eNote 6

Cluster Analysis

Indhold

6.1 Reading material

Both of the chemometrics books:

- Ron Wehrens (2012). Chemometrics With R: Multivariate Data Analysis in the Natural Sciences and Life Sciences. Springer, Heidelberg.(Chapter 6)
- K. Varmuza and P. Filzmoser (2009). Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press. (Chapter 6)

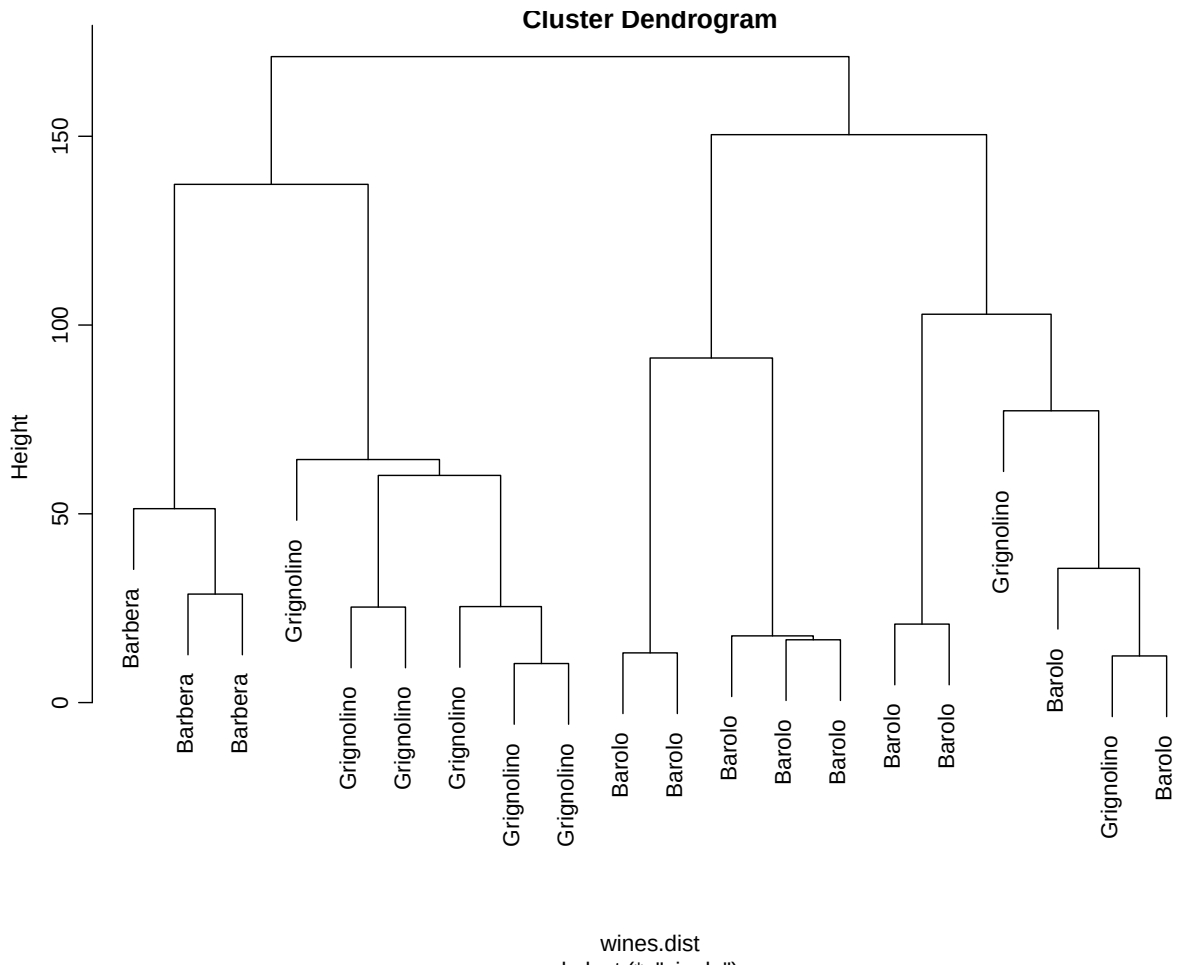
include a Chapter on clustering methods. We will primarily work with hierarchical clustering methods, so in the Wehrens book, read page 79-84 (including R-examples). In the Varmuza-book, section 6.4 (3 pages).

6.2 Example: Wine data

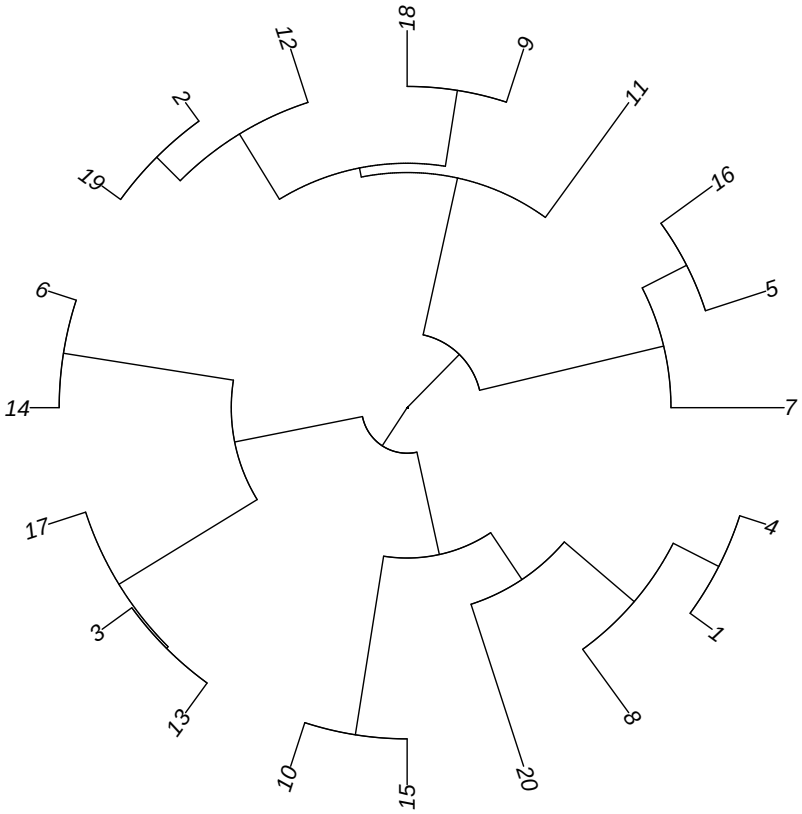
Check the help of the functions `dist` and `hclust`: (check the `method` option of each function to see the different possibilities of computing distances and making hierarchical clustering based on these two functions)

```
?dist
?hclust
```

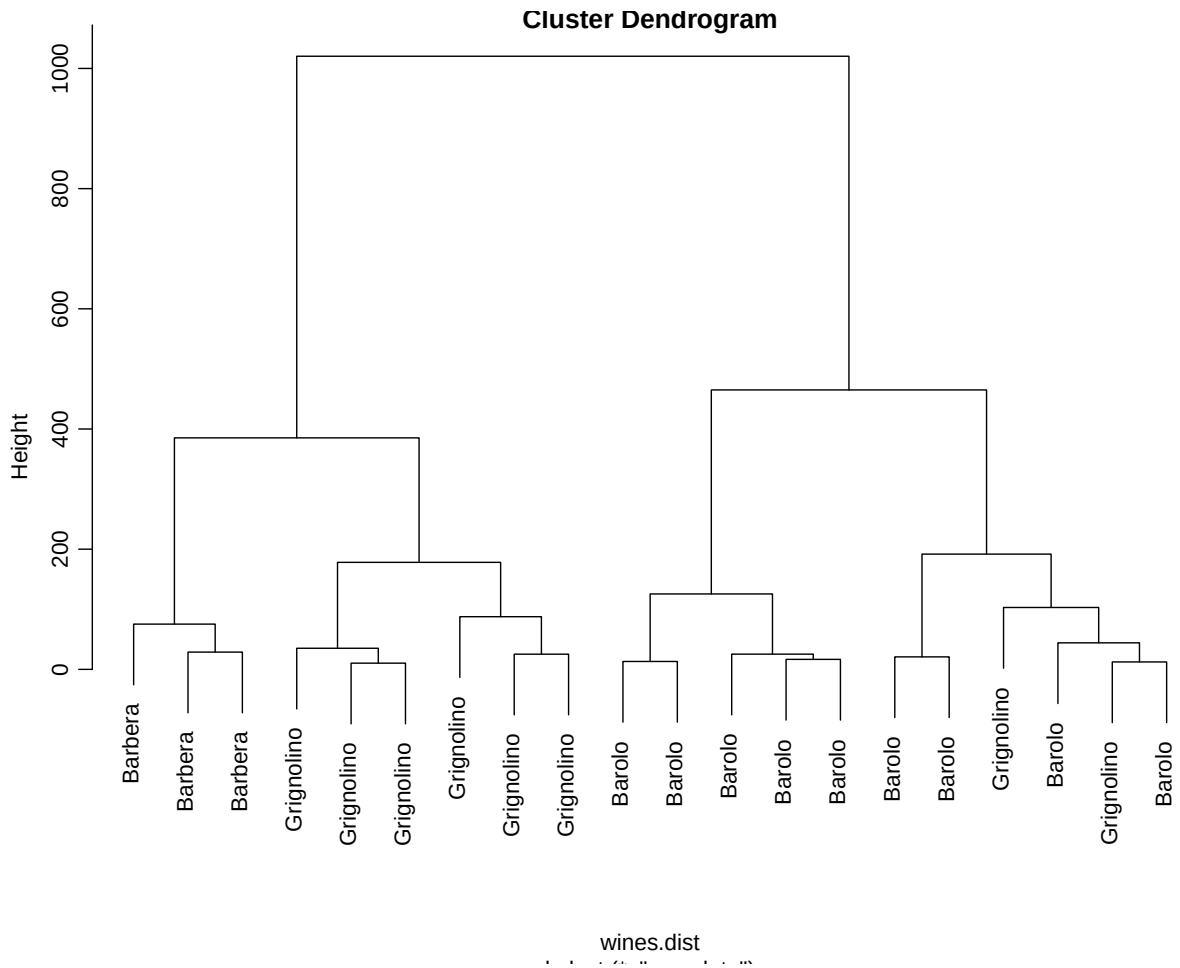
```
library(ChemometricsWithR)
data(wines)

## Distance matrix for a subset of data:
subset <- sample(nrow(wines), 20)
wines.dist <- dist(wines[subset,])

## Single linkage clustering:
wines.hcsingle <- hclust(wines.dist, method = "single")
plot(wines.hcsingle, labels = vintages[subset])
```



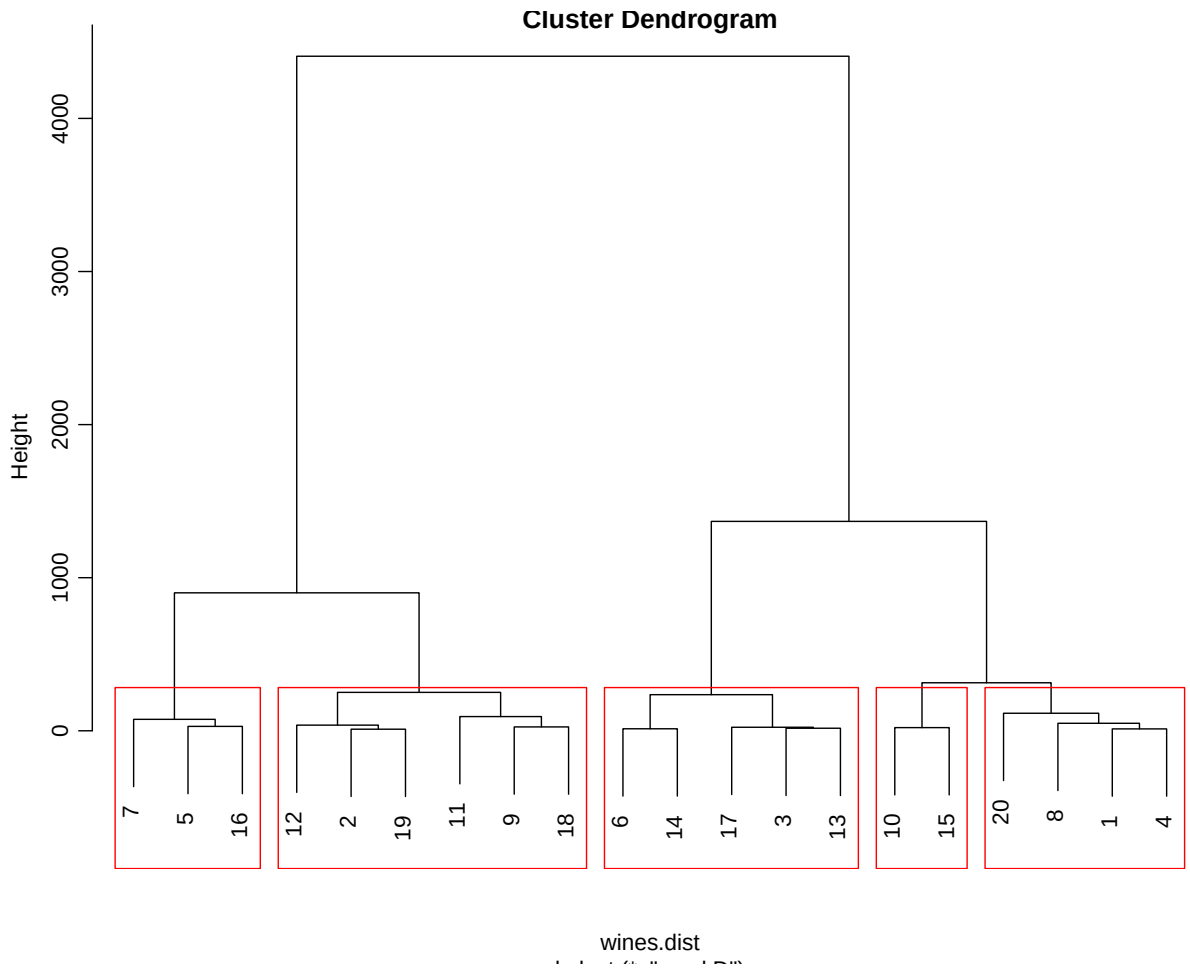
```
## Creates plot of phylogenetic tree in a fan form.
library(ape)
plot(as.phylo(wines.hcsingle), type = "fan", cex = 1)
```




```
## Complete Linkage clustering:
wines.hccomplete <- hclust(wines.dist, method = "complete")
plot(wines.hccomplete, labels = vintages[subset])
```




```r
# Ward Hierarchical Clustering
wines.hcward <- hclust(wines.dist, method="ward.D")
plot(wines.hcward) # display dendogram
groups <- cutree(wines.hcward, k=5) # cut tree into 5 clusters
# draw dendogram with red borders around the 5 clusters
rect.hclust(wines.hcward, k=5, border="red")
```




```
## Clustering, cf. chapter in book:
wines.cl.single <- cutree(wines.hcsingle, h = 3.3)
table(wines.cl.single, vintages[subset])
```

```
wines.cl.single Barbera Barolo Grignolino
              1        0      0          1
              2        0      0          1
              3        0      1          0
              4        0      1          0
              5        1      0          0
              6        0      1          0
              7        1      0          0
              8        0      1          0
              9        0      0          1
              10       0      1          0
```

```
             11       0      0          1
             12       0      0          1
             13       0      1          0
             14       0      1          0
             15       0      1          0
             16       1      0          0
             17       0      1          0
             18       0      0          1
             19       0      0          1
             20       0      0          1
```

```
wines.dist <- dist(wines)
wines.hcsingle <- hclust(wines.dist, method = "single")
table(vintages, cutree(wines.hcsingle, k = 3))
```

```
vintages      1  2  3
  Barbera    48  0  0
  Barolo     52  5  1
  Grignolino 71  0  0
```

```
wines.hccomplete <- hclust(wines.dist, method = "complete")
table(vintages, cutree(wines.hccomplete, k = 3))
```

```
vintages      1  2  3
  Barbera     0 21 27
  Barolo     42 16  0
  Grignolino  0 15 56
```

In addition, we can find a coefficient measuring the amount of cluster structure, the "agglomerative coefficient", ac:

```
library(cluster)
wines.agness <- agnes(wines.dist, method = "single")
wines.agnesa <- agnes(wines.dist, method = "average")
wines.agnesc <- agnes(wines.dist, method = "complete")
```

```
cbind(wines.agness$ac, wines.agnesa$ac, wines.agnesc$ac)
```

```
          [,1]      [,2]      [,3]
[1,] 0.9153982 0.9785233 0.9899342
```

Or we can compute the cophenetic correlation:

```
## Computing cophenetic correlation:

cor(wines.dist, cophenetic(wines.hcsingle))
```

```
[1] 0.7779217
```

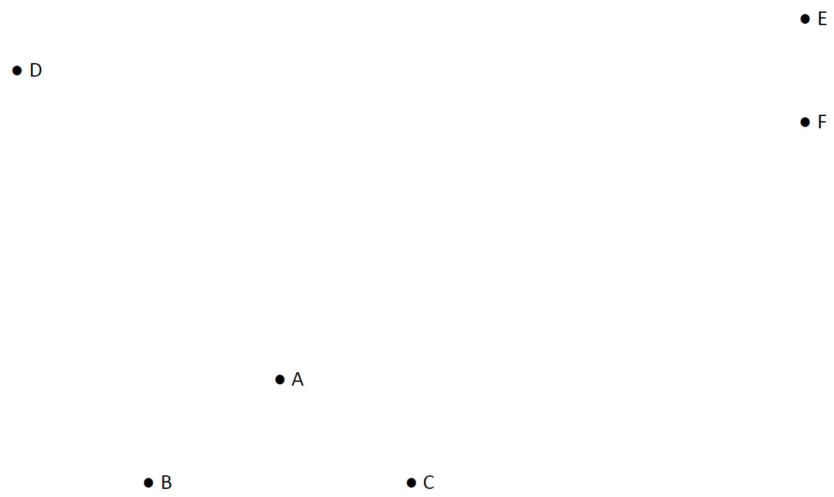
```
cor(wines.dist, cophenetic(wines.hccomplete))
```

```
[1] 0.7973172
```

6.3 Exercises

|||| Exercise 1 Cluster Analysis - hands-on

a) Make a cluster analysis on paper on the 6 points (A-F):



and write the dendrograms by hand, using UPGMA clustering (2 dimensional problem). First make the distance matrix, then draw the dendrograms.

b) Which distance measure do you actually use?

c) Make the same analysis using R!

|||| Exercise 2 EU data

a) Using the EU data (from correspondence analysis) make a cluster analysis using chi square distance and UPGMA clustering of the countries + the European Parliament and the European Commission. Do the Nordic countries and south European countries cluster separately and which country is closest to the EU parliament and EU Commission?

|||| Exercise 3 Wine data

a) For exam (to be presented): Here we use the wine data. Please cluster the 178 wine samples. Perform the cluster analysis using quantitative data using Euclidean distance and UPGMA clustering. First use the raw unstandardized data, then use standardized (autoscaled) data and see how many outliers you have in each wine type, and how many alien wines you have in each wine type, depending on standardization or not. Calculate the cophenetic correlation coefficient for these two clusterings. Which method is best for classifying the wines correctly? Does the cophenetic correlation tell anything about the best classification?

|||| Exercise 4 Wine data

a) For exam (to be presented): Again we use the wine data set (raw data), but this time we examine the variables. Cluster the variables using Euclidean distance and the `cor` (correlation). Do these distance measures give different clusterings of the variables? And which distance / correlation measure makes most sense to you?