ECS 271 Machine Learning Project Report

Taiwan Stock Market Analysis

Name: Shu-Wei Hsu

Student ID: 998569417

Introduction

Regarding to the significant amount of stocks and volatility of daily value change of Taiwan stock market, it is always challenging to do the stock data analysis and derive the pattern of values whether going up or down. It is a conventional closing price chart over different time scales in Figure 1. It is always believed that there are latent forces behind the trend of closing value of stock market. In this project, we selected 50 stocks (showed in Figure 2) in Computer and Related Peripherals sector of Taiwan stock market, and seven years of daily closing prices of each from 2005 to 2011 as training data (overall 1729 prices). Note that all data are normalized to 0 to 1 in the first place. We apply the unsupervised learning technique **Non-negative Matrix Factorization** with sum to one weight matrix H and trend matrix W to realize the clustering of 50 stocks. Note because it is for stocks clustering, the value of w in W is allowed to be negative.

Figure 1. A sample of stock chart for 2382 Quanta Computer Inc.

We utilize dynamic tuned learning rate α starting from 0.005 and grid searches for regularization of H and W, λ and γ, respectively. If the objective function converges, α is tuned to 1.25 times greater; if the objective function diverges, α is tuned to 0.5 less. The clustering performance is evaluated by

Silhouette value and mean intra/extra clustering distance ratio (similar to F-test) both on original training data set and testing data set extracted from daily closing prices of 50 stocks in 2012 (247 prices). Both the lacking data both in training set and testing set due to excluding right or dividend or other reasons are re-filled up with the mean of close prices on the table. We adopt **K-means Clustering** techniques with the same number of groups to compare to our method. Further analysis of closing price tendency in testing data set is measured by mean square error of differences of prices of stocks in the same cluster.

#	No.	Stock Compnay Name	#	No.	Stock Compnay Name
0	230	LITE-ON TECHNOLOGY CORP.	**2**	242	LUNG HWA ELECTRONICS CO., LTD.
1	230	Microtek International, Inc.	**2**	243	ENLIGHT CORPORATION
2	231	MITAC INTERNATIONAL CORP.	**2**	244	JEAN CO.,LTD
3	232	Compal Electronics, Inc.	**2**	246	LEADTEK RESEARCH INC.
4	233	Elitegroup Computer Systems Co.,Ltd.	**2**	247	CATCHER TECHNOLOGY CO., LTD.
5	235	Qisda Corporation	**3**	300	Ahoku Electronic Company
6	235	ACER INCORPORATED	**3**	300	Getac Technology Corporation I
7	235	INVENTEC CORPORATION	**3**	301	Chenming Mold Industrial Corp.
8	235	ASUSTEK COMPUTER INC.	**3**	301	Asia Vital Components Co., Ltd.
9	235	MAG TECHNOLOGY CO., LTD	**3**	302	ICP Electronics Inc.
10	236	MUSTEK SYSTEMS INC.	**3**	304	AOpen Inc.
11	236	CLEVO CO.	**3**	305	Promise Technology, Inc.
12	236	TWINHEAD INTERNATIONAL CORP.	**3**	306	MIN AIK TECHNOLOGY Co., Ltd.
13	236	KYE SYSTEMS CORP.	**3**	323	Wistron Corp.
14	237	GIGABYTE TECHNOLOGY CO.,LTD	**3**	370	Fic Global, Inc.
15	237	MICRO-STAR INTERNATIONAL CO.,LTD.	**4**	493	Pegatron Corp.
16	238	AVISION INC. I	**4**	611	In Win Development Inc.
17	238	QUANTA COMPUTER INC.	**4**	612	General Plastic Industrial Co., Ltd.
18	238	CHICONY ELECTRONICS CO. LTD.	**4**	616	Adlink Technology, Inc.
19	238	SUNREX TECHNOLOGY CORPORATION	**4**	617	Billionton Systems Inc.
20	239	ADVANTECH Co., Ltd. I	**4**	620	Flytech Technology Co., Ltd.
21	239	DFI Inc.	**4**	623	Waffer Technology Corp.
22	239	BIOSTAR MICROTECH INTERNATIONAL	**4**	627	Aten International Co., Ltd.
23	240	SHUTTLE INC. I	**4**	800	Lite-On It Corporation
24	241	AVerMedia Technologies, Inc.	**4**	991	Associated Industries China, Inc.

Figure 2. Selected 50 stocks in Computer and Related Peripherals sector in Taiwan stock market

Related Works

Wang et al. [1] has introduced using matrix factorization to do stock trend extraction. Afterwards, they utilized ward linkage algorithm and centroid algorithm to plot the dendrograms. They used Frobenius norm, J-divergence and Silhouette value to evaluate the results of clustering. Liu et al. [2] applied Non-Negative Matrix Factorization for stock market pricing. They used k-mean algorithm to cluster with various of k values. Gemulla et al. [3] Utilized stochastic gradient descent instead of batch gradient descent for better efficiency of large-scale matrix factorization.

Data Analysis

stkid	stknm	scid	scnm	buypic	sellpic	buymoney	sellmoney	buyrow	sellrow
2382	廣達	1020	合　庫	29250	25000	1714500	1459500	19	13
2382	廣達	1030	土　銀	21000	2000	1229500	117200	15	1
2382	廣達	1040	臺　銀	43000	4000	2526900	234300	29	4
2382	廣達	1090	工　銀	1000	5000	58100	292500	1	5
2382	廣達	1110	企　銀	23000	2000	1349100	118600	19	2
2382	廣達	1160	日　盛	122850	56000	7221900	3293900	92	40
2382	廣達	1230	彰　銀	10000	0	582000	0	1	0
2382	廣達	1260	宏　遠	36000	7000	2111400	410200	19	4
2382	廣達	1360	港麥格理	1071000	729000	62618300	43062000	422	348
2382	廣達	1380	港商里昂	217000	226000	12586000	13256900	32	114
2382	廣達	1440	美林	141000	159000	8349100	9316100	64	76
2382	廣達	1470	台灣摩根	59828	573000	3505024	33835400	41	296
2382	廣達	1480	美商高盛	30000	597000	1741500	34902800	3	250

Figure 3. Trading record of 2382 Quanta Computer (廣達 in Chinese)

A snippet of a sample dataset of the daily trading records regarding to a single security with respective brokers in Taiwan stock market across ten years through 2001 to 2011 is as shown below (Figure 3). The daily trading records of each security with respect to each broker are labeled with the captions:

- Stkid: stock id
- Stknm: stock name
- Scid: broker id
- Scnm: broker name
- Buypic: the volume for purchasing the stock
- Sellpic: the volume for selling the stock
- Buymoney: the total amount of money for purchasing the stock
- Sellmoney: the total amount of money for selling the stock
- Buyrow: remarks for purchasing the stock
- Sellrow: remarks for selling the stock

Here the Buyrow and Sellrow indicate the volume of transactions of purchasing and selling via the broker, respectively. Furthermore, the instantaneous value of purchasing is defined by the total amount of money for purchasing the stock divided by the volume for purchasing the stock.

Taking the a specifically daily report from Taiwan Stock Exchange Corp. (TWSE) in Figure 4 of Daily Trading Value/Volume of Security 2382 on March, 2011 for instance: the sum of buypic (the volume for purchasing the stock) is equal to the trade volume; the sum of each buymoney (the total amount of money for purchasing the stock) is equal to the trade value; and the sum of buyrow (transaction) is equal to the transaction.

2011/03 2382								(NT$,share)
Date	Trade Volume	Trade Value	Opening Price	Highest Price	Lowest Price	Closing Price	Change	Transaction
2011/03/01	7,675,135	460,905,853	58.90	61.00	58.80	60.80	+2.70	3,442
2011/03/02	7,294,495	428,098,110	60.00	60.00	58.00	58.00	-2.80	3,166
2011/03/03	8,818,906	524,300,260	59.30	60.20	58.50	60.10	+2.10	3,558
2011/03/04	7,054,955	434,166,238	60.80	62.70	60.50	61.00	+0.90	3,194
2011/03/07	1,840,954	111,440,406	60.60	61.00	60.10	60.30	-0.70	1,004
2011/03/08	3,588,343	214,940,436	59.80	60.30	59.50	60.00	-0.30	1,777
2011/03/09	4,715,852	277,400,013	60.00	60.00	58.30	58.60	-1.40	2,046
2011/03/10	6,646,912	380,513,484	58.00	58.00	56.90	57.00	-1.60	2,483
2011/03/11	12,847,882	710,588,751	56.00	56.30	54.70	54.70	-2.30	4,681
2011/03/14	6,739,771	359,363,314	53.10	54.10	52.70	53.60	-1.10	2,892
2011/03/15	11,629,913	599,450,769	52.80	54.00	49.90	51.90	-1.70	4,092
2011/03/16	7,201,164	381,686,051	53.50	53.50	52.20	52.70	+0.80	2,905
2011/03/17	6,232,935	322,165,768	50.50	52.90	50.50	52.70	0.00	2,590
2011/03/18	6,076,568	321,562,406	52.50	53.70	52.10	53.20	+0.50	2,359
2011/03/21	3,205,747	171,822,227	53.50	54.20	52.90	53.60	+0.40	1,565
2011/03/22	3,525,317	187,062,255	53.80	53.90	52.80	52.90	-0.70	1,733
2011/03/23	4,376,826	231,021,378	53.60	53.60	52.20	53.00	+0.10	2,154
2011/03/24	7,687,606	406,361,418	53.20	53.80	52.50	52.60	-0.40	3,169
2011/03/25	8,970,887	485,829,166	53.10	55.20	53.00	54.30	+1.70	3,763
2011/03/28	7,142,772	379,919,144	54.30	54.30	52.30	53.70	-0.60	2,968
2011/03/29	5,581,551	300,819,056	52.80	54.70	52.70	54.30	+0.60	2,537
2011/03/30	11,707,099	659,965,769	54.30	57.60	54.30	56.40	+2.10	5,103
2011/03/31	10,771,397	595,376,778	56.40	56.40	54.90	55.50	-0.90	2,925

Figure 4. Daily Trading Value/Volume of Security 2382 on March, 2011

Implementation

It is widely believed that the stock market value and volume are determined by some underlying factors. Matrix factorization may be used to extract underlying trends. It has been generally realized that stocks of the same sector/industry may not have the same sensitivity to the changes inside or outside the market. Nonetheless, matrix factorization may also further group stock portfolios into different clusters based on patterns from the history records. Recently, Non-negative matrix factorization (NMF), under the name of positive matrix factorization (PMF), is commonly applied to predict the individual stock trend due to being positive of all the coefficients.

The obstacle in modeling stock market and forecasting lies on being unable to discover the true data generating process and understand the underlying dynamics and reconstruction of stock returns. The underlying factors can be called the latent bases. Assume there are *n* stocks with *m* price records each in a time intervals. Note the prices of every individual stock should be normalized in the first place. There are *K* latent bases, *W1, W2, . . . , WK*, each is a row vector of size m. Each stock *Si* can be expressed by an aggregation of these underlying factors with coefficients:

$$S1 = h11W1 + h12W2 + ... + h1kWk + ... + h1KWK + N1$$

$$S2 = h21W1 + h22W2 + ... + h2kWk + ... + h2KWK + N2$$

...

$$Si = \cdot hi1W1 + hi2W2 + ... + hikWk + ... + hiKWK + Ni$$

...

$$Sn = \cdot hn1W1 + hn2W2 + ... + hnkWk + ... + hnKWK + Nn$$

h_{ik} is a nonnegative real number and indicates to which degree the stock i is associated with the basis W_k. N_i is an observation noise vector and can be seen as the approximation error when sum up all the product of h_{ik} and W_k. Therefore, The general formula for the i^{th} stock S_i is:

$$S_i = H_i W_i + N_i$$

Where H_i is a row vector of all h_{ik}, while W_i is a column vector of all W_k. This is, a vector of all stock S_i can be expressed as:

$$S = HW + N$$

Where $H \in R^{n \times K}$, $W \in R^{K \times m}$ and $N \in R^{n \times m}$. H is the weight matrix and W is trend matrix for a given S and K for this typical Non-negative Matrix Factorization if enforcing the non-negative constraint in trend matrix W. H is used to do the partition of stocks into K clusters later. We enforce a coefficient to regularize the objective function to make more smoothly in the presence of noise in the dataset by the Frobenius norm:

$$J1(W) = \|W\| = tr(WW^T)$$

Note that the initial values of each entry in W are set from 0 to 1. In the same fashion, the regularization by H can be written as a penalty term in the form:

$$J2(H) = \|H\| = tr(H^T H)$$

In the weight matrix H, the rows are used to determine weights of the components in each observed object. Therefore, we make a sum to one H. that is, we force the initial values in each row in H to 1. For disclosing the contributions of all of the components in W, we enforce that all of the elements in each row of H sum to one. Each row is as projection vector as basis. The objective/cost/loss function can be expressed as:

$$\text{argmin } J = \alpha \| H W - S\|^2 + \gamma\|W\| + \lambda\|H\|$$

Note for finding the minimum, α can be set to 1, then new $\gamma = \gamma/\alpha$ and new $\lambda = \lambda/\alpha$. To update the value of W and H to search the optimal, we take the partial derivatives of objective function with respect to H and W:

$$Dw = 2H^T HW - 2H^T S + 2\gamma W$$

$$D_H = 2HWW^T - 2SW^T + 2\lambda H$$

We update the value of W and H with batch gradient descent for better accuracy. Therefore, the update of w and h can be rewrote as:

$$w_t = w_{t-1} - \alpha (2H^T HW - 2H^T S + 2\gamma W)$$

$$h_t = h_{t-1} - \alpha (2HWW^T - 2SW^T + 2\lambda H)$$

The learning rate α is dynamic tuned starting from 0.005. We adopt grid searches for regularization of H and W, λ and γ, respectively. There are two convergence terms: 10000 iterations for grid search and 100000 iterations for clustering (fail convergence is not included); and when learning rate decreases to less than 1e-10. The convergence is checked by Root Mean Square Error (RMSE). RMSE is defined by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Where here the value in the parenthesis is defined by:

$$H_i W_i - S_i$$

We choose 50 stocks in computer and related peripherals sector in Taiwan stock market, and seven-year daily closing price from 2005 to 2011 to train the W and H by minimizing objective function (1729 prices). Furthermore, The stocks can be separated into k clusters of by selecting greatest weight value in H, or finding similarity from stock and w in W. Both are with the number of clusters limited by k latent dimension. That is, rows in trend matrix W are as centroids of clusters. We also utilize k-mean algorithm with individual vector of stock in original data set to compare with our algorithm. The clustering performance is evaluated by Silhouette value. It provides a succinct graphical representation of how well each data point lies within its cluster:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Considering being too strict of Silhouette value, we apply more straightforward evaluation technique for clustering: The mean intra/extra cluster distances ratio (similar to F-test), which is defined by:

Mean of each node's (mean intra-cluster distance / mean extra-cluster distance)

Where mean intra-cluster distance is:

The sum of distance of the node to the nodes in same cluster / (#nodes in that cluster-1)

And mean extra-cluster distance is:

The sum of distance of the node to the nodes in other clusters / (#nodes in other clusters)

Where the cluster with one or zero stock company in it is ignored. Afterwards, we iteratively use different constraint γ and λ to objective function, and see the in-cluster and inter-cluster distance discrepancy by Silhouette value and mean intra/extra cluster distance ratio. After finding the least values by a pair of γ and λ, we change the k value of k-mean clustering algorithm to see how the trend

matrix W can represent the whole data. Therefore, with analyzing 50 stocks' closing price and separating the stocks into clusters through machine learning techniques, we are able to utilize the intra cluster similar volatility to come up with a trend of stocks take it as a reference of diversification analysis for reducing investment losses in the stock market.

Furthermore, the clustering performance is also evaluated by Silhouette value and mean intra/extra clustering distance ratio on testing data set extracted from daily closing prices of 50 stocks in 2012 (247 prices). We also have done further analysis of closing price tendency in testing data set, which is measured by mean square error of differences of prices of each stock and other stocks in the same cluster.

Experimental Results

I. Grid Searches

Firstly, we start from the grid search of regularization of H, λ, and regularization of W, γ, and dimension 5 to 15 to find the convergence evaluated by Root Mean Square Error (RMSE). The reason we choose the range from 5 to 15 is: we do not expect the number of stock in one cluster to be too large, so the dimension is started from 5. On the other hand, it is meaningless if there are too many clusters for only 50 stocks. The result of RMSE of grid search is showed in Figure 5 and illustrated in visualization in Figure 10.

Dimension	RMSE	λ	γ
5	0.0783336886354	1e-06	0.001
6	0.069057314563	1e-05	0.1
7	0.0623982128914	1e-06	0.1
8	0.055962750299	0.0001	0.1
9	0.0518426978342	1e-06	0.1
10	0.0487466634979	1e-08	0.1
11	0.0464268869015	0.01	0.1
12	0.0481726646346	0.01	0.1
13	0.0502638228456	1e-09	0.1
14	0.0524835004697	0.1	0.1
15	0.0541417405491	1e-07	0.1

Figure 5. Best RMSE of objective function, with grid search of different latent dimensions and relative λ and γ

We also applied the Silhouette value and mean intra/extra cluster distance ratio to the grid search. The result of Silhouette values of the clustering by selecting greatest weight value in H (red line in Figure 11) and finding similarity from stock and w in W (blue line in Figure 11) are showed in Figure 6 and Figure 7, respectively, and illustrated in Figure 11. Note that for Silhouette value, the larger the value is, the better clustering performance is (if value > 0).

Dimension	**Silhouette Value**	**λ**	**γ**
5	0.174247272329	1e-05	0.01
6	0.12557974184	1e-07	γ
7	0.122336190186	1e-05	1e-06
8	0.12139250133	0.001	1e-10
9	0.115997623381	0.1	1e-10
10	0.0910716234585	1e-10	1e-05
11	0.128161035184	0.01	0.001
12	0.151395212107	1e-05	1e-09
13	0.108715525285	1e-07	1e-09
14	0.127918798173	0.001	1e-06
15	0.11765579667	1e-10	1e-09

Figure 6. Best Silhouette Value of stock clustering by selecting greatest weight value in H, with grid search of different latent dimensions and relative λ and γ

Dimension	**Silhouette Value**	**λ**	**γ**
5	0.174247272329	1e-05	0.01
6	0.120170054227	0.0001	0.1
7	0.0955534837311	0.0001	1e-08
8	0.126778266036	0.001	1e-10
9	0.0932862905188	1e-09	1e-10
10	0.105023104207	0.001	0.001
11	0.0721757769209	0.001	1e-10
12	0.0857011342293	0.1	1e-08
13	0.0668295247753	0.001	0.01
14	0.0627526059293	0.1	1e-10
15	0.0547920815513	1e-10	1e-09

Figure 7. Best Silhouette Value of stock clustering by finding similarity from stock and w in W, with grid search of different latent dimensions and relative λ and γ

Furthermore, the result of mean Intra/Extra cluster distance ratio of the clustering by selecting greatest weight value in H (red line in Figure 12) and finding similarity from stock and w in W (blue line in Figure 12) are showed in Figure 8 and Figure 9, respectively, and illustrated in Figure 11. Note that mean Intra/Extra cluster distance ratio, the smaller the value is, the better clustering performance is (when value < 1).

Dimension	Mean I/E C.D.R.	λ	γ
5	0.701448454805	1e-05	0.01
6	0.691626194631	0.0001	1e-09
7	0.693971429274	1e-05	1e-06
8	0.679486126046	1e-10	0.0001
9	0.674793403127	0.1	1e-05
10	0.669704634754	1e-10	1e-06
11	0.672980113989	0.01	0.001
12	0.656463310231	0.1	1e-06
13	0.664315643932	1e-08	1e-09
14	0.655756508847	0.01	1e-09
15	0.667479694806	1e-08	1e-10

Figure 8. Best Mean intra/extra cluster distance ratio of stock clustering by selecting greatest weight value in H, with grid search of different latent dimensions and relative λ and γ

Dimension	Mean I/E C.D.R.	λ	γ
5	0.69605355463	1e-08	0.1
6	0.712093431908	1e-09	1e-05
7	0.726915333074	0.0001	1e-08
8	0.695275349739	1e-05	1e-09
9	0.686989235356	1e-07	0.0001
10	0.698965493142	1e-10	1e-05
11	0.727958528746	1e-05	1e-07
12	0.694858605369	0.1	1e-08
13	0.679120082156	0.001	0.01
14	0.683322210442	0.1	1e-10
15	0.706481870386	1e-07	1e-10

Figure 9. Best Mean intra/extra cluster distance ratio of stock clustering by finding similarity from stock and w in W, with grid search of different latent dimensions and relative λ and γ

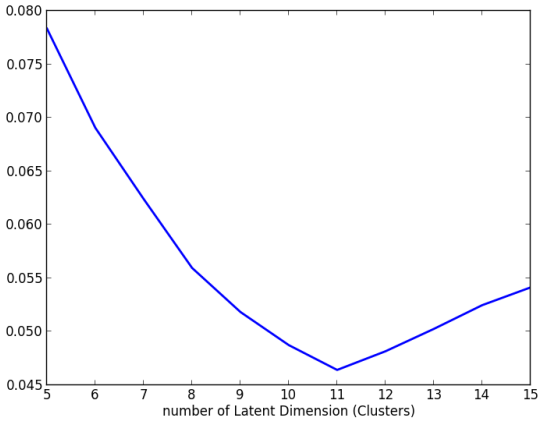

Figure 10. Best RMSE of objective function, with grid search of different latent dimensions

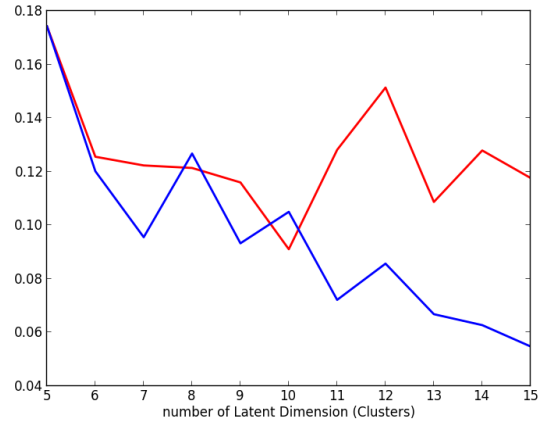

Figure 11. Best Silhouette Value of stock clustering by selecting greatest weight value in H (red) and finding similarity from stock and w in W (blue), with grid search of different latent dimensions

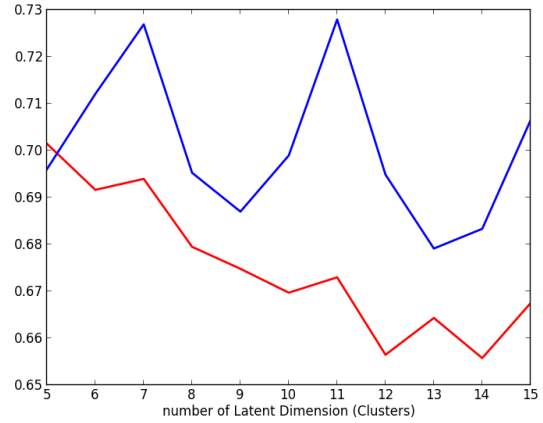

Figure 12. Best Mean intra/extra cluster distance ratio of stock clustering by selecting greatest weight value in H (red) and finding similarity from stock and w in W (blue), with grid search of different latent dimensions

Furthermore, actually we did all the grid search of latent dimension 5 to 15, regularization of H, λ, 0.1 to 1e-10, and regularization of W, γ, 0.1 to 1e-10, respectively. Here we only show the grid search of λ and γ with the latent dimension 5, 10, 15 and relative RMSE, Silhouette value, and mean intra/extra cluster distance ratio from Figure 13 to Figure 21.

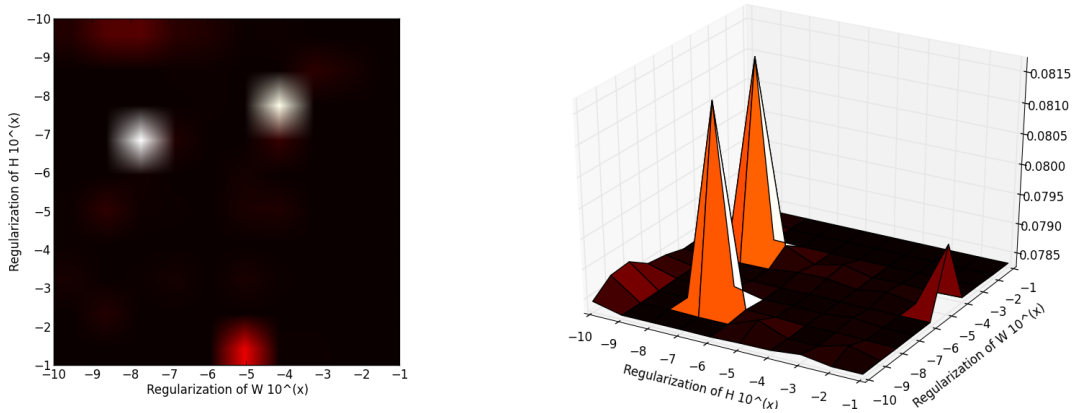

Figure 13. RMSE versus λ and γ when setting latent dimension 5

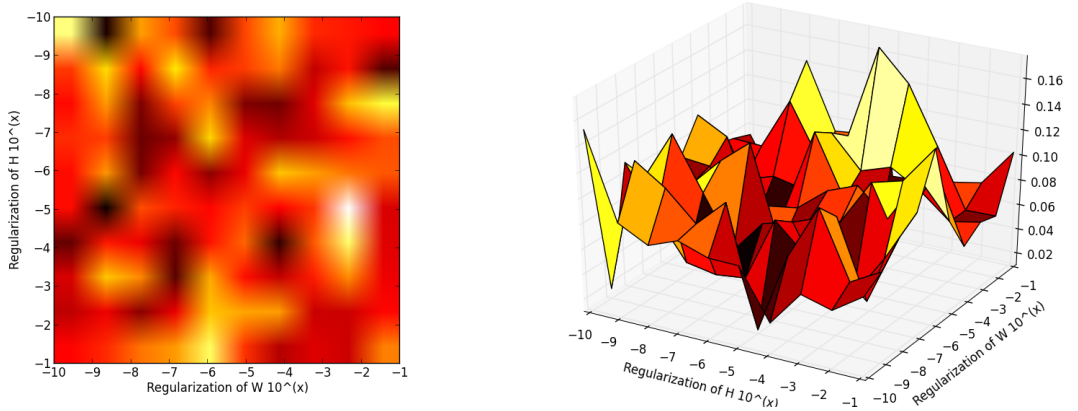

Figure 14. Silhouette value versus λ and γ when setting latent dimension 5

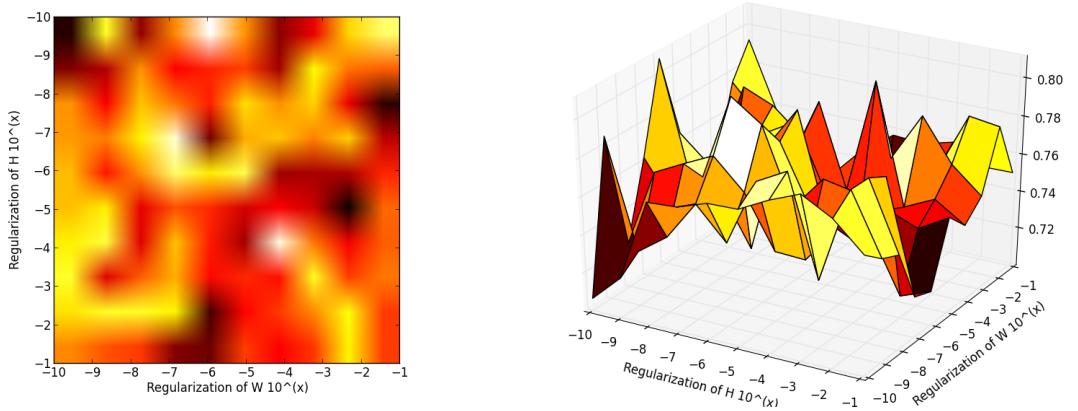

Figure 15. Mean I/E C.D.R. versus λ and γ when setting latent dimension 5

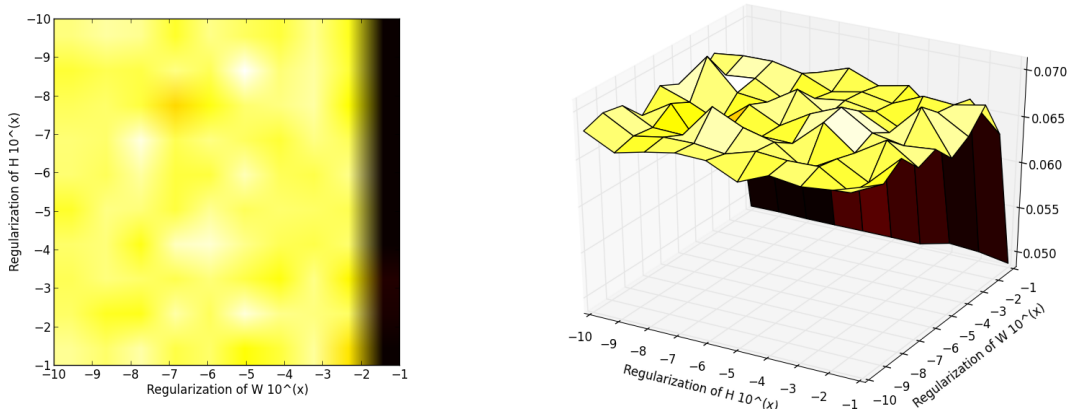

Figure 16. RMSE versus λ and γ when setting latent dimension 10

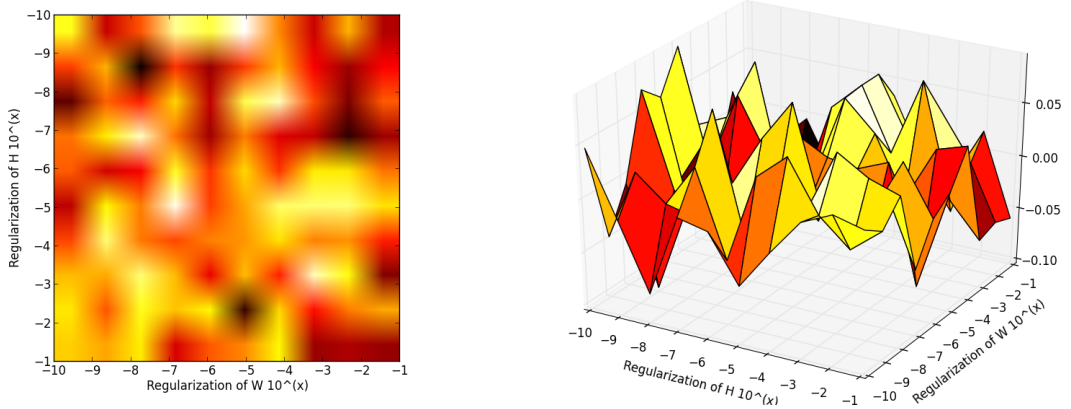

Figure 17. Silhouette value versus λ and γ when setting latent dimension 10

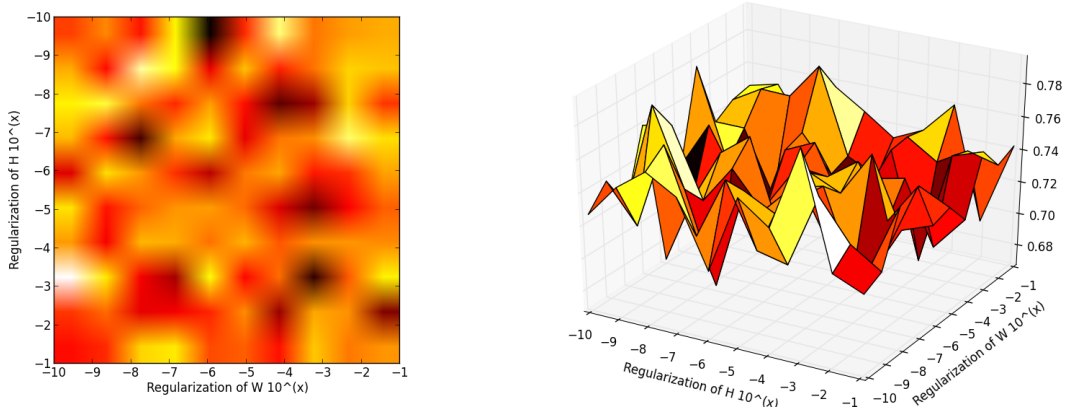

Figure 18. Mean I/E C.D.R. versus λ and γ when setting latent dimension 10

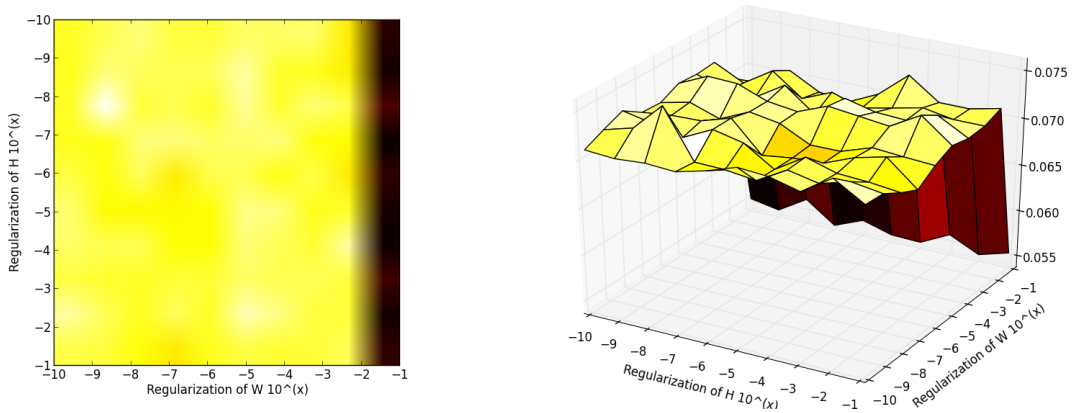

Figure 19. RMSE versus λ and γ when setting latent dimension 15

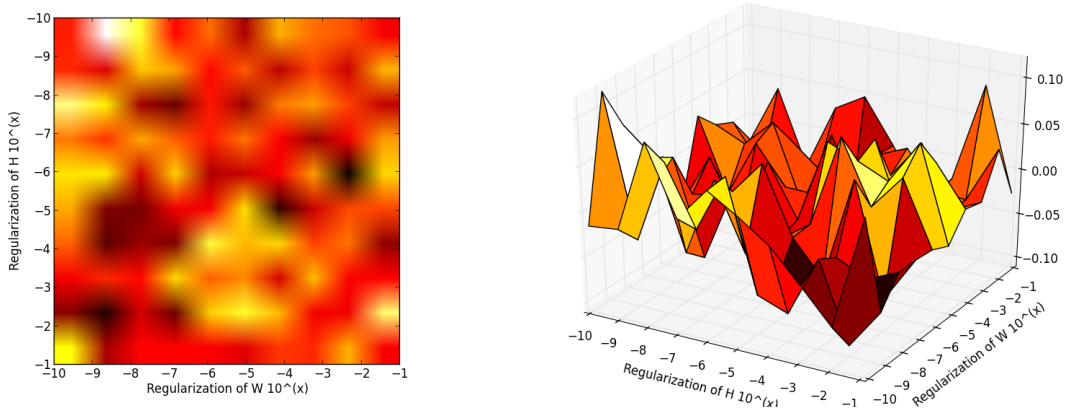

Figure 20. Silhouette value versus λ and γ when setting latent dimension 15

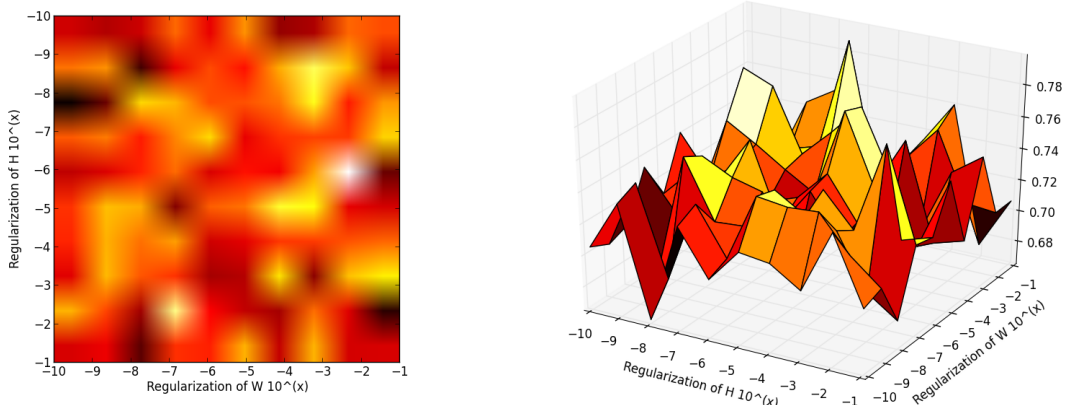

Figure 21. Mean I/E C.D.R. versus λ and γ when setting latent dimension 15

We found that for RMSE, except latent dimension 5, most regularization of W (γ) are selected to be 0.1 to have best RMSE values. The average of regularization of H (λ) is around 1e-5. The graphs of Silhouette value and mean intra/extra cluster distance ratio are fluctuating with no evidence of order. We choose regularization of W (γ) 0.1 and regularization of H (λ) 1e-5 for following clustering evaluations.

In Figure 10, the number of latent dimension 11 has least RMSE. Moreover, in Figure 11 and figure 12, the clustering of NMF using greatest weight value in H (red) is generally better than finding similarity from stock and w in W (blue) (higher Silhouette Value and lower mean intra/extra cluster distance ratio). The average best performance latent dimension is 12. Therefore, we are using greatest weight value in H for grid search of λ and λ for the evaluation of Silhouette value and mean intra/extra cluster distance ratio from Figure 13 to Figure 21. However, because the variances are comparative low (0.07/1 for Silhouette value, 0.05/1 for mean intra/extra cluster distance ratio), for convenience of implementations, we are still using the latent dimension 5, 10, and 15 for the following clustering evaluations.

We also apply K-means clustering technique to our stock data set and evaluate with Silhouette Value (red) and mean intra/extra cluster distance ratio (blue) showed in Figure 22. We notice that the mean intra/extra cluster distance ratio drops significantly when number of K increases. However, better clustering performance in training set does not promise better clustering performance evaluated by testing set prices. Moreover, it does not guarantee less variance of prices in the same cluster evaluated by mean square error. We will discuss these two terms in next two sections.

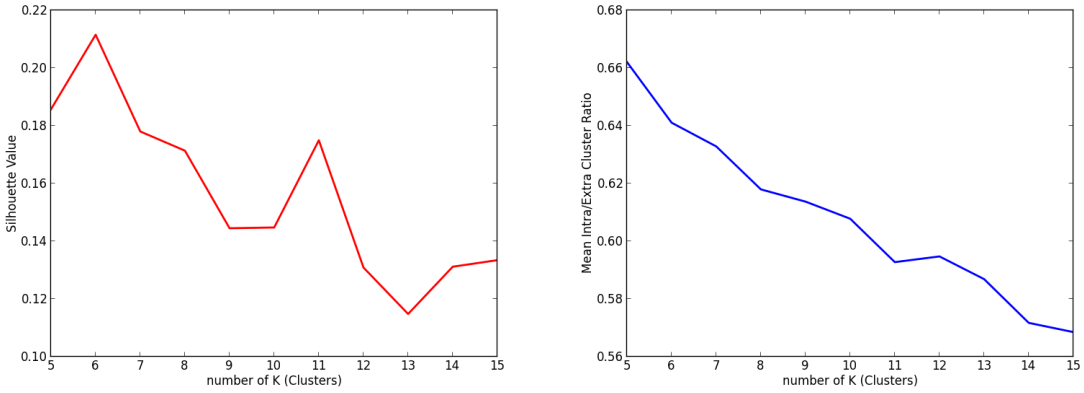

Figure 22. Silhouette Value (red) and mean intra/extra cluster distance ratio (blue) with respective k (clusters)

II. Evaluation with Testing Set

In this section, we adopt the testing data set of (247 prices) in 2012 to evaluate the clustering done by the training set also by Silhouette value and mean intra/extra clustering distance ratio. For Non-Negative Matrix Factorization, selecting greatest weight value in H (red) is no longer guaranteed to have higher Silhouette value and lower mean intra/extra clustering distance ratio. Note for these kind of hypothesis on stock market data trend with high volatility, the Silhouette value is too strict for applying to the testing set (all values are < 0), which makes our mean intra/extra clustering distance ratio more convincing.

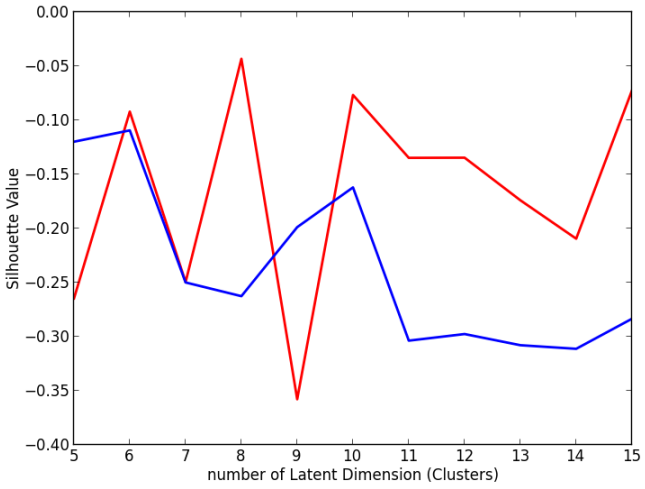

Figure 23. Silhouette Value calculated by testing set with stock clustering by selecting greatest weight value in H (red) and finding similarity from stock and w in W (blue), with grid search of different latent dimensions

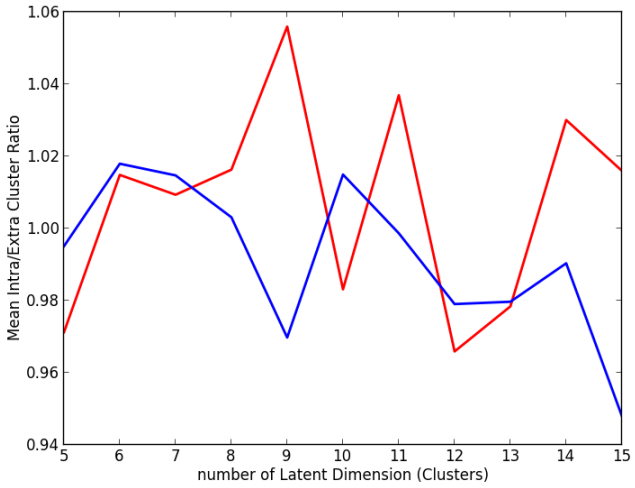

Figure 24. Mean intra/extra cluster distance ratio calculated by testing set with stock clustering by selecting greatest weight value in H (red) and finding similarity from stock and w in W (blue), with grid search of different latent dimensions

For K-means clustering algorithm showed in Figure 25, the performance of clustering is similar to our NMF algorithm.

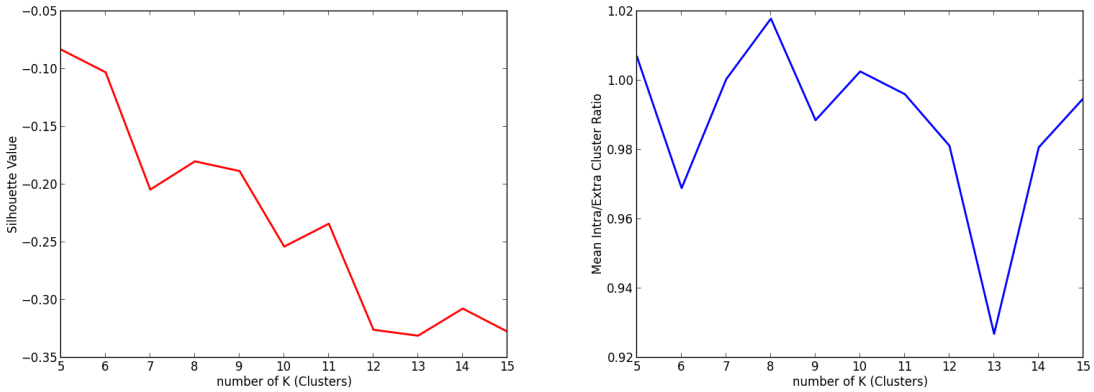

Figure 25. Silhouette Value (red) and mean intra/extra cluster distance ratio (blue) calculated by testing set with respective k (clusters)

III. In Cluster Evaluation

We utilize Mean Square Error using testing data set to measure the differences of prices of each stock and other stocks in the same cluster. We choose on single stock company in a cluster at one time, calculate the mean of all the individual value each day, get the error between two set of values, and get the average of all errors in the end.

In Figure 26 and Figure 27, by using selecting greatest weight value in H, the mean square error is lower than clustering by finding similarity from stock and w in W. The Average of mean square error by selecting greatest weight value in H is lower than clustering by K-means algorithm showed in Figure 28.

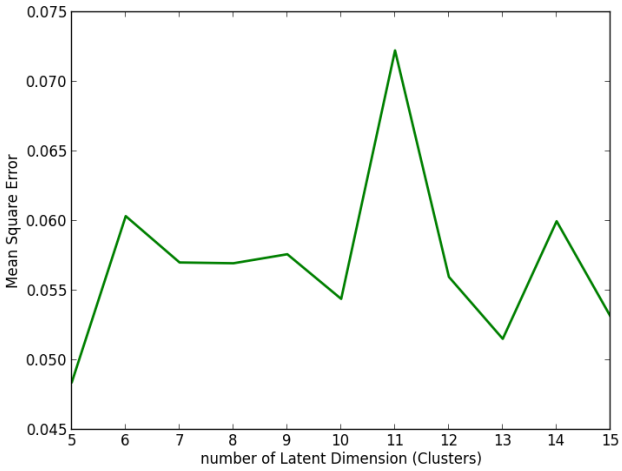

Figure 26. Mean Square Error of stock clustering by selecting greatest weight value in H

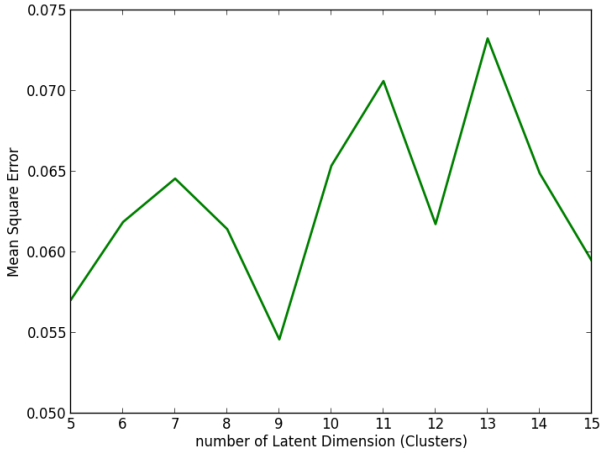

Figure 28. Mean Square Error of stock clustering by finding similarity from stock and w in W

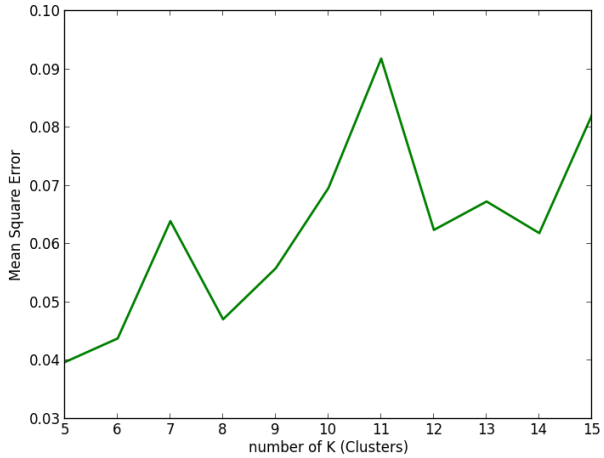

Figure 28. Mean Square Error of stock clustering by K-means

IV. Projection Vectors (Centroids)

We visualize all the 1729 stock closing price data in training set from 2005 to 2011 in Figure 29. After Applying Non-Negative Matrix Factorization, the stock closing price data can be represented by 5 projection vectors in Figure 30, 10 projection vectors in Figure 31, and 15 projection vectors in Figure 32. Note that all the training set data is normalized from 0 to 1. For stock market data, the entries in trend matrix is allowed as negative values regarding the previous implementation of NMF on stock data by Jie Wang [1].

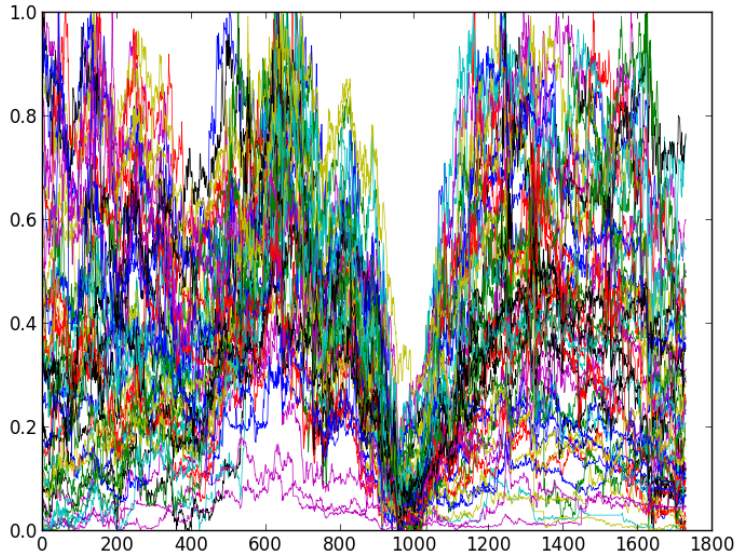

Figure 29. Visualization of all stock data in training set

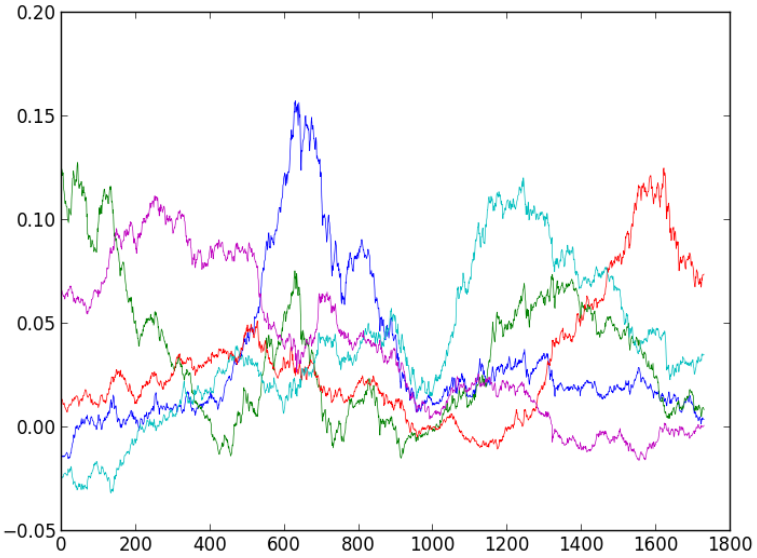

Figure 30. Visualization of all projection vectors in W for latent dimension 5

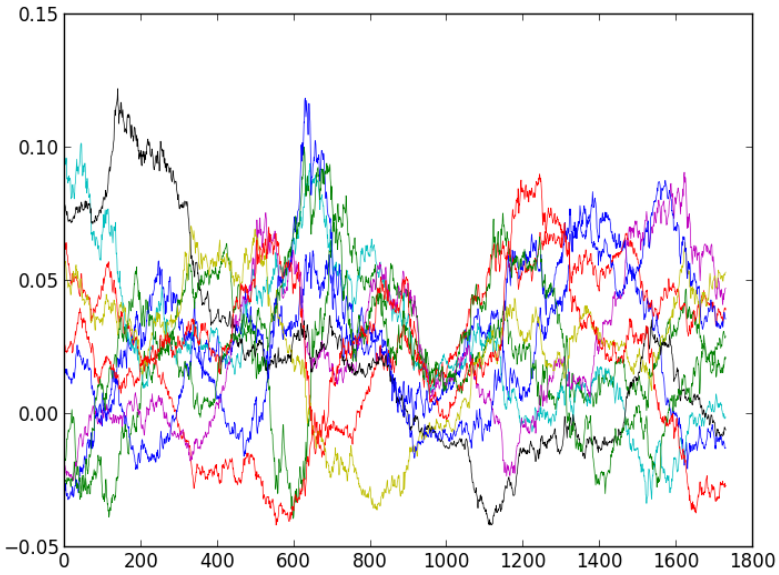

Figure 31. Visualization of all projection vectors in W for latent dimension 10

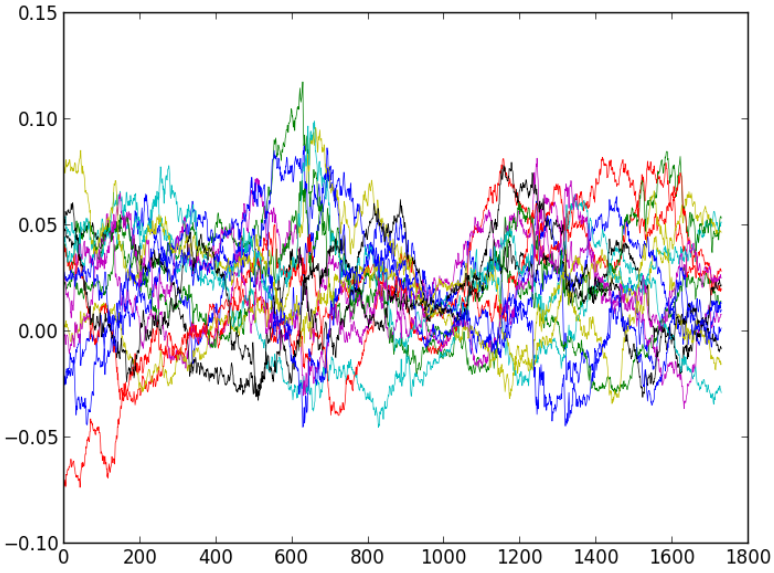

Figure 32. Visualization of all projection vectors in W for latent dimension 15

V. Clustering

In this section, we show the stocks closing price in one cluster on the left and relative projection vector on the right of two largest clusters in 50 stock companies with latent dimension 5, 10, and 15, respectively. The clustering proves that the stock closing price data can be represented by the projection vectors when selecting greatest weight value in H. The clustering results of different latent dimension and K in K-means algorithm are showed in Appendix A and Appendix B.

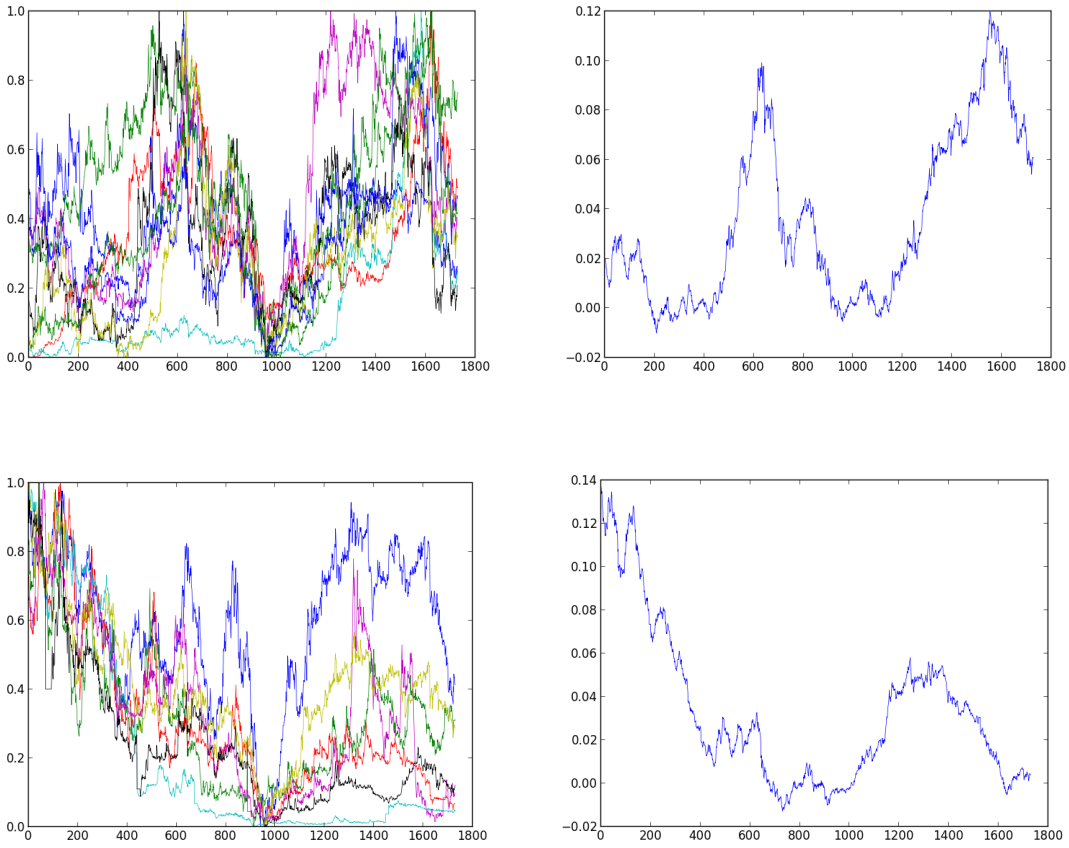

Figure 33. The stocks closing price in one cluster (left) and relative projection vector (right) in NMF with latent dimension 5

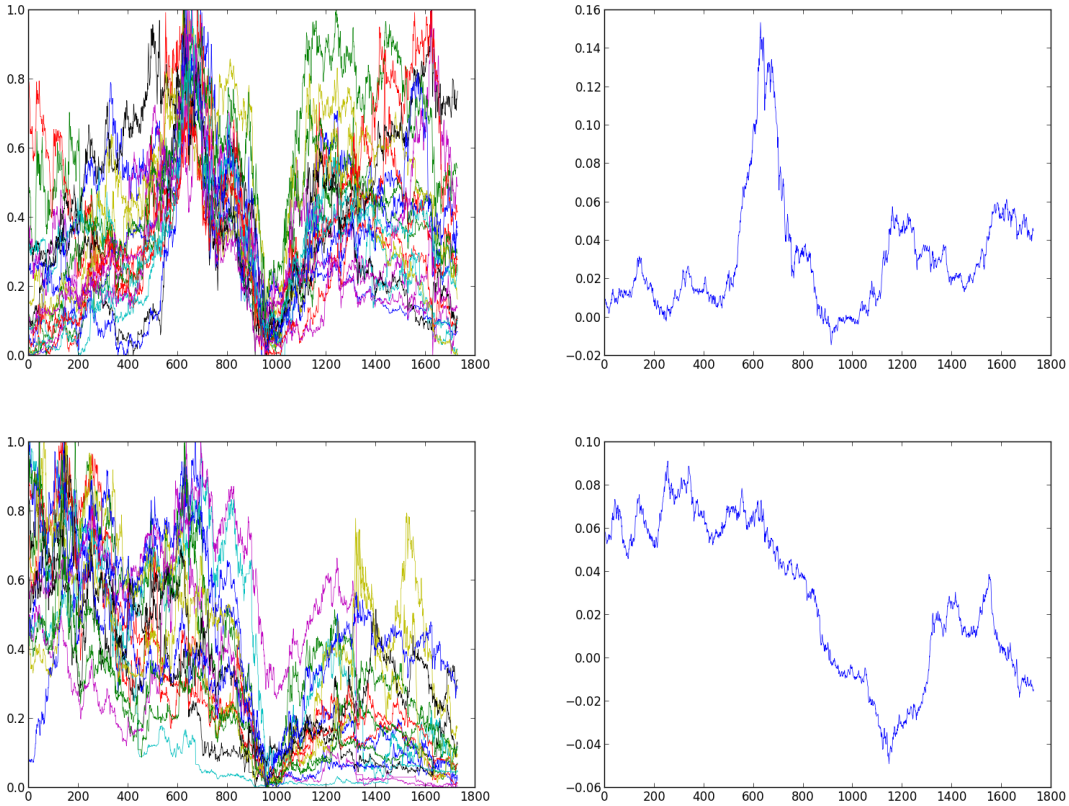

Figure 34. The stocks closing price in one cluster (left) and relative projection vector (right) in NMF with latent dimension 10

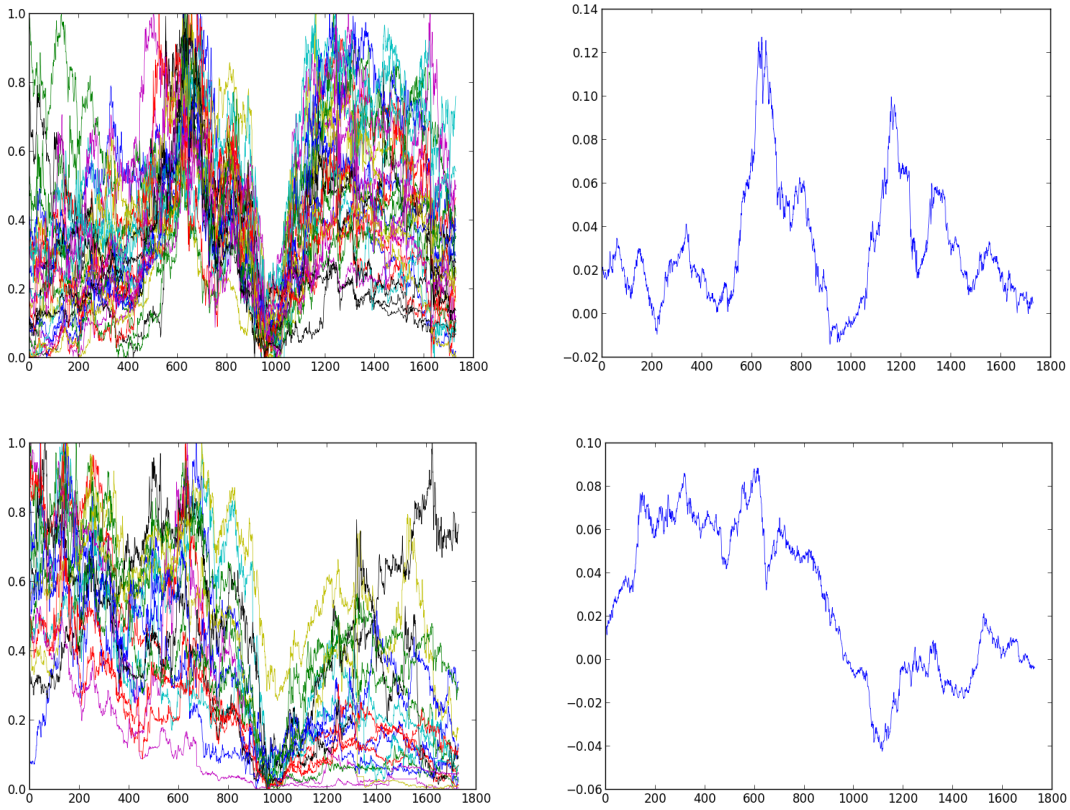

Figure 35. The stocks closing price in one cluster (left) and relative projection vector (right) in NMF with latent dimension 15

VI. Hypothesis of Testing Data

In this section, we try to predict the closing price data of one random chosen stock company compared with mean of closing price data of the rest stock companies in the same cluster both clustered by NMF or K-means. Firstly, Figure 36 and Figure 37 shows closing prices of two random chosen companies and mean of closing price data of the rest stock companies in their clusters, which are clustered by NMF and K-means with latent dimension/K 5, respectively. Secondly, Figure 38 and Figure 39 shows closing prices of two random chosen companies and mean of closing price data of the rest stock companies in their clusters, which are clustered by NMF and K-means with latent dimension/K 10, respectively. Thirdly, Figure 40 and Figure 41 shows closing prices of two random chosen companies and mean of closing price data of the rest stock companies in their clusters, which are clustered by NMF and K-means with latent dimension/K 15, respectively.

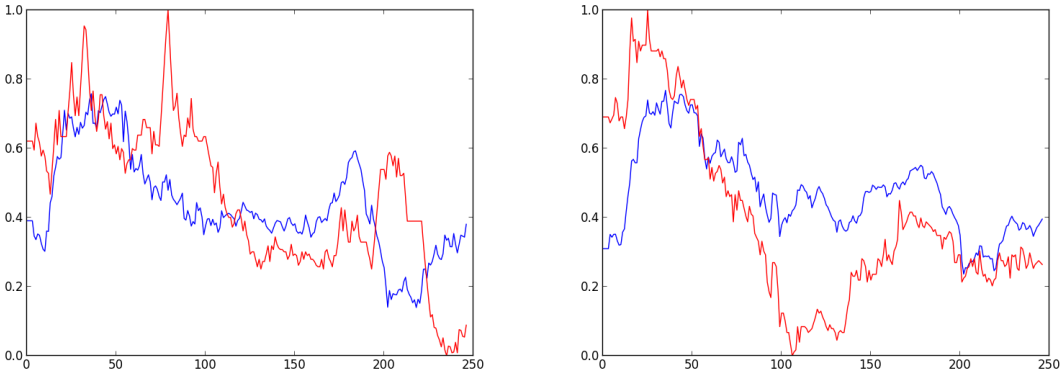

Figure 36. Two random chosen companies and mean of closing price data of the rest stock companies in their clusters, which are clustered by NMF with latent dimension "5"

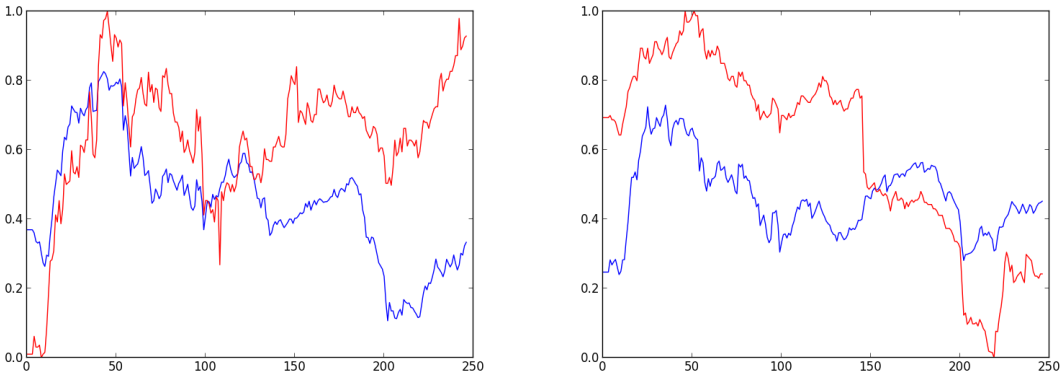

Figure 37. Two random chosen companies and mean of closing price data of the rest stock companies in their clusters, which are clustered by K-means with K "5"

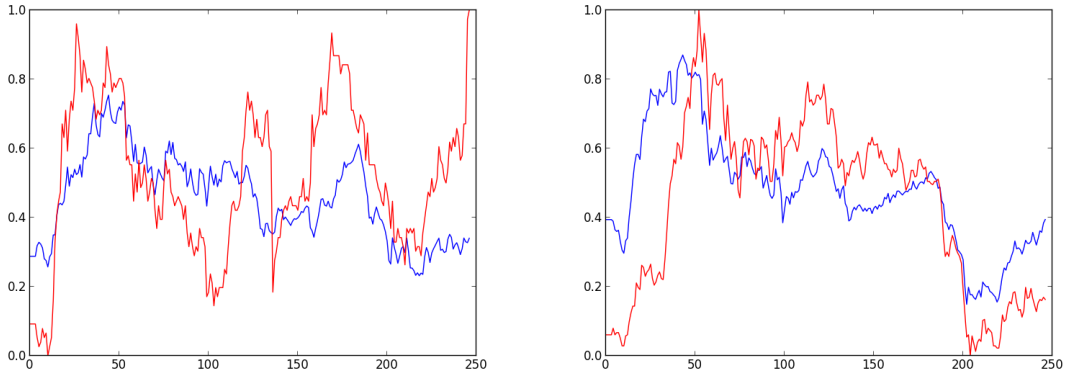

Figure 38. Two random chosen companies and mean of closing price data of the rest stock companies in their clusters, which are clustered by NMF with latent dimension "10"

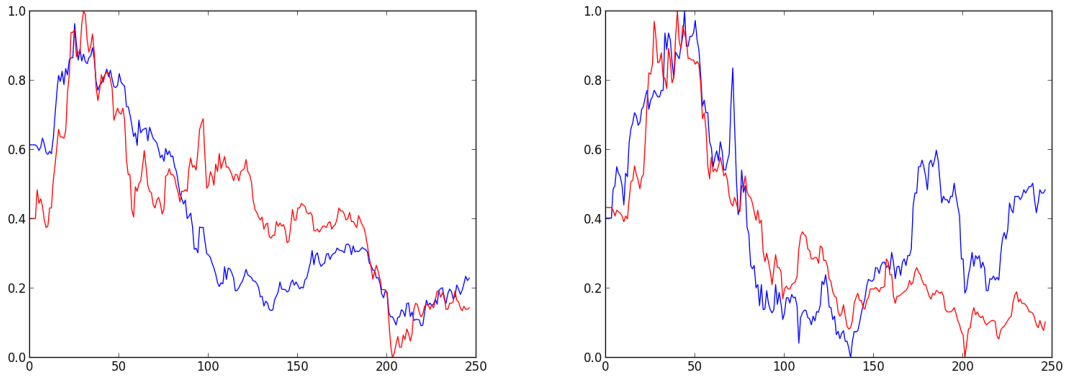

Figure 39. Two random chosen companies and mean of closing price data of the rest stock companies in their clusters, which are clustered by K-means with K "10"

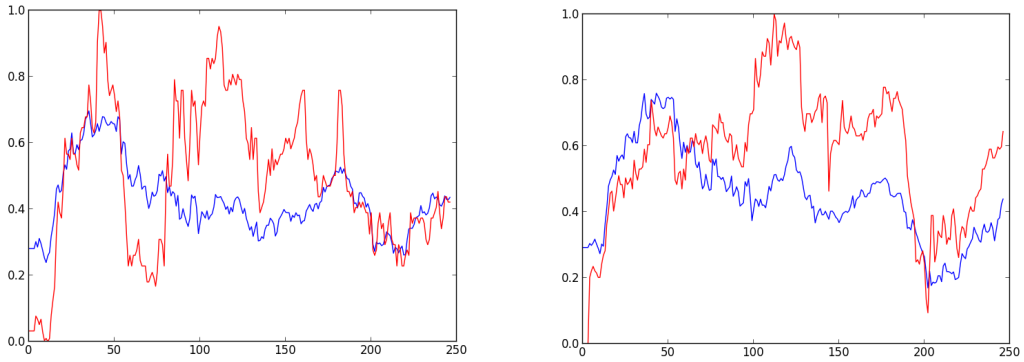

Figure 40. Two random chosen companies and mean of closing price data of the rest stock companies in their clusters, which are clustered by NMF with latent dimension "15"

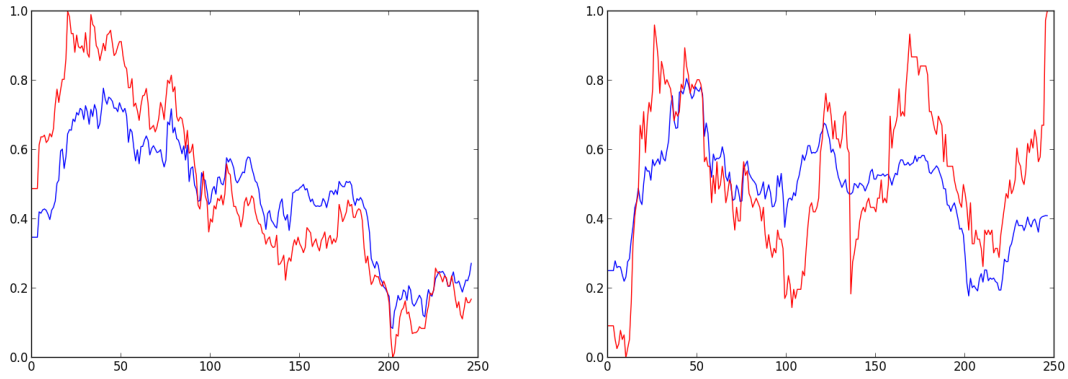

Figure 41. Two random chosen companies and mean of closing price data of the rest stock companies in their clusters, which are clustered by K-means with K "15"

Conclusion

In this work, we utilize Non-Negative Matrix Factorization to stock companies clustering. The latent dimension, weight matrix H and trend matrix W represents the latent forces and factors behind the great amount of data in Taiwan stock market. The Non-Negative Matrix Factorization is taken as a remodeling of a linear combination of 50 stocks over seven years closing prices. Through grid search for convergence and optimal, it is proved that the pattern of great amount stock data can be represented by projection vectors in W. The stock companies can be clustered by either selecting greatest weight value in H or finding similarity from stock and w in W. The clustering performance can be evaluated both by strict Silhouette value or loose mean intra/extra cluster distance ratio. Furthermore, by calculating the mean square error, we can also measure the differences of prices of stocks in the same cluster. After clustering of stock companies, hopefully we are able

to utilize the intra cluster similar volatility to come up with a trend of stocks take it as a reference of diversification analysis for reducing investment losses in the stock market.

Future Work

In the future, streaming data of new stock closing price can use online learning linear regression of individual stock or cluster may be applied to hypothesis of closing prices, optimize the algorithm, and adapt itself to new data. Further co-training of NMF and K-means clustering can be adopted to have better clustering performance in the future.

References

[1] Jie Wang. "Stock Trend Extraction via Matrix Factorization". In *ADMA 2012, LNAI 7713*, pp. 516–526, 2012

[2] Tang Liu. "Non-Negative Matrix Factorization for Stock Market Pricing". In *BMEI*, 2009

[3] Rainer Gemulla, Peter J. Haas, Erik Nijkamp, and Yannis Sismanis. "Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent". In *KDD'11,* August 21–24, ACM 978-1-4503-0813-7/11/08, 2011

Appendix A

5	6	7	8	9	10	11	12	13	14	15
3	1	0	11	17	4	0	1	39	3	22
6	3	14	32	20	5	2	12	48	11	31
9	6	32	43	32	10	4	28	49	18	7
11	9	43	1	33	28	5	36	1	42	3
17	11	13	2	37	35	8	41	3	0	6
18	12	21	8	42	39	10	0	6	2	11
23	17	24	10	43	44	16	16	9	4	17
32	18	29	13	0	46	26	20	12	16	18
33	19	20	15	1	49	27	29	14	20	32
38	22	47	19	2	25	28	47	23	25	38
42	23	1	20	3	48	35	6	27	26	42
2	32	3	21	7	0	39	35	30	28	1
7	33	6	24	11	3	40	44	31	30	9
8	34	7	27	12	13	44	17	33	35	12
40	36	9	29	13	19	46	14	36	41	14
0	37	11	30	15	22	49	39	41	44	23
1	38	12	31	18	24	31	49	42	17	27

14 16 20 29 34 36 37 43	41 42 43 45	17 18 19 22 23 30 31 33 36 38 41 42 45	36 40 41 45 47	19 21 22 24 27 29 30 31 34 36 38 41 45 47	38	1 37	2 4 5 8 10 15 21 26 27 40 46	2 7 8 40	29 32 43	28 30 33 34 41
	0 7 13 15 21 24 27 30 31		0 3 9 22 38 42		11 14 17 18 32 33 37 42 43	3 9 17 23 33 42		0	7 8 40	25 44 48
4 5 10 12 25 26 27 28 35 39 41 44 46 48 49								21		
								13 15 19 22 24 45	6 14 23 48	
			14 35 39 49		2 16 20 26 47	6 7 12 36 41	3 11 13 19 24 38 43			35 39 49
	20 29	2 4 5 8 10 15 16 25 26 27 28 35 39 40 44 46 48 49		6				20 29 32 34 37 43 47	1 9 12 33 36 37	2 4 5 8 26 40 46
	2 4 5 8 10 16 25 26 28 35 39 40 44 46 47 49		6 7 18 23	16 26	7 8	14 15 32 48	9 18 22 31 32 33 34 45			16 20 29 37 32
			17 33 34 37	14 25 48 49	21 29 30 40	11 13 15 18 19 21 22 24 34 38 45		4 5 10 16 25 26 28 35 44 46	10 13 15 19 21 22 24 27 31 34 38 45 47	
13 15 19 21 22 24 30 31 45 47			4 5 12 16 25 26 28 44 46 48	4 5 8 10 28 35 39 40 44 46	15 27 31					13 15 19 21 24 36 45
					1 6 9 12 23 36 41	20 29 30 43 47	25 37 48			0 10 47
	14 48	34 37		9 23	34 45		7 23 30 42	11 17 18 38	5 39 46 49	

Appendix B

5	6	7	8	9	10	11	12	13	14	15
1 9 12 32 34 36 37 43	0 1 7 12 14 16 20	11 18 33 38	2 4 8 26 40	10 27 28 31 47	32	5 25 35 39 44 46 48 49	29 32 36 37 43	32	1 12 25 35 44 48	27 31
0 2 4 8 16 20 26 40	29 32 34 36 37 43	0 7 14 19 20	10 13 15 21 24 27 28 30 31 47	2 4 8 26 40	29 34 37 43	3 6 7 17 19	10 21 28 47	1 12 30 36 41	2 8 40	3 11 14 17 18 19 33 38
5 10 25 28 35 39 44 46 48 49	3 6 9 11 17 18 19 22 23 33 38 42	5 25 35 39 44 46 48 49	1 12 14 16 48	29 32 34 37 43	2 4 8 26 40	4 14 16	2 4 8 26 40	37 43	9 23 42	26 46
3 6 7 11 14 17 18 19 23 33 38 42	10 13 15 21 24 27 30 31 41 45 47	1 10 12 13 15 21 22 24 27 28 30 31 36 41 45 47	5 25 35 39 44 46 49	0 7 16 20	9 13 15 21 22 23 24 27 30 31 36 41 45	11 18 33 38	14 0 16 20	5 25 35 39 44 48 49	30 6 10 28 47	2
13 15	2 4 8 26	3 6 9 17 23 42	32	13 15 21 22 24 30 45	14 16 48	13 15 21 24 27 31 45	9 23 42	10 15 21 27 28 31 47	3 6 14 17 33	5 25 35 39 44 49
		2	0 7 19 20	11 18	0 3 6 7 17 19 20	0 20	6 7	13 22 24 34 45	0 16 20	4 8 40
			3 6 11 17 18 23 33 38	1 9 12 23 36 41	1 12	2 8 26 40	1 12		30 36 41	0 7 16 20
				3 6 14 17 19 33		29 32 34 37 43	5 25 35 39 44 46 48 49	3 6 11 14 17 18 19	11 18	
						10 28	13 15 22 24 27 31 34		29 32 37 43	13 15 21 22 24 45
									13 15 22 24 34	9 23 42

21	40	4	42	38	10	47	45	33	45	1
22	5	8	9	42	28	1	3	38	10	12
24	25	16	22	5	47	12	11	0	28	41
27	28	26	29	25	5	36	17	7	39	48
29	35	40	34	35	25	41	18	16	49	29
30	39	29	36	39	35	9	19	20	7	32
31	44	32	37	44	39	22	33	29	19	34
41	46	34	41	46	44	23	38	4	38	43
45	48	37	43	48	46	30	30	26	27	36
47	49	43	35	49	49	42	41	46	31	37