CLUSTERING METHODS

Unsupervised learning

CLUSTERING

Finding *subgroups* or *clusters*

We partition datasets into distinct groups:
…the observations within each group are quite similar
…the observations in different groups are quite different

What it means for two observations to be similar or different?
…domain specific consideration

APPLICATIONS

World out there:
…targeted adds

Chemistry:
…grouping analytes together for an experiment
…grouping samples together for evaluating sample preparation efficiency
…discovering new tissue functionalities with MALDI imaging

The subgroups are **unknown** and we have to find them: we are discovering the structure.

METHODS

K-means clustering
…dividing dataset into predefined number of clusters

Hierarchical clustering
…no predefined cluster number
…end up with a tree like visual representation

K-MEANS CLUSTERING

K-MEANS CLUSTERING

1. Specify the number of clusters

Each observation belongs to at least on of the *K* clusters
The clusters are non-overlapping

MATH

Good clustering: the within cluster variation (W) is as small as possible:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}.$$

we want to partition the observations into *K* clusters such that the total within-cluster variation, summed over all *K* clusters, is as small as possible.
Within cluster variation can be defined as *squared Euclidean distance*

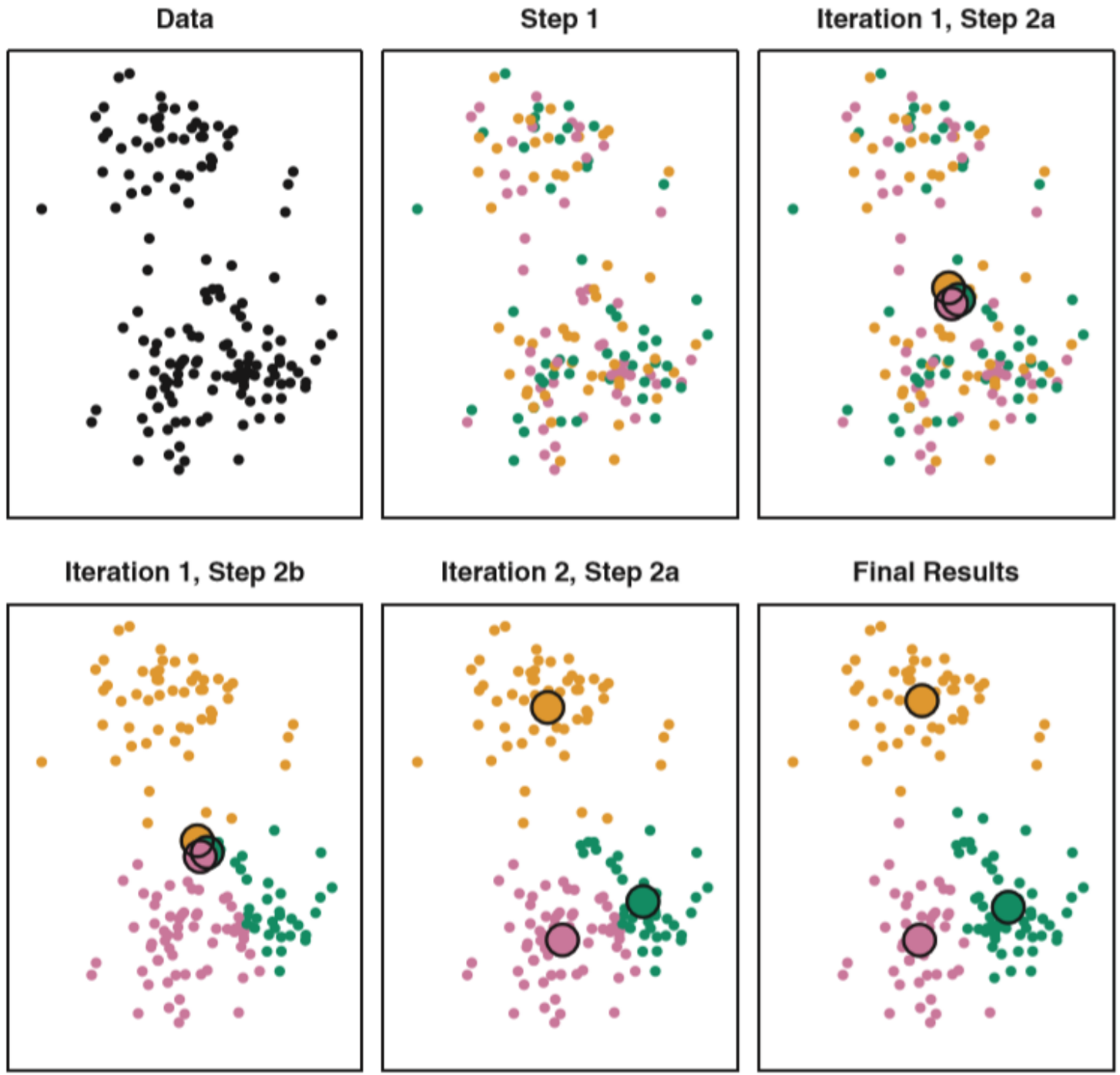
$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

ALGORITHM *K*-MEANS CLUSTERING

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 *k*th 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).



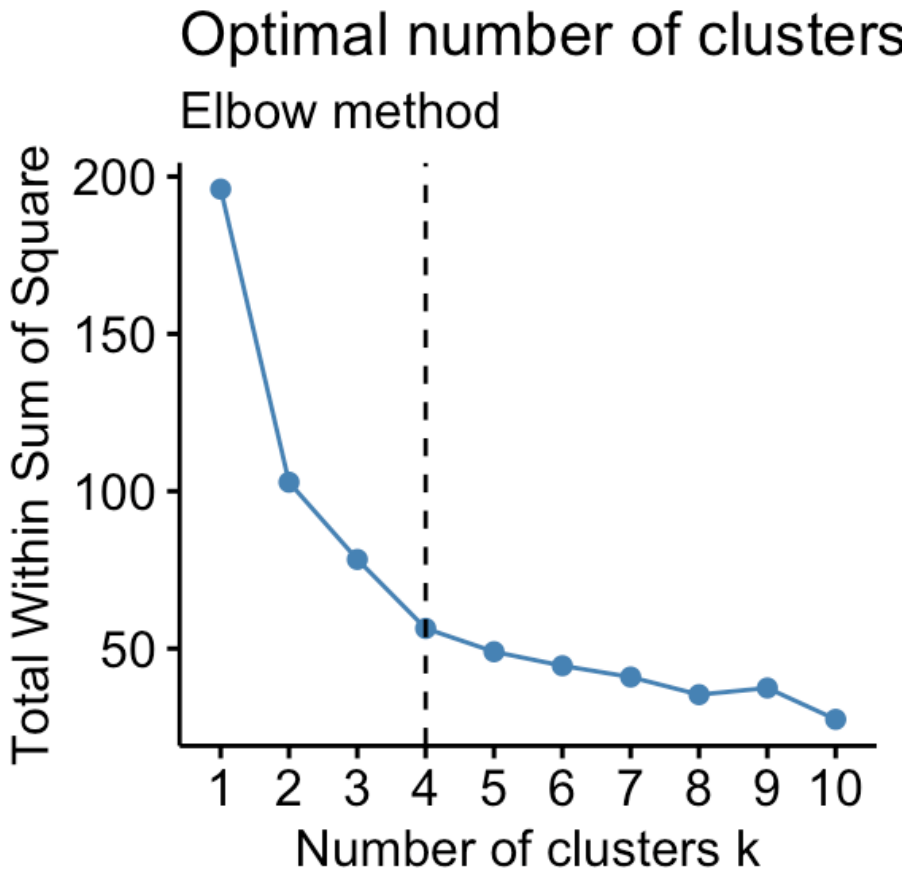
FINDING THE OPTIMA

K-means clustering finds the **local** optimum rather then a global
Result depends on the initial cluster assignment
Run the algorithm multiple times from different starting points
…then choose the best solution
…smallest within cluster variation over all clusters

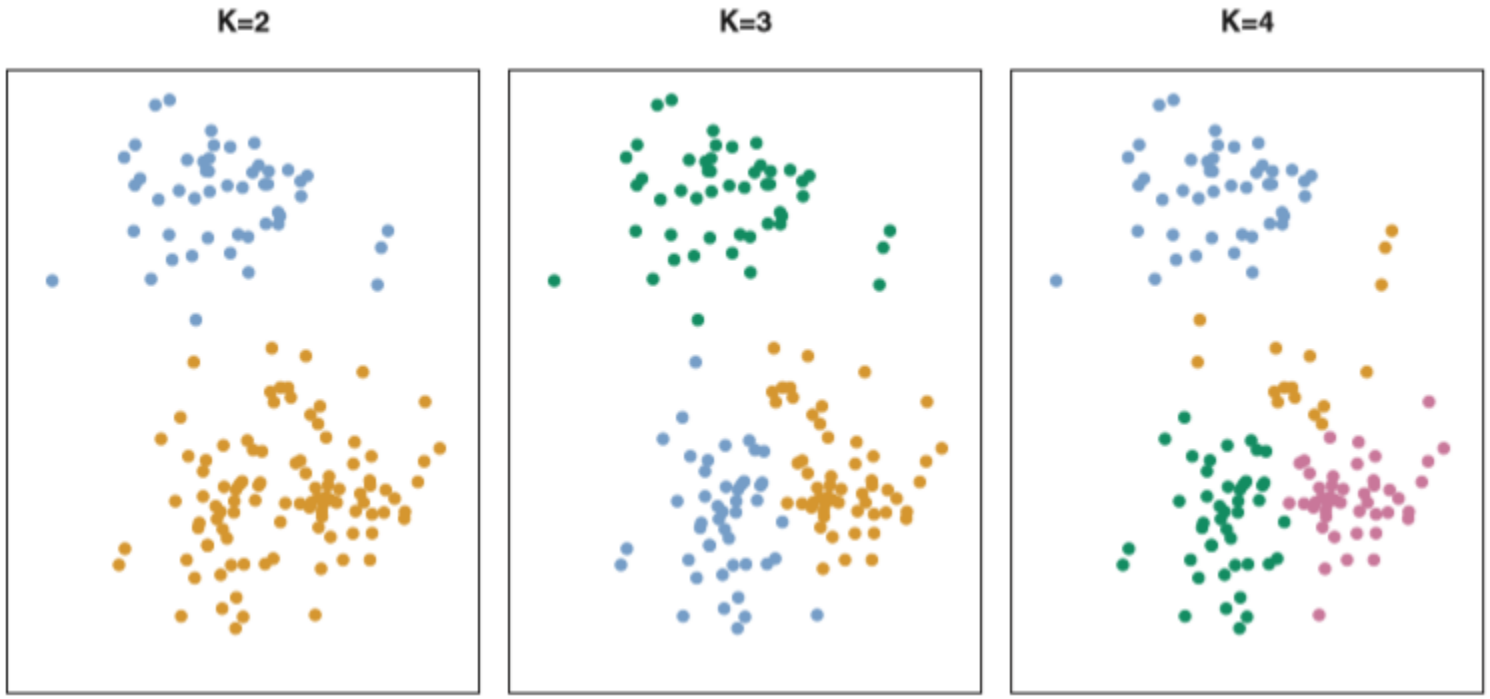

320.9
235.8
235.8
235.8
235.8
310.9

PROBLEMS

We need to decide on the number of clusters *K*
…scientific hypothesis
…testing several clusters
…Elbow method



NUMBER OF CLUSTERS



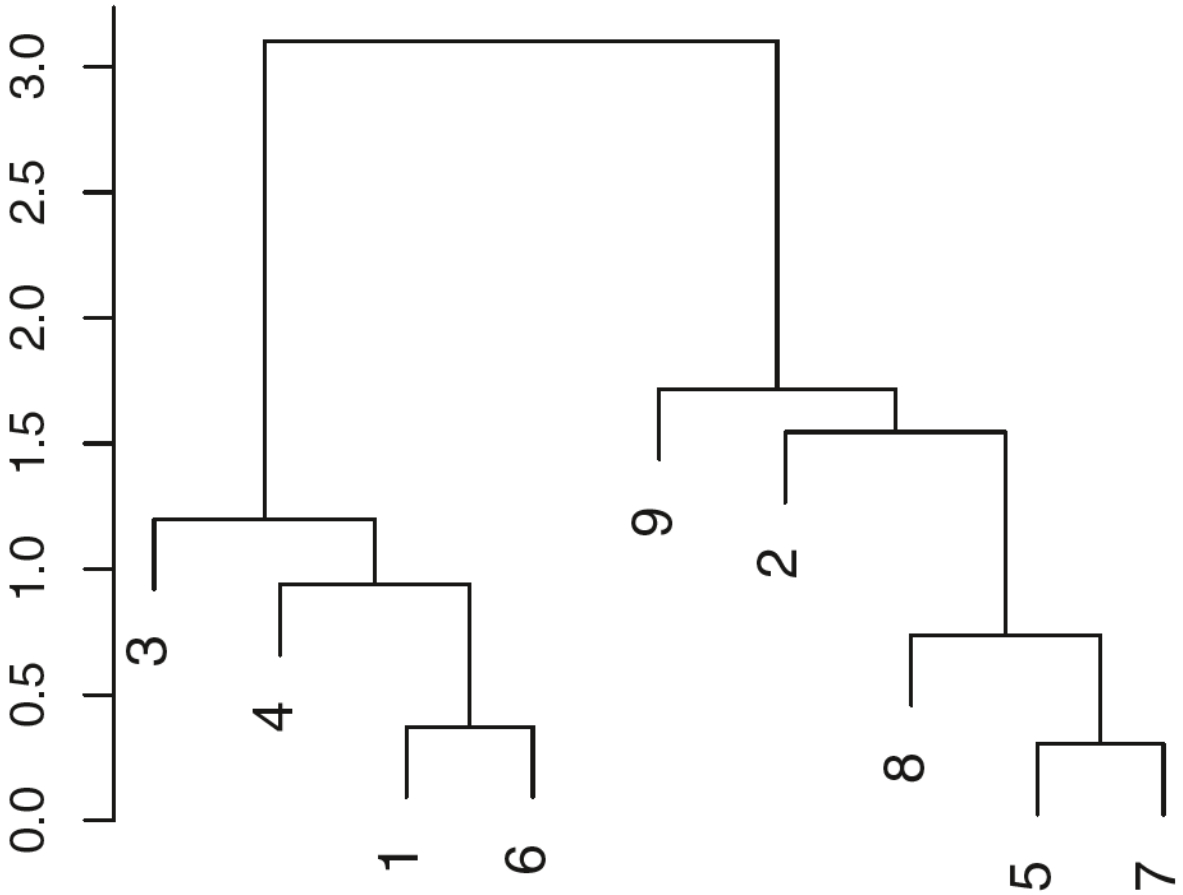
HIERARCHICAL CLUSTERING

HIERARCHICAL CLUSTERING

Aims at overcoming the need to specify *K*
Dendrogram a tree-based representation of the observations

Bottom-up or *agglomerative* – starting from the leaves nad combining clusters up to the trunk

DENDROGRAM



Anneli Kruve

DENDROGRAM

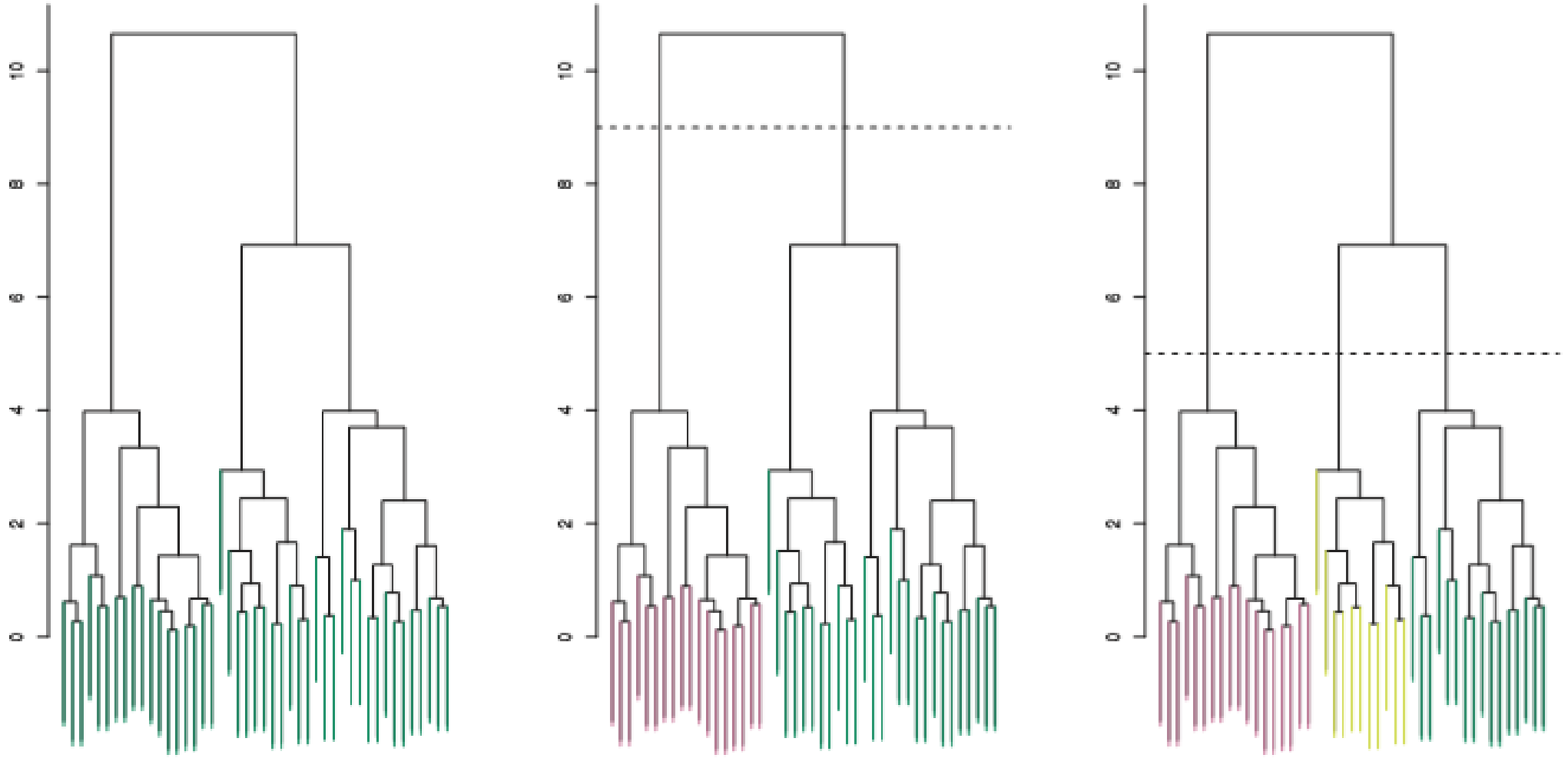
The earlier fusion occurs
…the **more similar** the observations are to each other

The height of this fusion (y axis) indicates **how different** the two observations are.

We **cannot** draw conclusions about the similarity of two observations based on their proximity along the x axis

Anneli Kruve

DENDROGRAM



Anneli Kruve

IDENTIFYING CLUSTERS

Cutting the dendrogram at a specified hight:
… 9 -> 2 clusters
… 5 -> 3 clusters
Any number of clusters from 1 to n can be obtained depending on the height of the cut
One dendrogram can be used to obtain any number of clusters
Sometimes “cutting the tree” is obvious by eye

Not always better then *K*-means, especially if several “real” clustering possibilities exist.

ALGORITHM: HIERARCHICAL CLUSTERING

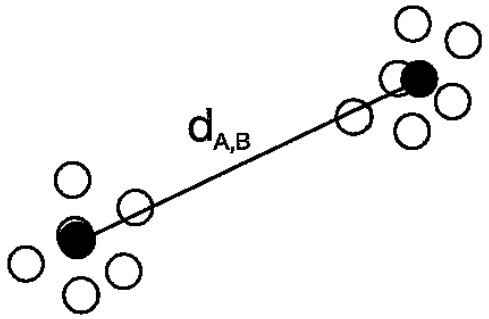
1. Begin with *n* observations and a measure (such as Euclidean distance) of all the pairwise dissimilarities. Treat each observation as its own cluster.
2. For $i = n, n-1, \ldots, 2$:
a. Examine all pairwise inter-cluster dissimilarities among the *i* clusters and identify the pair of clusters that 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.

LINKAGE TYPES

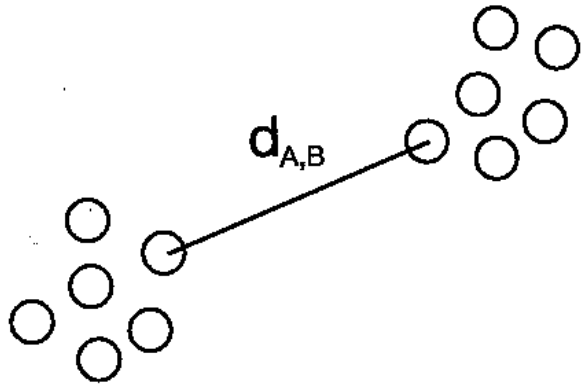
Linkage	Description
Complete	**Maximal intercluster dissimilarity**. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.
Single	**Minimal intercluster dissimilarity**. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time
Average	**Mean intercluster dissimilarity**. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversions.

LINKAGE TYPES

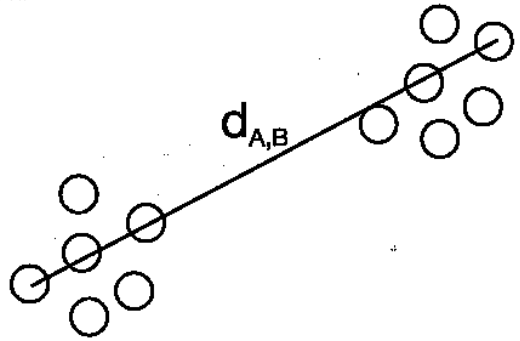
Centroid

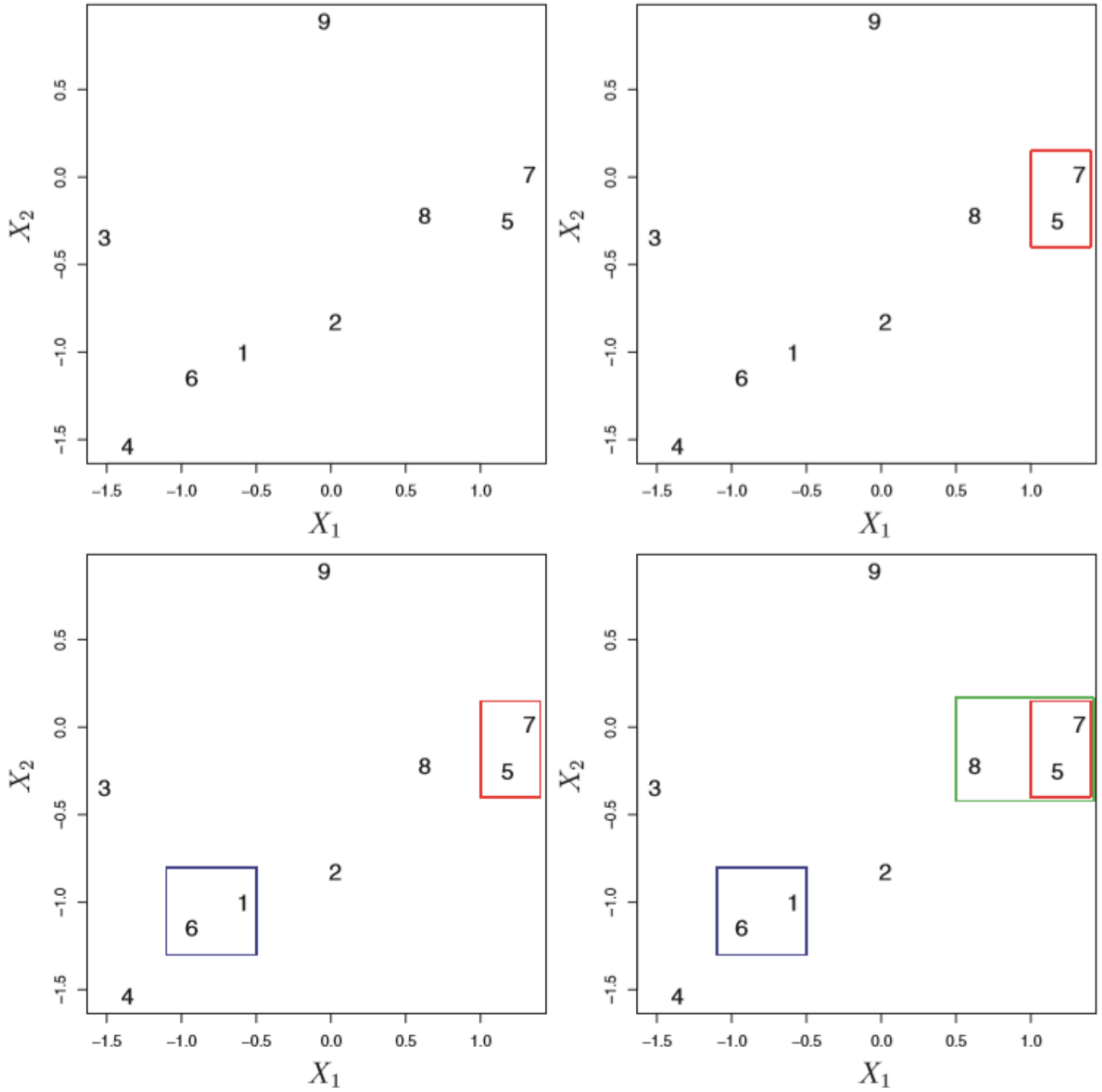


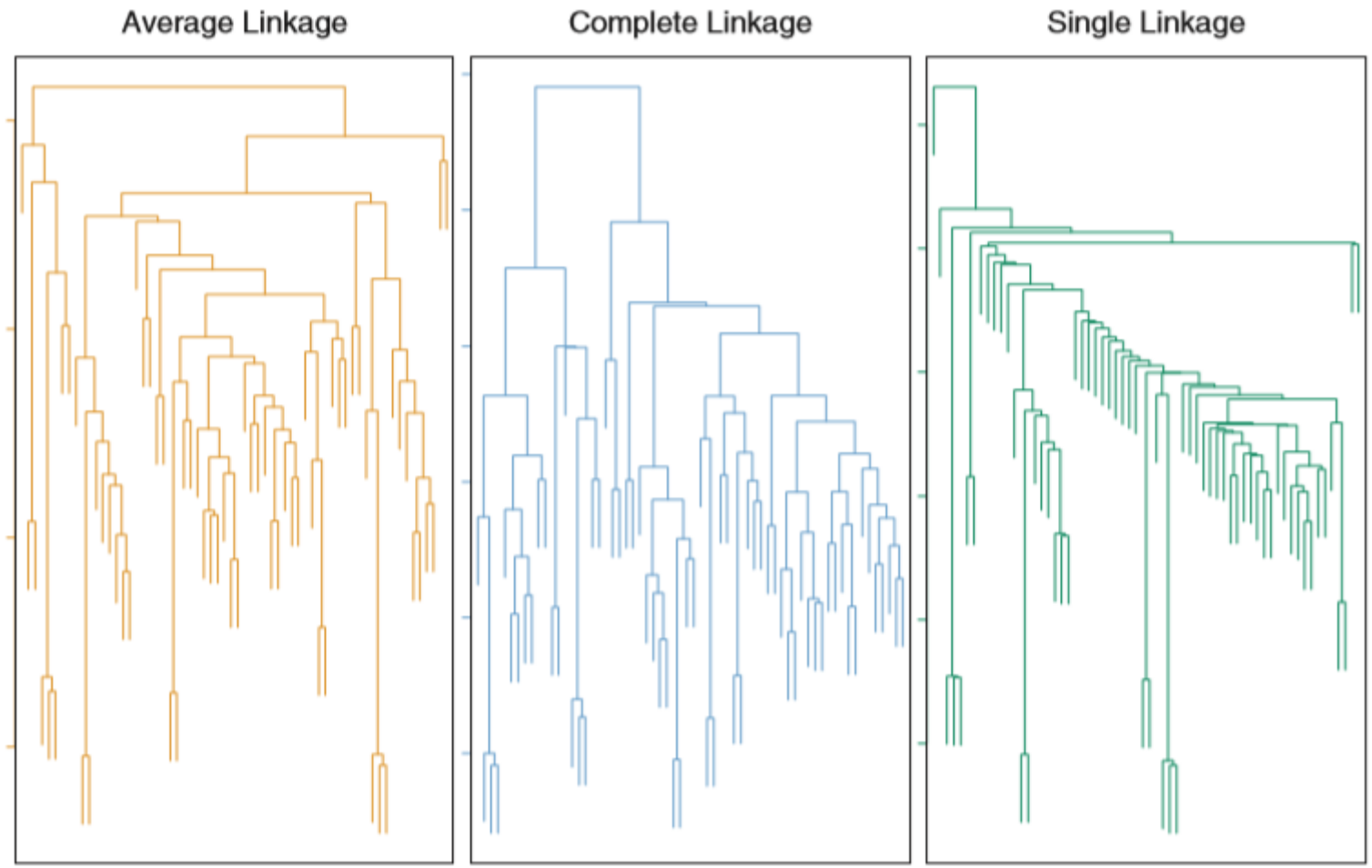
Single



Complete






Average Linkage
Complete Linkage
Single Linkage

QUESTIONS

Should the observations or features first be standardized in some way?
…scaling and centering

In the case of hierarchical clustering:
…what dissimilarity measure should be used?
…what type of linkage should be used?
…where should we cut the dendrogram in order to obtain clusters?

In the case of K-means clustering:
…how many clusters should we look for in the data?

IN PRACTICE

No single right answer
We try different choices
Analyse based on interpretability

VALIDATION

When we ask *K*-means clustering to find 2 clusters…
…it will always find us 2 clusters

Do the found clusters represent real groups?

Influence of small number of unknown subgroups

Often not very robust to adding/removing data…
…clustering subsets of data helps to evaluate the robustness

REMEMBER

These results should **not** be taken as **the absolute truth** about a data set.
Rather, they should constitute a **starting point** for the development of a **scientific hypothesis** and …
… further study, preferably on an **independent data set**.

CLUSTERING OF VARIABLES

We have viewed clustering based on samples
BUT clustering can be also done for the features
…in this case a good similarity metric is R^2 based

$$d_{i,k} = \sqrt{1 - R_{i,k}^2}$$