Multivariate Statistical Analysis

Homework 6

Lucas Fellmeth, Helen Kafka, Sven Bergmann

04/17/24

Problem 1

a)

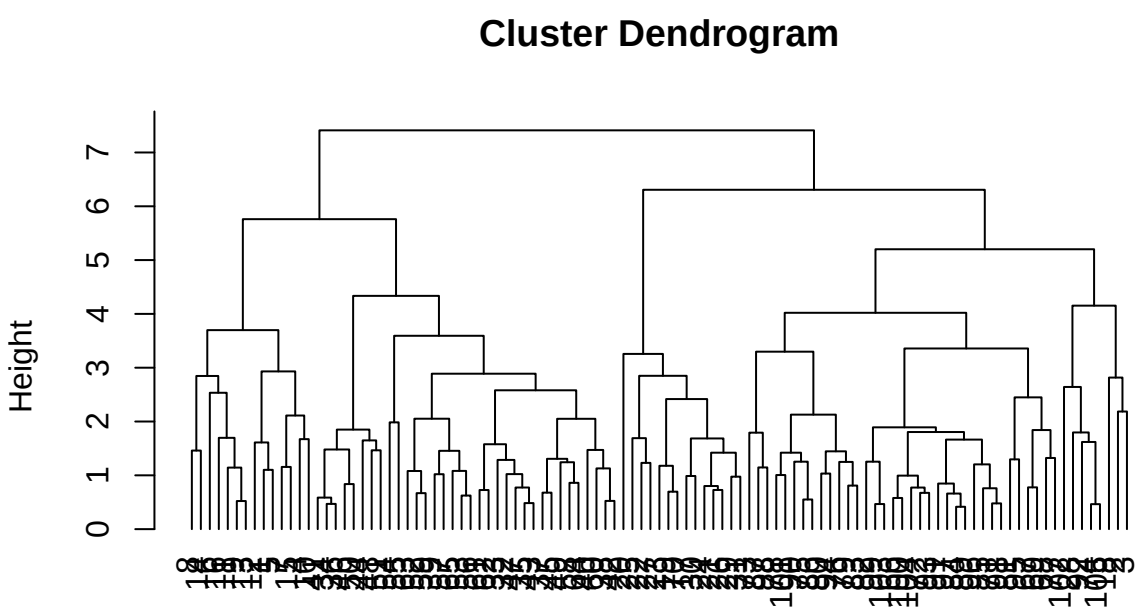
Carry out a hierarchical agglomerative cluster analysis of the `primate_scapulae` data using complete linkage. Use the variables AD.BD, AD.CD, EA.CD, Dx.CD, SH.ACR, EAD and β only. (Remember to scale the variables prior to clustering.) Does the dendrogram suggest a definite number of clusters as the best solution? If not, what would be reasonable solutions?

```
primate_sculpae <- read.csv(file = "../Data_csv/primate_scapulae.csv")
```

```
primate_sculpae$class <- factor(primate_sculpae$class)
primate_sculpae <- subset(primate_sculpae, select = c(AD.BD,
    AD.CD, EA.CD, Dx.CD, SH.ACR, EAD, beta, class))
primate_sculpae[, 1:7] <- scale(primate_sculpae[, 1:7])
head(primate_sculpae)
```

```
##        AD.BD     AD.CD      EA.CD       Dx.CD     SH.ACR       EAD       beta
## 1  0.7559403 2.2773065 -1.7108283 -0.18844057  0.9354432 0.4734640 -1.6063579
## 2 -0.1469654 0.5632474 -0.6876050  0.26337166  1.4835153 1.0373247 -1.1195828
## 3 -0.4403327 0.2518165 -0.4805998 -1.74641377  0.9004598 0.9433479 -0.7139369
## 4  1.0474586 3.5729544 -1.7607637 -0.26633923 -0.1373790 0.2855104 -1.7686163
## 5  0.6080240 1.7542926 -1.0444167  0.09199461  0.3057432 0.4734640 -1.6063579
## 6 -0.1586754 1.5308231 -1.1951310 -1.62177591 -1.1285734 2.4469765 -1.6063579
##       class
## 1 Hylobates
## 2 Hylobates
## 3 Hylobates
## 4 Hylobates
## 5 Hylobates
## 6 Hylobates
```

```
D <- dist(primate_sculpae[, 1:7])
hc <- hclust(D, method = "complete")
plot(hc, hang = -1)
```



The dendrogram does not suggest a definite number of clusters as best solution, instead, we would probably cut off at `height = 5` (k = 5 clusters) because it seems that the 5 clusters survive the longest.

b)

Decide on a reasonable solution in part (a) and assess that solution using a silhouette plot.

```
clusters <- cutree(hc, k = 5)
plot(cluster::silhouette(clusters, dist = D))
```

The average silhouette width is 0.37 which is not too good.

c)

Is there any correspondence between the clusters obtained above and the primate species as indicated by the variable class?

```
table(clusters, primate_sculpae$class)
```

```
##
## clusters Gorilla Homo Hylobates Pan Pongo
##        1       0    0        14   0     0
##        2       0    4         2   0     2
##        3       0    1         0   0    13
##        4      14    0         0  20     0
##        5       0   35         0   0     0
```

- Cluster 1 has almost all values of the group “Hylobates” and also just values of this group.
- Cluster 2 is kind of an intermediate cluster, no group really belongs there. Some values of the groups “Homo”, “Hylobates” and “Pongo” are in there.
- Cluster 3 has most of the group “Pongo” but also one value of the group “Homo”.
- Cluster 4 has two groups, “Gorillas” as well as “Pan”.
- Cluster 5 fully belongs to the group “Homo”.