Applied Machine Learning in Health Sciences 2023

-

Clustering

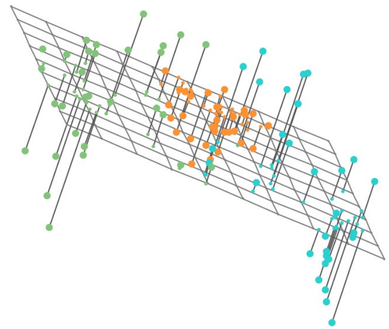
Peter Mondrup Rasmussen

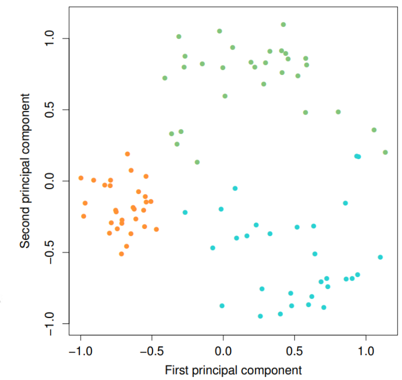
CFIN

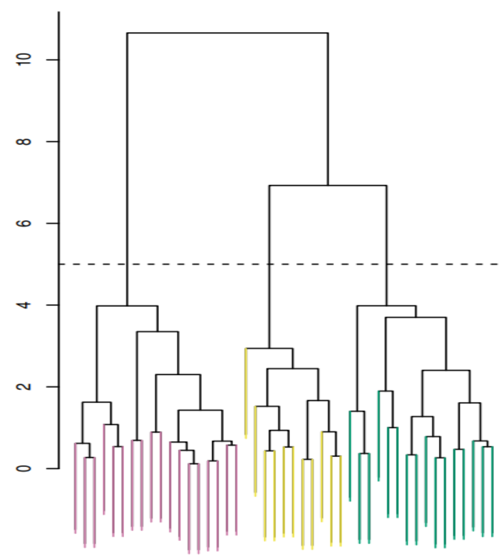
pmr@cfin.au.dk

Unsupervised learning

- In *supervised learning* we have an input X with p features $X_1, X_2, \ldots, X_p$ and a corresponding response Y and our goal is often to predict Y based on the input X.
- In *unsupervised learning* we only consider the features $X_1, X_2, \ldots, X_p$ and the goal is often to discover new *structure* in the data or to learn new *representations* of the data.
- Unsupervised learning is often used for *exploratory data analysis* where the analysis tend to be more subjective compared to supervised learning, where there often is a well-defined goal (predicting the response as good as possible).





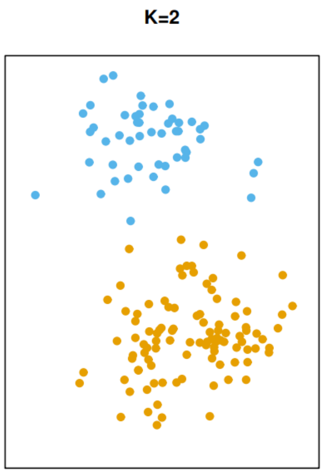


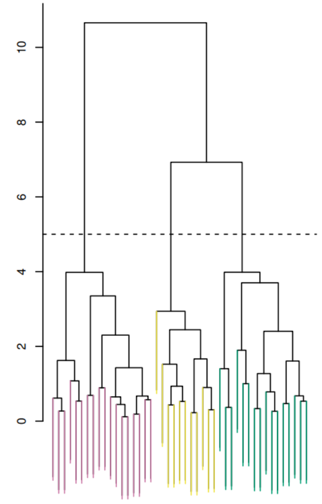
Clustering

K-means clustering
Hierarchical clustering

Clustering

- *Clustering* are techniques to find subgroups in data. We seek to find *distinct groups* so that data within individual groups are *similar*, while data in different groups are *different* from each other.
- Given a $(n \times p)$ data set $\boldsymbol{X}$ we could
  - Look for subgroups among observations based on the features.
  - Look for subgroups among features based on the observations.
- Many different clustering methods. We focus on
  - *K-means clustering* for partitioning the data into a pre-specified number of clusters.
  - *Hierarchical clustering* for building a tree-like *dendrogram* based on which we can perform clustering.





K-means clustering

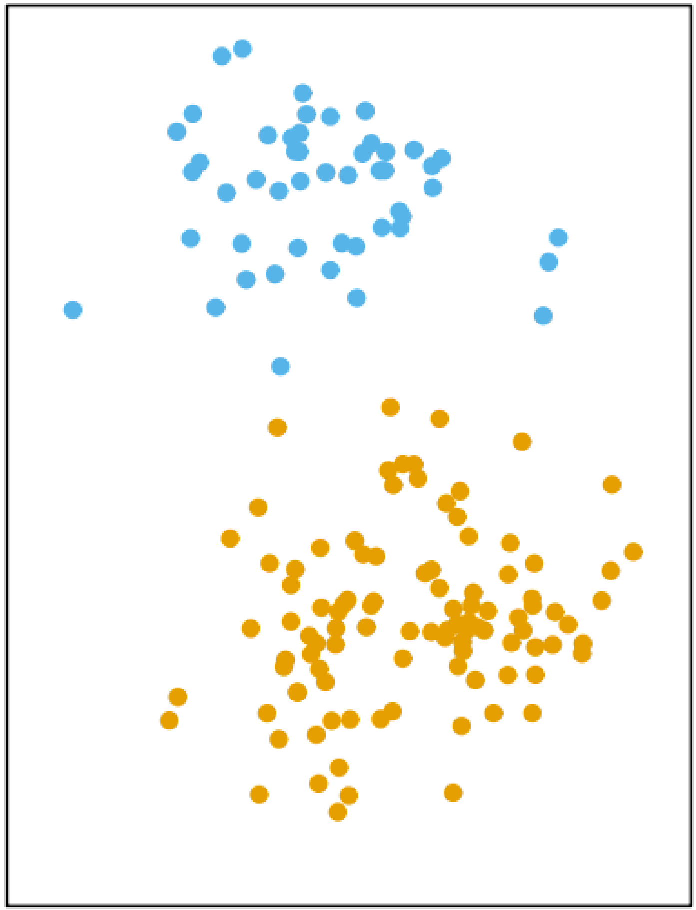
K-Means clustering

- K-means clustering partitions data into K distinct and non-overlapping clusters. Each observation belongs to one and only one cluster.
- $C_1, \dots, C_K$ denote sets containing indices of observations with individual clusters.
- K-means seeks to find clusters so that *within-cluster variation* is as small as possible

$$\min_{C_1,\dots,CK}\left\{\sum_{k=1}^{K} W(C_k)\right\}$$

- The mean squared Euclidian distance is often used to quantify *within-cluster variation*
- $W(C_k) = \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2$

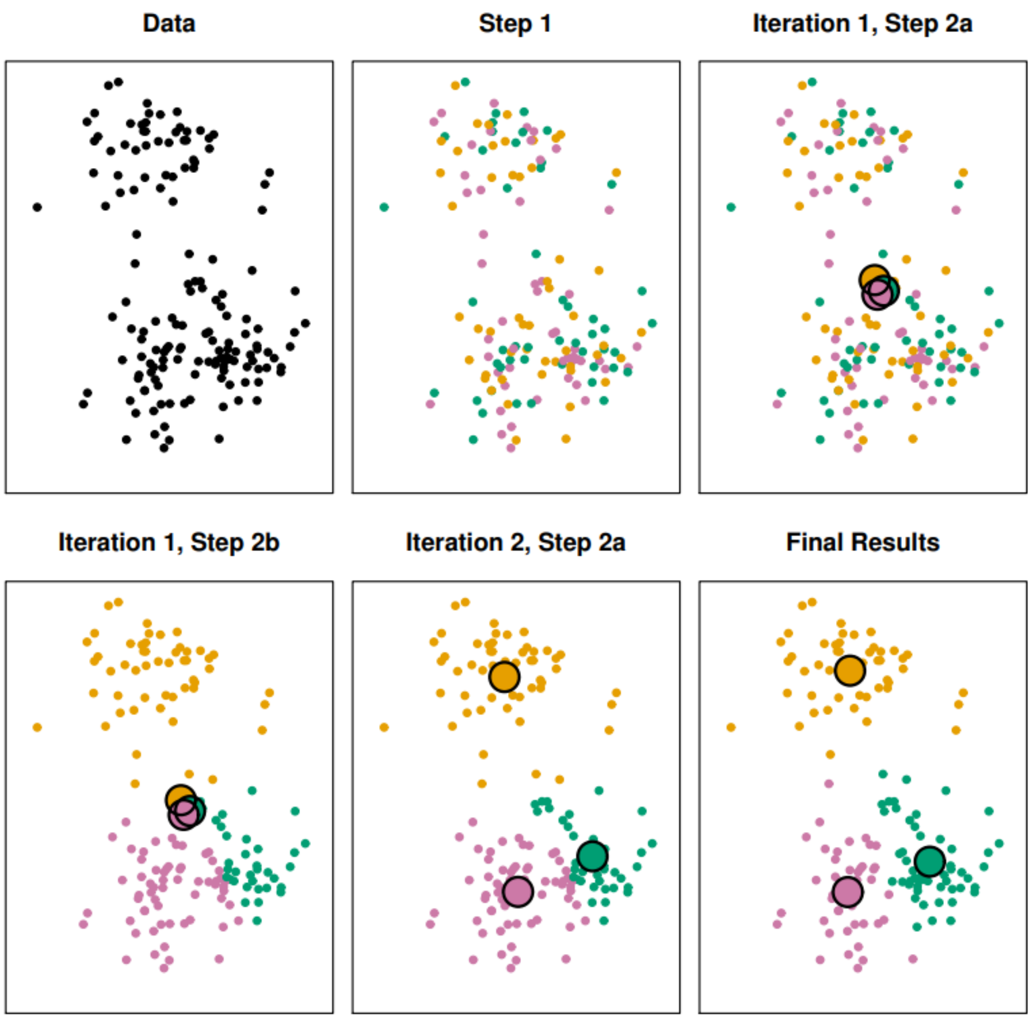
K=2



K-Means clustering

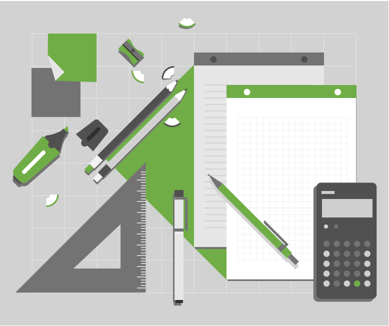
Iterative approach to minimize within-cluster variation (local minimum)

1. Randomly assign observations to the K clusters
2. Repeat until cluster assignment stop changes
   a) For each of the K clusters, compute the cluster centroid.
   b) Assign each observation to the cluster whose centroid is closest.



K-Means clustering

K-Means algorithm finds a *local* minimum. Run multiple times with different random initializations and select the *best* solution.



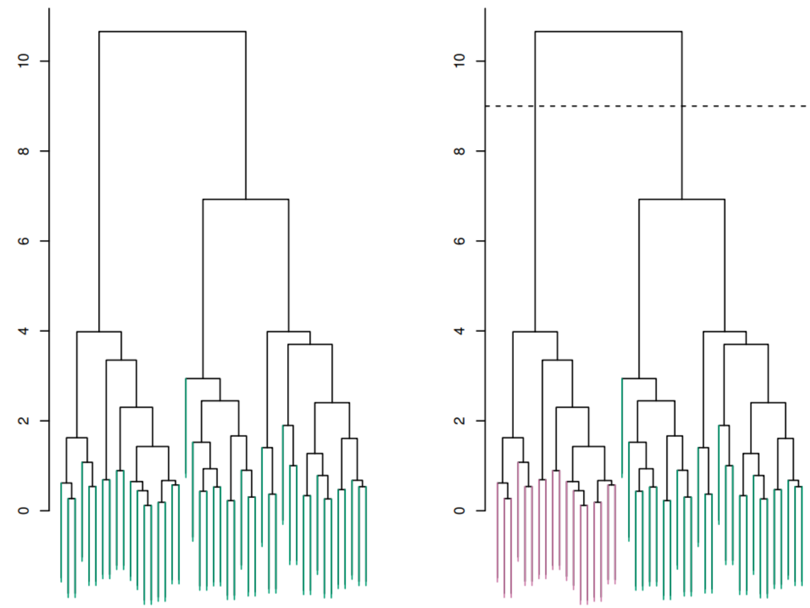
In-class exercises

10 K-means clustering

Hierarchical clustering

Hierarchical clustering

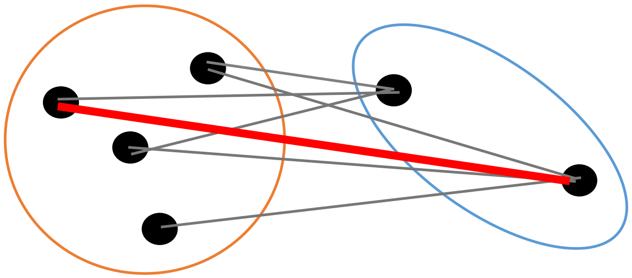
- Hierarchical clustering is a bottom-up approach resulting in a *dendrogram* – a tree-based representation of the data.
- Hierarchical clustering starts with *leaves* each representing individual data points. Based on their *dissimilarity*, leaves are *fused* into branches, which, in turn, are fused with other leaves or branches leading to a tree structure.
- Hierarchical clustering does not require a pre-determined number of clusters K.
- Clustering is performed by *cutting* the dendrogram along the vertical axis.



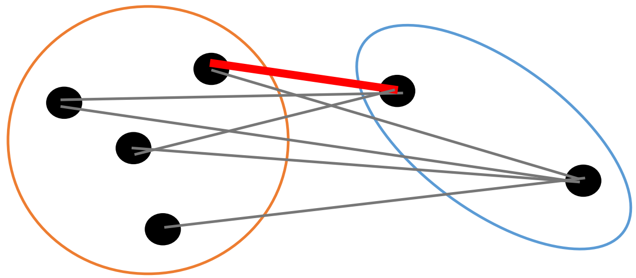
Hierarchical clustering

Linkage: Quantifying *dissimilarity* between *groups of observations.*

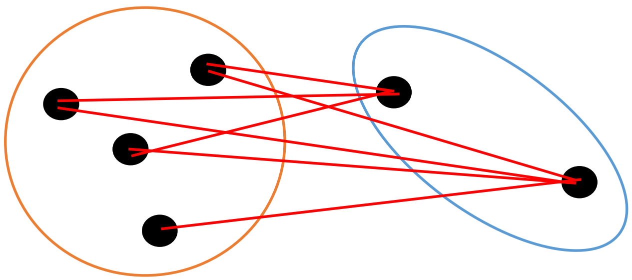
Complete linkage
maximal intercluster dissimilarity



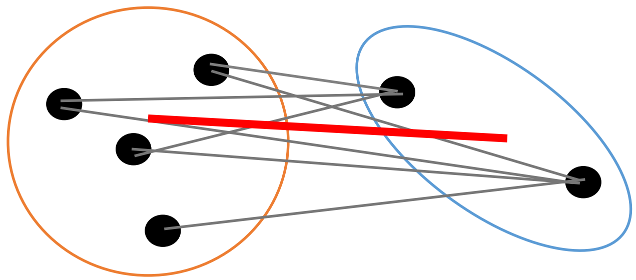
Single linkage
minimal intercluster dissimilarity



Average
Mean intercluster dissimilarity



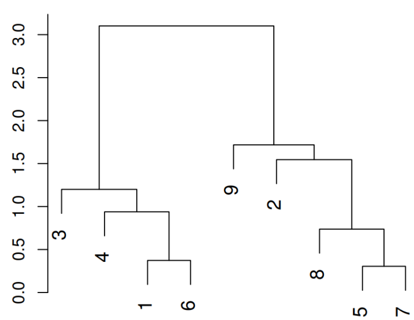
Centroid linkage
dissimilarity between cluster centroids



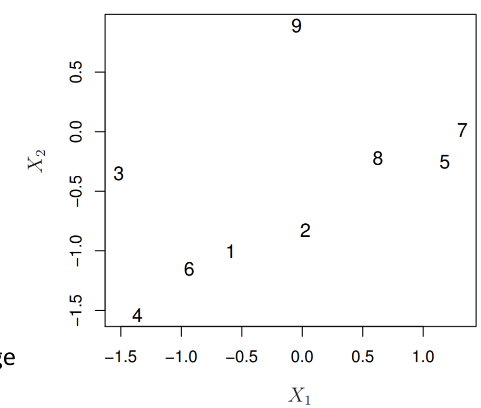
Hierarchical clustering

Hierarchical clustering algorithm:

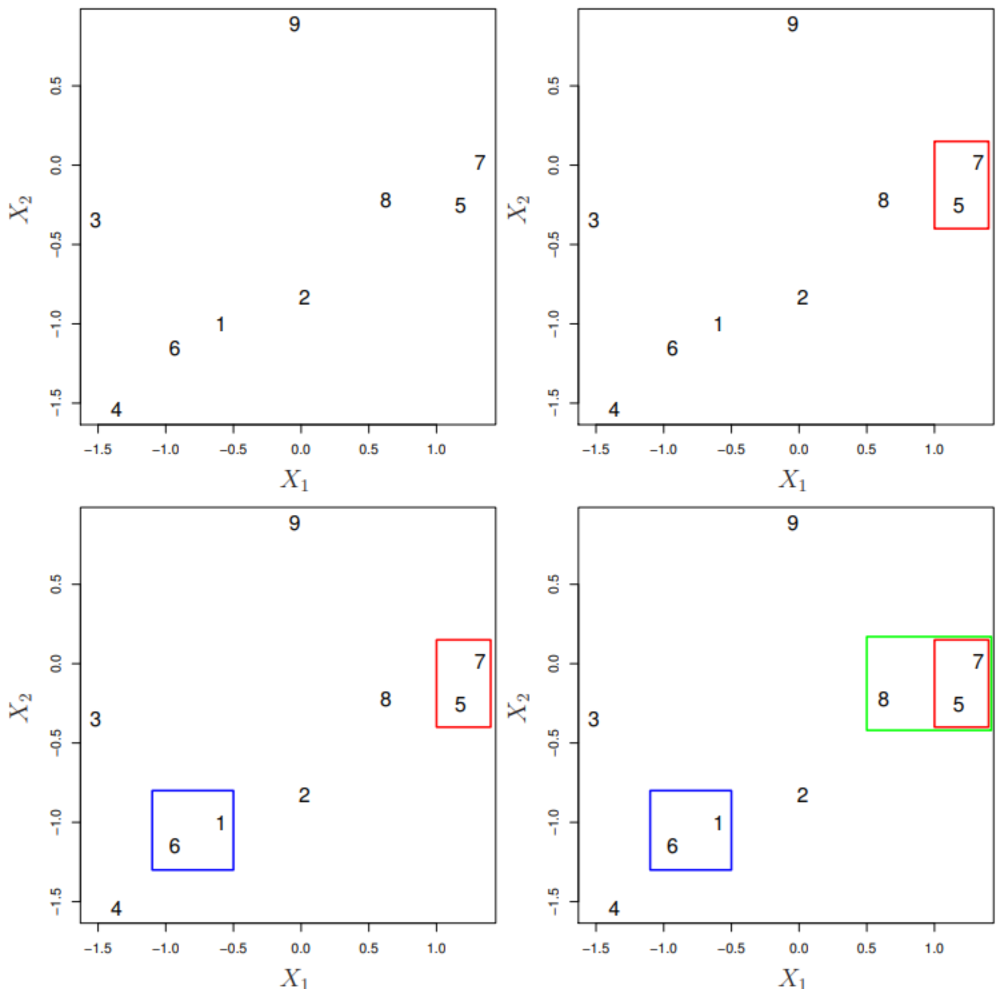
1. Compute *dissimilarity* (e.g. Euclidean distance) between all n data points. Consider individual data points as its own cluster.
2. Repeat
   a) Fuse the pair of clusters that are ***least*** dissimilar (***most similar***). The dissimilarity between the two clusters indicates the height in the dendrogram at which the clusters are fused.
   b) Compute the pairwise inter-cluster dissimilarity between the remaining clusters.

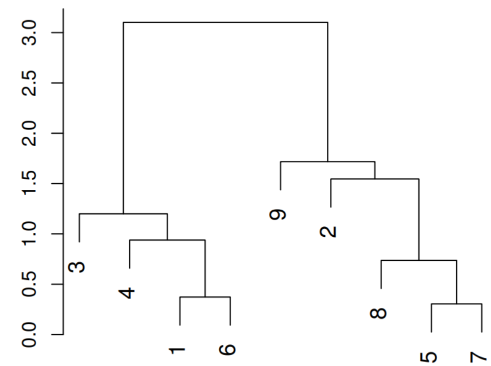


Computed using Euclidean distance and complete linkage



Hierarchical clustering

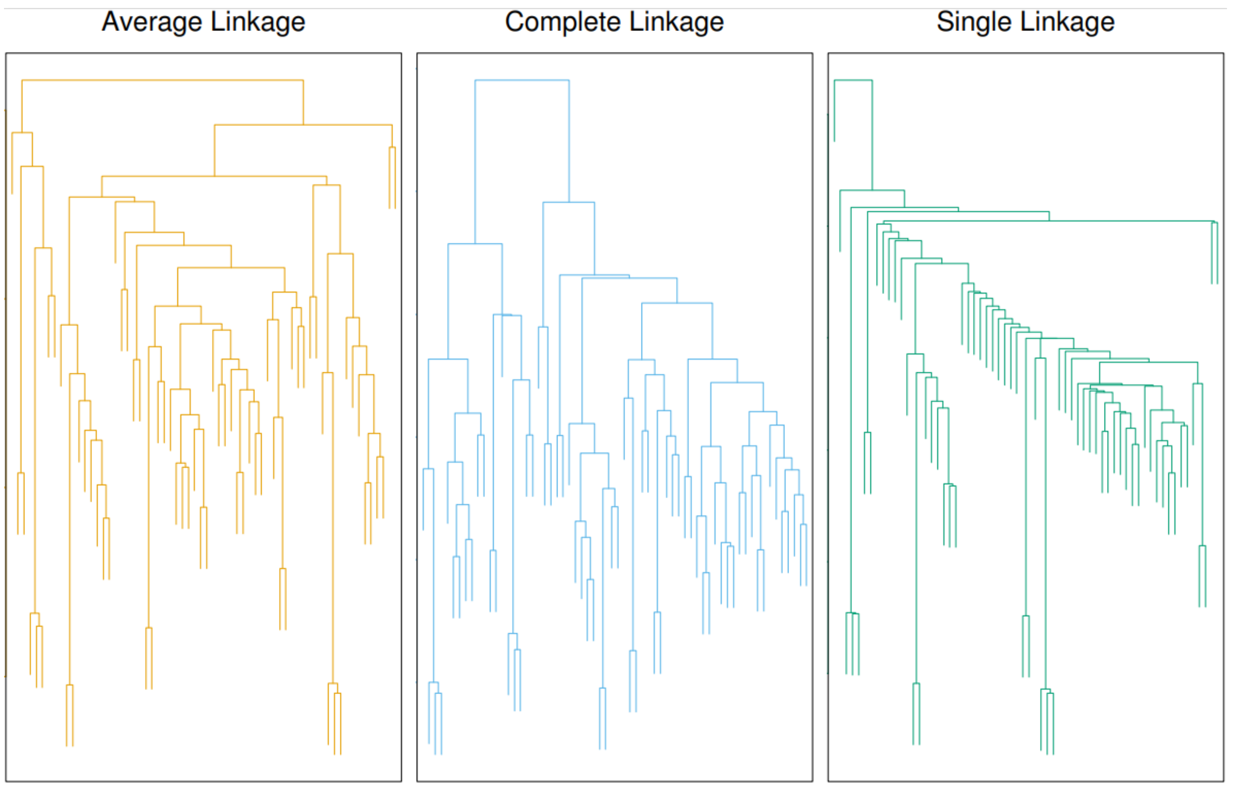




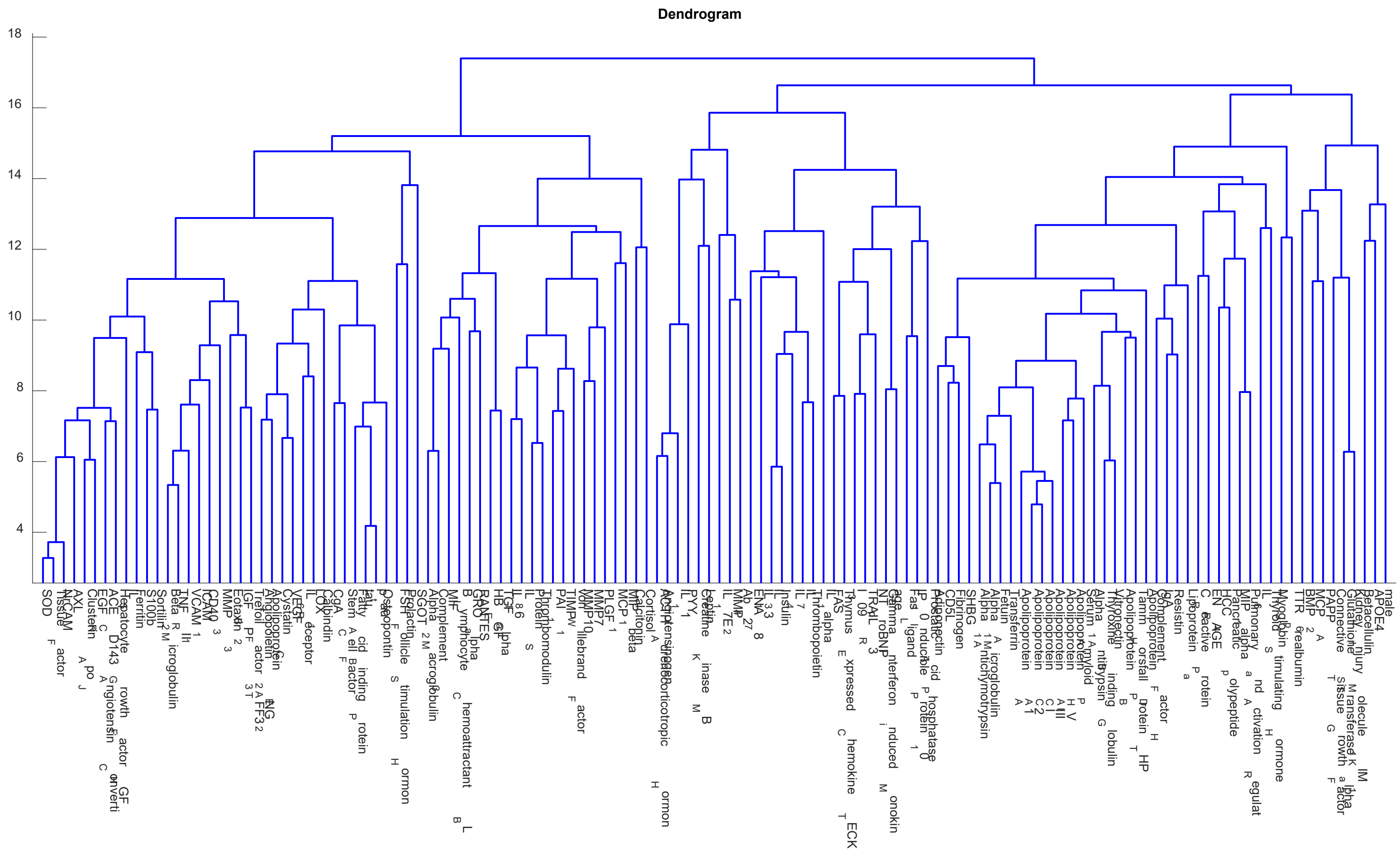
Euclidean distance and complete linkage

Hierarchical clustering

Usually, the dendrogram depends strongly on the chosen linkage type. Average- or complete linkage often result in the most balanced dendrograms. (Single linkage can produce “chain-effects” – but sometimes this structure is actually present in data instead of “clouds”)



Hierarchical clustering

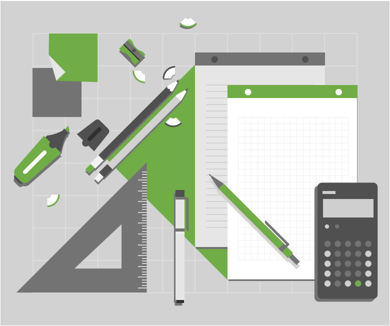


Clustering

Practical issues

Clustering – practical issues

- **Scaling/standardization:** Should the data be standardized before clustering?
- **K-means clustering:**
  - How many clusters K to look for?
  - Measure used to quantify inter-cluster variation.
- **Hierarchical clustering:**
  - Choice of dissimilarity metric.
  - Choice of linkage.
  - Where should the dendrogram be cut?
- **Interpretation:** We will always find clusters, are they meaningful?
- **Other methods:** Many other clustering methods, e.g. mixture models and spectral clustering.



In-class exercises

11 Hierarchical clustering

References

Figures from James et al. *An Introduction to Statistical Learning*, second edition, https://www.statlearning.com/resources-second-edition