**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

row vectors

row is the profile on the subject

**Distances**

Euclidean distance

maximum

Canberra

correlation

vectors

distances must be positive definite

distances must be symmetric

object is zero distance from itself

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

e.g. split the data into 2 clusters

centroid = cluster profile that represents variable means

C1 with observations of the variables

C2 with observations of the variables

start with k random clusters by randomly assigning each of the 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

Hierarchical Clustering

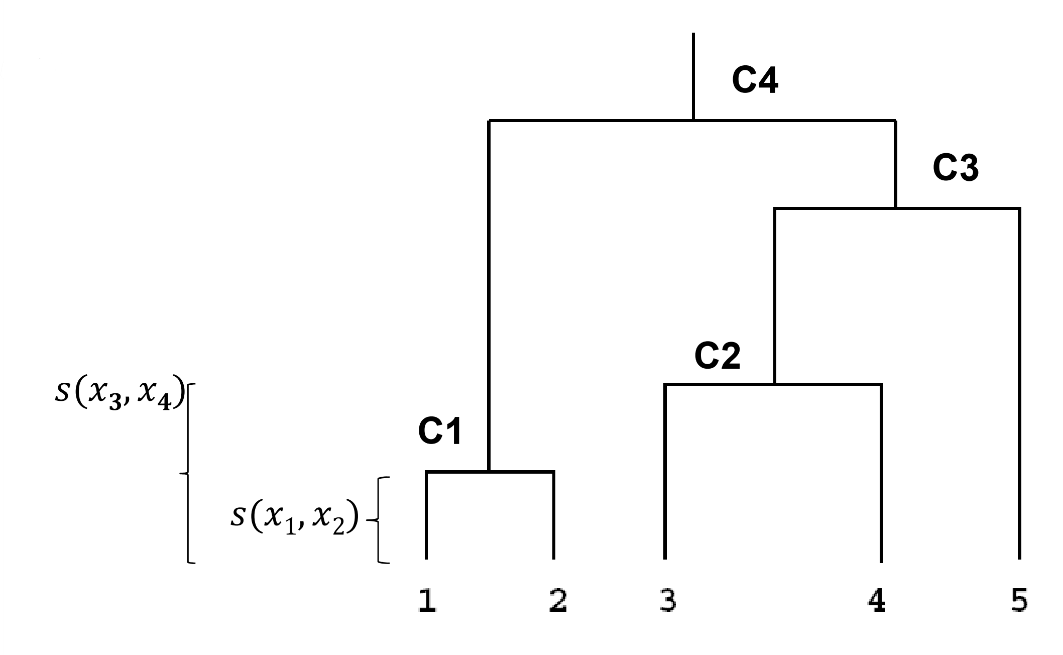
iteratively merges profiles into clusters using a simple search

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

dissimilarity matrix

only elements matter



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**

H0: There are no significant clusters.

H1: 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