Statistical Data Analysis

Assist. Prof. Dr. Zeyneb KURT

(Slides have been prepared by
Prof. Dr. Nizamettin AYDIN,
updated by Zeyneb KURT)

zeyneb@yildiz.edu.tr
http://avesis.yildiz.edu.tr/zeyneb/

Clustering Analysis

Introduction

- Linear regression models are used to predict the unknown values of the response variable.
  - In these models, the response variable has a central role;
    - the model building process is guided by explaining the variation of the response variable or predicting its values.
  - Therefore, building regression models is known as supervised learning.
- In contrast, building statistical models to identify the underlying structure of data is known as unsupervised learning.
  - An important class of unsupervised learning is clustering,
    - which is commonly used to identify subgroups within a population.
- In general, cluster analysis refers to the methods that attempt to divide the data into subgroups such that the observations within the same group are more similar compared to the observations in different groups.

Distance Measure

- 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,
    - which is usually defined for a pair of observations.
- The most commonly used distance measure is the **squared 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.

Similarity and Dissimilarity

- Similarity
  - is a numerical measure of how alike two data objects are
  - is higher when objects are more alike
  - often falls in the range [0,1]
- Dissimilarity
  - is a numerical measure of how two data objects are different
  - is lower when objects are more alike
    - Minimum dissimilarity is often 0
    - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Distance

- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^r \right)^{\frac{1}{r}}$$

where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

Distance

- In Minkowski Distance,
  - if $r = 1$ *dist* is City block (Manhattan, taxicab, L1 norm) distance.
  - if $r = 2$ *dist* is Euclidean distance
  - if $r = \infty$ *dist* is "supremum" (Lmax norm, L∞ norm) distance.
- In general, if we measure p random variables $X_1, \ldots, X_p$, the squared distance between two observations i and j in our sample is

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

- This measure of dissimilarity is called the **squared Euclidean distance**.

Example

- Suppose that we believe that while European countries are different with respect to their protein consumption, they could be divided into several groups such that countries within the same group can be considered similar to each other in terms protein consumption.
- Here, we use the *Protein* data set we discussed earlier.
  - It includes numerical measurements of the protein consumption from 9 different sources:
    - RedMeat, WhiteMeat, eggs, Milk, Fish, Cereals, Starch (starchy foods), nuts (pulses, nuts, and oil-seeds), and Fr.Veg (fruits and vegetables).
- To start, suppose that we want to group countries according to their consumption of red meat (redMeat) and fish (Fish).
- More information about the data can be found at
  - http://lib.stat.cmu.edu/DASL/Datafiles/Protein.html

Example

- In the *Protein* data set, the first two countries are Albania and Austria.
- Suppose we want to measure their degree of dissimilarity (i.e., their distance) in terms of their consumption of red meat and fish given in the following table.

Countries	RedMeat	Fish
Albania	10.1	0.2
Austria	8.9	2.1

Example

- The squared distance between these two countries
  - $(10.1 - 8.9)^2 = 1.44$ in terms of red meat consumption
  - $(0.2 - 2.1)^2 = 3.61$ in terms of fish consumption.
  - To find the overall distance between these two countries, we add the distances based on different variables:

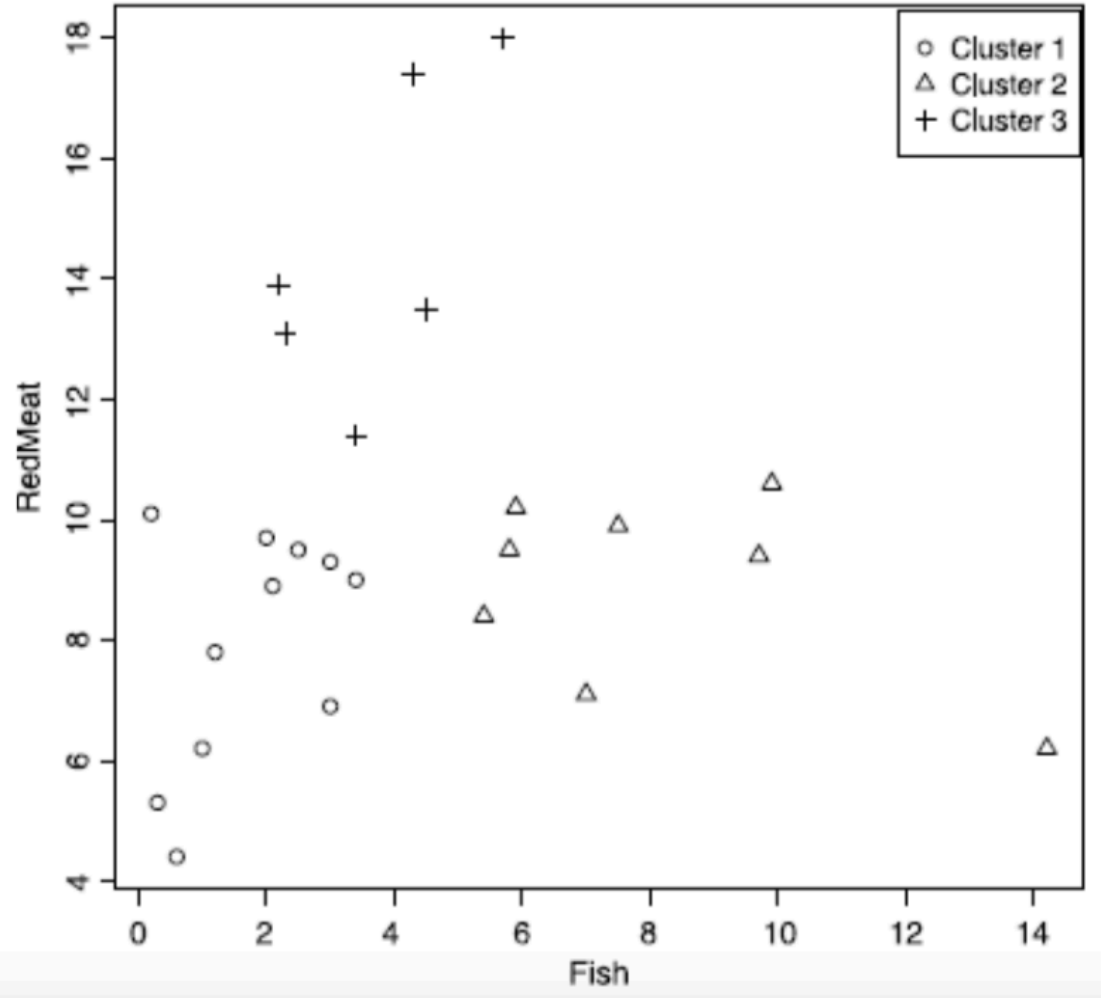
$$d = 1.44 + 3.61 = 5.05$$

K-means Clustering

- ***K*-means clustering** is a simple algorithm that uses the squared Euclidean distance as its measure of dissimilarity.
- After randomly partitioning the observations into *K* groups and finding the **center** or **centroid** of each cluster, the *K*-means algorithm finds the best clusters by iteratively repeating the following steps
  - For each observation, find its squared Euclidean distance to all *K* centers, and assign it to the cluster with the smallest distance.
  - After regrouping all the observations into *K* clusters, recalculate the *K* centers.
- These steps are applied until the clusters do not change
  - i.e., the centers remain the same after each iteration.

K-means Clustering

- An example of visualizing the results of *K*-means clustering with a scatterplot (with R-Commander).



- The three clusters are represented by circles, triangles, and crosses.
- They clearly partition the countries into
  - a group with a low consumption of fish and red meat,
  - a group with a high consumption of fish,
  - a group with a high consumption of red meat.

Hierarchical Clustering

- There are two potential problems with the *K*-means clustering algorithm.
  - It is a **flat** clustering method.
  - We need to specify the number of clusters *K* a priori.
- An alternative approach that avoids these issues is **hierarchical clustering**.
- The result of this method is a **dendrogram** (a tree).
  - The *root* of the dendrogram is its highest level and contains all *n* observations.
  - The *leaves* of the tree are its lowest level and are each a unique observation.

Hierarchical Clustering

- There are two general algorithms for hierarchical clustering:
  - **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 dissimilar.
      - Now we have two clusters.
    - We continue splitting existing clusters until every observation is its own cluster.

Hierarchical Clustering

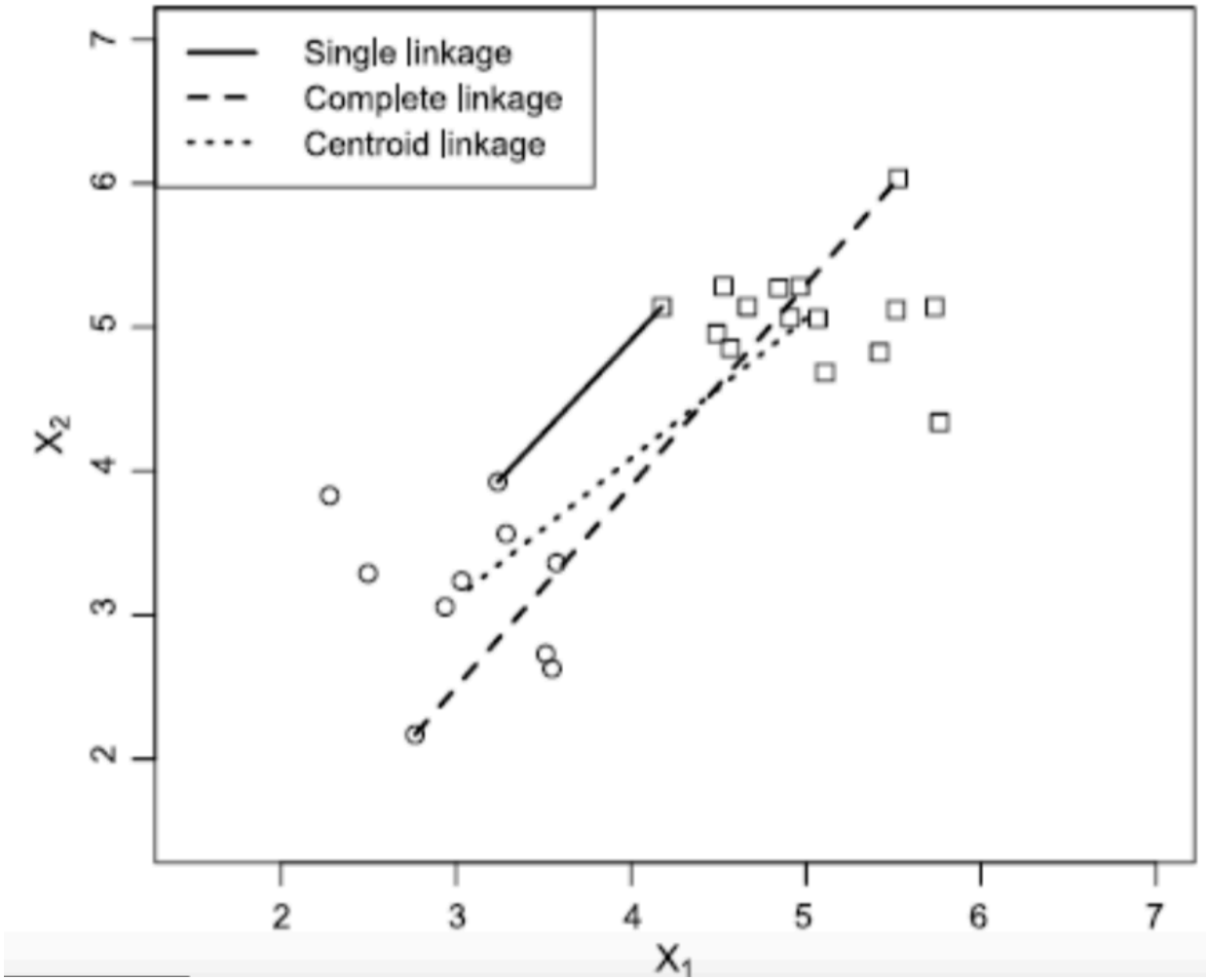
- **Agglomerative** (bottom-up):
  - We start at the bottom of the tree, where every observation is a cluster
    - i.e., there are n clusters.
  - Then we merge two of the clusters with the smallest degree of dissimilarity
    - i.e., the two most similar clusters.
    - Now we have $n - 1$ clusters.
  - We continue merging clusters until we have only one cluster (the root) that includes all observations.

Hierarchical Clustering

- We can use one of the following methods to calculate the overall distance between two clusters
  - Single linkage clustering uses the minimum d_{ij} among all possible pairs as the distance between the two clusters.
  - Complete linkage clustering uses the maximum d_{ij} as the distance between the two clusters.
  - 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.

Hierarchical Clustering

- The following figure illustrates the difference between the single linkage method, the complete linkage method, and the centroid linkage method to determine the distance d_{ij} between the two clusters shown as circles and squares.



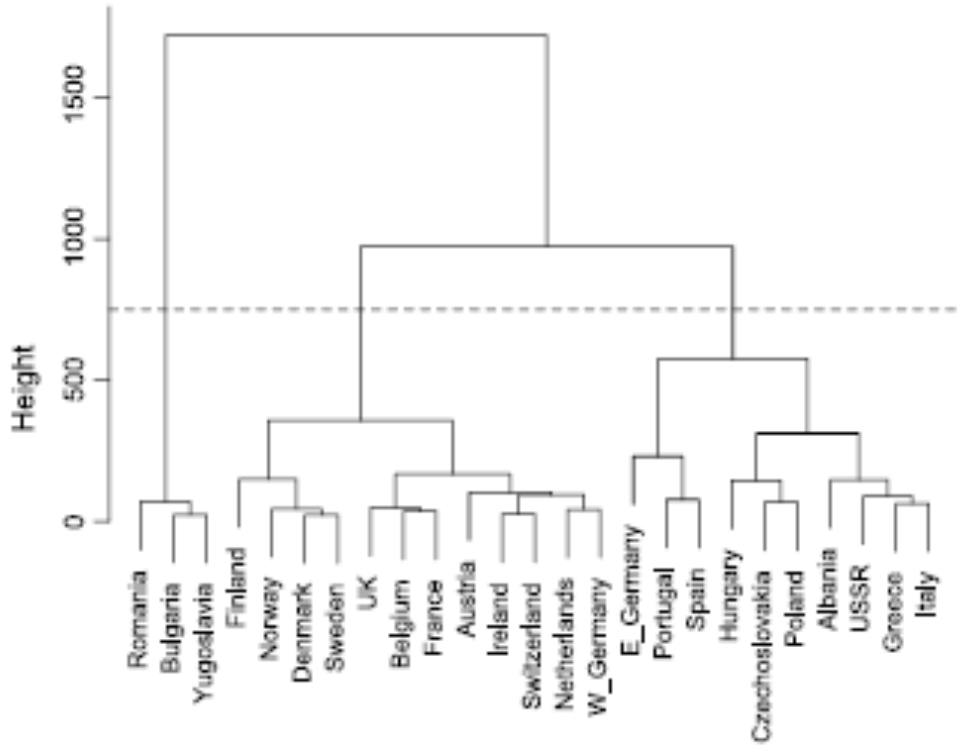
- Note that the dotted line connects the centers (as opposed to observations) of the two clusters.
- There are of course other ways for defining the distance between two clusters.
- However, the above measures are the most commonly used.

Hierarchical Clustering

- As an example, follow the following procedures in R-Commander to perform complete linkage clustering to create a dendrogram of countries based on their protein consumption.
- Click
  - *Statistics→Dimensional analysis → Cluster analysis → Hierarchical cluster analysis*.
- Select all nine food groups (hold the *control* key) for the *Variables*.
- Next, choose *Complete Linkage* as the *Clustering Method* and *Squared-Euclidean* as the *Distance Measure*.
- Lastly, make sure the option *Plot Dendrogram* is checked.
- R-Commander then creates a dendrogram similar to the one shown in the next slide

Hierarchical Clustering

The dendrogram resulting from complete linkage clustering of the 25 countries based on their protein consumption.



The *dashed line* shows where to cut the dendrogram to create three clusters

Hierarchical Clustering

- The clusters seemed to be defined by geographic location:
  - Balkan countries (Romania, Bulgaria, and Yugoslavia),
  - Scandinavian countries (Finland, Norway, Denmark, and Sweden),
  - Western European countries (UK, Belgium, France, Austria, Ireland, Switzerland, Netherlands, and West Germany),
  - Eastern European countries (East Germany, Hungary, Czechoslovakia, Poland, Albania, USSR)
  - the Mediterranean countries (Portugal, Spain, Greece, Italy).

Comparison between agglomerative and divisive methods

- **Divisive** clustering is conceptually more complex than **agglomerative** one since we need a second, flat clustering algorithm (e.g. k-means) as a "subroutine".
- **Divisive** algorithms can produce more accurate hierarchies than **agglomerative**.
- **Agglomerative** methods make clustering decisions based on local patterns without initially taking into account the global distribution. These early decisions cannot be undone.
- **Divisive** clustering benefits from complete information about the global distribution when making top-level partitioning decisions.