STAT 524
HW9

Satoshi Ido
34788706
12/02/2023

12.4. Show that the monotonicity property holds for the similarity coefficients 1, 2, and 3 in Table 12.1.
Hint: $(b + c) = p - (a + d)$. So, for instance,

$$\frac{a + d}{a + d + 2(b + c)} = \frac{1}{1 + 2[p/(a + d) - 1]}$$

This equation relates coefficients 3 and 1. Find analogous representations for the other pairs.

Given $(b+c) = p - (a+d)$,

Coefficients 1 : $\frac{a+d}{p} = \frac{a+d}{a+b+c+d}$

Coefficients 2 : $\frac{2(a+d)}{2(a+d)+b+c} = \frac{2a+2d}{2a+b+c+2d}$

Coefficients 3 : $\frac{a+d}{a+d+2(b+c)} = \frac{a+d}{a+2b+2c+d}$

The coefficients 1 express the ratio of matches to the total numbers of pairs. If a or d increase, the value of coefficients increases, showing monotonicity. Conversely, if b or c increase, the value decreases, preserving monotonicity.

We can apply the same pattern to the coefficient 2 and 3 with different weight for either matches or mismatches, meaning they also hold monotonicity.

12.7. Sample correlations for five stocks were given in Example 8.5. These correlations, rounded to two decimal places, are reproduced as follows:

	JP Morgan	Citibank	Wells Fargo	Royal DutchShell	Exxon Mobil
JP Morgan	1				
Citibank	.63	1			
Wells Fargo	.51	.57	1		
Royal DutchShell	.12	.32	.18	1	
ExxonMobil	.16	.21	.15	.68	1

Treating the sample correlations as similarity measures, cluster the stocks using the single linkage and complete linkage hierarchical procedures. Draw the dendrograms and compare the results.

Since all the similarities are nonnegative definite, we can convert this similarities matrix into a distance matrix by using

$d_{ik} = \sqrt{2(1-\tilde{s}_{ik})}$ where s_{ik} = similarity between i and k and d_{ik} = corresponding distance

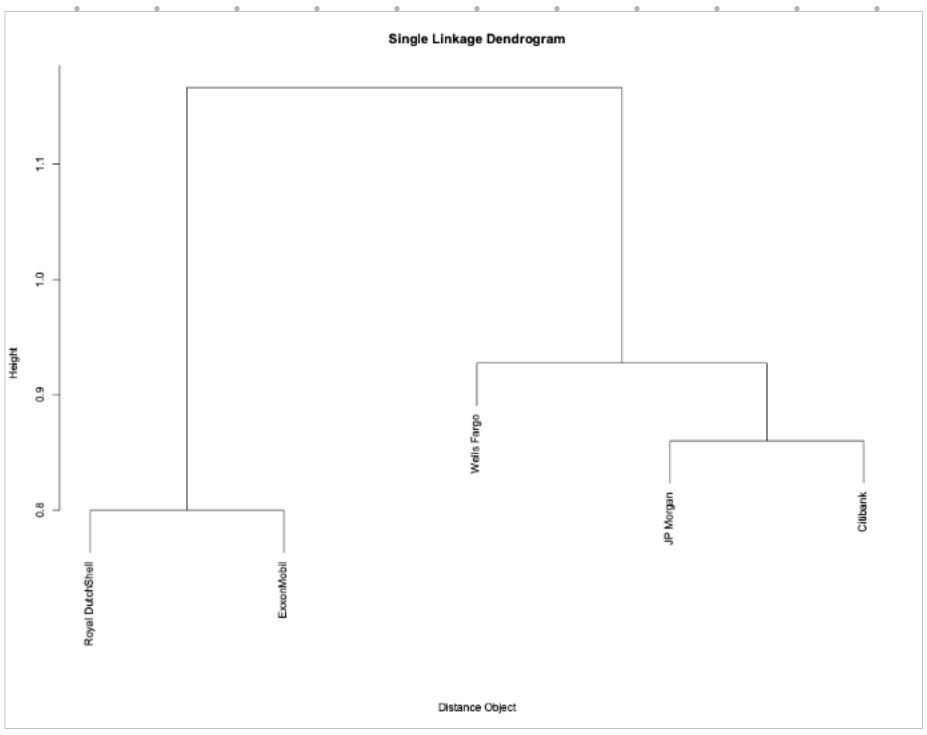
similarity matrix =
$$\begin{bmatrix} 1 & & & & \\ 0.63 & 1 & & & \\ 0.51 & 0.57 & 1 & & \\ 0.12 & 0.32 & 0.18 & 1 & \\ 0.16 & 0.21 & 0.15 & 0.68 & 1 \end{bmatrix}$$

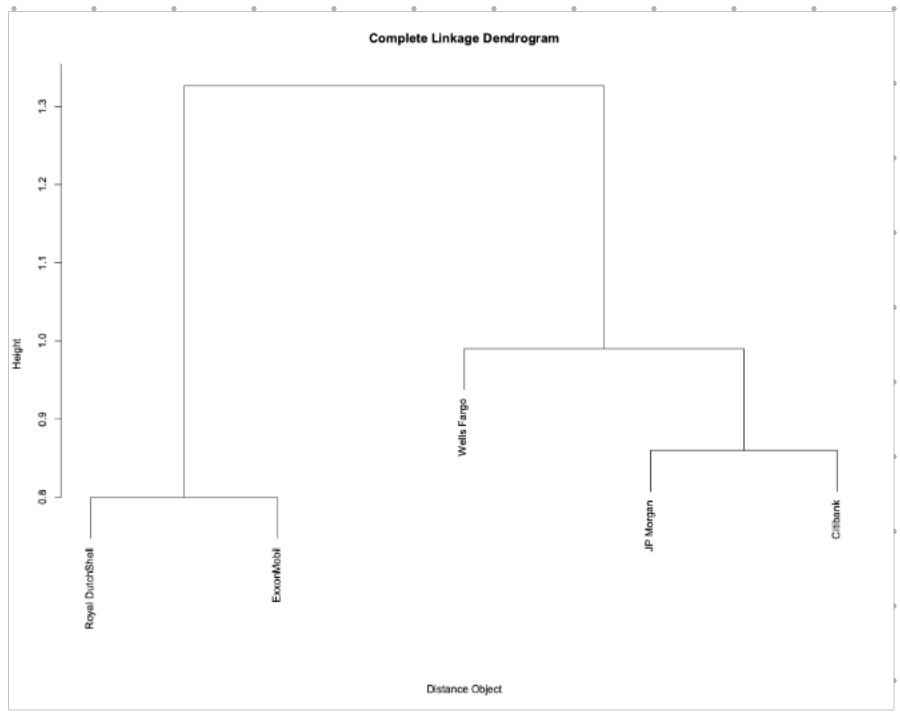
distance matrix =
$$\begin{bmatrix} 0 & & & & \\ 0.86 & 0 & & & \\ 0.99 & 0.93 & 0 & & \\ 1.32 & 1.17 & 1.28 & 0 & \\ 1.29 & 1.25 & 1.30 & 0.80 & 0 \end{bmatrix}$$

Furthermore, we convert this 'distance' matrix into 'dist' object.

'dist' object =
$$\begin{bmatrix} 0.86 & & & \\ 0.99 & 0.93 & & \\ 1.32 & 1.17 & 1.28 & \\ 1.29 & 1.25 & 1.30 & 0.80 \end{bmatrix}$$

Base on the information, we use R to perform single or complete linkage clustering and plot dendrograms for each.





Comparing these plots, the results are same. They both show pairs of {Royal Dutchshell, Exxon Mobil} and {JP Morgan, Citibank} are the closest initial pairs.

12.10. Use Ward's method to cluster the four items whose measurements on a single variable X are given in the following table.

Item	Measurements x
1	2
2	1
3	5
4	8

(a) Initially, each item is a cluster and we have the clusters

$$\{1\} \quad \{2\} \quad \{3\} \quad \{4\}$$

Show that ESS $= 0$, as it must.

(b) If we join clusters $\{1\}$ and $\{2\}$, the new cluster $\{12\}$ has

$$\text{ESS}_1 = \sum (x_j - \bar{x})^2 = (2 - 1.5)^2 + (1 - 1.5)^2 = .5$$

and the ESS associated with the grouping $\{12\}$, $\{3\}$, $\{4\}$ is ESS $= .5 + 0 + 0 = .5$. The *increase* in ESS (loss of information) from the first step to the current step in $.5 - 0 = .5$. Complete the following table by determining the increase in ESS for all the possibilities at step 2.

Clusters			Increase in ESS
{12}	{3}	{4}	.5
{13}	{2}	{4}	
{14}	{2}	{3}	
{1}	{23}	{4}	
{1}	{24}	{3}	
{1}	{2}	{34}	

(c) Complete the last two algamation steps, and construct the dendrogram showing the values of ESS at which the mergers take place.

(a) $ESS_{total} = \sum_{j=1}^{N} (X_j - \bar{X})^T (X_j - \bar{X})$

$ESS_{(1)} = (2-2)^2 = 0$

$ESS_{(2)} = (1-1)^2 = 0$

$ESS_{(3)} = (5-5)^2 = 0$

$ESS_{(4)} = (8-8)^2 = 0$

$\therefore ESS_{total} = 0 + 0 + 0 + 0 = 0$ Q.E.D

(b) For each scenario, the increase in ESS is the ESS of the new cluster minus the sum of ESS of the original clusters, which is zero since each item is in its own cluster initially

For $\{13\}\{2\}\{4\}$,

$$ESS = \sum (x_j - \bar{x})^2 = (2-3.5)^2 + (5-3.5)^2 + 0^2 + 0^2 = 4.5$$

For $\{14\}\{2\}\{3\}$, ESS = 18.0

For $\{1\}\{23\}\{4\}$, ESS = 8.0

For $\{1\}\{24\}\{3\}$, ESS = 24.5

For $\{1\}\{2\}\{34\}$, ESS = 4.5 //

(c) Given the previous, the smallest increase is for merging clusters $\{1\}$ and $\{2\}$ (0.5 increase). So, after the first step, the clusters are $\{12\}, \{3\}, \{4\}$.

The next step is to test the following algamations.

$\{12\}$ and $\{3\}$, $\{12\}$ and $\{4\}$, $\{3\}$ and $\{4\}$. Each ESS is

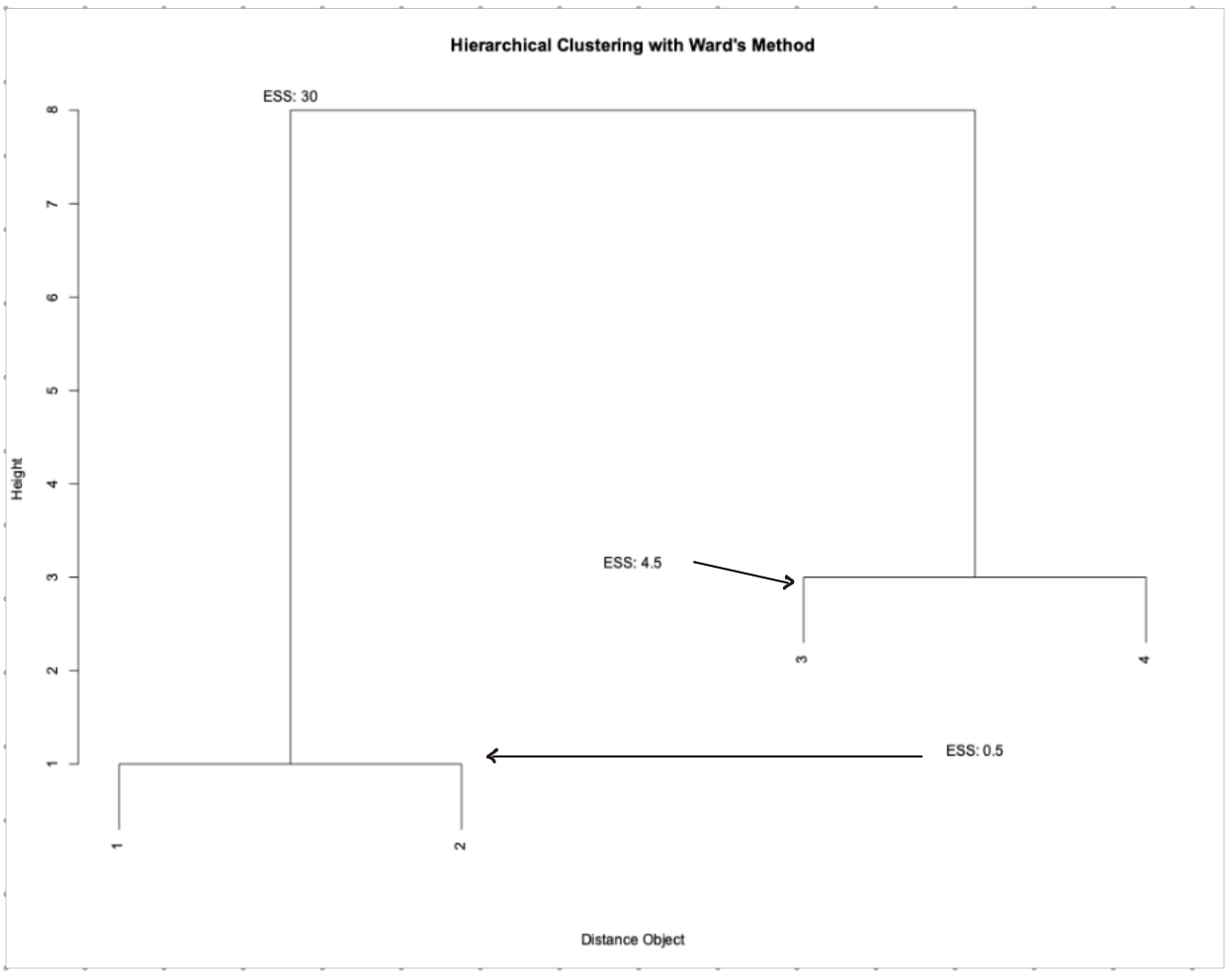
$= 6.6$ $= 25.67$ $= 4.5$

$\therefore$ algamation between $\{3\}$ and $\{4\}$ is chosen

Then, we finally calculate $\{12\}$ and $\{34\}$.

ESS = 30.0.

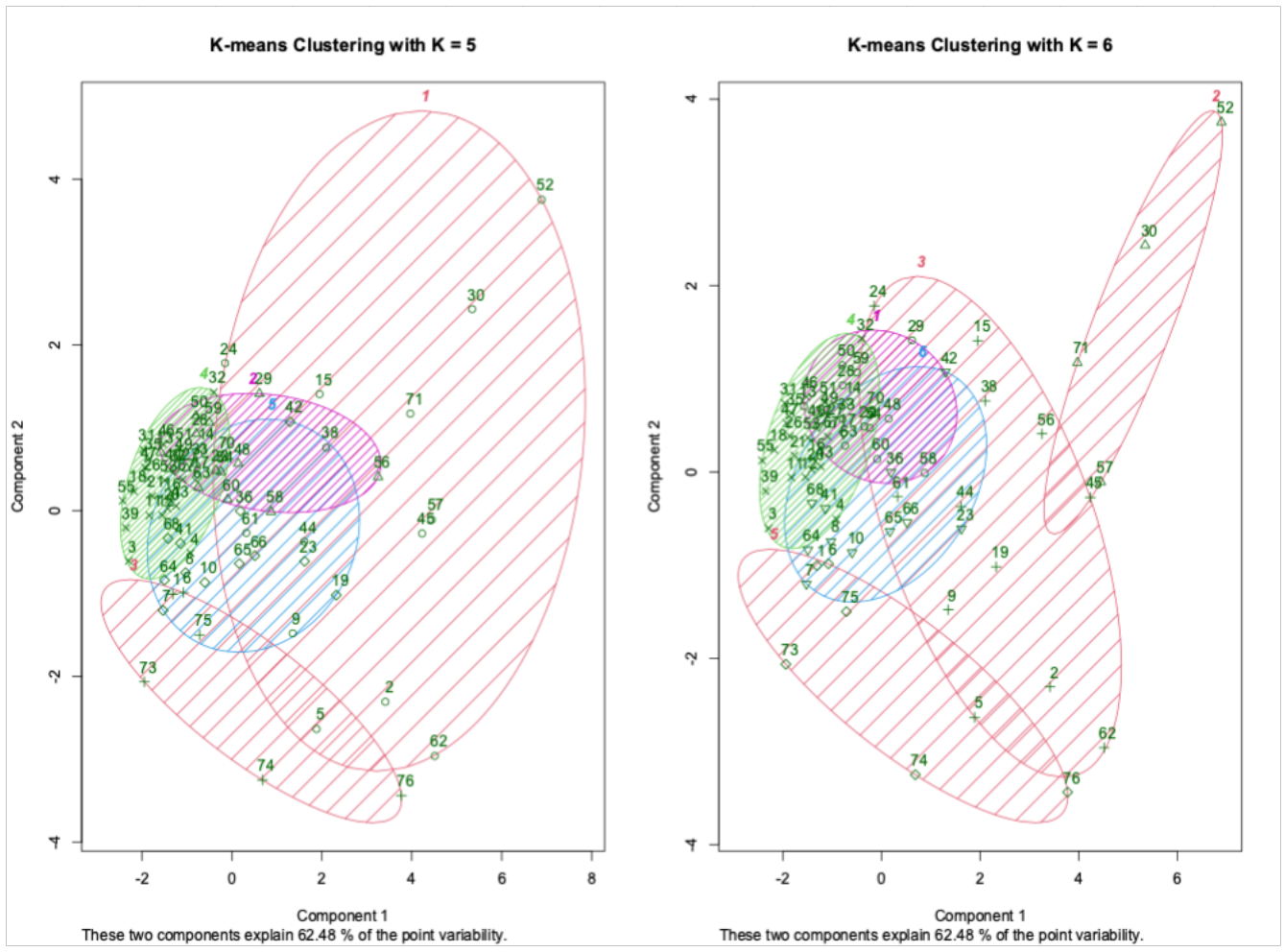
The dendrogram looks like below:



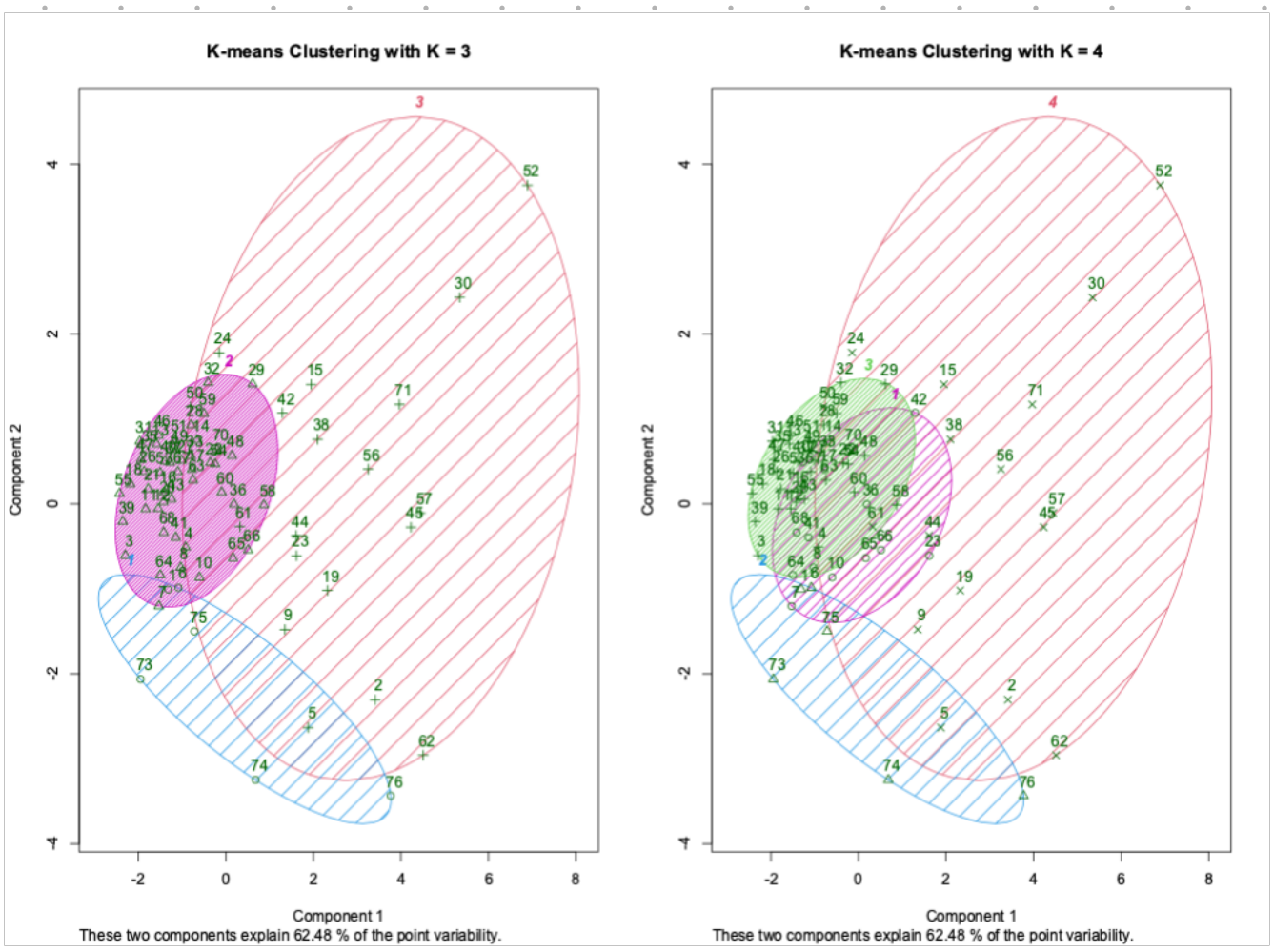
12.28. Using the Mali family farm data in Table 8.7 with the outliers 25, 34, 69 and 72 removed, cluster the farms with the K-means clustering algorithm for $K = 5$ and $K = 6$. Compare the results with those in Exercise 12.26. Is 5 or 6 about the right number of distinct clusters? Discuss.

From the R-outputs of k-mean clustering, we can say K=5 or K=6 are overaly clustered. Based on the Bivariate cluster plots and comparison wit the output of a hierachical method in Exercise 12.26, k=3 is appropriate enough.

Bivariate cluster plots with K=5 or 6.
We can see many clusters are overlapped with each other, meaning this clustering does not work effectively.
Hence, we can explore more k-means analyses with fewer clusters.



For the reference,
Bivariate cluster plots
with K=3 and 4.
They seem to divide
the groups well.



The dendrograms from
Exercise 12.26.
We can see their
algamation patterns act
similarly as k-means.
The both methods
show either k=3 or 4
seem a suitable number
of group.