STATS 235: Modern Data Analysis
Clustering

Babak Shahbaba

Department of Statistics, UCI

Clustering

- Building statistical models to identify the underlying structure of data without focusing on of a specific outcome variable is known as **unsupervised learning**.
- An important class of unsupervised learning is **clustering**, which is commonly used to identify sub-groups within a population.
- In general, cluster analysis refers to methods that attempt to divide the data into sub-groups such that the observations within the same group are more similar compared to the observations in different groups.
- In this lecture, we discuss clustering methods that are not based on any probabilistic model; clustering methods based some underlying probabilistic model are discussed in the next lecture

Clustering

- The core concept in any cluster analysis is the notion of similarity and dissimilarity. It is common to quantify the degree of dissimilarity based on a **distance** measure defined for a pair of observations.

- The most commonly used distance measure is the **squared Euclidean distance**:

$$d_{ij} = (x_i - x_j)^2$$

where d_{ij} refers to the distance between observations i and j, x_i is the value of random variable X for observation i and x_j is the value for observation j.

- In general, if we measure p random variables, $X_1, \ldots, X_p$, the squared Euclidean distance between two observations i and j in our sample is

$$d_{ij} = (x_{i1} - x_{j1})^2 + \ldots + (x_{ip} - x_{jp})^2$$

K-means Clustering

- *K-means clustering* is a simple algorithm that uses the squared Euclidean distance as its measure of dissimilarity.
- We start by specifying the number of clusters (groups) K. This is the number of groups that we believe exist in the population.
- Our goal is then to group the n observations into K clusters, such that the overall measure of dissimilarity is small within groups and large between groups.
- Initially, we divide the observations into K groups randomly.
- Then the algorithm iteratively improves the clusters.

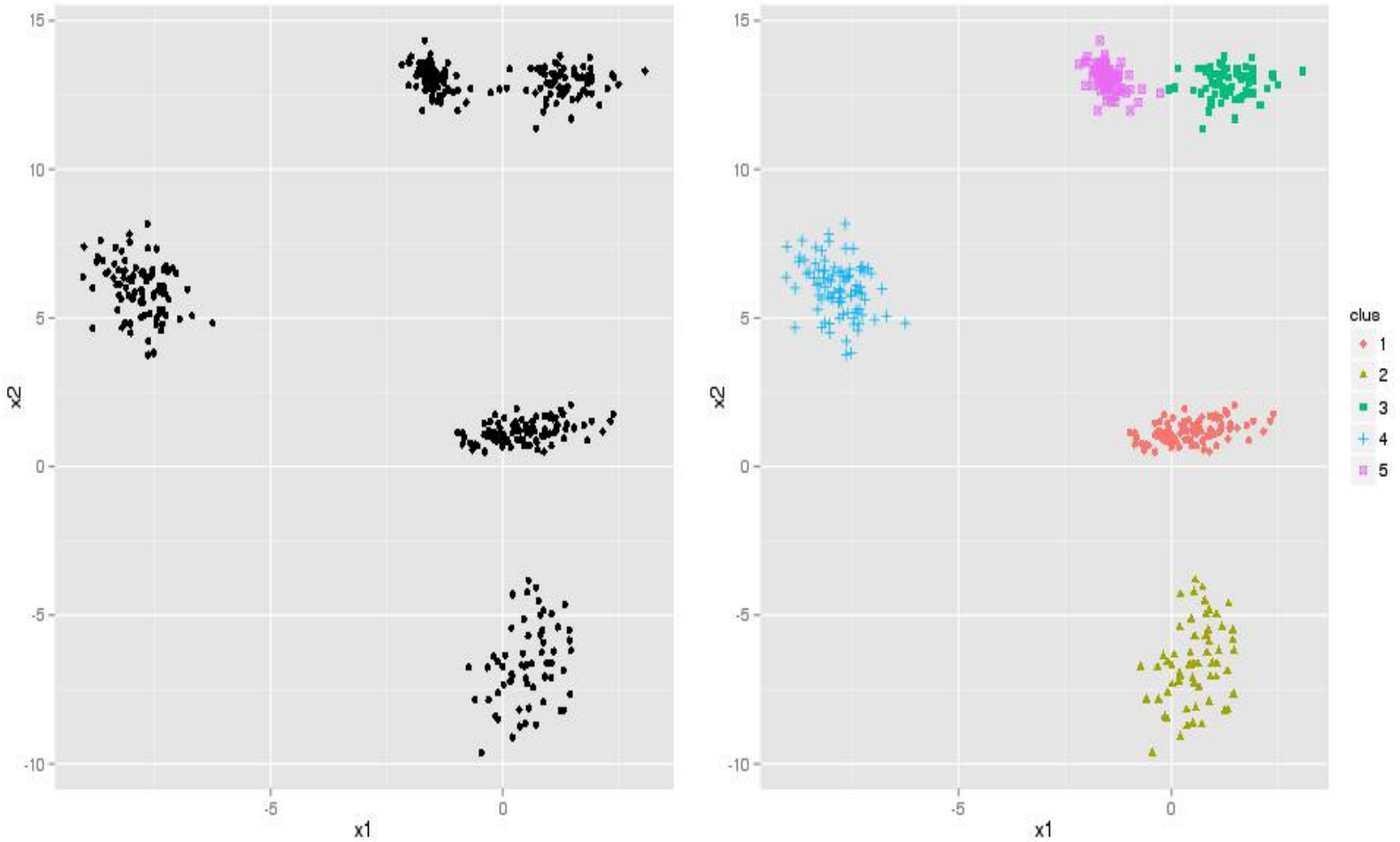
K-means Clustering

- For each cluster, we define the **center** or **centroid** as an imaginary observation, whose measurements are the sample average of all observations in that cluster.

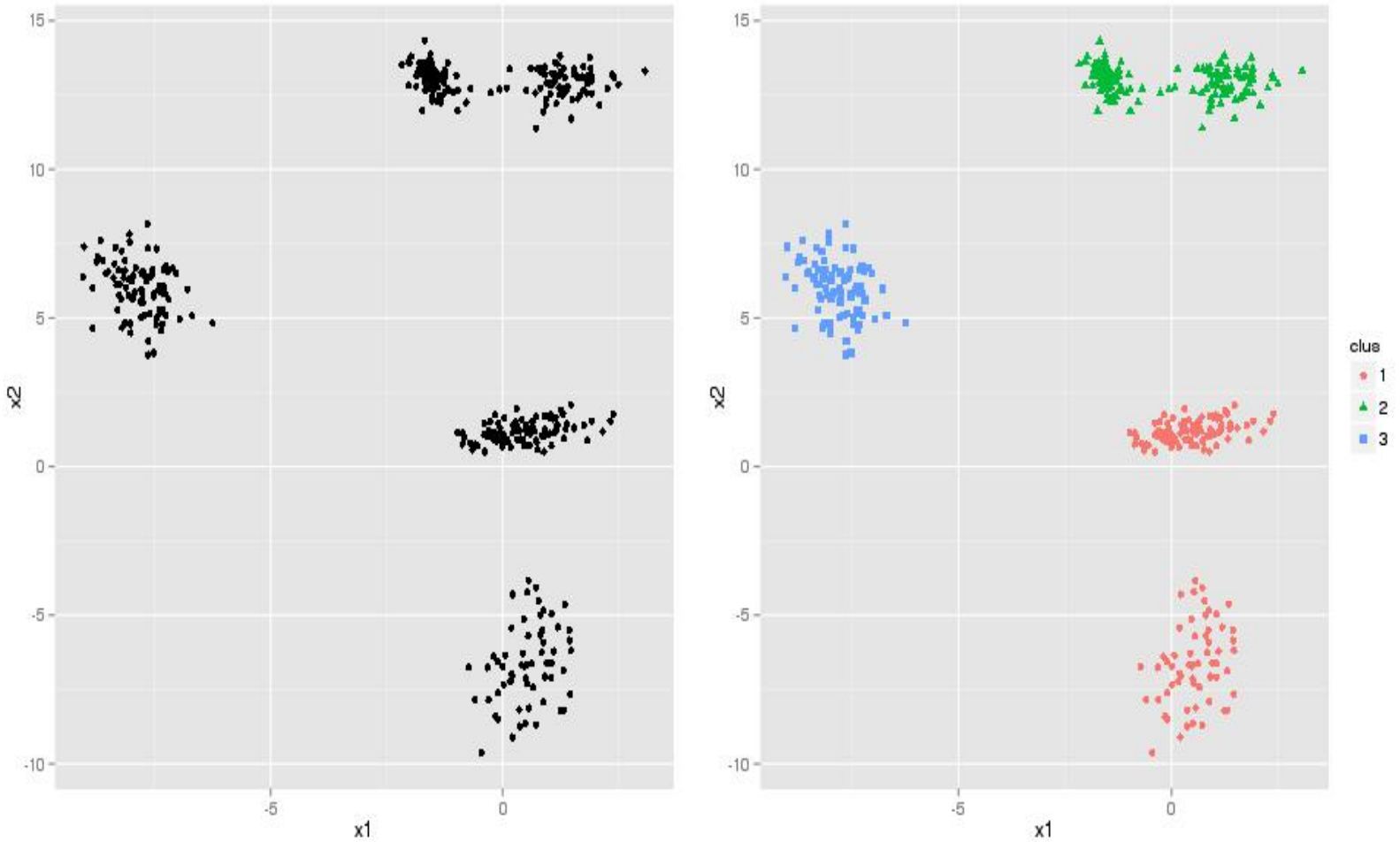
- After randomly partitioning the observations into K groups and finding the center of each cluster, the K-means algorithm finds the best clusters by iteratively repeating these steps:
  1. For each observation, find its squared Euclidean distance to all K centers, and assign it to the cluster with the smallest distance.
  2. After regrouping all the sampling units into K clusters, re-calculate the K centers.

  We repeat the above steps until the clusters do not change (i.e., the centers remain the same after each iteration).

Example: $K = 5$



Example: $K = 3$

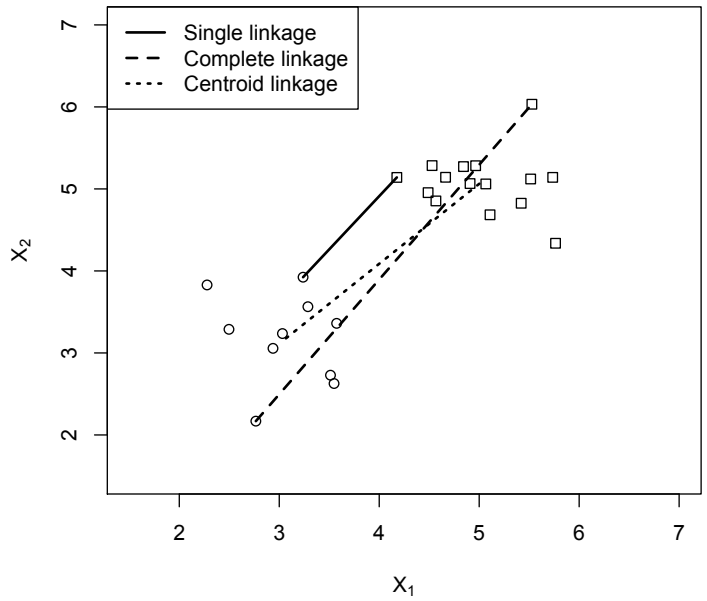


Hierarchical clustering

- There are two general algorithms for hierarchical clustering Hastie et. al.:
  - **Agglomerative** (bottom-up): We start at the bottom of the tree, where every observation is a cluster (i.e., there are n clusters). Then we merge two of the clusters with the smallest degree of dissimilarity (i.e., the two most similar clusters). Now we have $n-1$ clusters. We continue merging clusters until we have only one cluster (the root) that includes all observations.
  - **Divisive** (top-down): We start at the top of the tree, where all observations are grouped in a single cluster. Then we divide the cluster into two new clusters that are most dissimiliar. Now we have two clusters. We continue splitting existing clusters until every observation is its own cluster.

Hierarchical clustering

- Of the above two strategies, agglomerative algorithm is the most common.
- Both algorithms, however, require a measure of dissimilarity between two clusters.
- We need to specify a distance measure for two clusters analogous to the distance measure we defined for two observations.



Hierarchical clustering

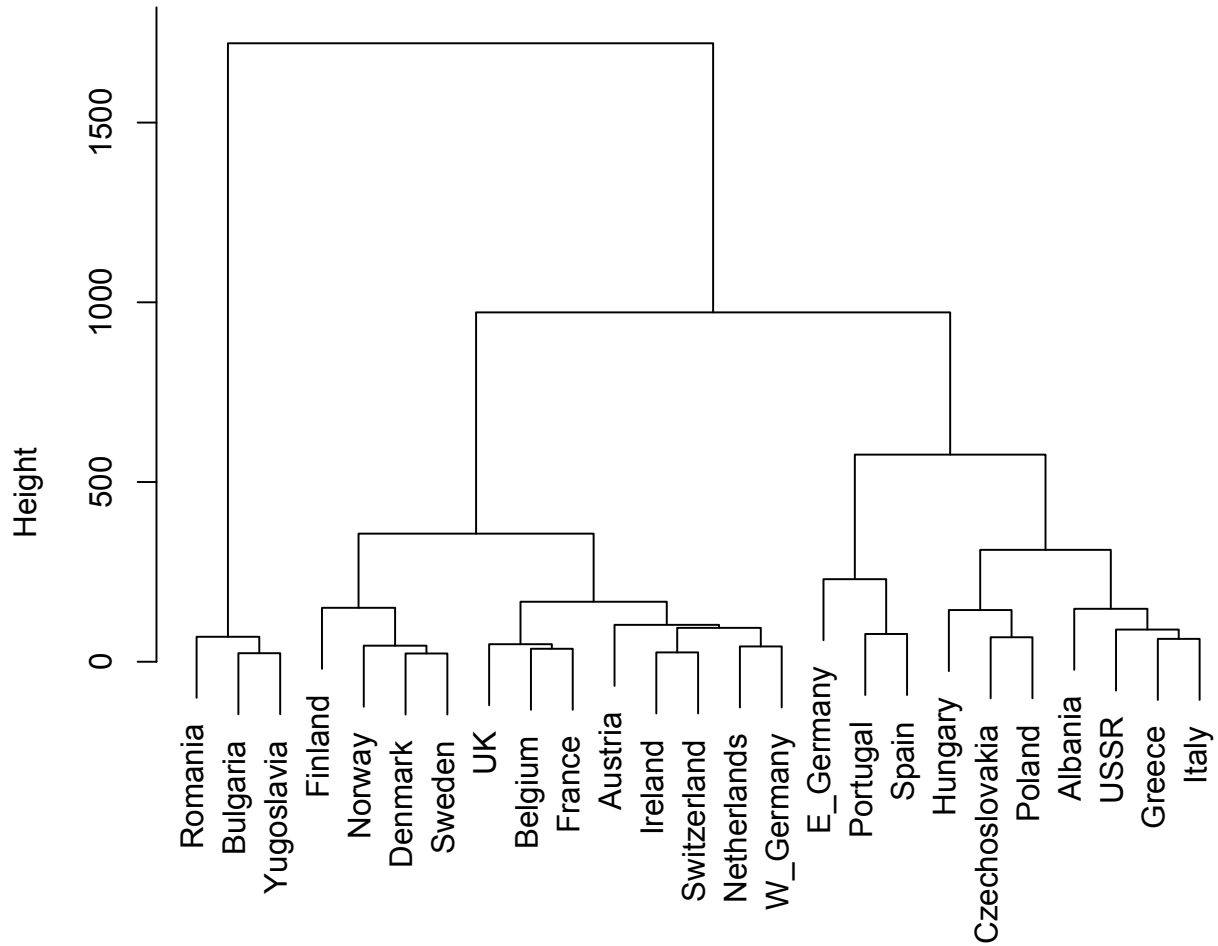
- These are some common methods to calculate the overall distance between two clusters:
  - *Single linkage* clustering use the minimum d_{ij} among all possible pairs as the distance between the two clusters. This is the distance between two observations, one from each cluster, that are closest to each other.
  - *Complete linkage* clustering uses the maximum d_{ij} as the distance between the two clusters. This is the distance between two observations, one from each cluster, that are furthest apart.
  - *Average linkage* clustering uses the average d_{ij} over all possible pairs as the distance between the two clusters.
  - *Centroid linkage* clustering finds the centroids of the two clusters and uses the distance between the centroids as the distance between the two clusters.

Example: Protein consumption in 25 European countries

Protein

	Country	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr.Veg
1	Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
2	Austria	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
3	Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
4	Bulgaria	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
5	Czechoslovakia	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
6	Denmark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
7	E.Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
8	Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
9	France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5

Example: Protein consumption



Example: Protein consumption