Module 10: Unsupervised learning (Overview/quizz lecture)

TMA4268 Statistical Learning V2023

Sara Martino, Department of Mathematical Sciences, NTNU

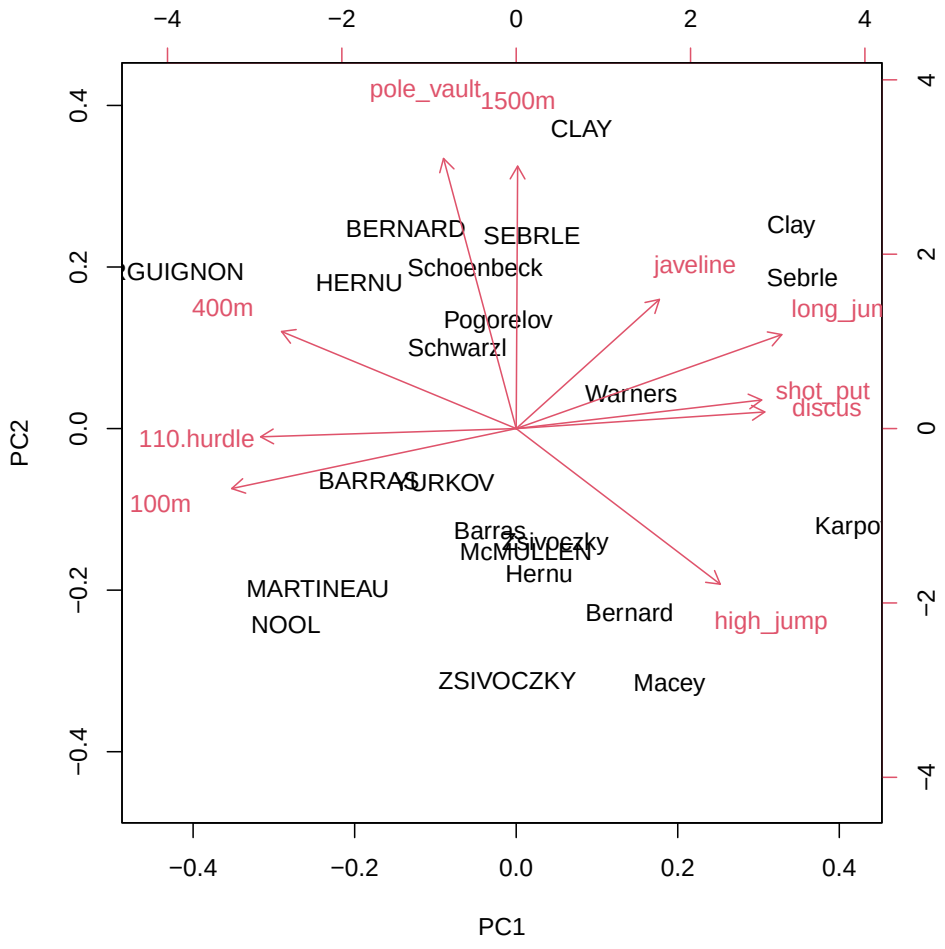
April 5, 2024

PCA example

- We study the `decathlon2` dataset from the `factoextra` package in R, where Athletes' performance during a sporting meeting was recorded.
- We look at 23 athletes and the results from the 10 disciplines in two competitions.

```
library(factoextra)
library(FactoMineR)
data("decathlon2")
decathlon2.active <- decathlon2[1:23, 1:10]
names(decathlon2.active) <- c("100m", "long_jump", "shot_put", "
    "400m", "110.hurdle", "discus", "pole_vault", "javeline", "1
```

```
##           100m long_jump shot_put high_jump  400m 110.hurdle discus pole_vault
## SEBRLE   11.04      7.58    14.83      2.07 49.81      14.69  43.75       5.02
## BERNARD  11.02      7.23    14.25      1.92 48.93      14.99  40.87       5.32
## YURKOV   11.34      7.09    15.19      2.10 50.42      15.31  46.26       4.72
##          javeline 1500m
## SEBRLE      63.19 291.7
## BERNARD     62.77 280.1
## YURKOV      63.44 276.4
```


pole_vault
1500m
CLAY
BERNARD
SEBRLE
Clay
GUIGNON
HERNU
Schoenbeck
javeline
Sebrle
400m
Pogorelov
long_jum
Schwarzl
Warners
shot_put
discus
110.hurdle
BARRAS
SURKOV
100m
Barras
Karpo
McMULLEN
Hernu
MARTINEAU
NOOL
Bernard
high_jump
ZSIVOCZKY
Macey
PC1
PC2

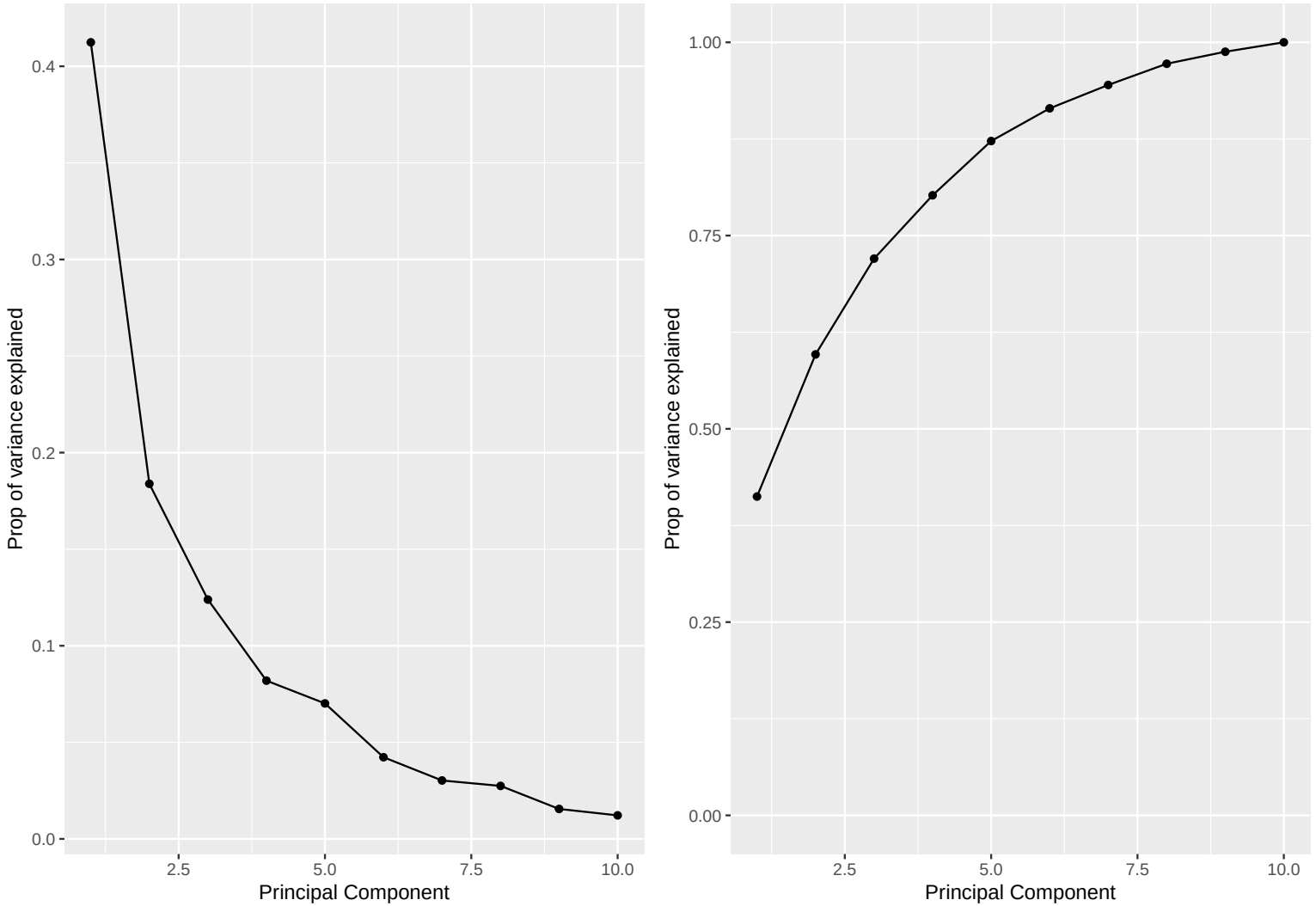
Proportion of varianced explained (PVE)

Recap: The PVE by PC m is given by

$$\frac{\sum_{i=1}^{m} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

Scree plot

A graphical description of the **proportion of variance explained (PVE)** by a certain number of PCs:

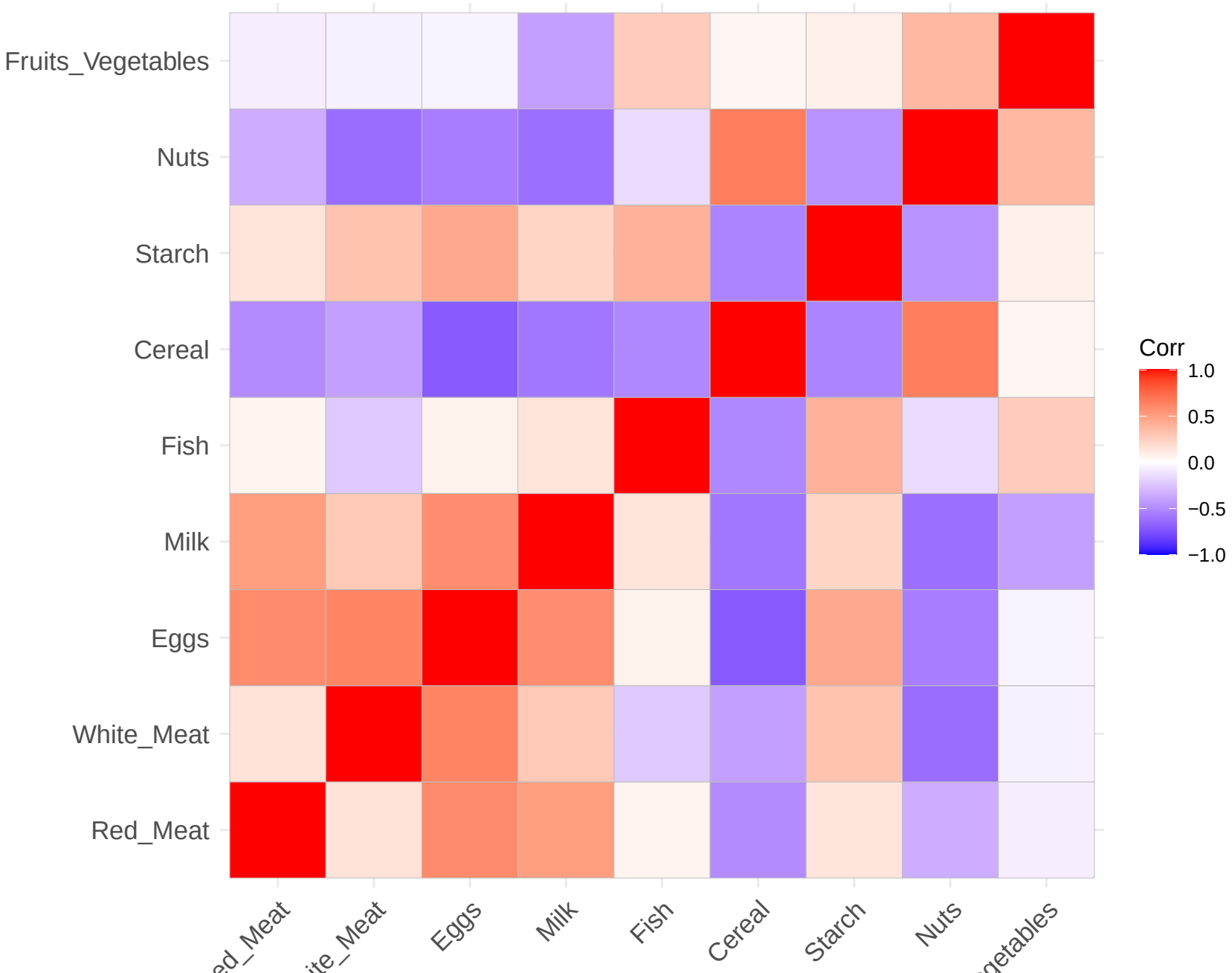


Another example

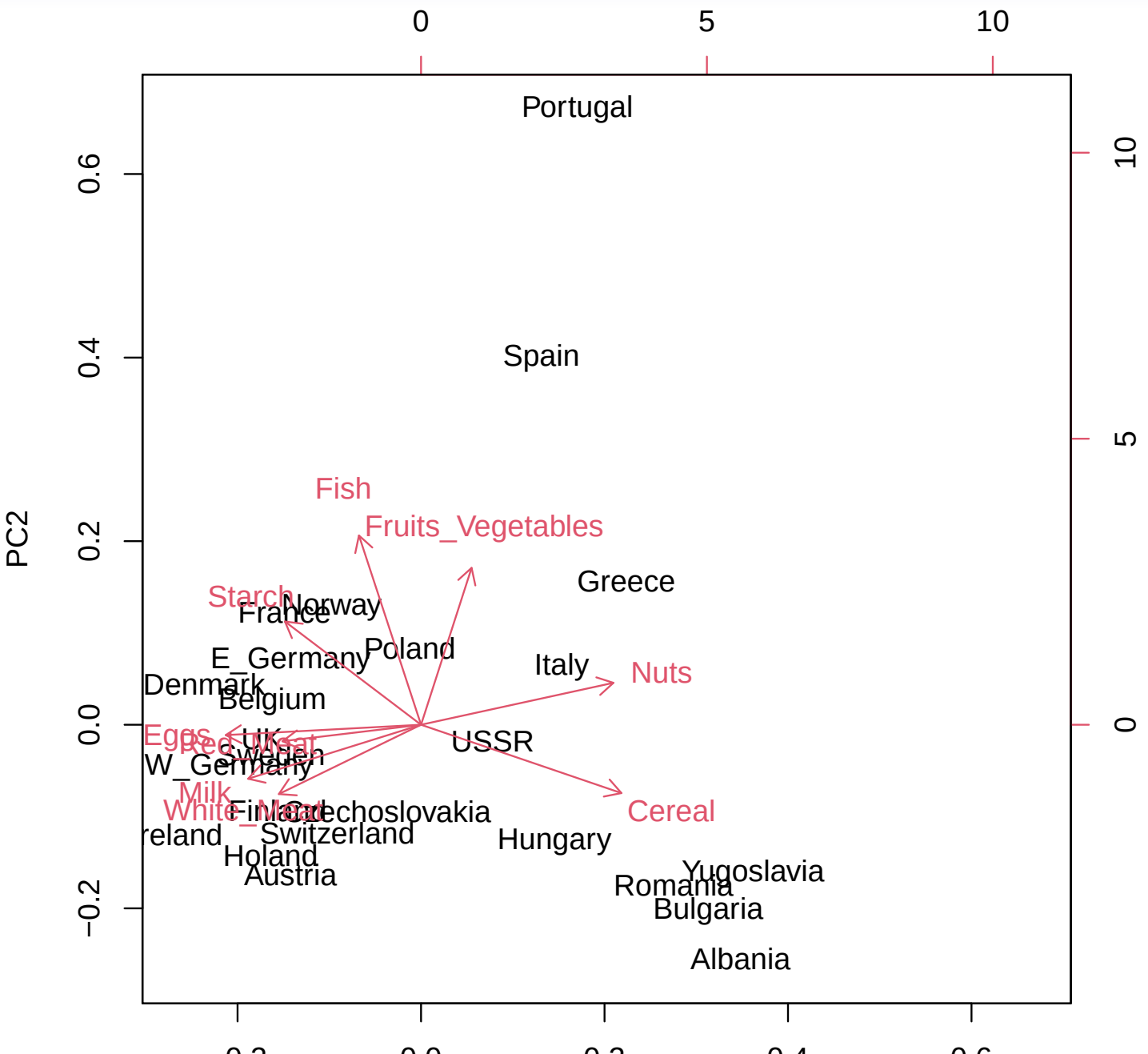
Protein consumption in twenty-five European countries for nine food groups.

```
##                 Red_Meat White_Meat Eggs Milk Fish Cereal Star
## Albania             10.1        1.4  0.5  8.9  0.2   42.3    0
## Austria              8.9       14.0  4.3 19.9  2.1   28.0    3
## Belgium             13.5        9.3  4.1 17.5  4.5   26.6    5
## Bulgaria             7.8        6.0  1.6  8.3  1.2   56.7    1
## Czechoslovakia       9.7       11.4  2.8 12.5  2.0   34.3    5
## Denmark             10.6       10.8  3.7 25.0  9.9   21.9    4
##                 Fruits_Vegetables
## Albania                       1.7
## Austria                       4.3
## Belgium                       4.0
## Bulgaria                      4.2
## Czechoslovakia                4.0
## Denmark                       2.4
```

Correlation Matrix



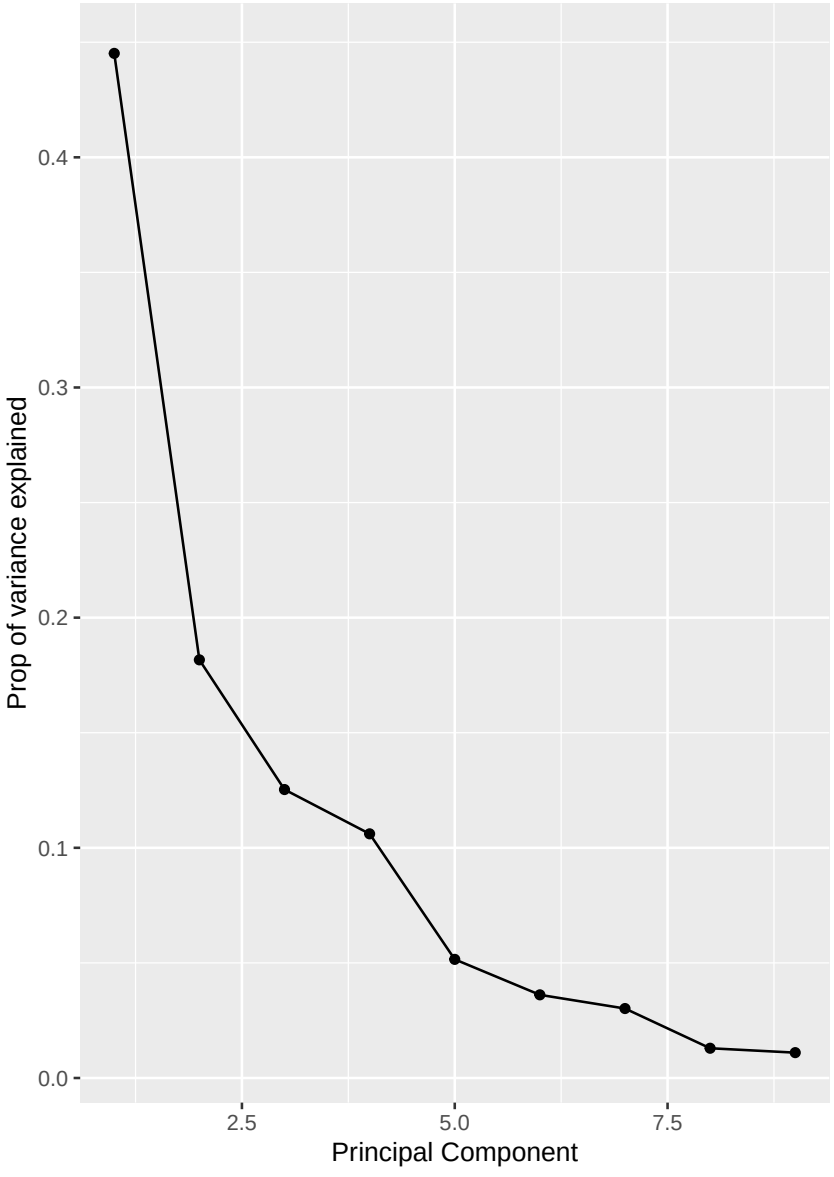
PCA

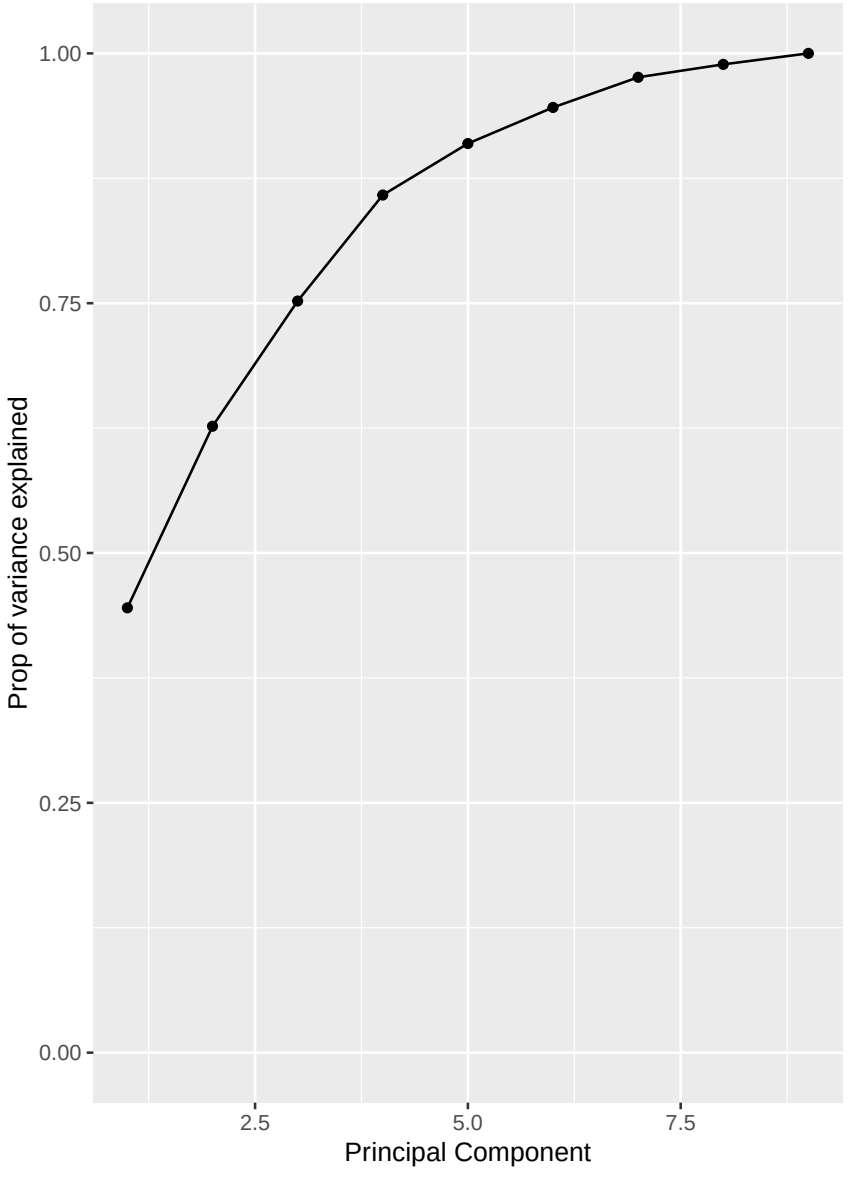

0
5
10
Portugal
Spain
Fish
Fruits_Vegetables
Greece
Starch
France
Norway
E_Germany
Poland
Italy
Nuts
Denmark
Belgium
Eggs
USSR
W_Germany
Milk
White_Meat
Cereal
Finland
Czechoslovakia
Switzerland
Hungary
Holland
Austria
Yugoslavia
Romania
Bulgaria
Albania
PC2
0.6
0.4
0.2
0.0
−0.2
10
5
0

Variance Explained

```
## Importance of components:
##                             PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     2.0016 1.2787 1.0620 0.9771 0.68106 0.57020 0.52116
## Proportion of Variance 0.4452 0.1817 0.1253 0.1061 0.05154 0.03613 0.03018
## Cumulative Proportion  0.4452 0.6268 0.7521 0.8582 0.90976 0.94589 0.97607
##                             PC8     PC9
## Standard deviation     0.34102 0.31482
## Proportion of Variance 0.01292 0.01101
## Cumulative Proportion  0.98899 1.00000
```

Variance Explained - Scree plot





Clustering

- The aim is to find *clusters* or *subgroups*.
- Clustering looks for homogeneous subgroups in the data.

Difference to PCA?

Clustering

- The aim is to find *clusters* or *subgroups*.
- Clustering looks for homogeneous subgroups in the data.

Difference to PCA?

$\rightarrow$ PCA looks for low-dimensional representation of the data.

K-means vs. hierarchical clustering

See menti.com

K-means clustering

- Fix the number of clusters K.
- Find groups such that the sum of the within-cluster variation is minimized.

K-means clustering - Algorithm

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Hierarchical clustering

Bottom-up agglomerative clustering that results in a *dendogram*.

Algorithm 12.3 *Hierarchical Clustering*

---

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

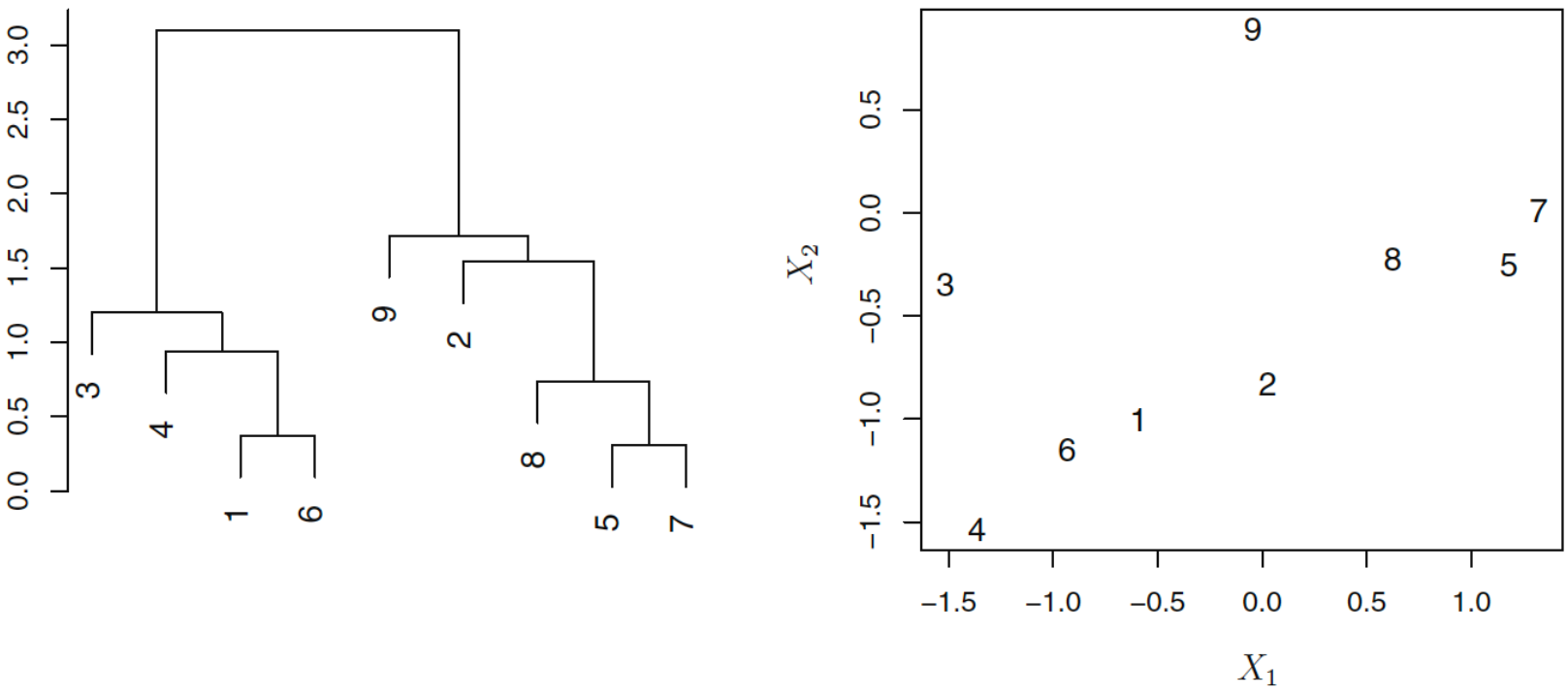
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

Important in hierarchical clustering

- *Linkage:* Complete, single, average centroid.
- *Dissimilarity measure:* Euclidian distance, correlation. *Other similarity/distance measures?* [1]

[1] Note: Correlation is actually a similarity measure, not a distance measure. Implication?

Hierarchical clustering – example



Note: The representation on the right is not possible in high-dimensional space (i.e., if we have $X_1, X_2, X_3, ..., X_p$).

Exercise 2 from the book

We have the following dissimilarity matrix:

$$\begin{bmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{bmatrix}$$

1. Sketch the dendogram using *complete* linkage, indicate on the plot the height at wich each fusion occurs, as well as the observations corresponding to each leaf in the dendogram
2. Repeat using *single* linkage
3. Suppose we cut the two dendograms such that 2 clusters result. Which observations are in each cluster?

Exercise 11 from the book

13. On the book website, `www.statlearning.com`, there is a gene expression data set (`Ch12Ex13.csv`) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

    (a) Load in the data using `read.csv()`. You will need to select `header = F`.

    (b) Apply hierarchical clustering to the samples using correlation-based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used?

Pros and cons of clusterization methods / practical issues

References