Unsupervised Learning

UNIVERSITY OF
SAN FRANCISCO

James D. Wilson

MATH 373

Plan for this Lecture

- What is Clustering?
- Applications
- Clustering Algorithms
  - k-Means Clustering
  - Hierarchical Clustering

Reference: ISL Sections 10.1; 10.2; 10.3

What is Clustering?

Setting: We observe $n \times p$ data matrix X only (no labels!)

Aim: To cluster or identify homogeneous subgroups in either the rows or columns of X. In particular, find the division $\pi = \{C_1, \ldots, C_k\}$ of objects into a small number of disjoint groups, called clusters, such that

- distances within clusters are small
- distances between clusters are large
- division is done without supervision

Clustering vs. Classification

Clustering: "Unsupervised Learning"

- Identify group structure in set of unlabeled objects.
- Special case of exploratory data analysis

Classification: "Supervised Learning"

- Given labeled samples $(X_1, Y_1), \ldots (X_n, Y_n)$, with $X \in \mathcal{X}$ and $Y_i \in \{1, \ldots, m\}$, look for a simple rule that classifies each unlabeled object $\mathbf{x}_o$.

Feature Vectors

Objects $x \in \mathcal{X}$ (i.e. rows or columns of X) typically represented by a feature vector

$$x = (v_1, \ldots, v_p)$$

Here v_j is a numerical/categorical measurement of interest:

- $v_j \in \mathbb{R}$ numerical feature
- $v_j \in \{a, b, \ldots\}$ categorical feature

For purposes of clustering, we need to identify object and its feature vector.

Applications

1. Medicine
   - Object = patient
   - Feature v_i = outcome of a diagnostic test on patient
2. Microarrays
   - Object = tissue sample
   - Feature v_i = measured expression level of gene i in that sample
3. Data Mining
   - Object = consumer
   - Features v_i = type, location, or amount of recent purchases

Basic steps in clustering

Objects $x_1, \ldots, x_n$

$\Downarrow$

Extraction of Features

$\Downarrow$

Dissimilarity matrix

$$D = \{d(x_i, x_j) : 1 \leq i, j \leq n\}$$

$\Downarrow$

Run Clustering Algorithm

$\Downarrow$

Partition $\pi = \{C_1, \ldots, C_k\}$ of $x_1, \ldots, x_n$

Distances between Features

1. Euclidean: $d(u, v) = \sqrt{\sum_i (u_i - v_i)^2}$
2. Manhattan: $d(u, v) = \sum_i |u_i - v_i|$
3. Correlation: $d(u, v) = 1 - \text{corr}(u, v)$
4. Hamming: $d(u, v) = \sum_i I\{u_i \neq v_i\}$
5. Mixtures of these

Unsupervised Methods

- **Principal Component Analysis**: looks to find a low-dimensional representation of the observations that explain a good fraction of the variance
- **Clustering Methods**:
  - Hierarchical: Candidate divisions of data described by a binary tree
  - Iterative: k-means, self-organizing maps
  - Model-based: Fit feature vectors with a finite mixture model
  - Spectral: Threshold eigenvectors of Laplacian of dissimilarity matrix

k-means Clustering

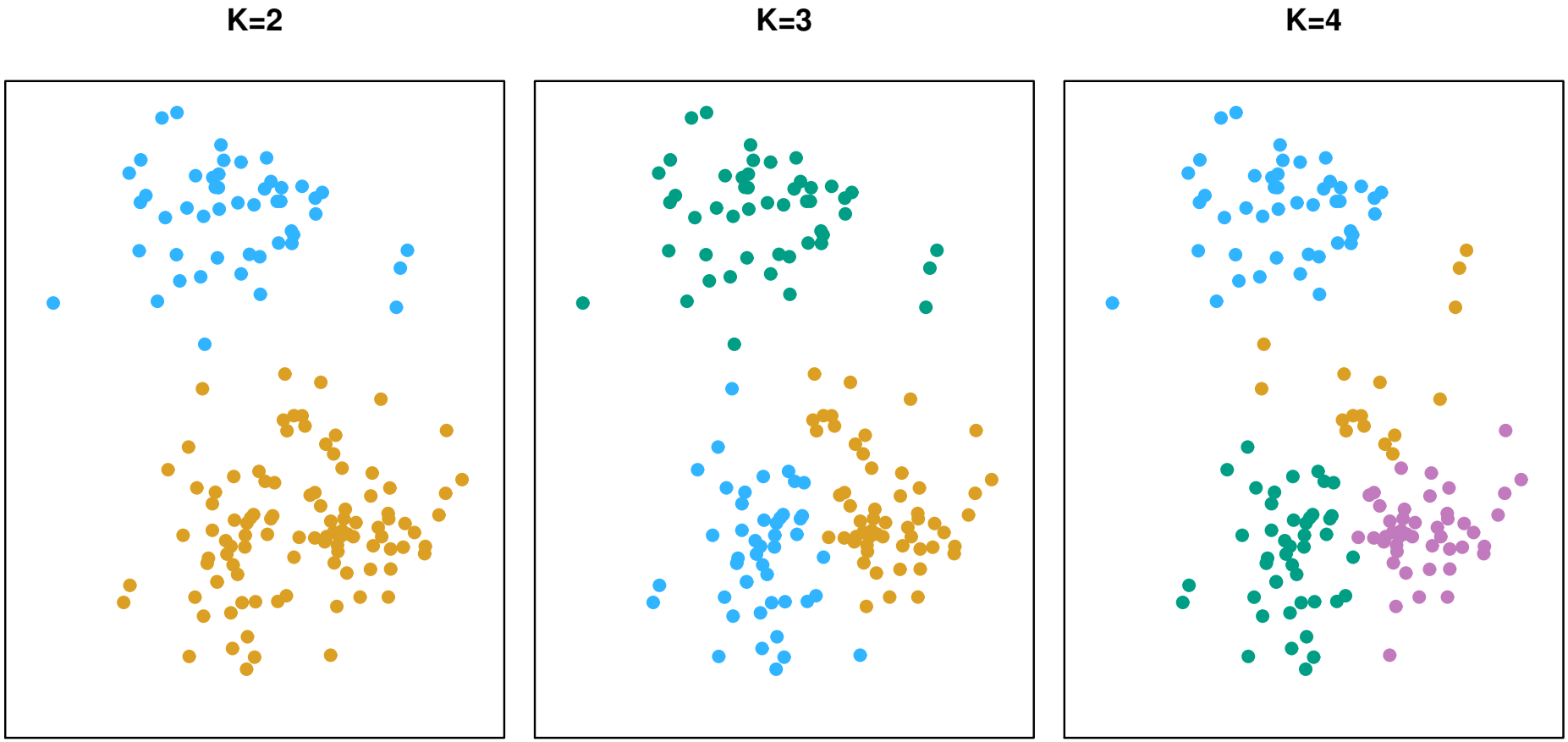
Clustering Problem: Divide $x_1, \ldots, x_n \in \mathbb{R}^p$ into k clusters $C_1, \ldots, C_k$ where

- $C_1 \cup C_2 \cup \ldots \cup C_k = \{1, \ldots, n\}$
- $C_j \cap C_\ell = \varnothing$ for all $j \neq \ell$

General Idea of k-means:

- Specify the number of clusters k
- Identify the k clusters that minimize the within-cluster variation. That is, find clusters such that objects in the same cluster are "close"

Example of k-means



k-means Clustering

Optimization: Find centers / centroids $c_1, \ldots, c_k$ to minimize the sum of squares (SoS) cost function

$$\text{Cost}(c_1, \ldots, c_k) = \sum_{i=1}^{n} \min_{1 \le j \le k} ||x_i - c_j||^2$$

i.e., the sum of squared distances from each point to its nearest center. Next we place observations into the cluster that contains its nearest center.

Problem: Optimization problem is computationally intractable.

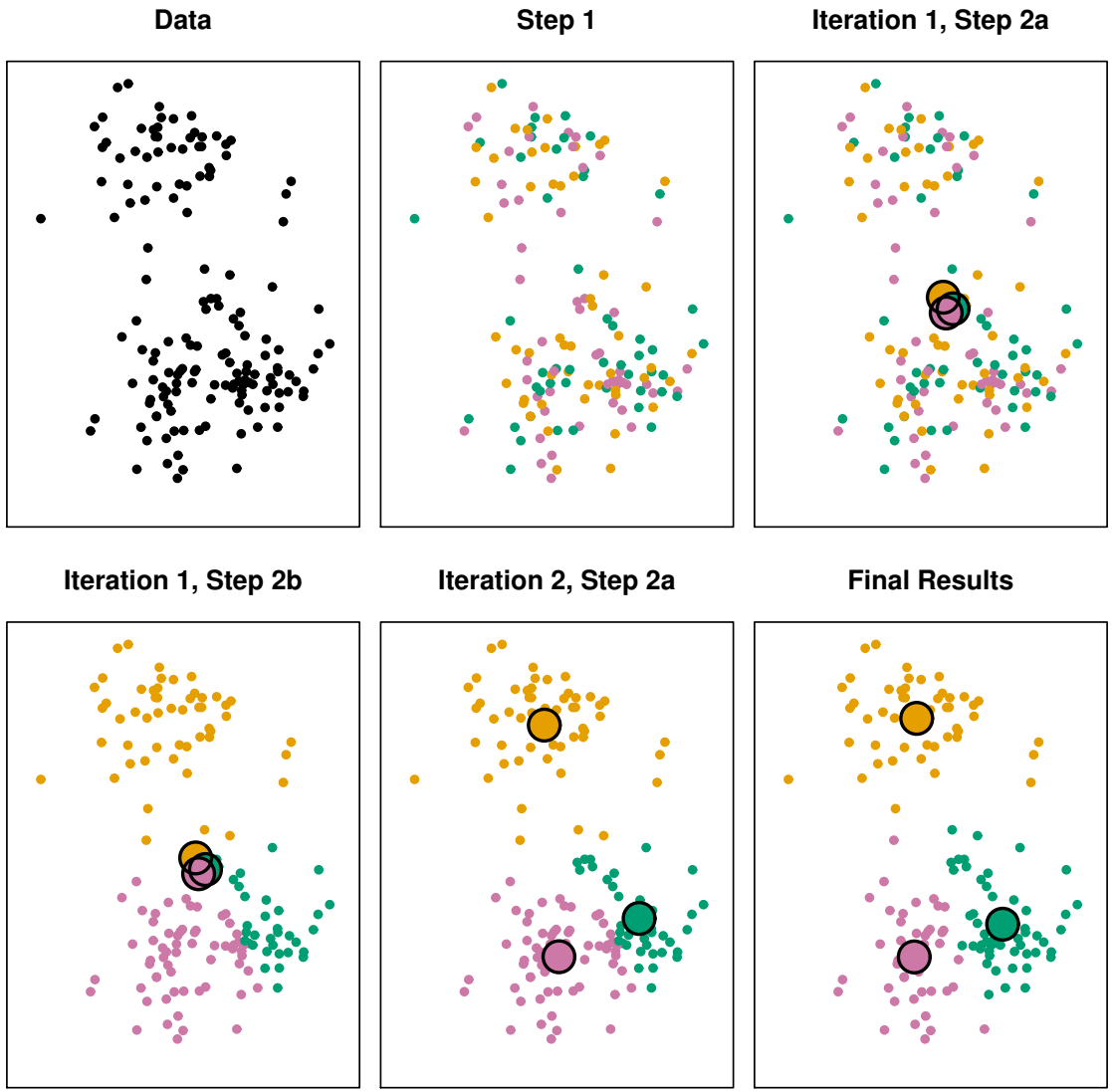
Solution: Iterative methods that find *local* optima of SoS cost.

k-means Clustering

Algorithm

1. *Randomly* assign a number, from 1 to k, to each observation. These serve as the initial cluster assignments.
2. Iterate until the cluster assignments stop changing:
    1. For each of the k clusters, compute the cluster centroid. This will be a vector of p feature means.
    2. Assign each obervation to the cluster whose centroid is closest according to Euclidean distance.

k-means Demonstration



k-means Clustering: Two Facts

Recall: Sum of Squares (SoS) cost function

$$\text{Cost}(c_1, \ldots, c_k) = \sum_{i=1}^{n} \min_{1 \le j \le k} ||x_i - c_j||^2$$

Mathematical Fact 1: (Pro) The SoS cost function decreases at each stage of the k-means algorithm.

Mathematical Fact 2: (Con) The SoS cost function will decrease as k increases.

In practice: The choice of k is usually made *a priori*. There are alternative means to choose k (such as the "elbow" method like that used in PCA), but none are widely agreed upon.

k-means Clustering in Practice

Feature 1: The k-means algorithm is initiated with a *random* seed / partition.

Feature 2: The algorithm will find a *local minimum* rather than the *global minimum*.

- The above features suggest that two different runs of k-means can lead to completely different solutions
- In practice, we run k-means many times (hundreds) with the same k and keep the clusterings that minimize the cost function

k-means Clustering in Practice

If clusters are present, their features can affect performance of different clustering procedures.

k-means clustering tends to perform best when clusters are:

- Spherical or elliptical in shape
- Similar in overall variance/spread
- Similar in size (number of points)

Example: `http://www.onmyphd.com/?p=k-means.clustering`

Hierarchical Clustering

- A direct similarity-based procedure for clustering
- Relies upon a dissimilarity measure, as well as a linkage to measure the distance between clusters
- Can be represented by a binary tree
- Given a linkage and a dissimilarity measure, this procedure is *deterministic*

Binary trees

Distinguished node called the “root” with zero or two children but no parent.

Every other node has one parent and zero or two children.

- Nodes with no children are called leaves
- Nodes with two children are called internal

Tree usually drawn upside-down, with root node at the top.

Linkage: distances between clusters

Goal: to measure the distance between two clusters C and C'. There are three common types:

1. Single Linkage:

$$d_s(C, C') = \min_{x_i \in C,\ x_j \in C'} d(x_i, x_j)$$

2. Average Linkage:

$$d_a(C, C') = \frac{1}{|C||C'|} \sum_{x_i \in C,\ x_j \in C'} d(x_i, x_j)$$

3. Total Linkage:

$$d_t(C, C') = \max_{x_i \in C,\ x_j \in C'} d(x_i, x_j)$$

Dendrograms

Binary tree associated with an agglomerative clustering procedure

Serves as a graphical record of the clustering process

Building a dendrogram: agglomerative approach

Initialize: Each singleton $\{x_i\}$ corresponds to a node at height 0

Update: If two clusters C, C' are combined, their respective nodes are joined to a parent node at *vertical* height $d(C, C')$

Dendrograms

- Each node of dendrogram corresponds to a set of objects
- Objects associated with two nodes merged when forming their parent
- Leaves correspond to individual objects
- The root corresponds to all objects

Dendrograms

A dendrogram represents many possible clusterings, one for each (rooted) subtree.

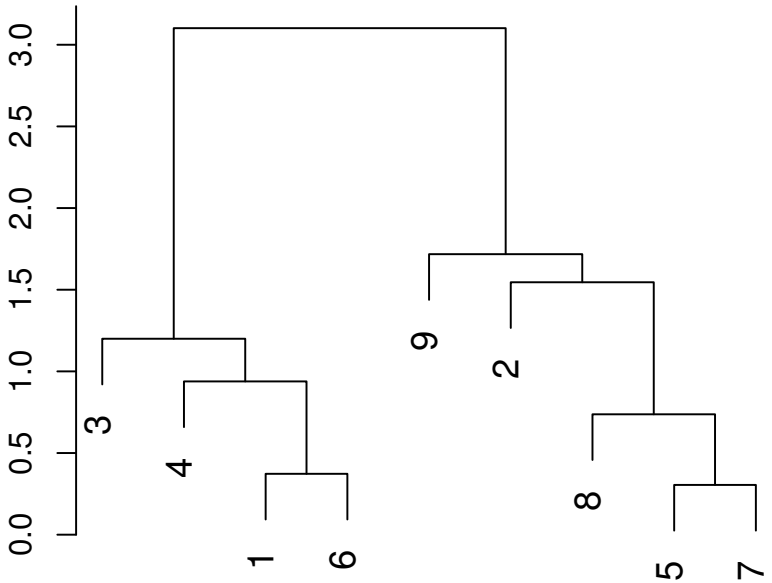
Methods for selecting a clustering/subtree

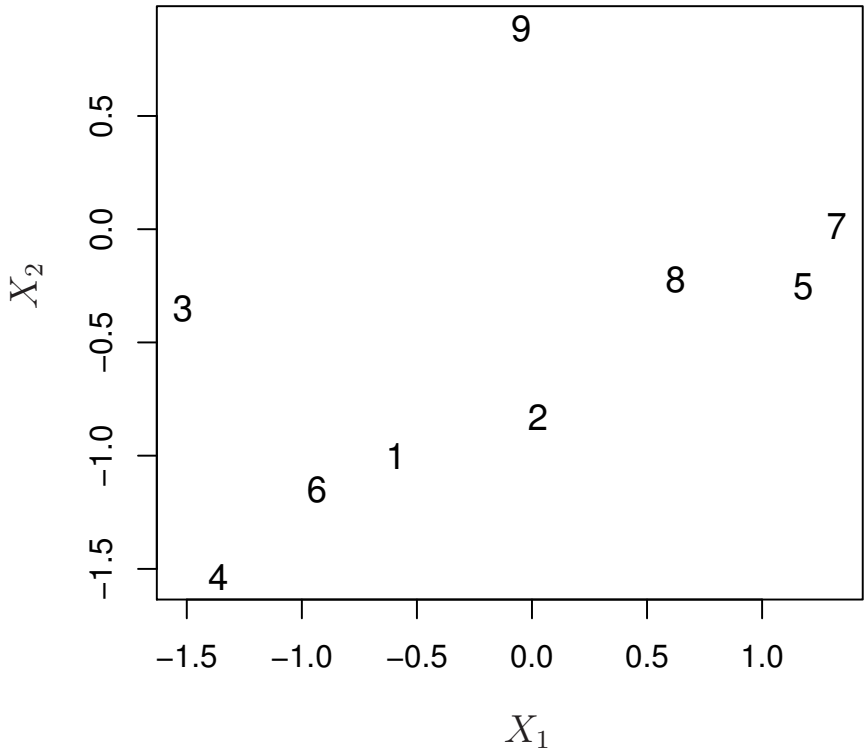
- Ad hoc selection (by eye)
- “Cutting” dendrogram at fixed level
- Penalized pruning

Visualization of clustering structure

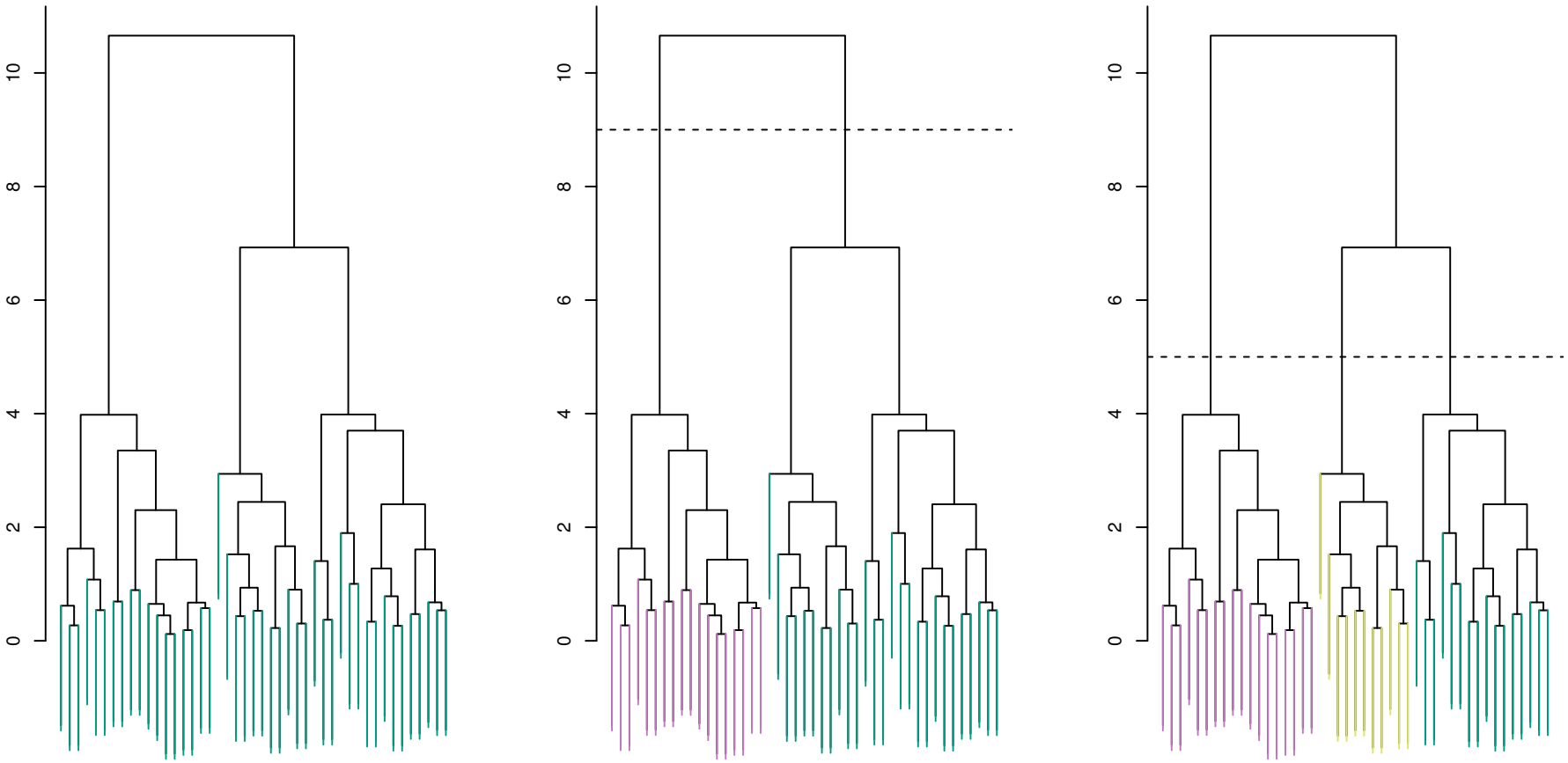
- Order objects in the same way as the leaves of the dendrogram
- Note: many orderings possible
- Vertical distance of tree has meaning!

Dendrogram Examples





Dendrogram Examples

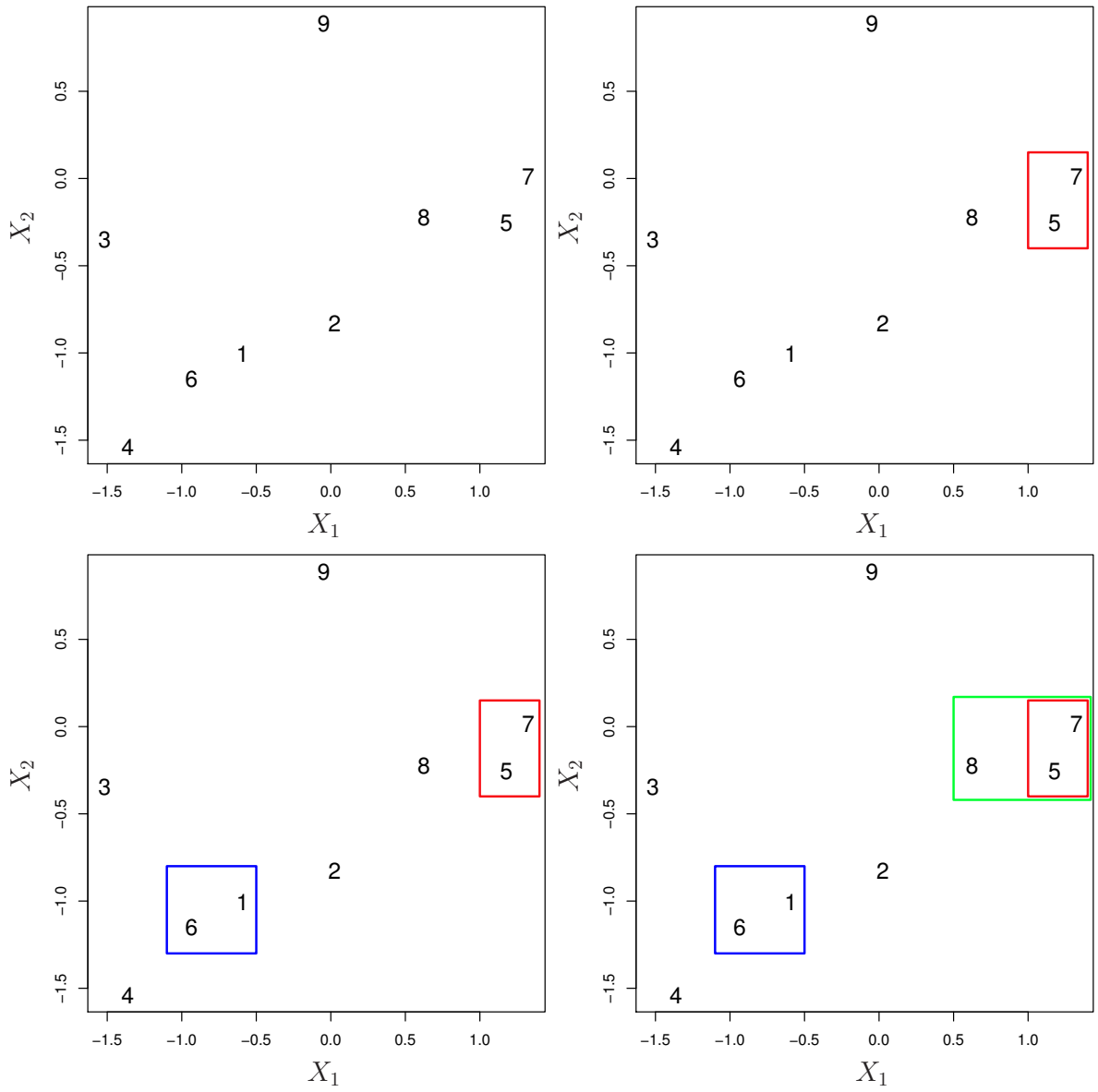


Hierarchical Clustering

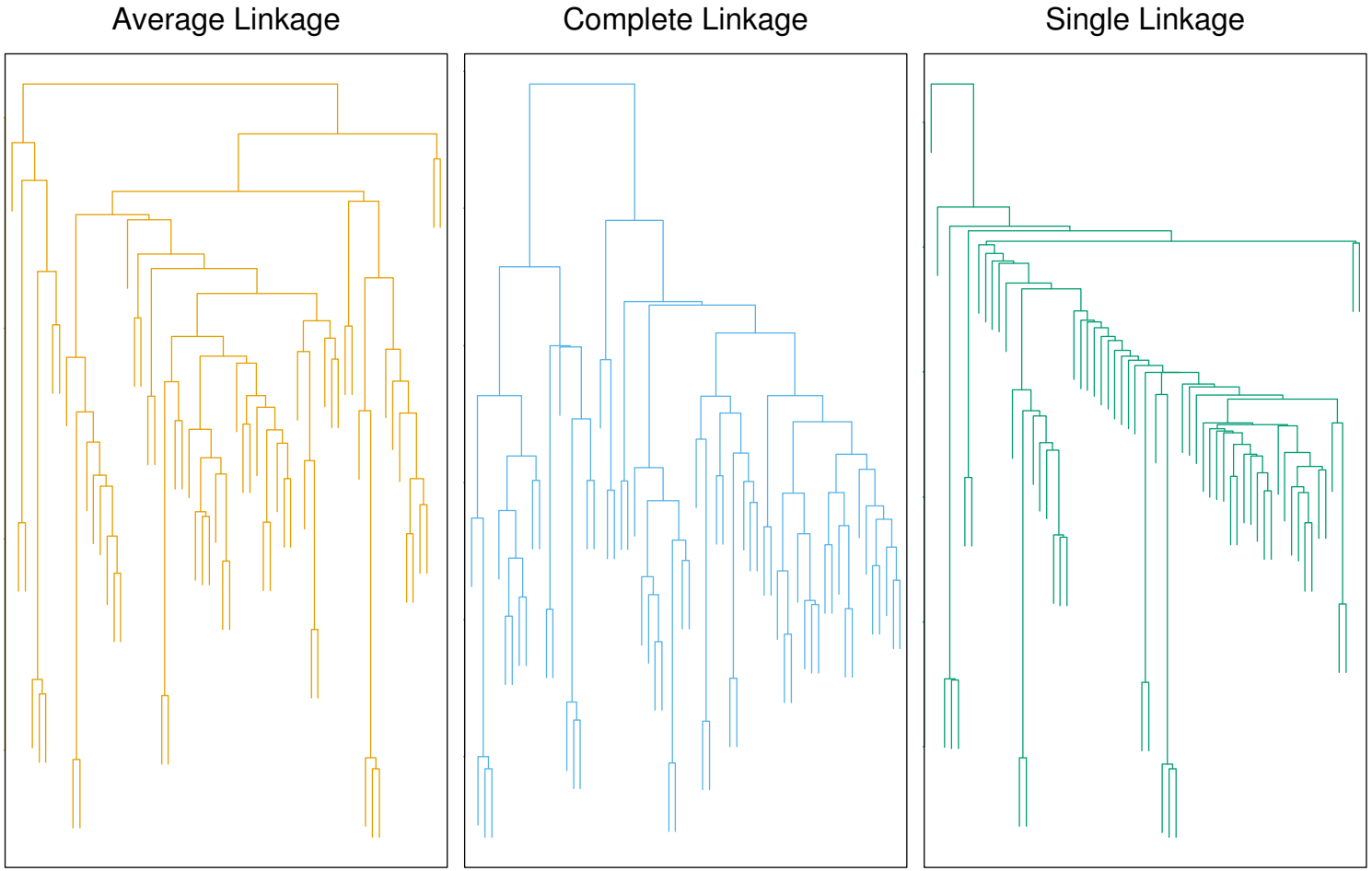
Algorithm

1. **Input:** n observations, a measure of dissimilarity, and a linkage
2. Measure all $\binom{n}{2}$ pairwise dissimilarities. Treat each observation as its own cluster
3. **Loop:** for $i = n, n-1, \ldots, 2$
   1. Examine all inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are the most similar. Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.
   2. Compute the new pairwise inter-cluster dissimilarities among the remaining $i-1$ remaining clusters.

Example of Hierarchical Clustering



Choice of Linkage Matters!



Generally, complete and average linkage give more balanced trees and are preferred.

Practical Considerations with Hierarchical Clustering

- Where do we "cut" the dendrogram?
- What linkage do we use?
- What dissimilarity measure do we use?

These choices all depend upon your observations (discrete vs. continuous), and the type of data you are working with (would you like the 'best' 4 clusters?, etc.)

Practical Considerations with Clustering

- What is the right number of clusters?
- What is right measure of distance?
- What is the best clustering method for the data?
- How robust is an observed clustering?
- What significance can be assigned to the observed clustering of the objects?
- Remeber: clustering is exploratory! It should be used to help develop hypotheses about data.