12.4 Clustering Methods

Clustering

- *Clustering* refers to a broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek to partition the data into distinct groups so that the observations within each group are similar to each other.
- To make this concrete, we need to define what it means for two observations to be *similar* or *different*.
  - This can be non-trivial!
  - Often a domain-specific consideration that needs to be made in conjunction with an expert on the data being studied.

PCA versus Clustering

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
- Clustering looks for homogeneous subgroups among the observations.

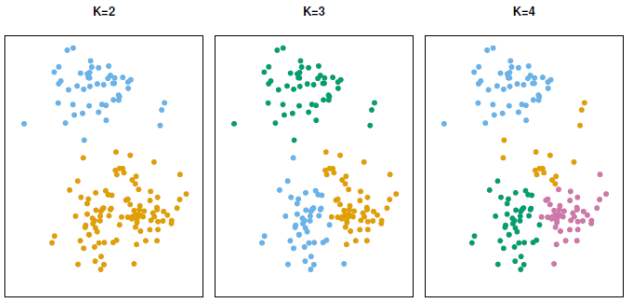
Clustering for Market Segmentation

- Suppose we have access to a large number of measurements (e.g. median household income, occupation, distance from nearest urban area, and so forth) for a large number of people.
- Our goal is to perform market segmentation by identifying subgroups of people who might be more receptive to a particular form of advertising, or more likely to purchase a particular product.
- The task of performing market segmentation amounts to clustering the people in the data set.

Two Clustering Methods

- In *K-means clustering*, we seek to partition the observations into a pre-specified number of clusters.
- In *hierarchical clustering*, we do not know in advance how many clusters we want.
  - Instead, we end up with a tree-like visual representation called a *dendrogram* which allows us to view the clusterings obtained for each possible number of clusters.

K-means Clustering



- Simulated data with 150 observations.
- Pictured: results from k-means clustering with 2, 3, and 4 clusters

Details of K-Means Clustering

Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster.

These sets satisfy two properties:

1. $C_1 \cup C_2 \cup \cdots \cup C_k = \{1, \ldots, n\}$ (Each observation belongs to at least one of the clusters.)
2. $C_1 \cap C_2 \cap \cdots \cap C_k = \emptyset$ for all $k \neq k'$ (The clusters are non-overlapping.)

That is, each observation is in exactly one cluster.

Details of K-Means Clustering

- Idea: a good clustering is one for which the *within cluster variation* is small.
- The within-cluster variation for cluster C_k is a measure $\mathrm{WCV}(C_k)$ of the amount by which the observations within a cluster differ from each other.
- So we want to solve

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \mathrm{WCV}(C_k) \right\}$$

- Basically, we want to sort the observations into k clusters such that the total within cluster variation is as small as possible.

Details of K-Means Clustering

How do we define WCV(C_k)?

- Typically use Euclidean distance:

$$\text{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

  where $|C_k|$ is the number of observations in the kth cluster.
- This leads to the optimization problem

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

K-Means Clustering Algorithm

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments.
2. Iterate until the cluster assignments stop changing:
   a. For each of the K clusters, compute the cluster *centroid*, The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   b. Assign each observation to the cluster whose centoid is closest (based on Euclidean distance)

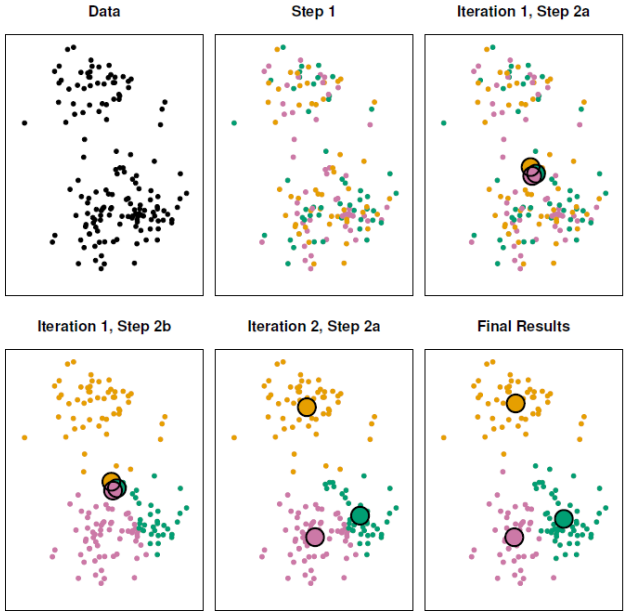
Properties

- This algorithm is guaranteed to decrease the value of the objective as each step:

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

- However, it is *not* guaranteed to reach the global minimum.

Example


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Example Details

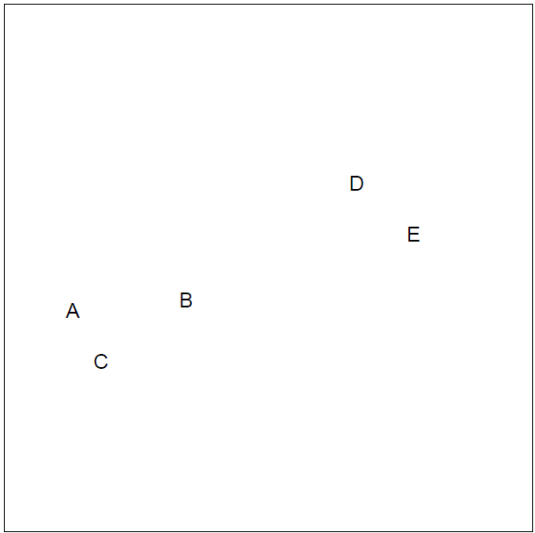
- Initially, the cluster centroids (disks) are almost completely overlapping.
- As observations are assigned to the nearest centroid, the centroids shift.
- Eventually, the centroids have settled into place and all observations are sorted.

Example: Different Starting Values

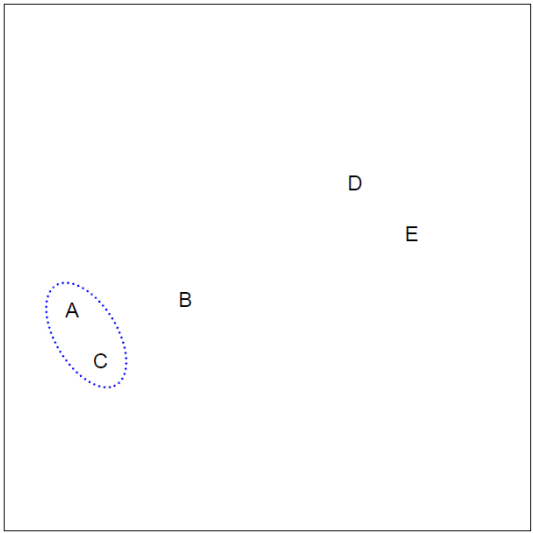
Hierarchical Clustering

- K-means clustering requires us to pre-specify the number of clusters K.
  - This can be a disadvantage.
  - Later we'll go over some strategies for choosing K
- *Hierarchical clustering* does not require us to commit to a particular choice of K.
- There are multiple approaches to hierarchical clustering, but we describe the most common: *bottom-up* or *agglomerative* clustering.

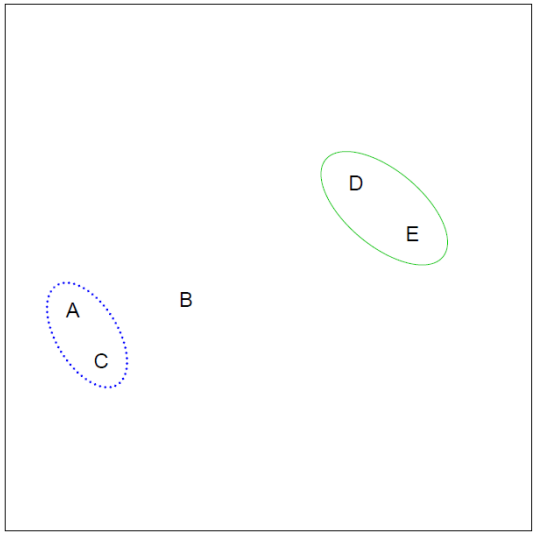
Hierarchical Clustering: Idea



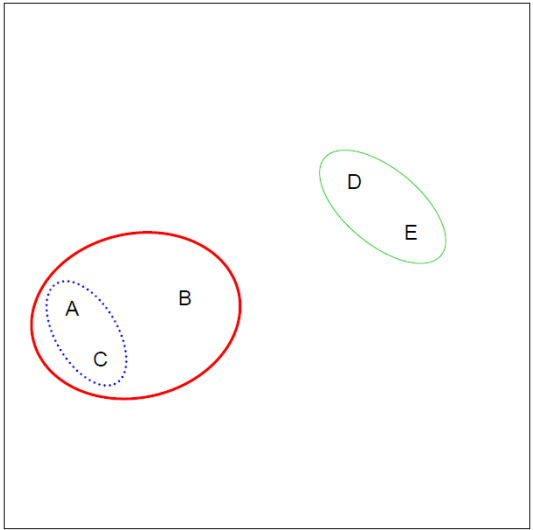
Hierarchical Clustering: Idea



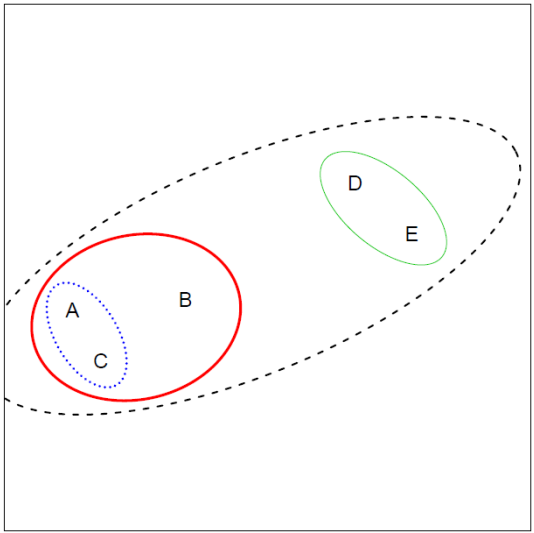
Hierarchical Clustering: Idea



Hierarchical Clustering: Idea



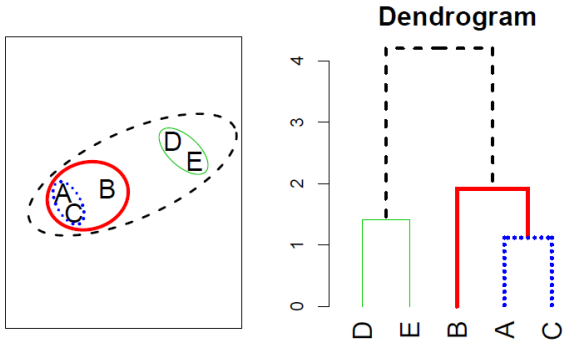
Hierarchical Clustering: Idea



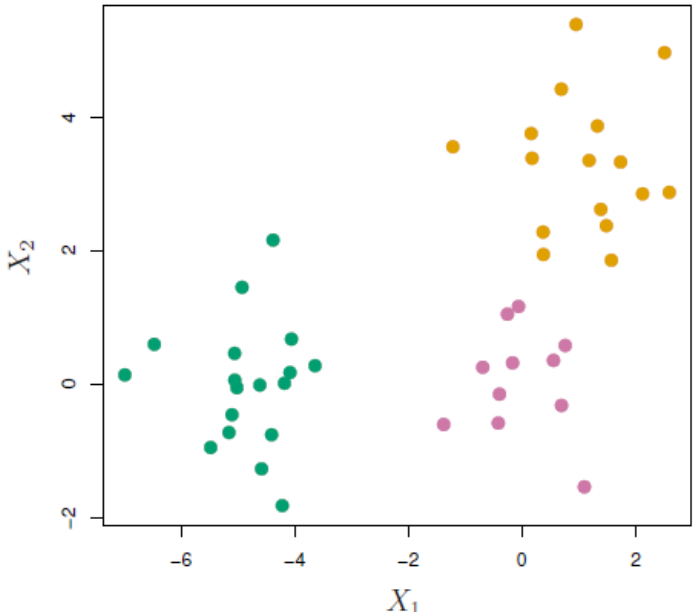
Hierarchical Clustering Algorithm

In words:

- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat until all points are in a single cluster.



Example



Example

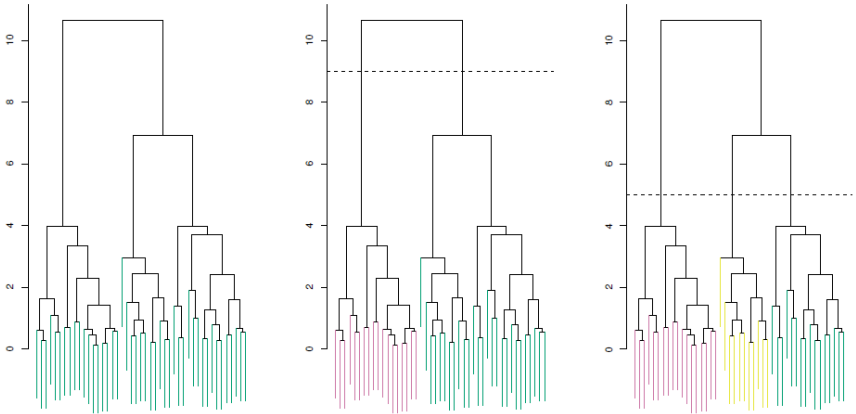


Figure Details

- Left: Dendrogram obtained from hierarchically clustering the data from previous slide, with complete linkage and Euclidean distance.
- Center: The dendrogram from the left-hand panel, cut at a height of 9 (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.
  - The height typically represents the linkage distance computed by the algorithm.
- Right: The dendrogram from the left-hand panel, now cut at a height of 5. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes.

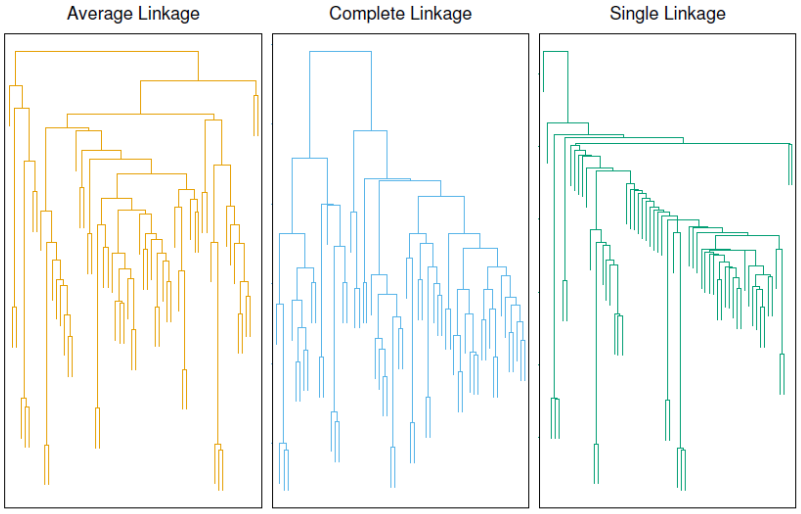
Linkages

- **Complete**: 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**: 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.

Linkages

- **Average**: 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**: Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversions. (Clusters fuse in ways that don't make sense in the dendrogram.)

Linkages



Choice of Dissimilarity Measure

- One alternative to Euclidean distance is *correlation-based distance*.
  - Considers two observations to be similar if their features are highly correlated.
- This is an unusual use of correlation, as we are computing between observations rather than between variables.

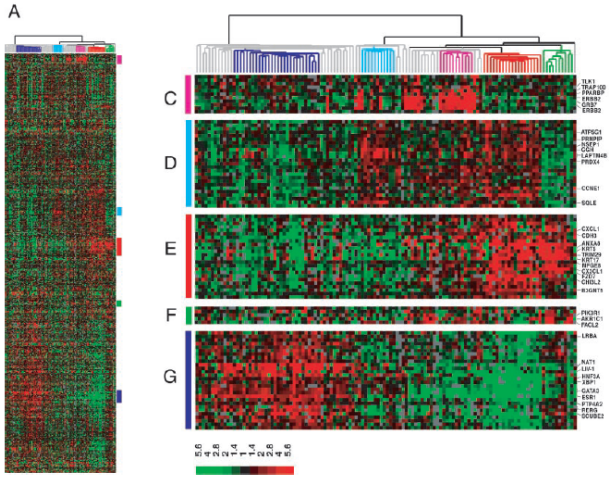
Practical Issues

- Scaling matters!
- Which dissimilarity measure to use, what type of linkage
- How many clusters to choose (no agreed upon method!)
- Which features should we use to drive the clustering?

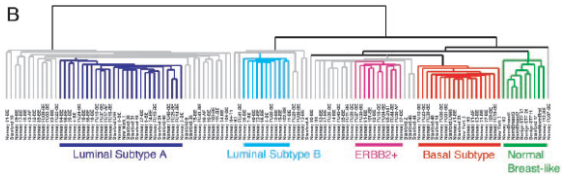
Example: Breast Cancer Microarray

- "Repeated observation of breast tumor subtypes in independent gene expression data sets." Sorlie at el. PNAS (2003).
- Gene expression measurements for approx 8000 genes, for each of 88 breast cancer patients.
- Average linkage, correlation metric.
- Clustered samples using 500 intrinsic genes: each woman was measured before and after chemotherapy. Intrinsic genes have smallest within/between variation.

Example: Breast Cancer Microarray



Example: Breast Cancer Microarray



Conclusions

- Unsupervised learning is important for understanding the variaton and grouping structure of a set of ununlabeled data, and can be a useful pre-processor for supervised learning
- It is intrinsically more difficult than supervised learning because there is no gold standard (like an outcome variable) and no single objective (like test set accuracy).
- It is an active field of research, with many recently developed tools such as self-organizing maps, independent components analysis and spectral clustering.