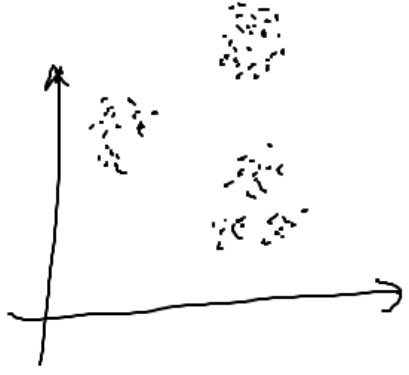
Unsupervised learning – continued

Clustering

Goal: partition the data into distinct groups (clusters) so that the observations within each group are (quite) similar to each other, while obs. in different groups are (quite) different from each other.



K-means clustering

We specify the desired number of clusters K.

Notation. $C_1, C_2, \dots, C_K \subset \{1, 2, \dots, n\}$ – subsets of indices of n observations $1, 2, \dots, n$

- $C_1 \cup C_2 \cup \dots \cup C_K = \{1, 2, \dots, n\}$
- $C_i \cap C_j = \emptyset$ for $i \neq j$

Then $C_1, C_2, \dots, C_K$ are <u>clusters</u> of (indices of) observations. $x_1, x_2, \dots, x_n \in \mathbb{R}^p$

Idea: the clustering is good if total within-cluster variation $W(C_j)$ is as small as possible,

i.e., we minimise $\sum_{j=1}^{K} W(C_j)$

Common choice: $W(C_j) = \frac{1}{|C_j|} \sum_{i,i' \in C_j} \underbrace{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}_{\|x_i - x_{i'}\|^2}$

This minimization problem is difficult to solve for K, n not very small.

K-means algorithm:

1) randomly assign a number from 1 to K to each of the observations (but each number to at least 1 observation). These serve as initial cluster assignment for the observations.

2) Iterate until assignments stop to change:

(a) for each of the clusters, find its <u>centroid</u>: $\bar{X}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i \in \mathbb{R}^p$

(b) (re)assign each observation to the cluster whose centroid is closest (w/r to Euclidean distance)

<u>Remark:</u>

Instead of 1), one can choose some $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k \in \mathbb{R}^p$ (e.g., some observations) and start from 2(b), and then iterate step 2) as before.

It turns out that iterating 2) decreases $\sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$

Indeed, $\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \sum_{j=1}^{p} \left[\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} (x_{ij} - x_{i'j})^2 \right] = \dots$

Denote $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$. Then

$$\sum_{i,i' \in C_k} (x_{ij} - x_{i'j})^2 = \sum_{i,i' \in C_k} (x_{ij} - \bar{x}_{kj} + \bar{x}_{kj} - x_{i'j})^2 = \sum_{i,i' \in C_k} (x_{ij} - \bar{x}_{kj})^2 + 2 \underbrace{\sum_{i \in C_k} (x_{ij} - \bar{x}_{kj})}_{\| \; \sum_i x_{ij} - |C_k| \cdot \bar{x}_{kj} = 0} \sum_{i' \in C_k} (\bar{x}_{kj} - x_{i'j})$$

$$+ \sum_{i,i' \in C_k} (\bar{x}_{kj} - x_{i'j})^2 = 2 |C_k| \cdot \sum_{i \in C_k} (x_{ij} - \bar{x}_{kj})^2$$

$$\dots = 2 \sum_{j=1}^{p} \left[\sum_{k=1}^{K} \sum_{i \in C_k} (x_{ij} - \bar{x}_{kj})^2 \right]$$

We need another observation:

$$\sum_{i \in C_k} \left(x_{ij} - \bar{x}_{kj}\right)^2 = \min_{a_j \in \mathbb{R}} \sum_{i \in C_k} \left(x_{ij} - a_j\right)^2$$

Indeed,

$$\sum_{i \in C_k} \left(x_{ij} - a_j\right)^2 = \sum_{i \in C_k} \left(x_{ij} - \bar{x}_{kj} + \bar{x}_{kj} - a_j\right)^2 =$$

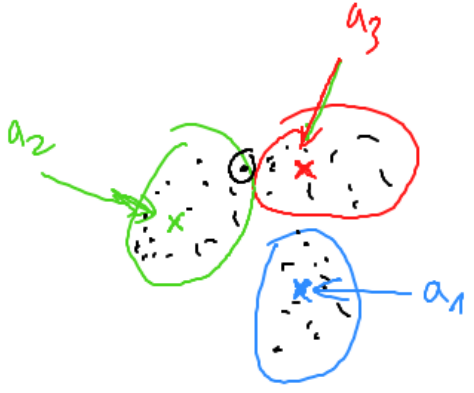
$$= \sum_{i \in C_k} \left(x_{ij} - \bar{x}_{kj}\right)^2 + 2 \underbrace{\sum_{i \in C_k} \left(x_{ij} - \bar{x}_{kj}\right)}_{=0} \underbrace{\left(\bar{x}_{kj} - a_j\right)}_{\text{does not depend on } i} + \sum_{i \in C_k} \left(\bar{x}_{kj} - a_j\right)^2$$

$$= \sum_{i \in C_k} \left(x_{ij} - \bar{x}_{kj}\right)^2 + \underbrace{|C_k| \cdot \left(\bar{x}_{kj} - a_j\right)^2}_{\geq 0} \geq \sum_{i \in C_k} \left(x_{ij} - \bar{x}_{kj}\right)^2$$

In step 2)

(a) compute the centroids $\bar{x}_k$

(b) reassign ...



we want to minimize: $2 \sum_{j=1}^{p} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i \in C_k} (x_{ij} - \underbrace{\bar{x}_{kj}}_{a_{kj}})^2$

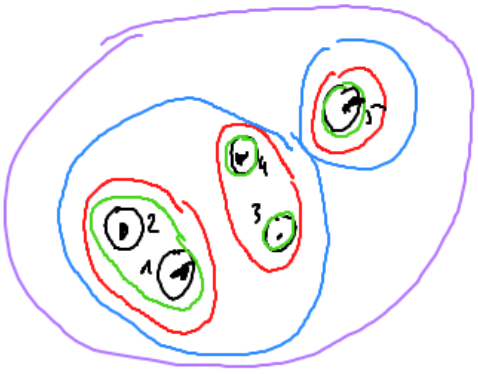
$\Rightarrow$ the sum $\sum_{k=1}^{K} W(C_k)$ decreases in step 2.

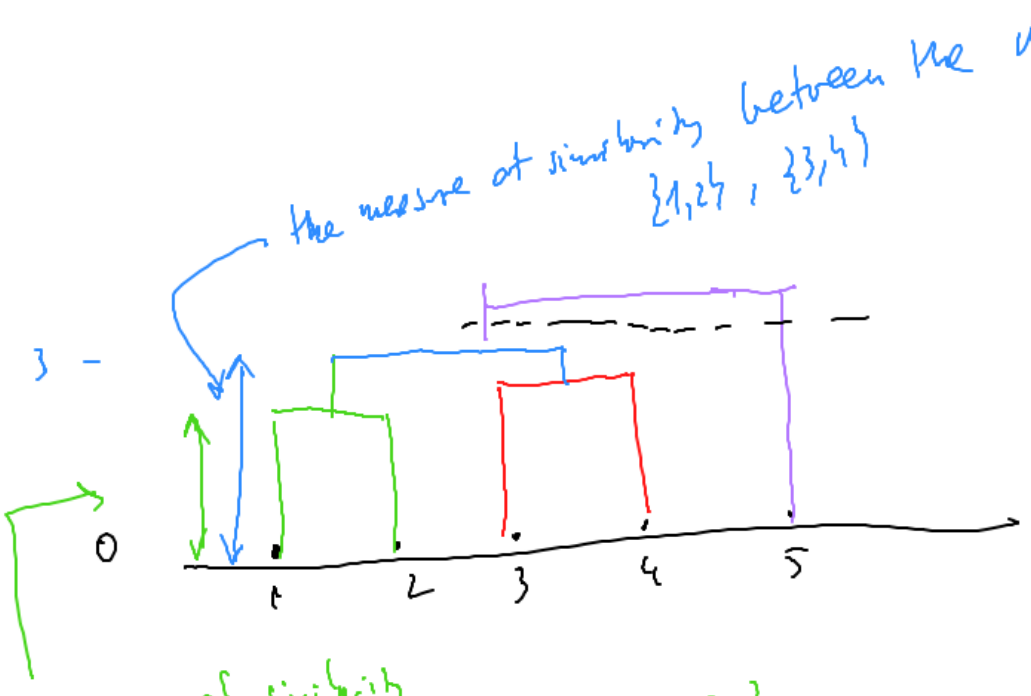
The clusters that we obtain do depend on the initial random assignment.

Therefore one usually does multiple runs of the K-means algorithm and checks the sum $\sum_{k=1}^{K} W(C_k)$ for each of the clusterings obtained.

The one with the smallest sum is the best solution found.

Hierarchical clustering





Algorithm

1) begin with n observations each in 1 cluster
   and a measure (e.g. Euclidean distance) of all the $\binom{n}{2}$ pairwise dissimilarities.

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 most similar.
   Fuse these two clusters.
   The dissimilarity between these two clusters is the height on the dendrogram
   at which fusion should be placed.
   (b) compute the new pairwise inter-cluster dissimilarities among the remaining i-1
   clusters

Common choices for the dissimilarity measure ("linkage")

Linkage

complete $$d(C_j, C_k) = \max_{\substack{i \in C_j \\ i' \in C_k}} \|x_i - x_{i'}\|$$

single $$d(C_j, C_k) = \min_{\substack{i \in C_j \\ i' \in C_k}} \|x_i - x_{i'}\|$$

average $$d(C_j, C_k) = \frac{1}{|C_j||C_k|} \sum_{\substack{i \in C_j \\ i' \in C_k}} \|x_i - x_{i'}\|$$

ward $$d(C_j, C_k) = \sum_{\substack{i \in C_j \\ i' \in C_k}} \|x_i - x_{i'}\|$$

Hierachical:

K=2

K=3

K-means

K=2

K=3