Intro to Cluster Analysis

Clutsering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set. When we cluster the observations, we partition the profiles into distinct groups so that the profiles are similar within the groups but different from other groups. To do this, we must define what makes observations similar or different.

PCA vs Clustering

Both clustering and PCA seek to simplify the data via a small number of summaries, but their mechanisms are different:

- PCA searches for a low-dimensional representation of the observations that explains a good fraction of the variance
- Clustering looks to find homogeneous subgroup among the observations

Notation

Input is data with p variables and n subjects

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ \cdots & \cdots & & & \cdots \\ x_{n1} & x_{n2} & & & x_{np} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \\ \\ x_n \end{pmatrix}$$

Row vectors

Row i is the **profile** of the ith subject

A distance between two vectors i and j must obey several rules:

- The distance must be positive definite, $d_{ij} >= 0$
- The distance must be symmetric, $d_{ij} = d_{ij}$, so that the distance from j to i is the same as the distance from i to j
- An object is zero distance from itself, $d_{ii} = 0$
- The triangle rule - When considering three objects i, j and k the distance from i to k is always less than or equal to the sum of the distance from i to j and the distance from j to k $d_{ik} <= d_{ij} + d_{jk}$

Clustering Procedures

- **Hierarchical clustering**: Iteratively merges profiles into clusters using a simple search. Start with each profile/cluster and end with one 1. The clustering procedure is represented by a *dendrogram*.

- **K-mean clustering:** Start with a per-specified number of clusters and random allocation of profiles to clusters. Iteratively move profiles from one cluster to the other to optimize some criterion. End up with the same number of clusters.

K-Means Clustering

We partition the K clusters so that we maximize the similarity within clusters and minimize the similarity between clusters.

We can represent the data with a vector of means, which is the overall profile or cluster:

$$\mu = \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 & \dots & \mu_p \end{pmatrix}; \ \mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$TotSS = \sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2$$

Supposed we split the data into 2 clusters, C1 with □□1 observations of the p variables, and C2 with □□2 observations of the p variables. We would have two different vectors of means representing the centroids of the clusters. The total sum of squares within clusters would be:

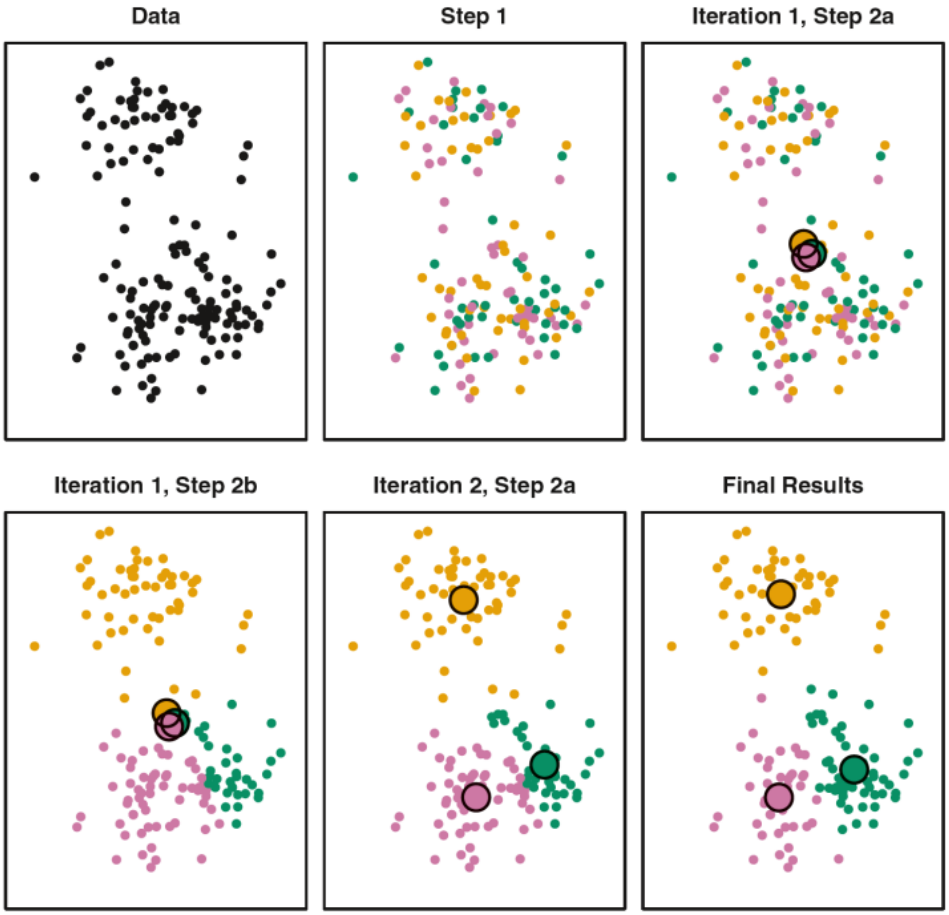
$$WSS = \sum_{x_i in C1}\sum_{j=1}^{p}(x_{ij} - \mu_1)^2 + \sum_{x_i in C2}\sum_{j=1}^{p}(x_{ij} - \mu_2)^2 \leq TotSS$$

And we seek to keep WSS small, but it is NOT guaranteed to give the minimum WSS so ideally one should start from different initial values.

- Start with K "random" clusters by assigning each of the n profiles to one of the K clusters at random
- Iterate until no more changes are possible:
  - For each of the K clusters, compute the cluster profile (centroid)
  - Assign each observation to the cluster whose centroid is closest (using Euclidean distance)

$$d_E(x_1, x_2) = \sum\nolimits_j (x_{1j} - x_{2j})^2$$

Example with K = 3, p = 2:



Step 1: random start

Iteration 1
Step 2a: compute centroid of each cluster
Step 2b: compute the distance of each point from the 3 centroids, and assign each point to the cluster with minimum distance

Iteration 2
Step 2a: compute centroid of each cluster
Step 2b: compute the distance of each point from 3 centroids, and assign each point to the cluster with minimum distance

Stop when no relabeling occurs

Standardization

When the variables are measured on different scales the measurement units may bias the cluster analysis. The Euclidean distance is not scale invariant.

Hierarchical Clustering

This is an alternative approach that does not require a fixed number of clusters. The algorithm essentially rearranges profiles so that similar profiles are displayed next to each other in a tree (dendrogram) and the dissimilar profiles are displayed in different branches.

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ \cdots & \cdots & & & \cdots \\ x_{n1} & x_{n2} & & & x_{np} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \\ \\ x_n \end{pmatrix}$$

$n \times n$ **Dissimilarity matrix**

$$\begin{matrix} s(x_1,x_1) & s(x_1,x_2) & s(x_1,x_3)\cdots \\ s(x_2,x_1) & s(x_2,x_2) & s(x_2,x_3)\cdots \\ s(x_3,x_1) & s(x_3,x_2) & s(x_3,x_3)\cdots \end{matrix}$$

Only $n \times (n-1)/2$ elements matter

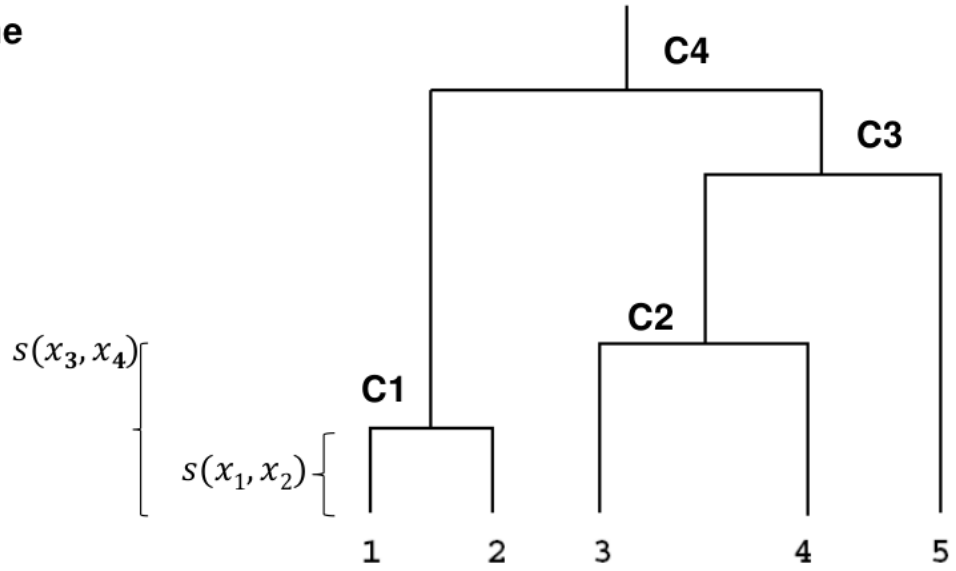
$$s(x_1,x_2) < s(x_3,x_4) < s(x_3,x_5) < s(x_4,x_5) < s(x_1,x_4) < s(x_2,x_4)$$

Tree structure representation of the clustering steps.

Profiles 1 and 2 are the most similar

Profiles 3 and 4 are second best most similar.

Profile 5 is more similar to 3 and 4 rather than 1 and 2



The branch length is an indication of the distance

We do this we define the similarity (distance) between:

- Two profiles: s(xi, xi)
- A profile and a cluster: s(xc, xi)
- Two clusters: s(xc, xc)

Similarity Between Clusters

Complete-Linkage Clustering

- Also known as the maximum or furthest-neighbor method
- The distance between two clusters is calculated as the greatest distance between members of the relevant clusters
- This method tends to produce very compact clusters of elements and the clusters are often similar in size

Single-Linkage Clustering

- Referred to as the minimum or nearest neighbor method
- The distance between two clusters is calculated as the minimum distance between members of the relevant clusters
- This method tends to produce clusters that are 'loose' because clusters can be joined if any two members are close together. In particular, this method often results in 'chaining' or sequential addition of single samples to an existing cluster and produces trees with many long, single-addition branches.

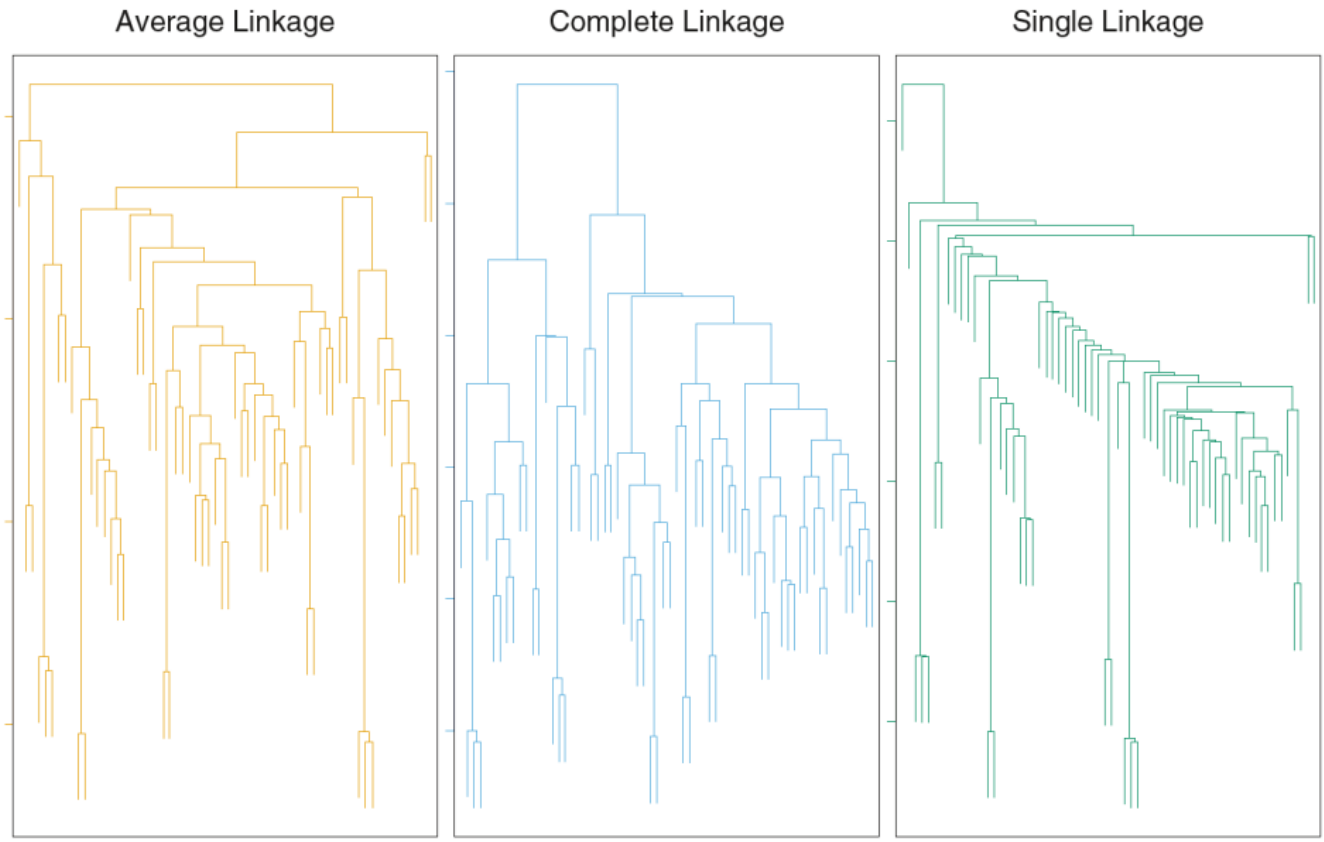
Average-Linkage Clustering

- The distance between clusters is calculated using average values. There are various methods used for calculating this average. The most common is the unweighted pair-group method average (UPGMA), where each average is calculated from the distance between each point in a cluster and all other points in another cluster and the 2 clusters with the lowest average distance are joined together into a new cluster.

Centroid Clustering

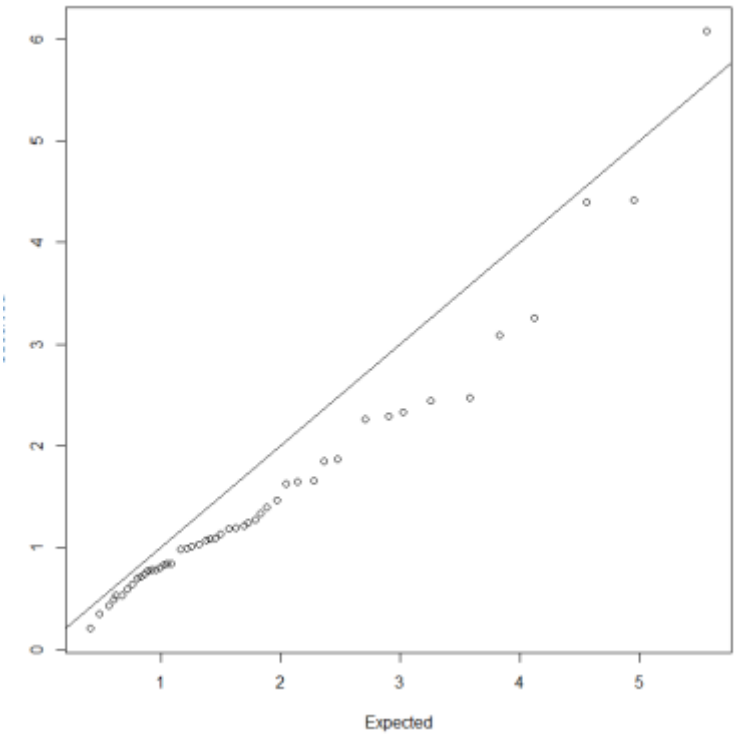
- Related methods substitute the centroid for the average.

Complete and average linkage are similar, but complete linkage is faster because it does not require recalculation of the similarity matrix at each step.



Detection of Clusters

Inspection of the QQ-plot would inform about the existence of clusters in the data.



- Observed and “expected” distances are statistically indistinguishable would suggest that there are no clusters in the data
- Departure of the QQ-plot from the diagonal line would suggest that there are clusters

We can also use a 95%-tile to detect the number of clusters using extreme percentiles of the reference distribution. The idea is to color the entry of the data set, so that the colors represent the standardized difference of the cell intensity from a baseline. Typically, columns are samples and rows are variables.

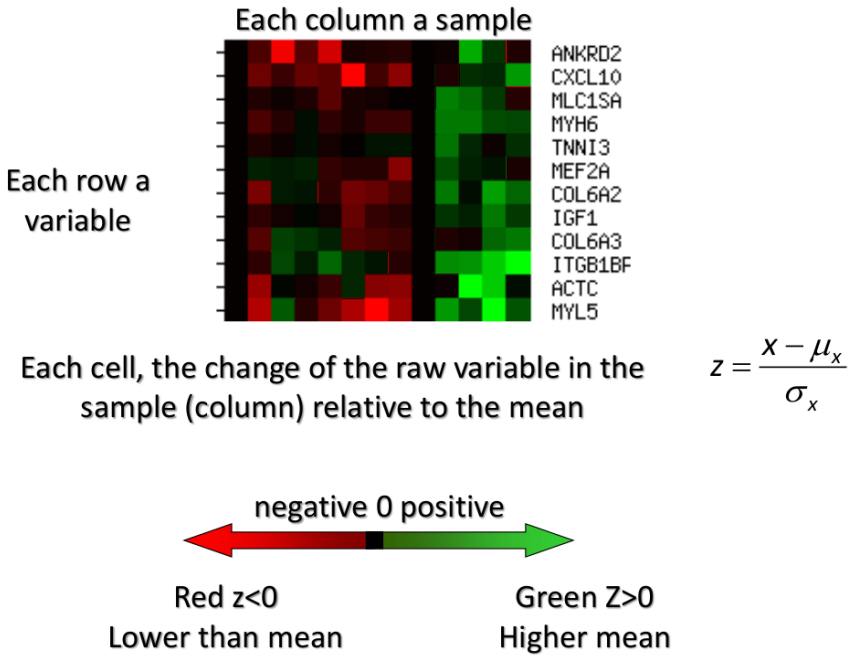
Samples

variables

$$\begin{matrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \end{matrix}$$

p x n

$$\begin{matrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\text{var}(x_1)}} & \frac{x_{12}-\bar{x}_1}{\sqrt{\text{var}(x_1)}} & \frac{x_{13}-\bar{x}_1}{\sqrt{\text{var}(x_1)}} & \frac{x_{14}-\bar{x}_1}{\sqrt{\text{var}(x_1)}} \\ \frac{x_{21}-\bar{x}_2}{\sqrt{\text{var}(x_2)}} & \frac{x_{22}-\bar{x}_2}{\sqrt{\text{var}(x_2)}} & \cdots & \cdots \\ \frac{x_{31}-\bar{x}_3}{\sqrt{\text{var}(x_3)}} & \frac{x_{32}-\bar{x}_3}{\sqrt{\text{var}(x_3)}} & \cdots & \cdots \\ \frac{x_{41}-\bar{x}_4}{\sqrt{\text{var}(x_4)}} & \cdots & \cdots & \cdots \end{matrix}$$



Revision #2
Created 16 November 2022 01:15:01 by Elkip
Updated 16 November 2022 15:32:49 by Elkip