Day3 exercise solutions

Ali Movasati

Sept. 30rd, 2024

```r
# Set global code chunk options
knitr::opts_chunk$set(warning = FALSE)
```

```r
# load required libraries
library(ggplot2)
library(magrittr)
library(dplyr)
library(tibble)
library(maps)
library(fields)


# define functions
`%notin%` <- Negate(`%in%`)
```

Problem 1

```r
# load the data and inspect it

protein <- read.table("/Users/alimos313/Documents/studies/phd/university/courses/stat-modelling/day2/da

str(protein)
```

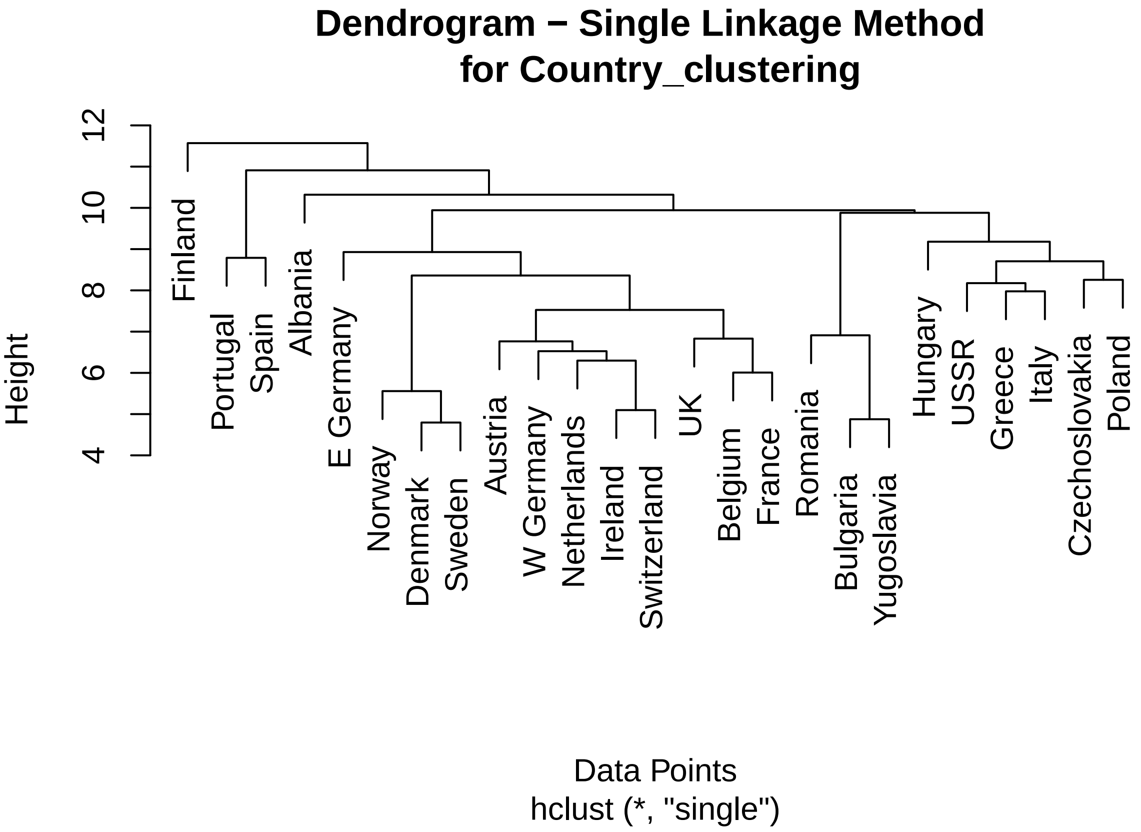
```
## 'data.frame':    25 obs. of  10 variables:
##  $ Country  : chr  "Albania" "Austria" "Belgium" "Bulgaria" ...
##  $ RedMeat  : num  10.1 8.9 13.5 7.8 9.7 10.6 8.4 9.5 18 10.2 ...
##  $ WhiteMeat: num  1.4 14 9.3 6 11.4 10.8 11.6 4.9 9.9 3 ...
##  $ Eggs     : num  0.5 4.3 4.1 1.6 2.8 3.7 3.7 2.7 3.3 2.8 ...
##  $ Milk     : num  8.9 19.9 17.5 8.3 12.5 25 11.1 33.7 19.5 17.6 ...
##  $ Fish     : num  0.2 2.1 4.5 1.2 2 9.9 5.4 5.8 5.7 5.9 ...
##  $ Cereals  : num  42.3 28 26.6 56.7 34.3 21.9 24.6 26.3 28.1 41.7 ...
##  $ Starch   : num  0.6 3.6 5.7 1.1 5 4.8 6.5 5.1 4.8 2.2 ...
##  $ Nuts     : num  5.5 1.3 2.1 3.7 1.1 0.7 0.8 1 2.4 7.8 ...
##  $ Fr.Veg   : num  1.7 4.3 4 4.2 4 2.4 3.6 1.4 6.5 6.5 ...
```

```r
# reshape data set for clustering

row.names(protein) <- protein$Country
protein <- as.matrix(protein[,-1])

protein_t <- t(protein)
```

1.A)

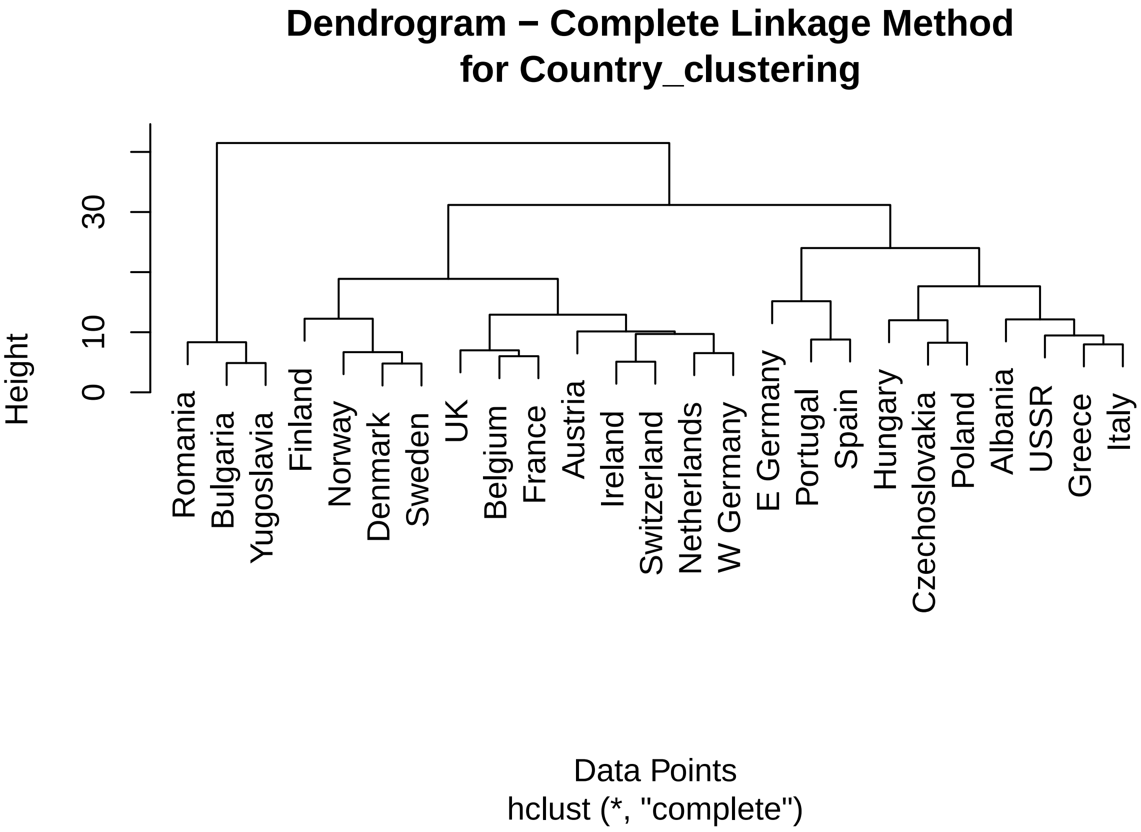
Single linkage hierarchical clustering

```r
# perform clustering



for (i in c("Country_clustering")){
    if (i == "Country_clustering"){
        data <- protein
    } else if (i == "Protein_clustering"){
        data <- protein_t
    }
    distance_matrix <- dist(data, method = "euclidean")
    hc_single <- hclust(distance_matrix, method = "single")
    plot(hc_single, main = paste0("Dendrogram - Single Linkage Method \n for ", i), xlab = "Data Points

}
```



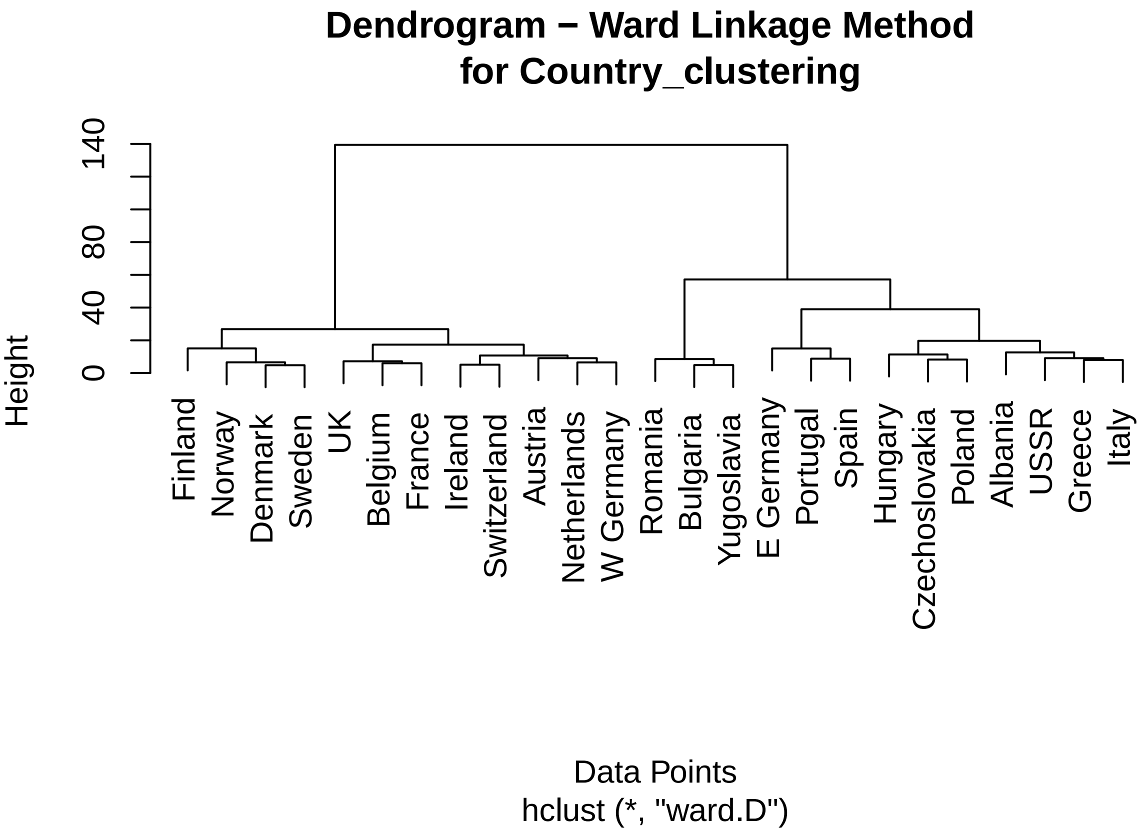
Complete linkage hierarchical clustering

```r
# perform clustering

for (i in c("Country_clustering")){
    if (i == "Country_clustering"){
        data <- protein
    } else if (i == "Protein_clustering"){
        data <- protein_t
    }
    distance_matrix <- dist(data, method = "euclidean")
    hc_single <- hclust(distance_matrix, method = "complete")
    plot(hc_single, main = paste0("Dendrogram - Complete Linkage Method \n for ", i), xlab = "Data Poin

}
```



Ward linkage hierarchical clustering

```
# perform clustering

for (i in c("Country_clustering")){
    if (i == "Country_clustering"){
        data <- protein
    } else if (i == "Protein_clustering"){
        data <- protein_t
    }
    distance_matrix <- dist(data, method = "euclidean")
    hc_single <- hclust(distance_matrix, method = "ward.D")
    plot(hc_single, main = paste0("Dendrogram - Ward Linkage Method \n for ", i), xlab = "Data Points",

}
```



« **Comments** » We performed hierarchical clustering with three different linkage methods (single, complete, and Ward) on dissimilarity values (Euclidean distances).

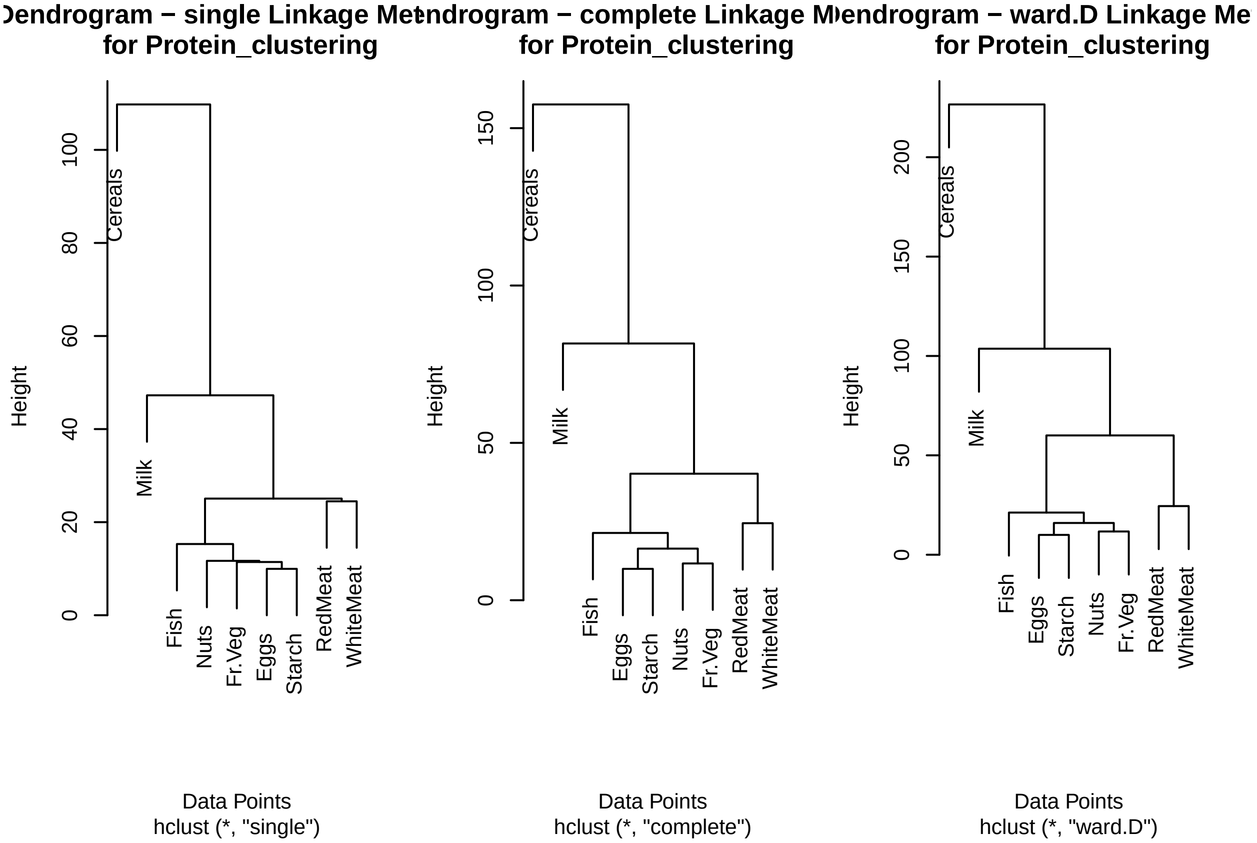
Each of these methods result in a different clustering since they use different criteria to calculate the distance between two neighboring clusters.

The choice of which of these methods should be selected depends on the nature of the data.

According to the obtained clusters we see that Single-linkage method does not result in a clear-cut distinction between clusters, probably due to the closeness of data points, while the other two methods perform better.

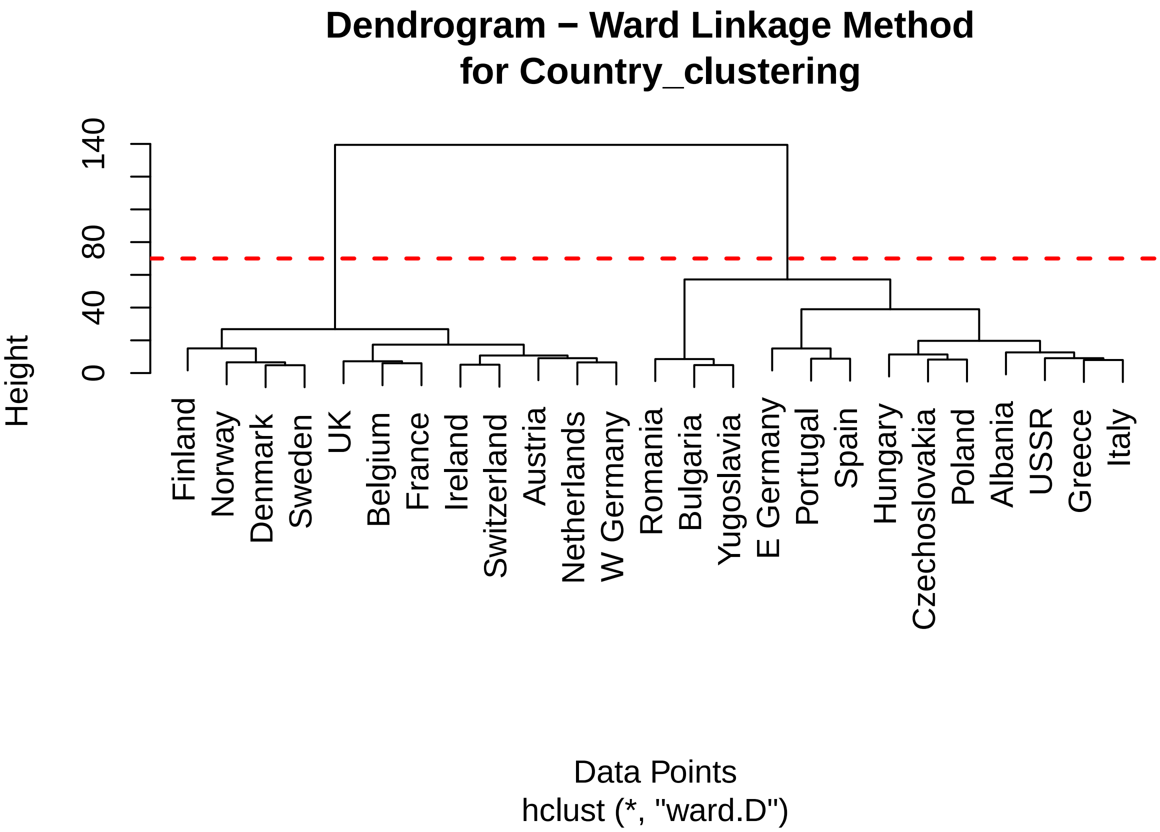
Hirechical clustering could also be done on the different protein groups to identify patterns of protein consumption. As the dendograms below show, it seems like for this analyses different types of linkage methods have little impact on the dendogram. Also we can appreciate that the consumption of white and red meat are very closely related, while cereal seems to be an outgroup.

```r
par(mfcol = c(1,3))

for (i in c("Protein_clustering")){
    if (i == "Country_clustering"){
        data <- protein
    } else if (i == "Protein_clustering"){
        data <- protein_t
    }
    for (clustering_method in c("single", "complete", "ward.D")){
        distance_matrix <- dist(data, method = "euclidean")
        hc_single <- hclust(distance_matrix, method = clustering_method)
        plot(hc_single, main = paste0("Dendrogram - ",clustering_method," Linkage Method \n for ", i), 
    }
}
```



1.B)

```
# perform clustering

for (i in c("Country_clustering")){
    if (i == "Country_clustering"){
        data <- protein
    } else if (i == "Protein_clustering"){
        data <- protein_t
    }
    distance_matrix <- dist(data, method = "euclidean")
    hc_single <- hclust(distance_matrix, method = "ward.D")
    plot(hc_single, main = paste0("Dendrogram - Ward Linkage Method \n for ", i), xlab = "Data Points",
    abline(h = 70, col = "red", lwd = 2, lty = 2)
}
```



« **Comments** » We chose the clustering method Ward. If we cut the dendogram at distance of 70 we will end up with two clear-cut clusters. One of these clusters mostly contain the advanced European countries while the other one mostly contain the developing European countries. Therefore, countries protein consumption patterns tend to reflect their underlying socio-economic.

1.C)

« **Comments** » While PCA's main application is to reduce dimensionality and redundancy in terms of variable co-variation, PCA and clustering can both be used to identify patterns and grouping in the data. As we can observe both approaches more or less show the same pattern in the data. When the main aim is to group the data however, clustering provides us with more freedom in choosing the right algorithm based on the nature of the underlying data. Also with clustering, we can pinpoint the relationship between each group and how they relate to each other, while with PCA we can only roughly see their closeness. Also the clustering approach provides us with measures to evaluate the quality of the resulting dendogram.

An interesting aspect when comparing the two results is that "Cereal" consumption is clustered as an outgroup to the rest of the protein groups in clustering, while it contributed the most to the loading of the first component in PCA analysis. This could be due to the fact that we have the most variation in this variable and it is not closely related to any other protein group. # Problem 2

```
vec <- c(1,2,1,3,2,6,5,7,6,12)
k_no <- 3

i <- 0
```

```r
my.kmean <- function(vec, k_no){
    set.seed(1)

    k_means <- sample(vec, size = k_no, replace = F)

    j <- 0
    while (j <= 10){

        vec_dist_df <- data.frame()

        for (i in 1:3){
            vec_dist_df <- rbind(vec_dist_df, abs(vec - k_means[i]))
        }

        assignments <- apply(vec_dist_df, 2, which.min)

        names(vec) <- as.numeric(k_means[assignments])

        k_means_new <- c()
        for (k in k_means){
            k_means_new <- c(k_means_new, mean(vec[names(vec) == k]))
        }

        k_means <- k_means_new
        j <- j + 1

    }

    vec_dist_df <- data.frame()

    for (i in 1:3){
        vec_dist_df <- rbind(vec_dist_df, abs(vec - k_means[i]))
    }

    assignments <- apply(vec_dist_df, 2, which.min)

    names(vec) <- as.numeric(k_means[assignments])

    return(vec)
}


k_assignments <- my.kmean(vec, k_no)


print(k_assignments)
```

```
## 1.8 1.8 1.8 1.8 1.8   6   6   6   6  12
##   1   2   1   3   2   6   5   7   6  12
```