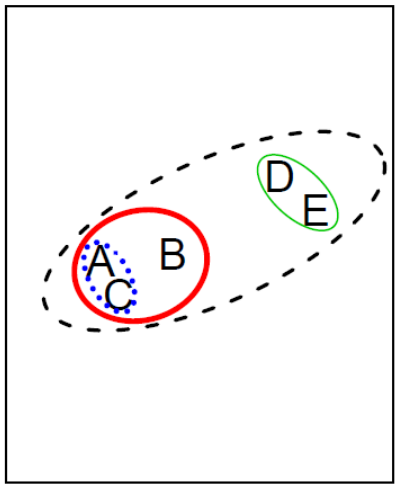
HIERARCHICAL CLUSTERING

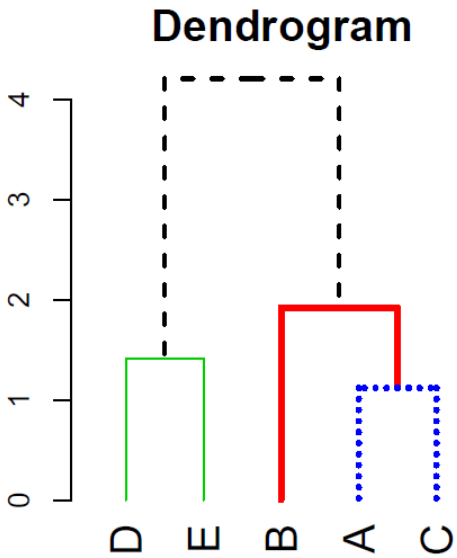
HIERARCHICAL CLUSTERING

- K-means clustering requires to pre-specify the number of clusters K.
- This can be a disadvantage
- Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K
- We describe bottom-up or agglomerative clustering
- This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk

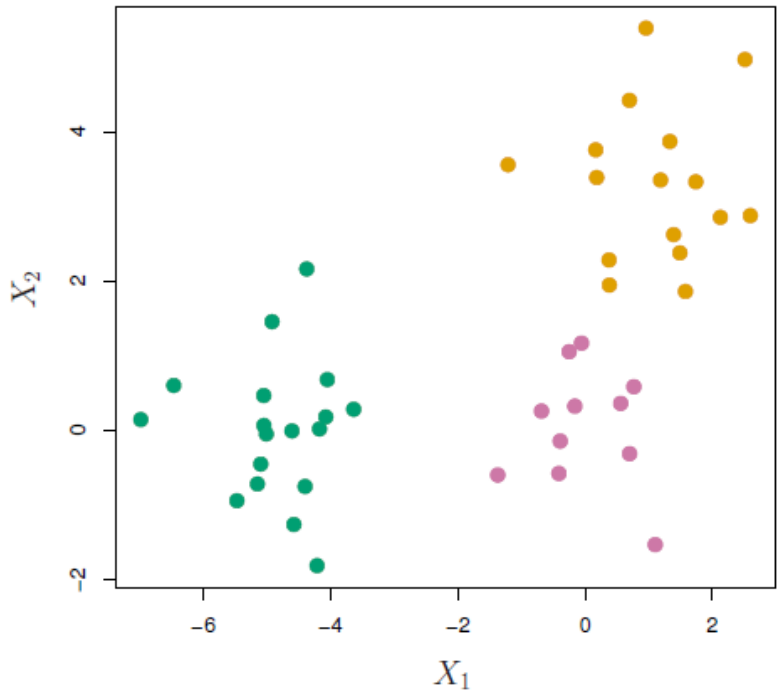
HIERARCHICAL CLUSTERING ALGORITHM

- Start with each point as a separate cluster (n clusters)
- Calculate the measure of dissimilarity between all points/clusters
- Fuse two clusters that are most similar so that there are now $n - 1$ clusters
- Fuse next two most similar clusters so there are now $n - 2$ clusters
- Continue until there is only 1 cluster





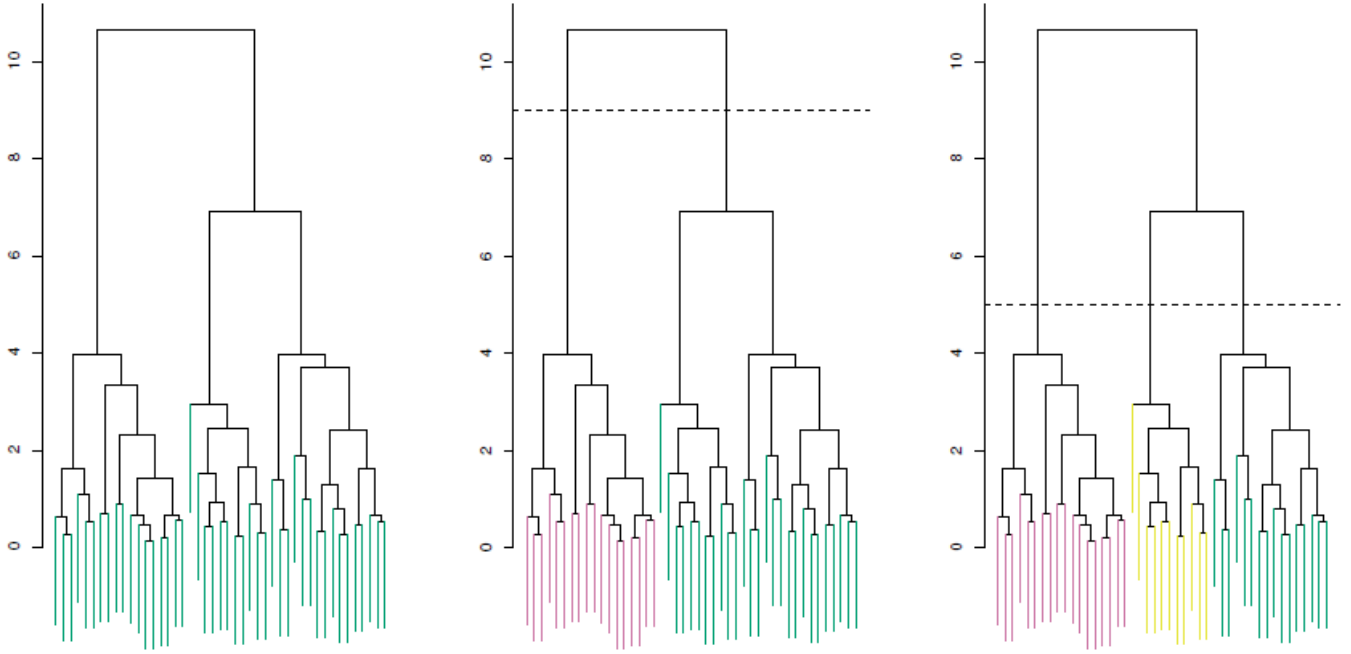
EXAMPLE



- 45 observations generated in 2-dimensional space. In reality there are three distinct classes, shown in separate colors
- However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data

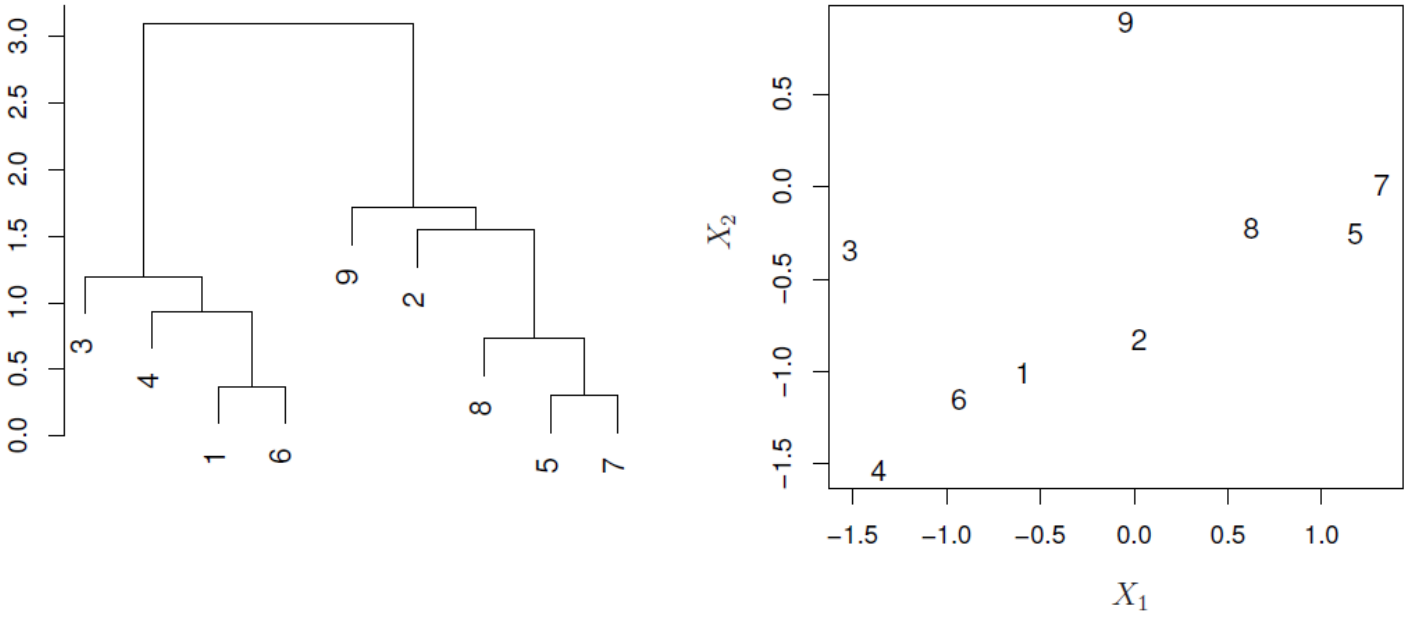
APPLICATION OF HIERARCHICAL CLUSTERING

- To choose clusters we draw lines across the dendrogram
- We can form any number of clusters depending on where we draw the break point
- We draw conclusions about the similarity of two observations based on the location on the *vertical axis* showing *dissimilarity* between *objects*



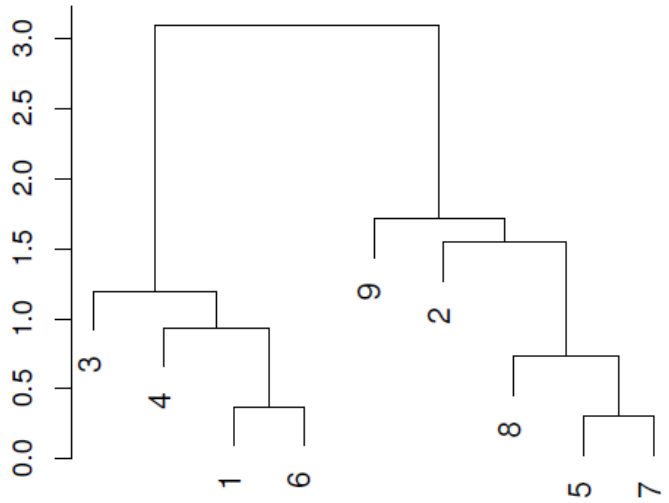
One Cluster | Two Clusters | Three Clusters

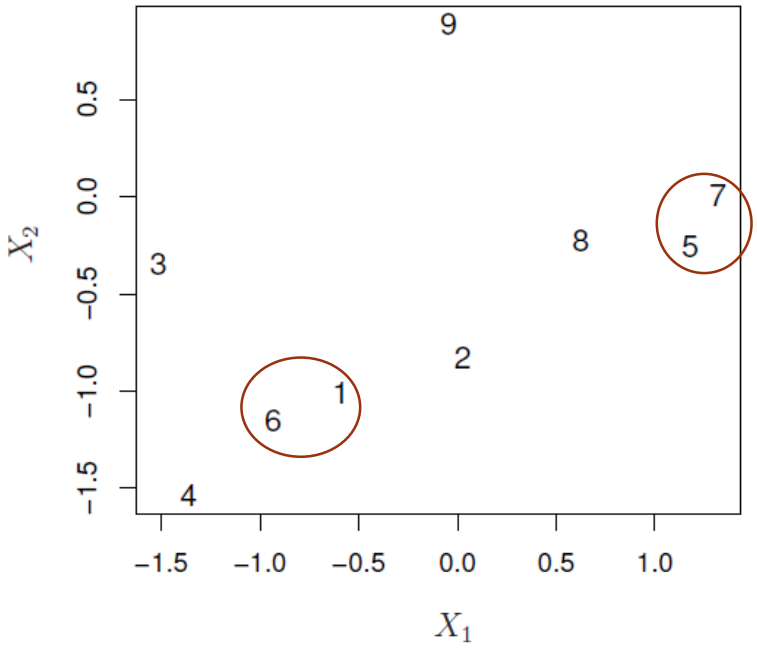
ANOTHER EXAMPLE



- An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space
- The raw data on the right was used to generate the dendrogram on the left

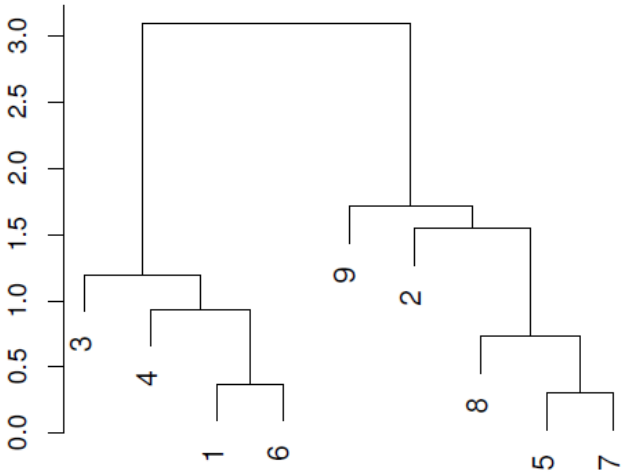
ANOTHER EXAMPLE

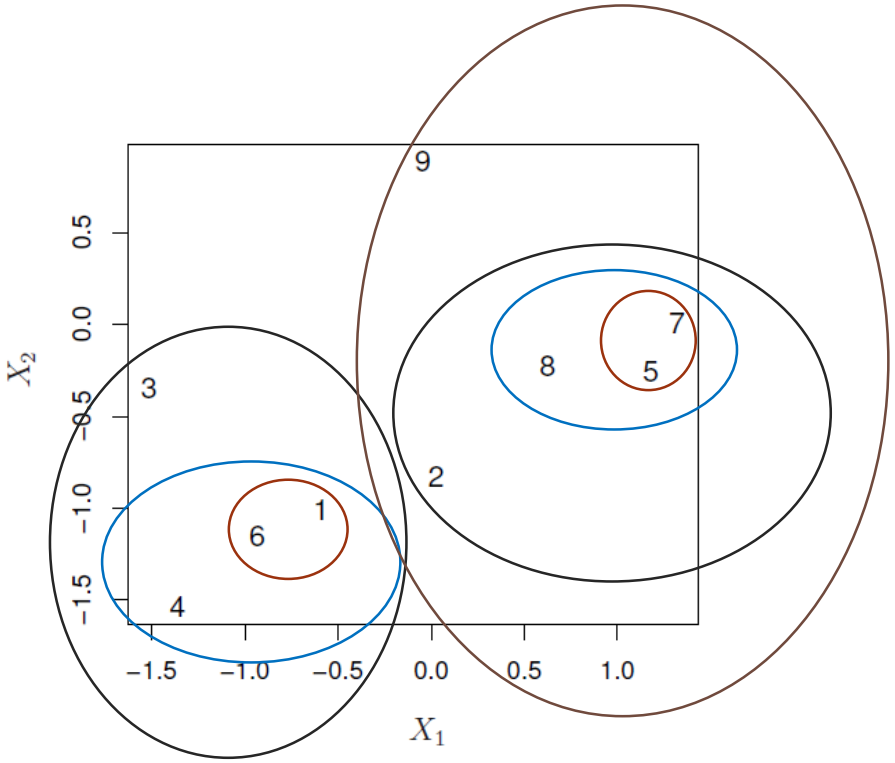




- Observations 5 and 7 are quite similar to each other, as are observations 1 and 6

ANOTHER EXAMPLE





- Observation 9 is no more similar to observation 2 than it is to observations 8; 5; and 7, even though observations 9 and 2 are close together in terms of horizontal distance
- This is because observations 2, 8, 5, and 7 all fuse with observation 9 at the same height, approximately 1.8 (dissimilarity measure)

LINKAGE: DISTANCE BETWEEN CLUSTERS

- Implementing hierarchical clustering involves one obvious issue
- How do we define the dissimilarity, or linkage, between the fused (5,7) cluster and 8?
- There are four options:
  - Complete Linkage
  - Single Linkage
  - Average Linkage
  - Centroid Linkage

COMPLETE LINKAGE

- Maximal inter-cluster 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 LINKAGE

- Minimal inter-cluster dissimilarity
- Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities

AVERAGE LINKAGE

- Mean inter-cluster dissimilarity
- Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities

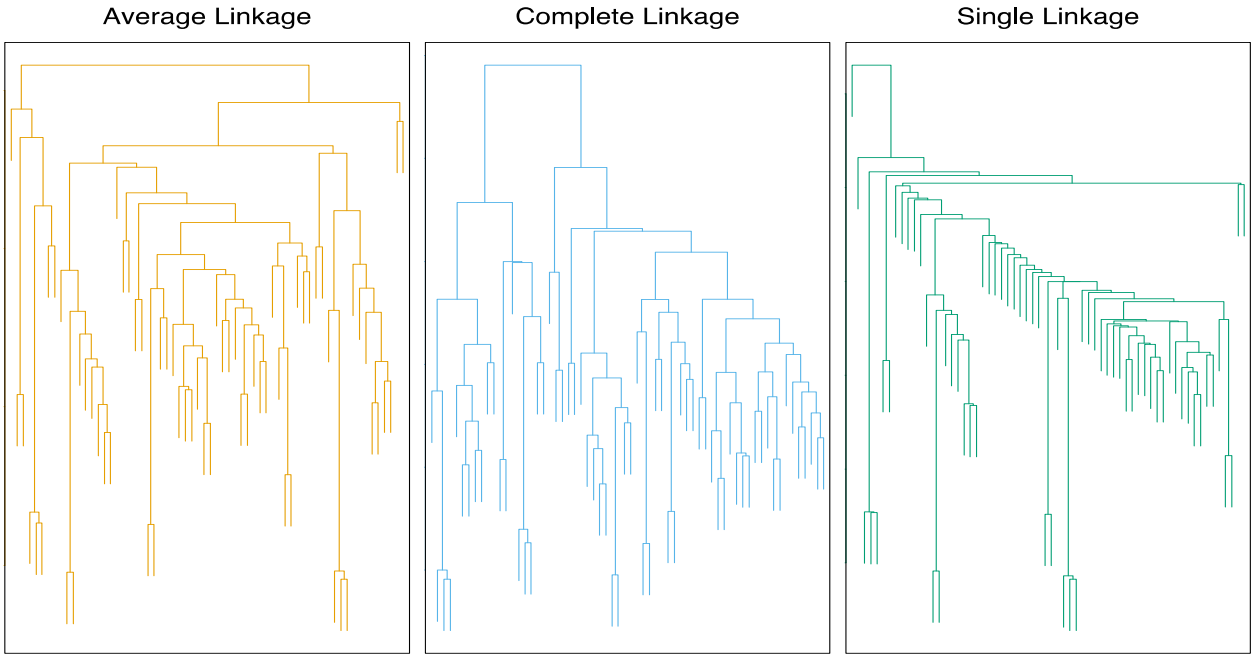
CENTROID LINKAGE

- Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B
- Centroid linkage is often used in genomics

SUMMARY

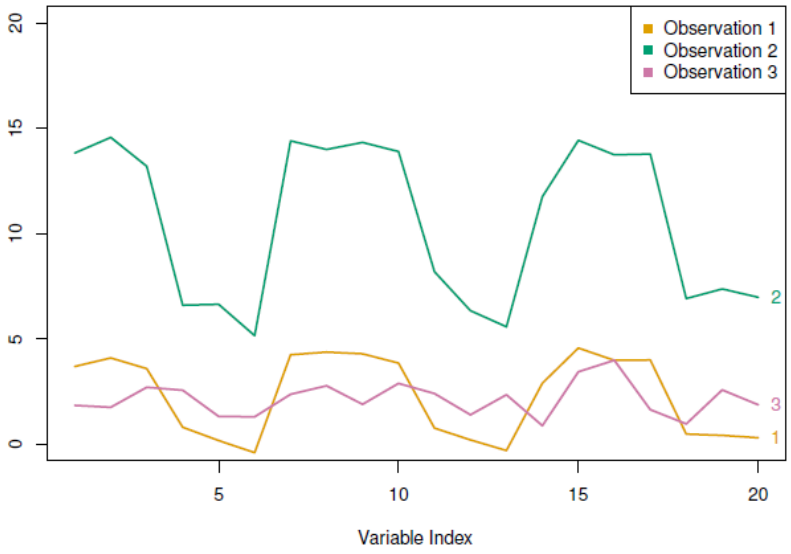
- Average and complete linkage are generally preferred over single linkage, as they tend to yield more balanced dendrograms

LINKAGE CAN BE IMPORTANT



- Three clustering results for the same data
- The only difference is the linkage method but the results are very different
- Complete and average linkage tend to yield evenly sized clusters whereas single linkage tends to yield extended clusters to which single leaves are fused one by one

CHOICE OF DISSIMILARITY MEASURE



- So far, we have considered using Euclidean distance as the dissimilarity measure
- However, an alternative measure that could make sense in some cases is the correlation based distance
- This is an unusual use of correlation, which is normally computed between variables; here it is computed between the observation proles for each pair of observations

CONCLUSION

- In order to perform clustering, some decisions must be made:
  - Should the features first be standardized? i.e. Have the variables centered to have a mean of zero and standard deviation of one.
  - In 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 case of K-means clustering:
    - How many clusters should we look for the data?

CONCLUSION

- In practice, we try several different choices, and look for the one with the most useful or interpretable solution
- There is no single right answer!
- Most importantly, one must be careful about how the results of a clustering analysis are reported
- These results should not be taken as the absolute truth about a data set
- Rather, they should constitute a starting point for the developments of a scientific hypothesis and further study, preferably on independent data