**Introduction**

Remarkable advances in data analyzing technologies, algorithms based on linear algebra and statistical technique have accelerated all filed of researches. As master of data science student, I want to apply mathematical algorithm and statistical technique into real data set. Dow Jones Index is one of a good data set because aDow Jones is consisted of major 30 companies, so it reflects whole market flow. This dataset contains weekly price moving and dividend during the first two quarter in 2011 for each stocks. Hence, it can help to see market flow in half year of 2011. Therefore, I planned this project.

The goal of this project is ﻿petitioning Dow Jones Index into two groups, less risk stock and more risk stock, with linear algebra technique and statistical theory. In this project, there is three major steps; Clean data, SVD and K-MEAN Clustering, and make OLS model in order to predict stock price.

a *Dow Jones Index consisted 30 companies which is in traditional heavy industry. The average of market capitalization is 175bilion dollar. Apple, Disney, Goldman Sachs, Nike, Verizon, Pfizer, Exxon and 23 other companies in this list.*

**Method & Result**

**More Risk Group**

**Future Price**

**Less Risk Group**

**Future Price**

**K-Means**

**K-Means**

**2:** **V.T contains columns(companies) price feature,**

**so it is suitable values for clustering with K-means**

**SVD**

**Transpose**

**OLS**

**OLS**

3. Predict from Regressing Model

**<6x2>**

**<24x2>**

**3:** **Predict future price with linear regression model.**

**Predict**

**Predict**

**More Risk**

**Group**

**Less Risk**

**Group**

1. Read & Clean Data

**<30x30>**

**<750x16>**

**Dow Jones**

**Raw data**

**<2x2>**

**V.T**

**<2x2>**

**<30x2>**

***Modify column;***

***mean & std***

**Clean data**

**U**

2. Clustering

***1. Read & Clean Data:***

**1:** **Focus on weekly mean price**

**std. for cleaning data**

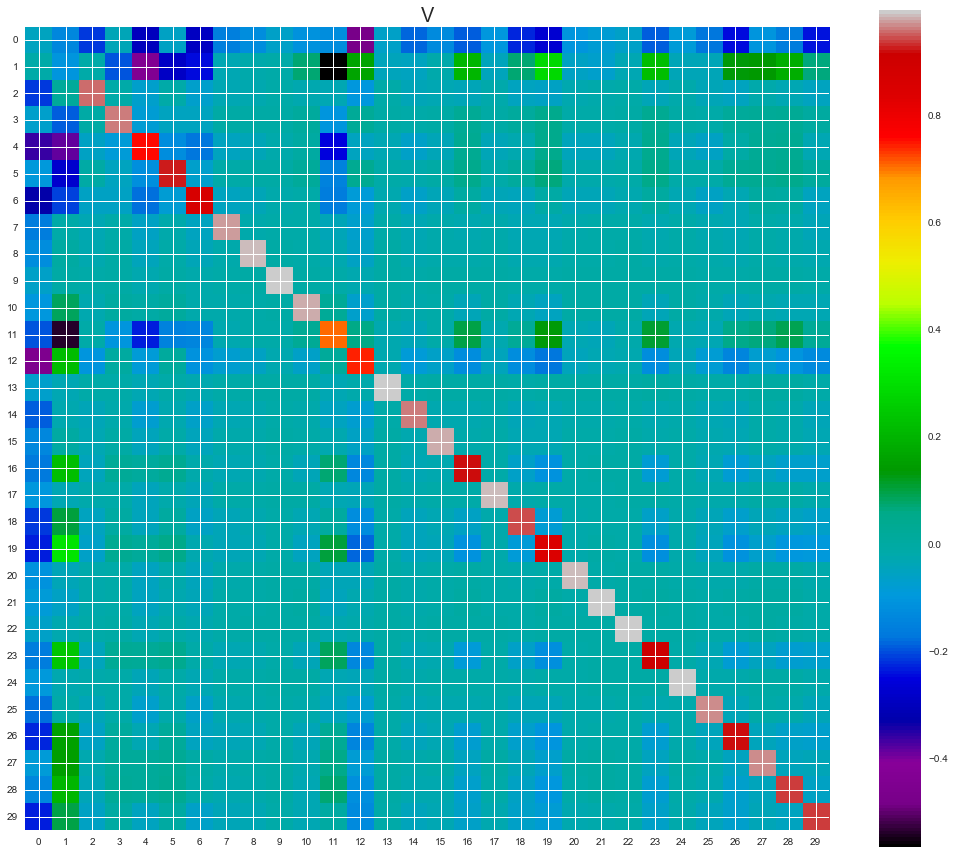
The shape of raw data is 750 x 16 matrix. It is consisted of 25 weekly (first 2 quarter in 2011) price high, low, open, close, volume and other relative columns. However, in order to separate into two group based on risk, the data need to clean avoid the noise from other data, so I calculate average price and standard deviation for each week, then combine weekly data into half year data for each companies. Hence, I get 30 x 2 matrix for clean data; 30 name of companies in rows and half year average price and standard deviation in columns.

***2. Clustering:***

In clustering, I used two kinds of mathematical algorithm; bSVD and cK-Mean clustering.

The first, I got three matrix U, Σ and V.T from clean data with SVD (used Transpose of clean data). Since the U and V.T matrix consist of orthonormal eigenvectors each rows or columns, they are independent. It means that they have a different feature in each rows or columns. The 2x2 matrix U contains feature information about rows of (clean data).T; half year average price and standard deviation, and the 30x30 matrix V.T contains feature information about columns of (clean data).T; meaning V.T shows 30 companies features. And since the eigenvalue, Σ (344.11957825, 4.16425378), so I decided that the first column of U and the first row of V.T contribute to matrix A. Therefore, the first orthonormal vector in U (first column) has an important feature for half year average price and standard deviation, and the first orthonormal vector in V.T (first row) has an important feature about half year average and std price for 30 companies.

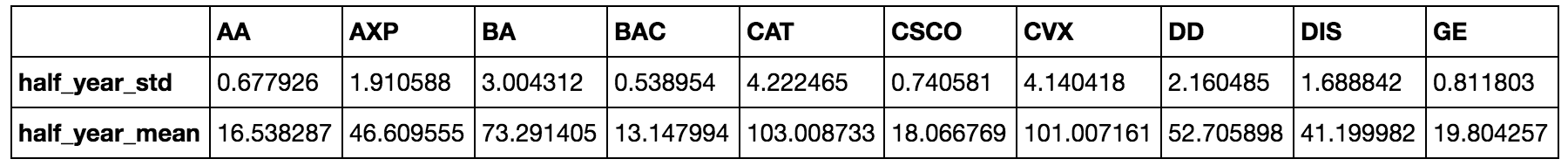
The second, I used K-mean clustering algorithm with matrix V to separate into less risk group and more risk group. Since V.T contains feature of 30 companies about price variance and mean for half year, I believe that it is good values to separate groups as risk level.



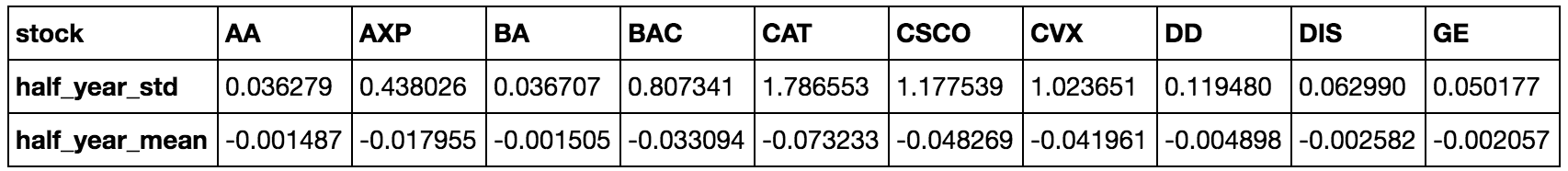
<Figure1>

Eigenvalue σ1 in Σ about an order of magnitude greater than σ2, indicating that orthonormal eigenvector in the first rows of V are much more important than eigenvectors in the other row. In fact, I could closely reproduce A using just the first rows of V and first columns of U. In the first row, I there is 30 companies feature about price of mean and std, so I can tell that companies which are similar color is in the same group from K-means clustering.

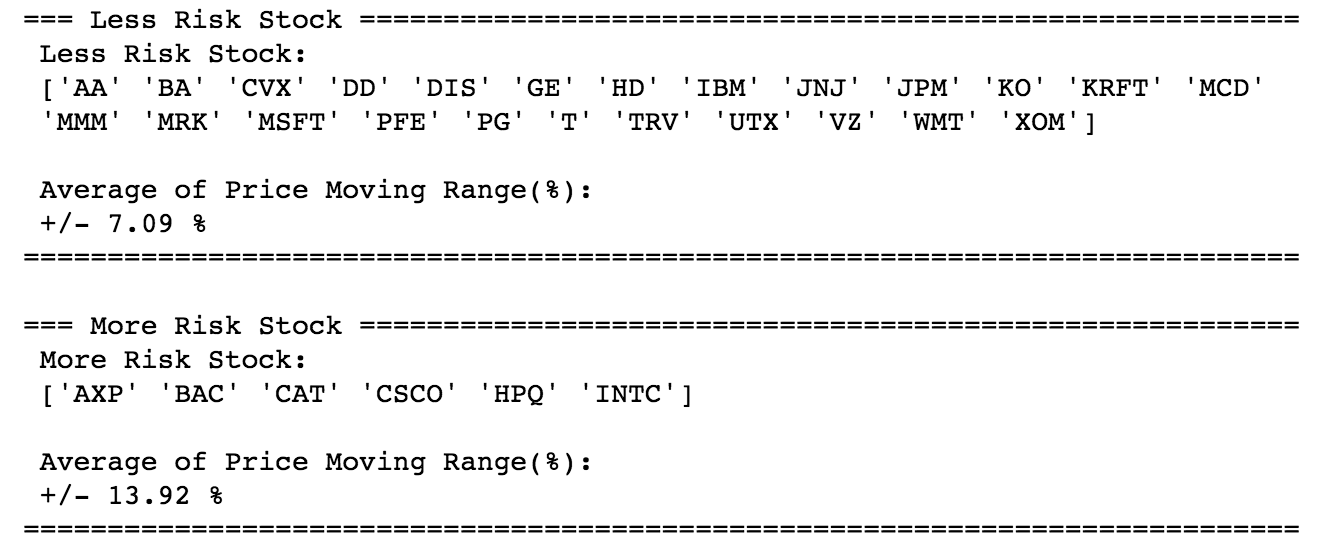
<Figure2-1; Multiplication from the First Orthonormal Vector, U[ : , :1]) \* Σ [:1]) \* V.T[ :1, : ])>



<Figure2-2; Error from Actual Value>



Although the error of standard deviation for some stocks (mostly in more risk group) are large number of difference, the error for mean in all stocks is less than 1%, so I can say that the matrix in ‘Figure2-1’ (first rows of orthonormal eigenvector in V and first columns of orthonormal eigenvector U) can reproduce A, and it contains price features of original matrix A.



<Figure3>

As the result from K-mean, there is 24 companies in less risk group, and 6 companies are in more risk group which is based on eigenvector or feature each company. Also, I calculate risk variance for each group in order to check K-means clustering results, and it is an appropriate.

<Figure3; Result from K-Mean>

b *SVD (Single Value Decomposition)*

*Any matrix A can be decomposed to three matrices U, Σ and V such that A=UΣV, this is called singular value decomposition. The columns of U and V are consisted of orthonormal eigenvectors and Σ is diagonal, which is consisted of eigenvalues. Assume A is mxn matrix with column vectors a1, a2, ..., an. In SVD of A, U will be mxm, Σ will be mxn and V will be nxn. The column vectors of U as u1, u2, ..., um and V as v1, v2, ..., vn, similarly to A. The diagonal of Σ as σ1, σ2, ..., σn (σ1>σ2>...>σn). After decompose matrix A, there are only rank(A) non-zero values in Σ (σ1, …, σr 0; σr+1, ..., σn =0), so columns of U beyond the rth column and rows of V beyond the rth row do not contribute to A and are usually omitted because it is not important values.*

*c K-MEAN Clustering*

*K-mean clustering algorithm is to petition the best division of n points (data) in k groups. In other word, the group’s members, which share centroid (or mean point), have the nearest distance between centroid and each member.*

**Algorithm** K\_Means\_Clustering **is**

**Input:** n points (x1, x2, …, xn)

**1. Initialize** cluster centroids (μ1, μ2, …, μn) Randomly

**2. Repeat** until convergence: {

#assign xi to μi whose nearest mean centroid

**for** each data point xi **to** n

K-means Eqation

#recalculate the new position of μj (centroid)

**for** each cluster point μj **to** k

**return** Sk

**Output:** k disjoint sets (S1, S2, …, Sk)

***K-Mean Algorithm:***

***algorithm:***

***K-Mean Equation:***

||xi – μi|| = Euclidean distance function

μi = centroid for clusters

xi = case i

n = number of cases

k = number of clusters

***3. Predict from Regression Model:***

In order to predict stock price in each groups, I make OLS model for y=predict price and x=half\_year\_std, risk\_var (risk\_var is calculated based on the whole Dow Jones Index data set, and Risk Variance tells that '+' or '-' price movement from Mean\_Price). I take care of residuals normality to predict stock more accurately so I applicate log value to y=predict price, then convert to real price values after calculate with coefficient of variables.

|  |  |  |
| --- | --- | --- |
| <Figure3> | Less Risk Group | More Risk Group |
| Average Percentage Change (%) | 1.69 | 3.72 |

This is the values can explain that the companies classified less risk group are +/- 1.69% price range from current average price in next half year, but the companies classified more risk group are +/- 3.72% price range from current average price in next half year. Therefore, Dow Jones data with SVD and K-means clustering is successfully classified into two groups.