**Q1.**

**a**) Using the formula:

$$cov(X_j, X_k) = (\frac{1}{n-1}) \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k) \tag{1}$$

or otherwise, obtain an **estimate for the variance-covariance matrix S** for the data tabulated below. The random variables X1, X2 and X3 denote thumb size, index finger length and height of ear, respectively for a few people.

| Person | X1 | X2 | X3 |
|---|---|---|---|
| 1 | 7 | 9 | 6 |
| 2 | 8 | 10 | 7 |
| 3 | 5 | 7 | 8 |
| 4 | 7 | 11 | 7 |

**Solution:**

*Notations:*

The notations which will be used throughout this question are given as follows:

$$\begin{aligned} N &: \text{Number of values in given data} \\ X1_i, X2_i \& X3_i &: \text{Individual value in the first, second and third set of values} \\ \overline{X1}, \overline{X2} \& \overline{X3} &: \text{Average of N values in the first, second \& third data set} \\ (X_i - \overline{X}) &: \text{Deviation from the average} \\ {\sigma_i}^2 &: \text{ Variance} \\ \sigma_{ij} &: \text{Covariance} \end{aligned}$$

*To find:*

The variance-covariance matrix **S**:

$$\begin{bmatrix} {\sigma_x}^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & {\sigma_y}^2 & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & {\sigma_z}^2 \end{bmatrix}$$

*Finding Average:*

$$
\begin{aligned}
\overline{X1} &= (\frac{7+8+5+7}{4}) \\
&= \frac{27}{4} \\
&= 6.75 \\
\overline{X2} &= (\frac{9+10+7+11}{4}) \\
&= \frac{37}{4} \\
&= 9.25 \\
\overline{X3} &= (\frac{6+7+8+7}{4}) \\
&= \frac{28}{4} \\
&= 7
\end{aligned}
$$

*Finding Variance:*

$$
\begin{aligned}
{\sigma_x}^2 &= \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{n-1} \\
&= \frac{\sum_{i=1}^{4}(X1_i - \overline{X1})^2}{4-1} \\
&= \frac{(7-6.75)^2 + (8-6.75)^2 + (5-6.75)^2 + (7-6.75)^2}{3} \\
&= \frac{(0.25)^2 + (1.25)^2 + (-1.75)^2 + (0.25)^2}{3} \\
&= \frac{0.0625 + 1.5625 + 3.0625 + 0.0625}{3} \\
&= \frac{4.75}{3} \\
&= 1.58333
\end{aligned}
$$

$$
\begin{aligned}
{\sigma_y}^2 &= \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{n-1} \\
&= \frac{\sum_{i=1}^{4}(X2_i - \overline{X2})^2}{4-1} \\
&= \frac{(9-9.25)^2 + (10-9.25)^2 + (7-9.25)^2 + (11-9.25)^2}{3} \\
&= \frac{(-0.25)^2 + (0.75)^2 + (-2.25)^2 + (1.75)^2}{3} \\
&= \frac{0.0625 + 1.5625 + 0.5625 + 3.0625}{3} \\
&= \frac{8.75}{3} \\
&= 2.916667
\end{aligned}
$$

$$
\begin{aligned}
{\sigma_z}^2 &= \frac{\sum_{i=1}^{N}(Z_i - \overline{Z})^2}{n-1} \\
&= \frac{\sum_{i=1}^{4}(X3_i - \overline{X3})^2}{4-1} \\
&= \frac{(6-7)^2 + (7-7)^2 + (8-7)^2 + (7-7)^2}{3} \\
&= \frac{(-1)^2 + (0)^2 + (1)^2 + (0)^2}{3} \\
&= \frac{1+0+1+0}{3} \\
&= \frac{2}{3} \\
&= 0.666667
\end{aligned}
$$

**Therefore,** we have the variance as below:

$$
\begin{aligned}
{\sigma_x}^2 &= 1.583333 \\
{\sigma_y}^2 &= 2.916667 \\
{\sigma_z}^2 &= 0.666667
\end{aligned}
$$

*Finding Covariance:*

$$
\begin{aligned}
\sigma_{xy} &= \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \\
&= \frac{\sum_{i=1}^{N}(X1_i - \overline{X1})(X2_i - \overline{X2})}{4-1} \\
&= \frac{(7-6.75)(9-9.25)+(8-6.75)(10-9.25)+(5-6.75)(7-9.25)+(7-6.75)(11-9.25)}{3} \\
&= \frac{(0.25)(-0.25)+(1.25)(0.75)+(-1.75)(-2.25)+(0.25)(1.75)}{3} \\
&= \frac{-0.0625+0.9375+3.9375+0.4375}{3} \\
&= \frac{5.25}{3} \\
&= 1.75
\end{aligned}
$$

Since the Variance-Covariance matrix is symmetric,

$$\sigma_{xy} = \sigma_{yx}$$

[Note: Checking the value of $\sigma_{yx}$

$$
\begin{aligned}
\sigma_{yx} &= \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})(X_i - \overline{X})}{n-1} \\
&= \frac{\sum_{i=1}^{N}(X2_i - \overline{X2})(X1_i - \overline{X1})}{4-1} \\
&= \frac{(9-9.25)(7-6.75)+(10-9.25)(8-6.75)+(7-9.25)(5-6.75)+(11-9.25)(7-6.75)}{3} \\
&= \frac{(-0.25)(0.25)+(0.75)(1.25)+(-2.25)(-1.75)+(1.75)(0.25)}{3} \\
&= \frac{-0.0625+0.9375+3.9375+0.4375}{3} \\
&= \frac{5.25}{3} \\
&= 1.75]
\end{aligned}
$$

$$
\begin{aligned}
\sigma_{xz} &= \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Z_i - \overline{Z})}{n-1} \\
&= \frac{\sum_{i=1}^{N}(X1_i - \overline{X1})(X3_i - \overline{X3})}{4-1} \\
&= \frac{(7-6.75)(6-7)+(8-6.75)(7-7)+(5-6.75)(8-7)+(7-6.75)(7-7)}{3} \\
&= \frac{(0.25)(-1)+(1.25)(0)+(-1.75)(1)+(0.25)(0)}{3} \\
&= \frac{(-0.25)+(0)+(-1.75)+(0)}{3} \\
&= \frac{-0.25-1.75}{3} \\
&= \frac{-2}{3} \\
&= -0.666667
\end{aligned}
$$

Since the Variance-Covariance matrix is symmetric,

$$\sigma_{xz} = \sigma_{zx}$$

[ Note:Checking the value of $\sigma_{zx}$

$$
\begin{aligned}
\sigma_{xz} &= \frac{\sum_{i=1}^{N}(Z_i - \overline{Z})(X_i - \overline{X})}{n-1} = \frac{\sum_{i=1}^{N}(X3_i - \overline{X3})(X1_i - \overline{X1})}{4-1} \\
&= \frac{(6-7)(7-6.75)+(7-7)(8-6.75)+(8-7)(5-6.75)+(7-7)(7-6.75)}{3} \\
&= \frac{(-1)(0.25)+(0)(1.25)+(1)(-1.75)+(0)(0.25)}{3} \\
&= \frac{(-0.25)+(0)+(-1.75)+(0)}{3} \\
&= \frac{-0.25-1.75}{3} \\
&= \frac{-2}{3} \\
&= -0.666667]
\end{aligned}
$$

$$\begin{aligned}
\sigma_{yz} &= \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})(Z_i - \overline{Z})}{n-1} \\
&= \frac{\sum_{i=1}^{N}(X2_i - \overline{X2})(X3_i - \overline{X3})}{4-1} \\
&= \frac{(9-9.25)(6-7)+(10-9.25)(7-7)+(7-9.25)(8-7)+(11-9.25)(7-7)}{3} \\
&= \frac{(-0.25)(-1)+(0.75)(0)+(-2.25)(1)+(1.75)(0)}{3} \\
&= \frac{(0.25)+(0)+(-2.25)+(0)}{3} \\
&= \frac{0.25-2.25}{3} \\
&= \frac{(-2)}{3} \\
&= -0.666667
\end{aligned}$$

Since the Variance-Covariance matrix is symmetric,

$$\sigma_{yz} = \sigma_{zy}$$

[Note:Checking the value of $\sigma_{zy}$

$$\begin{aligned}
\sigma_{zy} &= \frac{\sum_{i=1}^{N}(Z_i - \overline{Z})(Y_i - \overline{Y})}{n-1} = \frac{\sum_{i=1}^{N}(X3_i - \overline{X3})(X2_i - \overline{X2})}{4-1} \\
&= \frac{(6-7)(9-9.25)+(7-7)(10-9.25)+(8-7)(7-9.25)+(7-7)(11-9.25)}{3} \\
&= \frac{(-1)(-0.25)+(0)(0.75)+(1)(-2.25)+(0)(1.75)}{3} \\
&= \frac{(0.25)+(0)+(-2.25)+(0)}{3} \\
&= \frac{0.25-2.25}{3} \\
&= \frac{(-2)}{3} = -0.666667]
\end{aligned}$$

Finally,

we got the required Variance-Covariance matrix S:

$$\begin{bmatrix} {\sigma_x}^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & {\sigma_y}^2 & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & {\sigma_z}^2 \end{bmatrix} = \begin{bmatrix} 1.58333 & 1.75000 & -0.666667 \\ 1.75000 & 2.916667 & -0.66667 \\ -0.66667 & -0.66667 & 0.66667 \end{bmatrix}$$

**b**) Use R to perform a principal component analysis on scaled features, and answer the following:

- provide the R output of the laodings for the first principal component?
- compute the proportion of variance explained (PVE) by the **first two** principal components.

**Solution:**

Firstly, obtaining the variance-covariance matrix and correlation matrix on scaled version of the given data as follows:

```
Console   Terminal ×   Background Jobs ×
R 4.2.2 · ~/
> Scaled_DF <- scale(DF)
> Scaled_DF
             x1        X2        X3
[1,]  0.1986799 -0.146385 -1.224745
[2,]  0.9933993  0.439155  0.000000
[3,] -1.3907590 -1.317465  1.224745
[4,]  0.1986799  1.024695  0.000000
attr(,"scaled:center")
  x1   X2   X3
6.75 9.25 7.00
attr(,"scaled:scale")
       x1        X2        X3
1.2583057 1.7078251 0.8164966
> cov(Scaled_DF)
           x1         X2         X3
x1  1.0000000  0.8143451 -0.6488857
X2  0.8143451  1.0000000 -0.4780914
X3 -0.6488857 -0.4780914  1.0000000
> cor(Scaled_DF)
           x1         X2         X3
x1  1.0000000  0.8143451 -0.6488857
X2  0.8143451  1.0000000 -0.4780914
X3 -0.6488857 -0.4780914  1.0000000
```

Figure 1: **Scaled features of the given data**

Providing the R output of the loadings for the first principal component as follows:

```
Console   Terminal ×   Background Jobs ×
R 4.2.2 · ~/
> pca<-prcomp(DF, scale. =TRUE)
> names(pca)
[1] "sdev"     "rotation" "center"   "scale"    "x"
> pca$rotation
          PC1       PC2        PC3
x1 -0.6233082 0.1654420  0.7642747
X2 -0.5811197 0.5559861 -0.5942890
X3  0.5232464 0.8145603  0.2504091
> pca$rotation[,1] # the loadings for the first principal component(PC1)
        x1         X2         X3
-0.6233082 -0.5811197  0.5232464
```

Figure 2: **Loadings of the First Principal Component**

Calculating the proportion variance explained by first two PCs

```
Console   Terminal ×   Background Jobs ×
R 4.2.2 · ~/
> (PVE=pca$sdev^2 / sum(pca$sdev^2))
[1] 0.76798138 0.18062719 0.05139144
```

Figure 3: **PVE by first two PCs**

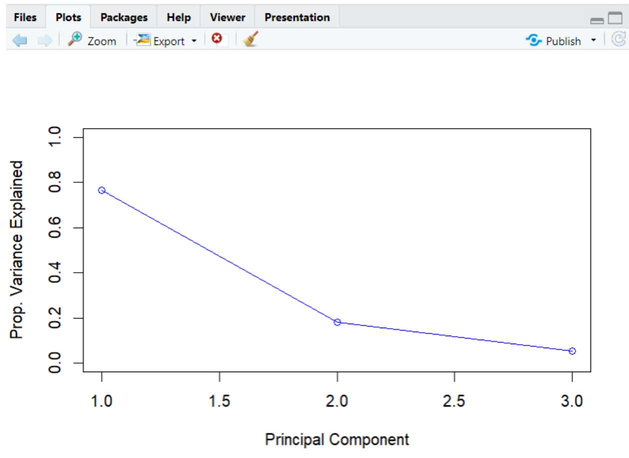


Figure 4: **Plot of PVE against PCs**

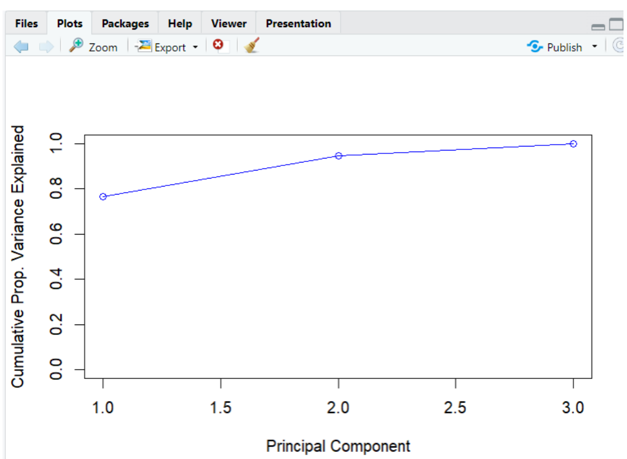


Figure 5: **Plot of cumulative PVE against PCs**

# Chapter 1

# Qn2

Perform **complete-linkage** agglomerative cluster analysis on the dissimilarity matrix below, and draw the associated dendrogram.

| | obs1 | obs2 | obs3 | obs4 | obs5 | obs6 | obs7 | obs8 |
|---|---|---|---|---|---|---|---|---|
| obs1 | 0 | 5.8 | 5.5 | 6.2 | 3 | 8.6 | 2.6 | 5.7 |
| obs2 | 5.8 | 0 | 6.9 | 3.4 | 8.2 | 2.1 | 5.6 | 12.4 |
| obs3 | 5.5 | 6.9 | 0 | 5.2 | 10.2 | 7.1 | 3.5 | 4.6 |
| obs4 | 6.2 | 3.4 | 5.2 | 0 | 14.3 | 8.4 | 10.4 | 9.7 |
| obs5 | 3 | 8.2 | 10.2 | 14.3 | 0 | 2.3 | 14.6 | 5.3 |
| obs6 | 8.6 | 2.1 | 7.1 | 8.4 | 2.3 | 0 | 5.1 | 7.3 |
| obs7 | 2.6 | 5.6 | 3.5 | 10.4 | 14.6 | 5.1 | 0 | 4.8 |
| obs8 | 5.7 | 12.4 | 4.6 | 9.7 | 5.3 | 7.3 | 4.8 | 0 |

**Solution**:

Step 1: given dissimilarity matrix

| | obs1 | obs2 | obs3 | obs4 | obs5 | obs6 | obs7 | obs8 |
|---|---|---|---|---|---|---|---|---|
| obs1 | 0 | | | | | | | |
| obs2 | 5.8 | 0 | | | | | | |
| obs3 | 5.5 | 6.9 | 0 | | | | | |
| obs4 | 6.2 | 3.4 | 5.2 | 0 | | | | |
| obs5 | 3 | 8.2 | 10.2 | 14.3 | 0 | | | |
| obs6 | 8.6 | 2.1 | 7.1 | 8.4 | 2.3 | 0 | | |
| obs7 | 2.6 | 5.6 | 3.5 | 10.4 | 14.6 | 5.1 | 0 | |
| obs8 | 5.7 | 12.4 | 4.6 | 9.7 | 5.3 | 7.3 | 4.8 | 0 |

Step 2:

*a) Identifying the pair of clusters that has the least distance:*

obs2 and obs6 has the least distance of **2.1**. So obs 2 and obs6 form a new cluster. We can use **(obs2,obs6)** to represent such a new cluster.

*(b) updating the dissimilarity matrix based on* ***complete-linkage***

Calculating the distance affected by the new cluster:

$$
\begin{aligned}
d((obs2, obs6), ob1)) &= max\{d(obs2, obs1), d(obs6, obs1)\} \\
&= max5.8, 8.6 \\
&= 8.6 \\
d((obs2, obs6), ob3)) &= max\{d(obs2, obs3), d(obs6, obs3)\} \\
&= max6.9, 7.1 \\
&= 7.1 \\
d((obs2, obs6), ob4)) &= max\{d(obs2, obs4), d(obs6, obs4)\} \\
&= max3.4, 8.4 \\
&= 8.4 \\
d((obs2, obs6), ob5)) &= max\{d(obs2, obs5), d(obs6, obs5)\} \\
&= max8.2, 2.3 \\
&= 8.2 \\
d((obs2, obs6), ob7)) &= max\{d(obs2, obs7), d(obs6, obs7)\} \\
&= max5.6, 5.1 \\
&= 5.6 \\
d((obs2, obs6), ob8)) &= max\{d(obs2, obs8), d(obs6, obs8)\} \\
&= max12.4, 7.3 \\
&= 12.4
\end{aligned}
$$

| | obs1 | (obs2,obs6) | obs3 | obs4 | obs5 | obs7 | obs8 |
|---|---|---|---|---|---|---|---|
| obs1 | 0 | | | | | | |
| (obs2,obs6) | 8.6 | 0 | | | | | |
| obs3 | 5.5 | 7.1 | 0 | | | | |
| obs4 | 6.2 | 8.4 | 5.2 | 0 | | | |
| obs5 | 3 | 8.2 | 10.2 | 14.3 | 0 | | |
| obs7 | 2.6 | 5.6 | 3.5 | 10.4 | 14.6 | 0 | |
| obs8 | 5.7 | 12.4 | 4.6 | 9.7 | 5.3 | 0 | |

Step 3

*a) Identifying the pair of clusters that has the least distance:*

obs1 and obs7 has the least distance of **2.6**. So obs 2 and obs6 form a new cluster. We can use **(obs1,obs7)** to represent such a new cluster.

*(b) updating the dissimilarity matrix based on **complete-linkage***

Calculating the distance affected by the new cluster:

$$\begin{aligned}
d((obs1, obs7), (obs2, obs6)) &= max\{d(((obs2, obs6), obs1), (obs2, obs6), obs7))\} \\
&= max8.6, 5.6 \\
&= 8.6 \\
d((obs1, obs7), ob3)) &= max\{d(obs1, obs3), d(obs7, obs3)\} \\
&= max5.5, 3.5 \\
&= 5.5 \\
d((obs1, obs7), ob4)) &= max\{d(obs1, obs4), d(obs7, obs4)\} \\
&= max6.2, 10.4 \\
&= 10.4 \\
d((obs1, obs7), ob5)) &= max\{d(obs1, obs5), d(obs7, obs5)\} \\
&= max3, 14.6 \\
&= 14.6 \\
d((obs1, obs7), ob8)) &= max\{d(obs1, obs8), d(obs1, obs8)\} \\
&= max5.7, 4.8 \\
&= 5.7
\end{aligned}$$

| | (obs1,obs7) | (obs2,obs6) | obs3 | obs4 | obs5 | obs8 |
|---|---|---|---|---|---|---|
| (obs1,obs7) | 0 | | | | | |
| (obs2,obs6) | 8.6 | 0 | | | | |
| obs3 | 5.5 | 7.1 | 0 | | | |
| obs4 | 10.4 | 8.4 | 5.2 | 0 | | |
| obs5 | 14.6 | 8.2 | 10.2 | 14.3 | 0 | |
| obs8 | 5.7 | 12.4 | 4.6 | 9.7 | 5.3 | 0 |

Step 4

*a) Identifying the pair of clusters that has the least distance:*

obs3 and obs8 has the least distance of 4.6. So obs 2 and obs6 form a new cluster. We can use **(obs3,obs8)** to represent such a new cluster.

*(b) updating the dissimilarity matrix based on* ***complete-linkage***

Calculating the distance affected by the new cluster:

$$\begin{aligned}
d((obs3, obs8), (obs1, obs7)) &= max\{d(((obs1, obs7), obs3), ((obs1, obs7), obs8))\} \\
&= max5.5, 5.7 \\
&= 5.7 \\
d((obs3, obs8), (ob2, obs6))) &= max\{d((obs3, (obs2, obs6)), (obs8, (obs2, obs6)))\} \\
&= max7.1, 12.4 \\
&= 12.4 \\
d((obs3, obs8), ob4)) &= max\{d(obs3, obs4), d(obs8, obs4)\} \\
&= max5.2, 9.7 \\
&= 9.7 \\
d((obs3, obs8), ob5)) &= max\{d(obs3, obs5), d(obs8, obs5)\} \\
&= max10.2, 5.3 \\
&= 10.2
\end{aligned}$$

| | (obs1,obs7) | (obs2,obs6) | (obs3,obs8) | obs4 | obs5 |
|---|---|---|---|---|---|
| (obs1,obs7) | 0 | | | | |
| (obs2,obs6) | 8.6 | 0 | | | |
| (obs3,obs8) | 5.7 | 12.4 | 0 | | |
| obs4 | 10.4 | 8.4 | 9.7 | 0 | |
| obs5 | 14.6 | 8.2 | 10.2 | 14.3 | 0 |

Step 5

*a) Identifying the pair of clusters that has the least distance:*

(obs3,obs8) and (obs1,obs7) has the least distance of **5.7**. So (obs3,obs8) and (obs1,obs7) form a new cluster. We can use **((obs3,obs8),(obs1,obs7))** to represent such a new cluster.

*(b) updating the dissimilarity matrix based on* ***complete-linkage***

Calculating the distance affected by the new cluster:

$$\begin{aligned} d(((obs3, obs8), (obs1, obs7)), (obs2, obs6)) &= max\{d(((obs1, obs7), (obs2, obs6)), ((obs2, obs6), (obs3, ob \\ &= max8.6, 12.4 \\ &= 12.4 \\ d(((obs3, obs8), (ob1, obs7)), obs4) &= max\{d((obs4, (obs3, obs8)), (obs4, (obs1, obs7)))\} \\ &= max10.4, 9.7 \\ &= 10.4 \\ d(((obs3, obs8), (ob1, obs7)), obs5)) &= max\{d((obs5, (obs3, obs8)), (obs5, (obs1, obs7)))\} \\ &= max14.6, 10.2 \\ &= 14.6 \end{aligned}$$

| | [(obs1,obs7),(obs3,obs8)] | (obs2,obs6) | obs4 | obs5 |
|---|---|---|---|---|
| [(obs1,obs7),(obs3,obs8)] | 0 | | | |
| (obs2,obs6) | 2.4 | 0 | | |
| obs4 | 10.4 | 8.4 | 0 | |
| obs5 | 14.6 | 8.2 | 14.3 | 0 |

Step 6

*a) Identifying the pair of clusters that has the least distance:*

(obs2,obs6) and obs5 has the least distance of **8.2**. So (obs 2,obs6) and obs5 form a new cluster. We can use **((obs2,obs6),obs5)** to represent such a new cluster.

*(b) updating the dissimilarity matrix based on **complete-linkage***

Calculating the distance affected by the new cluster:

$$\begin{aligned} d((((obs3,obs8),(obs1,obs7)),(obs2,obs6)),obs5) &= max\{d((((obs3,obs8),(obs1,obs7)),(obs2,obs6)), \\ &= max12.4, 14.6 \\ &= 14.6 \\ d(((obs2,obs6),obs4),obs5) &= max\{d(((obs2,obs6),obs4),((obs2,obs6),obs5))\} \\ &= max8.4, 14.3 \\ &= 14.3 \end{aligned}$$

| | [(obs1,obs7),(obs3,obs8)] | {(obs2,obs6),obs5} | obs4 |
|---|---|---|---|
| [(obs1,obs7),(obs3,obs8)] | 0 | | |
| {(obs2,obs6),obs5} | 14.6 | 0 | |
| obs4 | 10.4 | 14.3 | 0 |

Step 7

*a) Identifying the pair of clusters that has the least distance:*

((obs1,obs7),(obs3,obs8)) and obs4 has the least distance of **10.4**. So (obs 2,obs6) and obs5 form a new cluster. We can use **(((obs1,obs7),(obs3,obs8)),obs4)** to represent such a new cluster.

*(b) updating the dissimilarity matrix based on **complete-linkage***

Calculating the distance affected by the new cluster:

$$
\begin{aligned}
d((((obs3, obs8), (obs1, obs7), obs4), (((obs2, obs6), obs5), obs4)) &= max\{d((((obs3, obs8), (obs1, obs7)), ( \\
&= max 14.6, 14.3 \\
&= 14.6
\end{aligned}
$$

| | {[(obs1,obs7),(obs3,obs8)],obs4} | {(obs2,obs6),obs5} |
|---|---|---|
| {[(obs1,obs7),(obs3,obs8)],obs4} | 0 | |
| {(obs2,obs6),obs5} | 14.6 | 0 |