Statistical Observation and Analysis of Dow Jones Returns Data

Doungporn Wiwatanapataphee

20 October 2020

Consider the Dow Jones returns data and use the 30 stocks as observations.

i. The data matrix is 30×1200 with 30 observations (stocks) of 1200 variables.
ii. The size of the covariance matrix of these observations is 1200×1200.
iii. The rank of the covariance matrix of these observations is 29.
iv. The largest eigenvalue of this covariance matrix is 0.04707259.
v. The 30th eigenvalue significantly drops and is also much closer to zero as compared to the 29th eigenvalue. For HDLSS data such as Dow Jones returns, this eigenvalues' characteristic is expected as at most $r = 29$ PCS are of interest, while all other PCs are significantly close to zero. This is confirmed by the rank calculated in iii, which serves as an upper bound for the number of PCs.

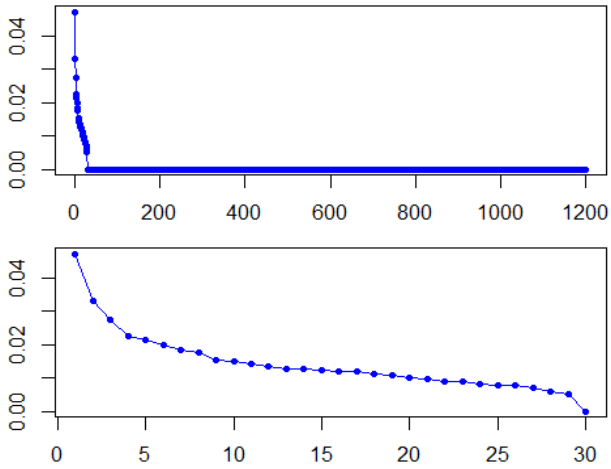

Figure 1 shows the eigenvalues plots of the covariance matrix of raw Dow Jones returns data set where all 1200 eigenvalues (top) and the first 30 eigenvalues (bottom) are visualised.

From the top, the eigenvalues of the Dow Jones data show a rapid initial decrease toward zero and then turn almost flat afterwards. Since only the first 30 PCs of Dow Jones data are meaningful, we examine the bottom plot to get further information.

For the bottom plot, the eigenvalues show a rapid initial decrease for the very first few eigenvalues. The decreasing rate significantly drops as we move forward. Then, the eigenvalues seem to decline steadily between the 5th and the 29th eigenvalues, and it reencounters sudden drop at the 30th eigenvalue. The eigenvalues of the covariance matrix are essentially the variability of the Dow Jones data in an orthogonal basis that captures as much of the data's variability as possible in the first few components. Hence, the trend of eigenvalues is shown in **Figure 1**.

Use the 30 stocks as observation and perform k-means clustering.

```
DJclus.out = kmeans(t(DJ1201), centers=2, nstart=25)
DJclus.out$size

## [1]  4 26
```

According to the output above, the sizes of the two clusters after performing k-means clustering with $k = 2$ and `nstart=25` are 4 and 26. The smaller cluster consists of `HWP`, `INTC`, `IBM`, and `MSFT`, which are all Tech stocks.

Use the 1200 daily returns as observation and perform k-means clustering.

```
set.seed(1910)
DJclus = data.frame()
for(k in 2:12){
  DJclus.out = kmeans(DJ1201, centers=k, nstart=25, iter.max = 50)
  clustered.df = data.frame( k = k,
                             cluster = DJclus.out$cluster,
                             WSS = DJclus.out$tot.withinss,
                             BSS = DJclus.out$betweenss,
                             totalSS = DJclus.out$tot.withinss)
  DJclus = rbind( DJclus, clustered.df )
}
```

For each $k = 2, \ldots, 12$ and for each of the $\mathcal{C}_k = 2, \ldots, k$ clusters, calculate the number of observations and display the results in a cluster table.

Clus	Level										
	2	3	4	5	6	7	8	9	10	11	12
1	602	195	345	351	279	263	4	183	4	114	133
2	598	362	394	239	220	117	182	118	181	206	11
3	0	643	285	117	101	4	80	163	104	28	141
4	0	0	176	311	102	88	159	220	118	74	125
5	0	0	0	182	247	280	241	31	131	176	50
6	0	0	0	0	251	237	154	112	38	113	102
7	0	0	0	0	0	211	246	211	160	134	66
8	0	0	0	0	0	0	134	158	76	249	50
9	0	0	0	0	0	0	0	4	199	91	4
10	0	0	0	0	0	0	0	0	189	4	242
11	0	0	0	0	0	0	0	0	0	11	181
12	0	0	0	0	0	0	0	0	0	0	95

Table 1 shows the number of observations containing in each cluster $\mathcal{C}_k$ for each number of clusters k up to 12.

For $k = 2$, there is a split of about half of the data into the two clusters. In addition, there are four singletons appear at level 7, which potentially be the dates in which the Dow Jones plunged as stated in Lecture 9. At level 11, there are another 11 days split from the rest of the observations, which possibly share some interesting characteristics.

For each $k = 2, \ldots, 12$ show a plot of the within-cluster variability W, the between-cluster variability B and the total sum of squares against the index k. Please see **Figure 2** below.

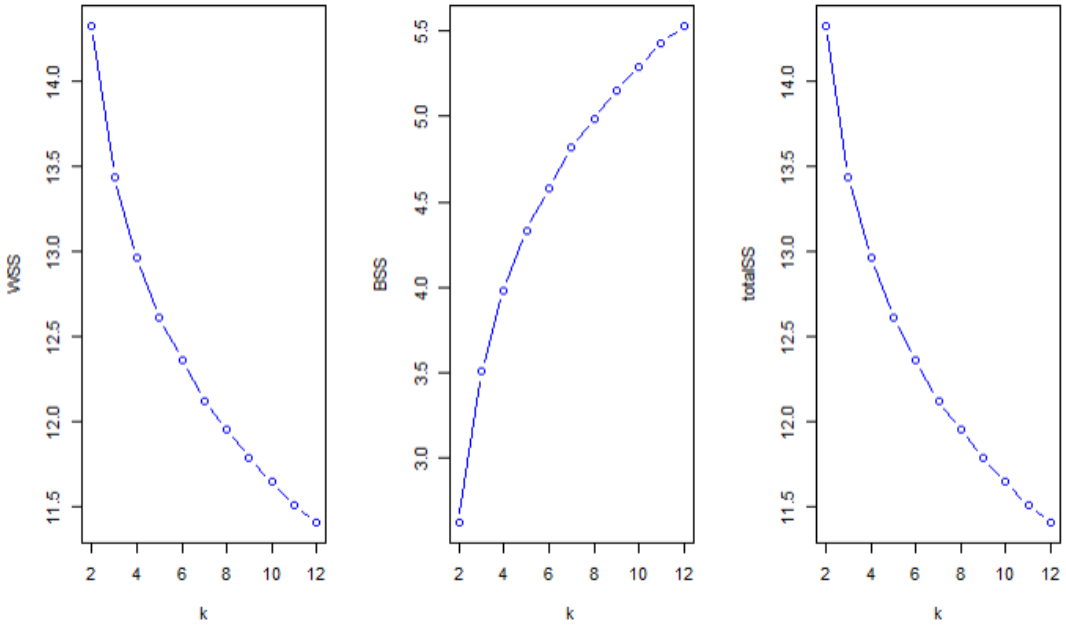

Figure 2 displays the within-cluster variability W (left), the between-cluster variability B (middle), and the total sum of squares against the index k (right).

The within-cluster variability and the total sum of squares are monotonically decreasing as the number of clusters increases and reach the minimum value at $k = 12$, while the between-cluster variability is monotonically increasing as the number of clusters increases and reaches the maximum value at $k = 12$

iii. Based on the calculations and graphs in parts i. and ii., state what you think is the right number of clusters for these data and give a reason for your choice. Comment on your results.

From Figure 2, it is quite challenging to justify which number of clusters is the most appropriate the Dow Jones data. Ideally, we would like to minimise the variability within clusters W and maximise the variability between clusters B. However, intuitively choosing the number of clusters which gives a desired W and B is not feasible as we would end up selecting a very large k almost every time. Then, it is more sensible to examine **Table 1** to see how days cluster.

In my opinion, the choice of two clusters is quite sensible as one of these two clusters probably belongs to the days which DJ rises, and another cluster belongs to the days which DJ falls. Also, **Table 1** displays four and eleven singletons at level 7 and level 11, respectively. These singletons probably share some interesting characteristics, which we might investigate further. Hence, either $k = 2$ and $k = 11$ clusters are great choices for the number of clusters.

Use the 30 stocks of DJ1201 as observation, calculate the first two principal components and show PC1/PC2 score plots

```
pc.DJ = data.frame(prcomp(t(DJ1201))$x)
```

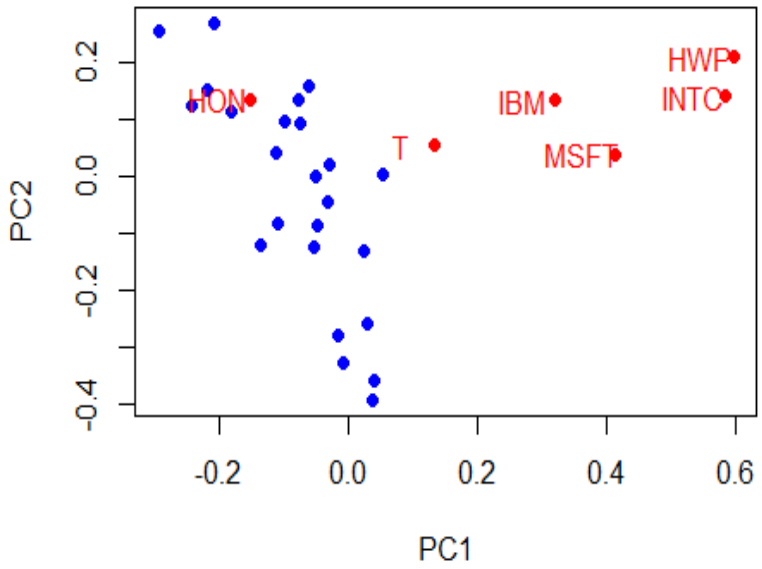

Figure 3 illustrates a PC1/PC2 score plot of the 30 stocks of the raw Dow Jones returns over the five years from 2 October 1995 to 30 June 2000. Tech stocks, namely, T, IBM, HWP, INTC, MSFT and HON, are displayed and labelled in red for convenient inspection.

Over the five years, the non-Tech stocks (blue) are much closer to each other as compared to the Tech stocks. In addition to that, the four Tech stocks HWP, INTC, IBM and MSFT are also quite close to each other and separate from the other stocks, while HON is obviously among the non-Tech stocks. Though T seems to be closer to the non-Tech stocks, it does not show a clear separation from the other four Tech stocks.

Next, further comparisons can be carried out by investigating shorter timeframes in **Figure 4**.

```
pc.DJ1 = data.frame(prcomp(t(DJ1201[1:300,]))$x)
pc.DJ2 = data.frame(prcomp(t(DJ1201[301:600,]))$x)
pc.DJ3 = data.frame(prcomp(t(DJ1201[601:900,]))$x)
pc.DJ4 = data.frame(prcomp(t(DJ1201[901:1200,]))$x)
```

Since there are 1200 days in the `DJ1201` data set, the principal components are calculated for each quarter consisting of 300 days each.

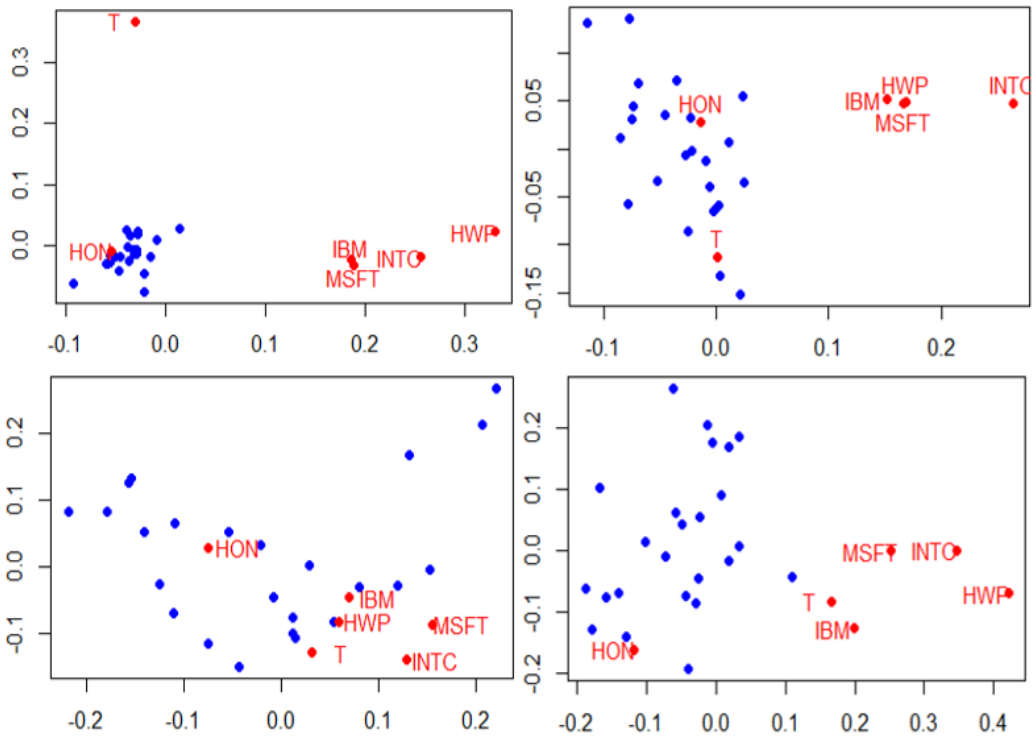

Figure 4 shows PC1/PC2 score plots of the 30 stocks of the raw Dow Jones returns for the 1st (top left), the 2nd (top right), the 3rd (bottom left) and the 4th (bottom right) quarter of the five years.

From the top left corner, the PC score plot corresponds to the first 300 days of the `DJ1201` data starting from 2 October 1995 to late 1996. This score plot represents the Dow Jones returns between early 1996 to mid-1998. Comparing this plot to that of **Figure 3**, there are apparent significant movements. Now, the non-Tech stocks are much closer to each other, and the four Tech stocks seem to form a group far away from the remaining stocks. It can also be seen that T is essentially an outlier with relatively high PC2 score. From further investigation on the stock price history of the Tech stocks in 1996, T is the only Tech stock that encountered a fall in stock price, while the remaining Tech stocks experienced a considerably high rise, which could potentially be the reason to its high PC2 score.

Moving to the top right corner, the PC score plot corresponds to the 2nd quarter between early 1997 to early 1998. This score plot looks almost similar to **Figure 3,** but with a few movements. The four Tech stocks are much closer together, especially HWP, IBM and MSFT. However, HON and T are among the non-Tech stocks, which are now clearly separate from the other four Tech stocks.

The bottom left PC score plot corresponds to the 3rd quarter between mid-1998 to mid-1999. This score plot does not look similar to any of the score plot we have seen before. However, there are some significant movements which are worth discussing. Firstly, the plot does not show a clear separation between the non-Tech and Tech stocks like before, but we can see that T has moved towards the four Tech stocks, while HON remains among the non-Tech stocks. From further investigation on the movement of T stock in 1998, there was a stock split for T around mid-1998 to boost the stock's liquidity, which is probably the reason of the changes in its characteristics.

Lastly, the bottom right PC score plot corresponds to the 4th quarter between mid-1999 to 30 June 2000. Now, T appears to be much closer to the Tech stocks, while HON still remains among the non-Tech stocks.

Apply agglomerative hierarchical clustering to the stocks based on the complete linkage and the Euclidean distance. Show the dendrogram of this cluster analysis

```
hc.complete = hclust(dist(t(DJ1201)), method="complete")
```

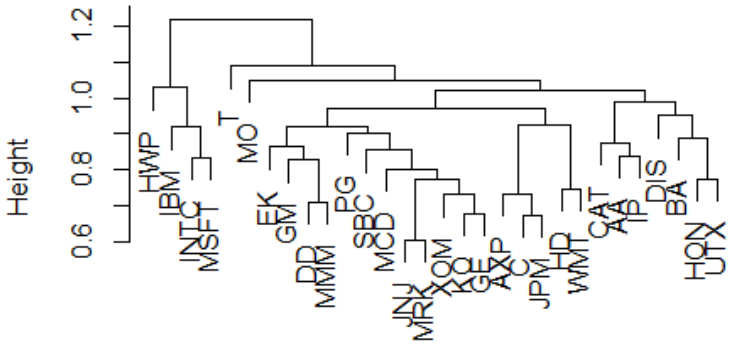

Figure 5 shows the dendrogram resulting from the hierarchical cluster analysis of the Dow Jones data based on the complete linkage and the Euclidean distance with 30 stocks as observations.

Show a cluster table by levels up to 12 levels

```
DJclus2 = data.frame()
for(k in 1:12){
  clustered.df = data.frame(k = k, cluster = cutree(hc.complete,k))
  DJclus2 = rbind( DJclus2, clustered.df )}
```

Clus	Level											
	1	2	3	4	5	6	7	8	9	10	11	12
1	30	26	25	24	24	7	3	3	3	3	3	3
2	0	4	1	1	1	17	17	5	5	3	3	3
3	0	0	4	4	1	1	1	1	1	1	1	1
4	0	0	0	1	3	1	4	4	3	3	3	3
5	0	0	0	0	1	3	1	12	12	12	8	8
6	0	0	0	0	0	1	3	1	1	1	4	4
7	0	0	0	0	0	0	1	3	3	2	1	1
8	0	0	0	0	0	0	0	1	1	3	2	2
9	0	0	0	0	0	0	0	0	1	1	3	2
10	0	0	0	0	0	0	0	0	0	1	1	1
11	0	0	0	0	0	0	0	0	0	0	1	1
12	0	0	0	0	0	0	0	0	0	0	0	1

Table 2 displays the number of observations containing in each cluster for each level up to 12 based on agglomerative hierarchical clustering carried out in 3(c). From the dendrogram and the level table, the four Tech stocks, namely, HWP, IBM, INTC and MSFT form a cluster at level 2, while T and HON are part of the cluster 1 with 26 stocks. T appears as a singleton at level 3, while HWP appears as a singleton at level 5. The cluster dendrogram and level table obtained from this analysis show that only 3 Tech stocks form a cluster at level 2, and IBM is part of the big cluster.

Use the daily returns of the DJ1201 as observations.

```
hc.complete2 = hclust(dist(DJ1201), method="complete")
```

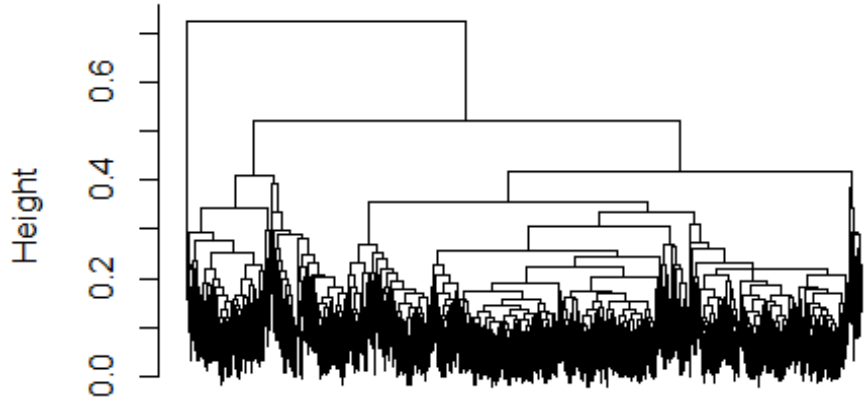

Figure 6 shows the dendrogram resulting from the agglomerative hierarchical cluster analysis of the Dow Jones returns data based on the complete linkage and the Euclidean distance with 1200 daily returns as observations. As there are a significantly large number of observations, the labels of the returns are suppressed as they are indistinguishable.

Show a cluster table by levels up to level 12

```
DJclus3 = data.frame()
for(k in 1:12){
  clustered.df = data.frame(k = k, cluster = cutree(hc.complete2,k))
  DJclus3 = rbind(DJclus3, clustered.df)}
```

Clus	Level											
	1	2	3	4	5	6	7	8	9	10	11	12
1	1200	1196	911	911	888	888	888	888	741	741	741	282
2	0	4	285	285	285	139	137	137	137	137	137	459
3	0	0	4	3	23	146	146	146	147	147	147	137
4	0	0	0	1	3	23	23	21	146	138	138	147
5	0	0	0	0	1	3	3	2	21	21	19	138
6	0	0	0	0	0	1	1	3	2	2	2	19
7	0	0	0	0	0	0	2	1	3	3	3	2
8	0	0	0	0	0	0	0	2	1	8	8	3
9	0	0	0	0	0	0	0	0	2	1	2	8
10	0	0	0	0	0	0	0	0	0	2	1	2
11	0	0	0	0	0	0	0	0	0	0	2	1
12	0	0	0	0	0	0	0	0	0	0	0	2

Table 3 displays the number of observations containing in each cluster for each level up to 12 based on agglomerative hierarchical clustering carried out in 3(d).

According to **Figure 6** and **Table 3**, 4 singletons can be seen clearly on the far left of the dendrogram. At level 2, these four singletons correspond to the dates in which four significant events occurred. These four dates are essentially part of the five dates split at level 2 for the 2527 observations. The only date that is left out is 2000-10-12, which is because the `DJ1201` data range is only up to 2000-06-30.

There are obviously two distinct groups if these four singletons on the far left are ignored. The left-hand group consists of about one-third of the `DJ1201` observations, as seen in level 3 of **Table 3**. On the other hand, the right-hand group seems to consist of one big cluster and a group of 23 singletons, as seen in level 5. This big cluster remains it size until level 8, split again at level 9, and come across one-third division again at level 12. At this level, there is essentially one big cluster accompanied by four small clusters and seven groups of singletons.