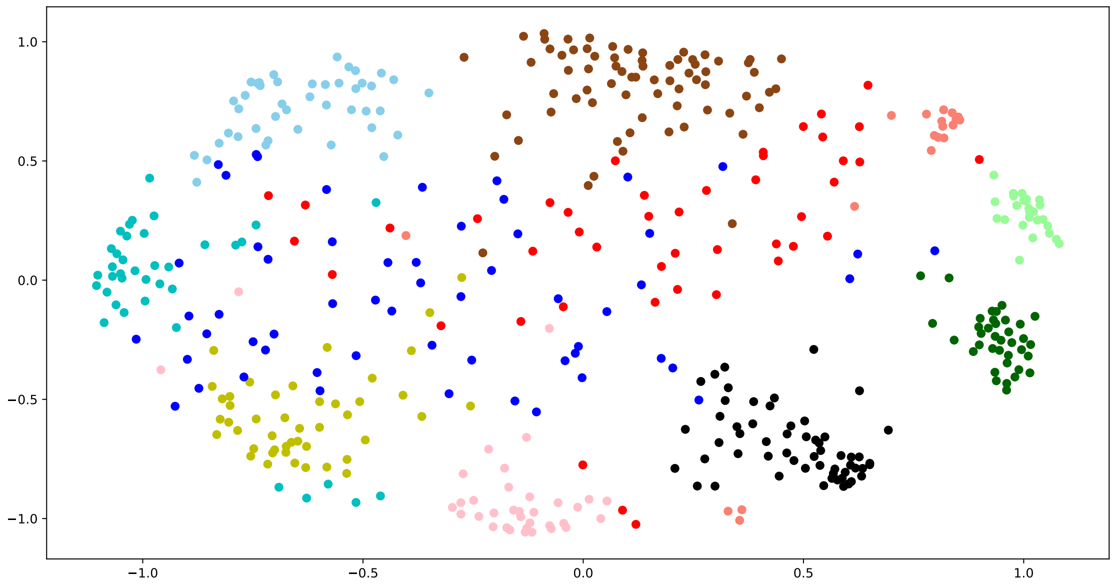

Fig. 2: Data points with each market sector indicated.

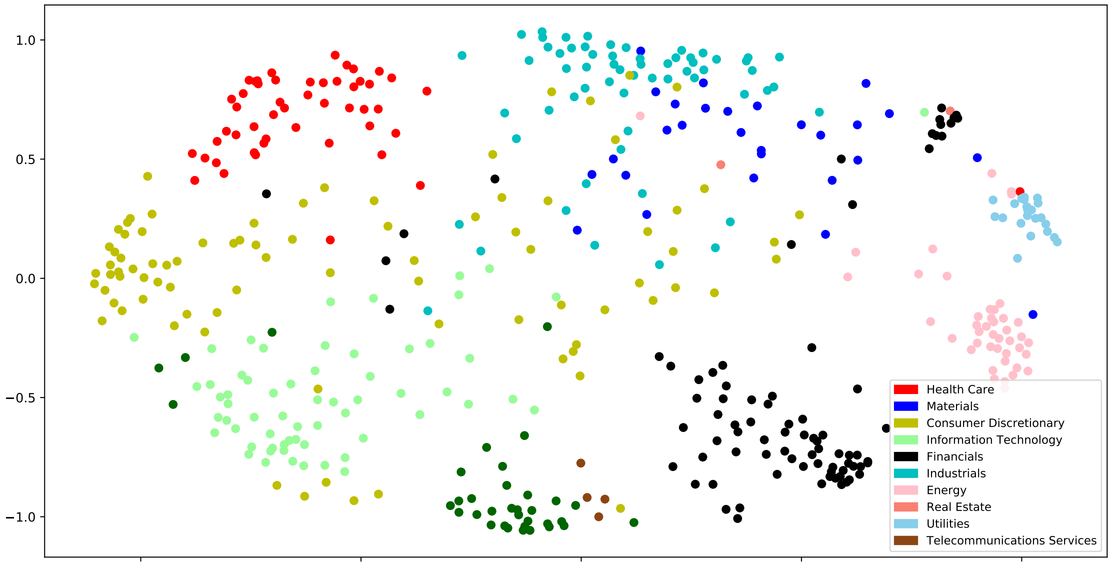

Fig. 3: Data points after clustering.

our procedure and also points are in a 2517-dimensional space and it is definitely impossible to evaluate the result; however, Multidimensional Scaling (MDS) is a simple way lead to a better visualization and make the output of our algorithms comparable with ground truth. It can decrease the dimension of the problem to two which makes comparing so straight. There are more information about how exactly MDS works in [2].

In Fig. 1,Fig. 2 and Fig. 3 we can see Distribution of all 473 data points in two dimensional projection of the T matrix space which has obtained from eigen-vectors of the Normalized T matrix. In Fig. 3 the colors show the real industry which each company belongs to and Fig. 2 it is the output of our spectral clustering algorithms.

By comparing the Fig. 1 and Fig. 2 we can observe that our clustering is generally in accordance with grouping companies by their real industry. This can let us accept our hypothesis which claims the fluctuation in the prices of companies are related to their industry and they are highly interdependent for companies comes from one particular industry. Although we have accepted our hypothesis based on some strong evidence, if we investigate the results more carefully, some deviations from our hypothesis can be observed. There is another hypothesis which should be investigated and can explain these deviations. There are some financial institution such as banks and insurance institutes which invest in different areas of industry so their stocks price variation would be connected to mare than one area and maybe these clustering can also help us find out what that areas are.

VII. Conclusion

To conclude, the main goal was to test the hypothesis that the stocks belonging to the same industry have almost the same price fluctuation and risk of investment. First, we assumed that we can change the risk measurement problem into a classification one. In order to classify the companies, we have extracted some features from their corresponding stock price time series. Next, we tried to obtain a similarity measure between companies based on their correlation coefficients and then cluster them into classes by using graph laplacian method. Then, we evaluated our classification using the ground truth provided by a real data set. The observations made some strong evidence to accept our hypothesis and deviation from this hypothesis lead to a new hypothesis which can be evaluated in future works.

References

[1] U. Von Luxburg, A tutorial on spectral clustering, Statistics and computing, vol. 17, no. 4, pp. 395416, 2007.

[2] Wikipedia page for more explanation about MDS https://en.wikipedia.org/wiki/Multidimensional_scaling

[3] Link for financial data set: https://finance.yahoo.com.

[4] For more information on GEs revenue by segment, visit: http://www.dividend.com/how-to-invest/how-does-general-electric-make-money-ge