Clustering

techniques for finding subgroups/clusters in a dataset

looks for homogeneous subgroups among the observations

partition the profiles into distinct groups so the profiles within each group are very similar to each other and profiles in different groups are very different from each other

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

$$x_1 = [x_{11} \quad x_{12} \quad \cdots \quad x_{1p}], x_2 = [x_{11} \quad x_{12} \quad \cdots \quad x_{1p}], \ldots x_n = [x_{n1} \quad x_{n2} \quad \cdots \quad x_{np}]$$

row vectors

row i is the profile on the i^{th} subject

Distances

Euclidean distance	$d_E(x_1, x_2)$	$\sum (x_{1i} - x_{2i})^2$
maximum	$d_{max}(x_1, x_2)$	$max_i(x_{1i} - x_{2i})^2$
Canberra	$d_c(x_1, x_2)$	$\sum \frac{\lvert x_{1i} - x_{2i} \rvert}{\lvert x_{1i} + x_{2i} \rvert}$
correlation	$s(x_1, x_2)$	$\frac{\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum (x_{1i} - \bar{x}_1)^2 \sum (x_{2i} - \bar{x}_2)^2}$

vectors a, b, c

$d_{ab} \geq 0$	distances must be positive definite
$d_{ab} = d_{ba}$	distances must be symmetric
$d_{aa} = 0$	object is zero distance from itself
$d_{ac} \leq d_{ab} + d_{bc}$	triangle rule

Clustering Procedures

K-Mean Clustering

partition data into k clusters to maximize similarity within cluster and minimize similarity between clusters

vector of means represents the overall profile

$$\mu = [\mu_1 \quad \mu_2 \quad \mu_3 \quad \cdots \quad \mu_p]$$

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

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

e.g. split the data into 2 clusters

centroid = cluster profile that represents variable means

C1 with n_1 observations of the p variables

$$\mu_1 = [\mu_{11} \quad \mu_{12} \quad \mu_{13} \quad \cdots \quad \mu_{1p}]$$

$$\mu_{1j} = \frac{1}{n_1} \sum_{x_i \ in\ C_1}^{n} x_{ij}$$

C2 with n_2 observations of the p variables

$$\mu_2 = [\mu_{21} \quad \mu_{22} \quad \mu_{23} \quad \cdots \quad \mu_{2p}]$$

$$\mu_{2j} = \frac{1}{n_2} \sum_{x_i \ in\ C_2}^{n} x_{ij}$$

$$WSS = \sum_{x_i \ in\ C_1} \sum_{j=1}^{p} (x_{ij} - \mu_1)^2 + \sum_{x_i \ in\ C_2} \sum_{j=1}^{p} (x_{ij} - \mu_2)^2 \leq TotSS$$

start with k random clusters by randomly assigning each of the n profiles to one of the k clusters
calculate the centroid for each of the k clusters
assign each observation to the cluster whose centroid has the closest Euclidean distance
repeat until no more changes are possible
decreases WSS and increases BSS at each step
doesn't always find the minimum WSS, so start from different initial values

WSS tends to decrease as the number of clusters increases
choose the k when the WSS stops decreasing sharply

when the variables are measured on different scales, measurement units can bias the cluster analysis because Euclidean distance isn't scale invariant

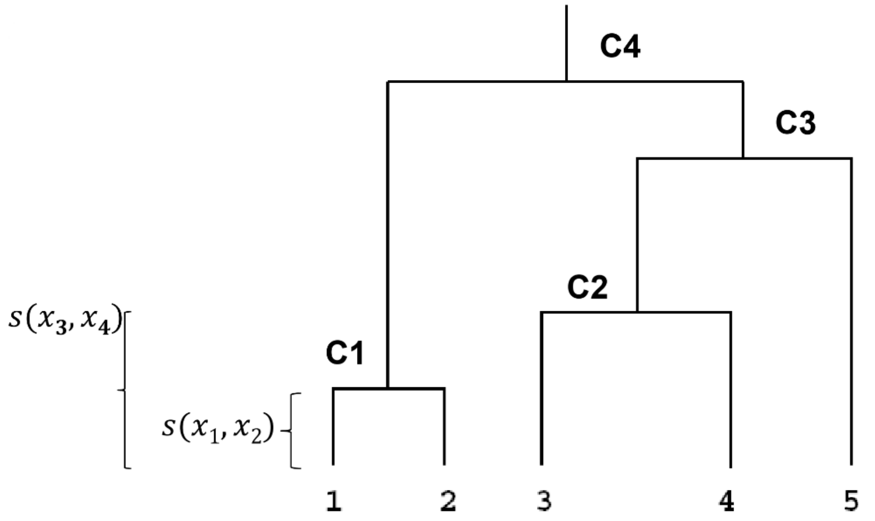
<u>Hierarchical Clustering</u>

iteratively merges profiles into clusters using a simple search
start with each profile being a cluster and end with only one cluster
similar profiles are displayed in the same branch and dissimilar profiles are displayed in different branches of the dendrogram tree

$n \times n$ dissimilarity matrix

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

only $n \times \frac{n-1}{2}$ elements matter



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

profiles 1 and 2 are the most similar
profiles 3 and 4 are the second most similar
profile 5 is more similar to 3&4 than 1&2

branch length is an indication of the distance
centroid = cluster profile that's a summary of the data allocated to the same cluster

Complete-Linkage Clustering
maximum/furthest neighbor method
distance between two clusters is calculated as the greatest distance between members of the relevant clusters
produces compact clusters of elements
clusters are often very similar in size

Single-Linkage Clustering
minimum/nearest neighbor method
distance between two clusters is calculated as the minimum distance between members of the relevant clusters
produces loose clusters because clusters can be joined if any 2 members are close together
results in sequential addition of single samples to an existing cluster
produces trees with many long, single-addition branches

Average-Linkage Clustering
distance between clusters is calculated using average values
average distance is the distance between each point in a cluster and all the other points in another cluster
the two clusters with the lowest average distance are joined together to form a new cluster

complete and average linkage are similar
complete linkage is faster because it doesn't require recalculation of the similarity matrix at each step

Detection of Significant Clusters

H_0: There are no significant clusters.
H_1: At least one of the clusters is significant.

generate many dendrograms from data with no clusters by permutating data
generate a reference distribution of heights under the null hypothesis
if the observed and expected distances are statistically indistinguishable, it suggests no clusters
departure from the diagonal line on the QQ-plot indicates clusters

Heatmaps

colors of the dataset represent the standardized difference of the cell intensity from a baseline
columns are samples and rows are variables
rescale variables to standardize them