Module 10: Unsupervised learning (Overview/quizz lecture)

TMA4268 Statistical Learning V2023

Stefanie Muff, Department of Mathematical Sciences, NTNU

March 23, 2023

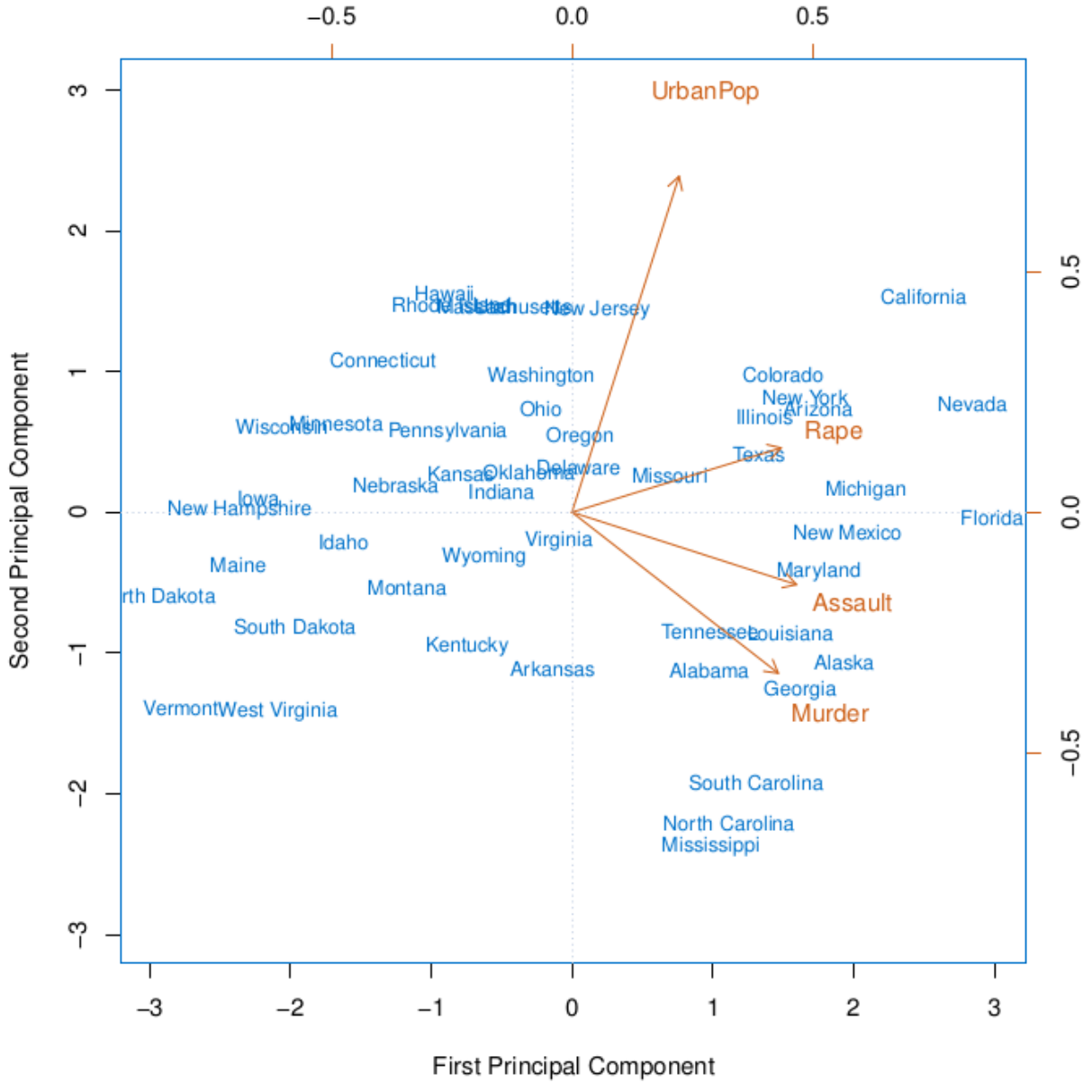
PC loadings vectors Φ

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

(Table 2.1)

Loadings vectors $\Phi_i = (\Phi_{1j}, \Phi_{2j}, \ldots, \Phi_{pj})^\top$: How much does the respective covariate contribute to PC_j?

The biplot

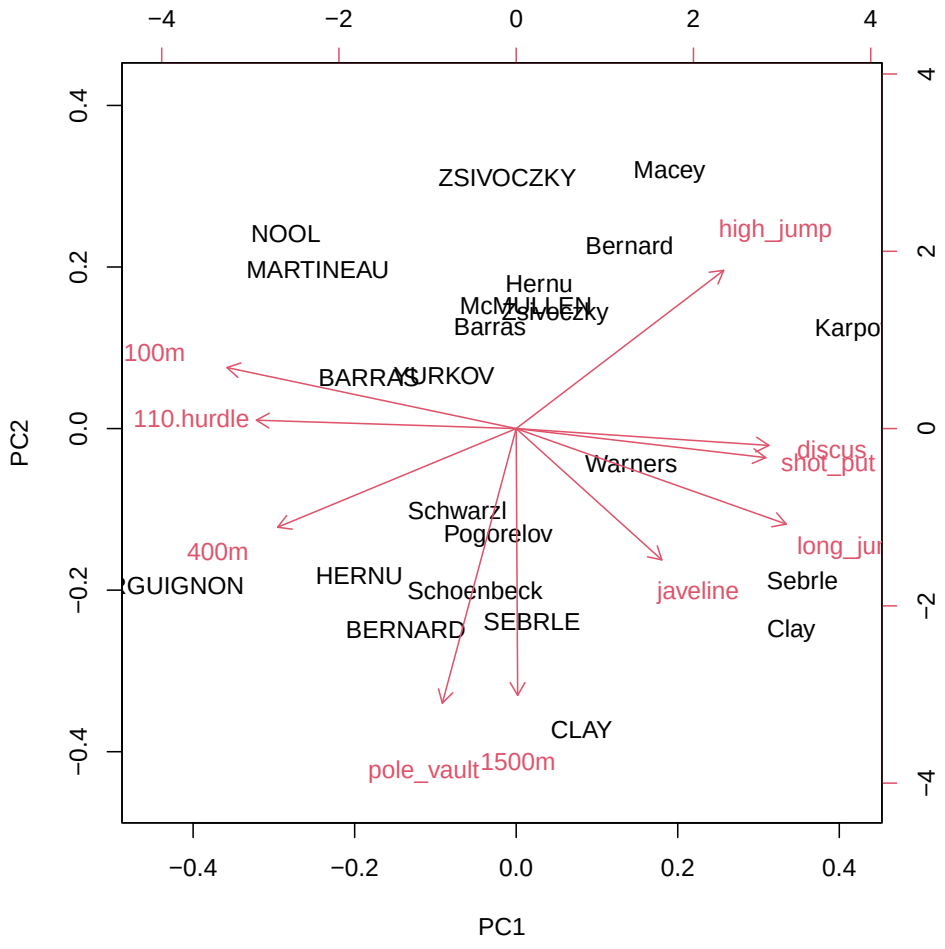

UrbanPop
Rape
Assault
Murder
California
Nevada
Florida
New Jersey
Connecticut
Washington
Colorado
New York
Arizona
Illinois
Ohio
Wisconsin
Minnesota
Pennsylvania
Oregon
Texas
Delaware
Kansas
Oklahoma
Missouri
Nebraska
Indiana
Michigan
Iowa
New Hampshire
Idaho
Virginia
New Mexico
Wyoming
Maine
Maryland
Montana
South Dakota
Tennessee
Louisiana
Kentucky
Alaska
Arkansas
Alabama
Georgia
Vermont
West Virginia
South Carolina
North Carolina
Mississippi
Second Principal Component
First Principal Component
−0.5
0.0
0.5
−3
−2
−1
0
1
2
3

Example from Compulsory 3, 2020

- 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.

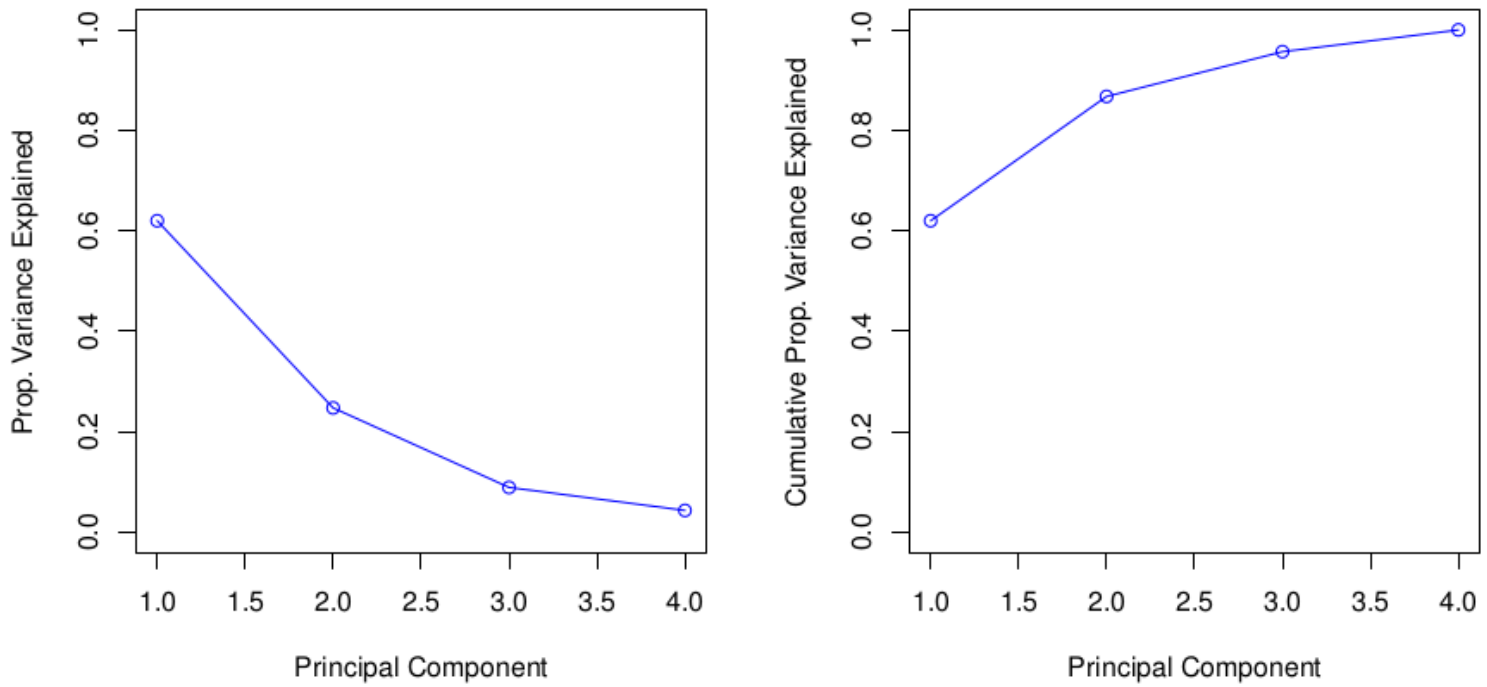
```
decathlon2.active[c(1, 3, 4), ]
```

```
##          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
```


−4
−2
0
2
4
0.4
0.2
0.0
−0.2
−0.4
PC2
PC1
ZSIVOCZKY
Macey
NOOL
MARTINEAU
Bernard
high_jump
Hernu
Barras
Karpo
100m
BARRAS
110.hurdle
discus
shot_put
Warners
Schwarzl
Pogorelov
400m
HERNU
Schoenbeck
javeline
Sebrle
BERNARD
SEBRLE
Clay
CLAY
pole_vault
1500m

Scree plot

A graphical description of the **proportion of variance explained (PVE)** by a certain number of PCs (see Fig 12.3 from James et al. (2021)):



Proportion of varianced explained (PVE)

Recap: The PVE by PC m is given by

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

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.
- Algorithm?

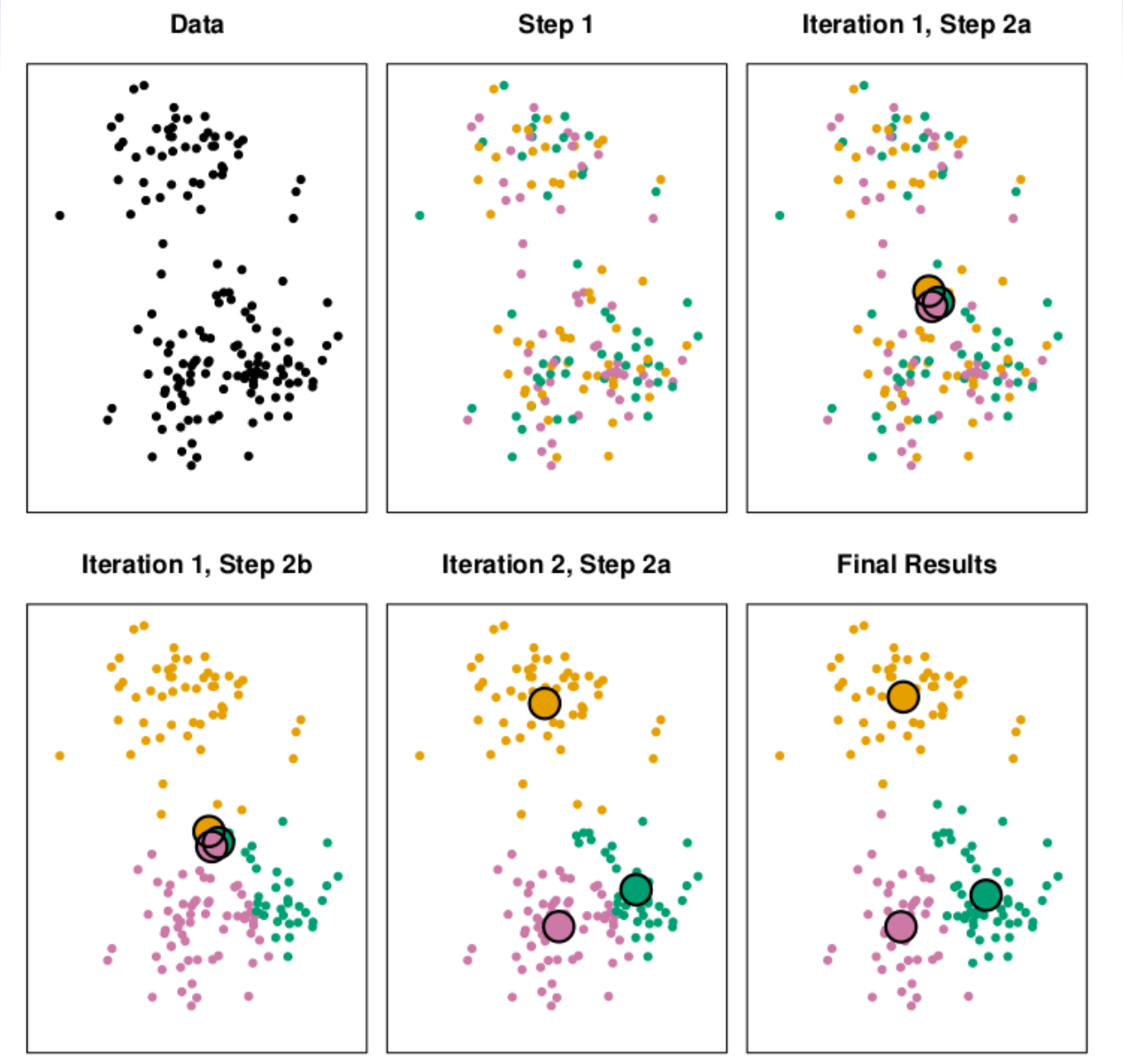
Algorithm 12.2 *K-Means Clustering*

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).



(Fig 12.8 from course book)

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.

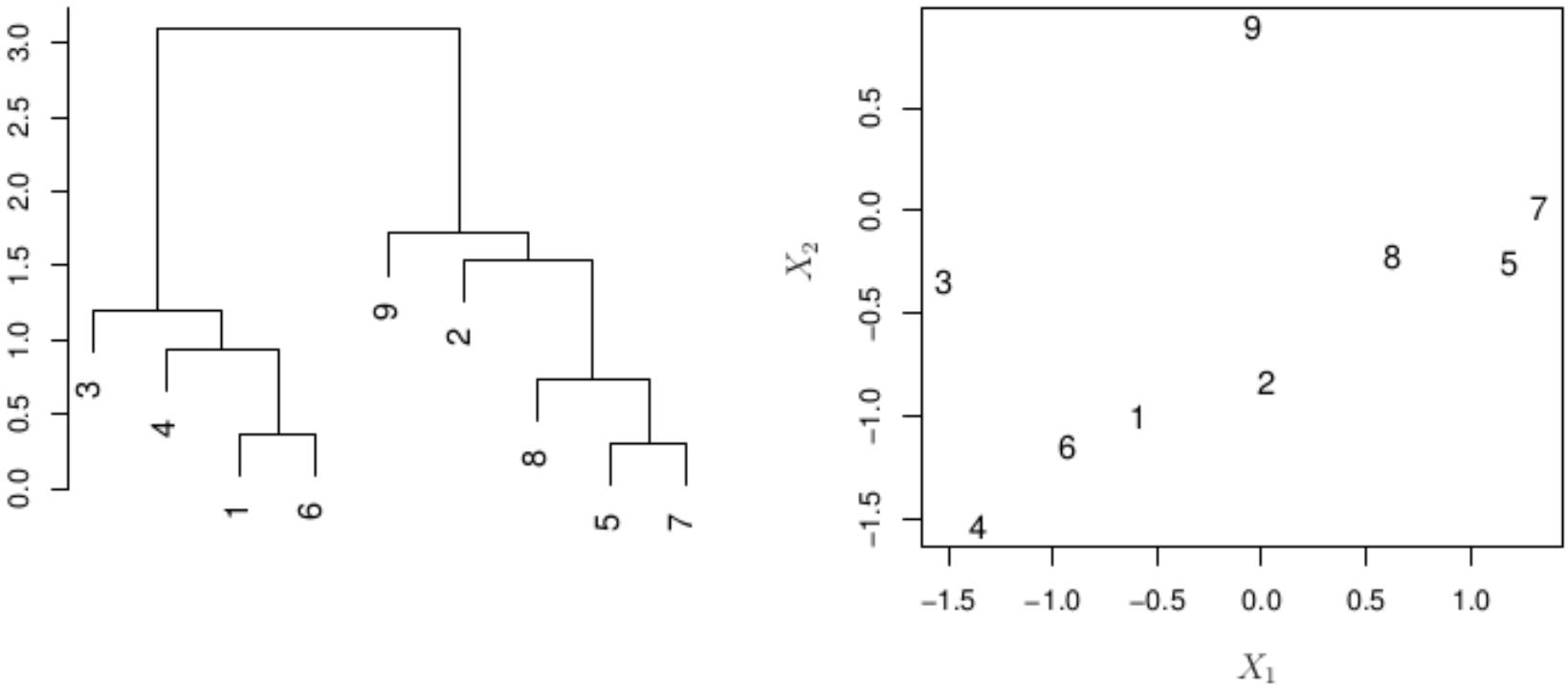
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

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 $X_1, X_2, X_3, ..., X_p$).

Hierarchical clustering – example

An exam question from 2022:

We have four observations for which we know the distance matrix in Euclidean space:

$$\begin{bmatrix} 0 & 3 & 5 & 7 \\ 3 & 0 & 6 & 4 \\ 5 & 6 & 0 & 5.5 \\ 7 & 4 & 5.5 & 0 \end{bmatrix}.$$

Based on this dissimilarity matrix, sketch the dendogram that results from hierarchical clustering using complete linkage. On the plot, indicate the height where each fusion occurs, as well as the observastions that correspond to the leafs in the dendogram (enumerated as 1, 2, 3, 4).

Pros and cons of clusterization methods / practical issues

References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. Springer.