

# RECSM Summer School: Machine Learning for Social Sciences

## Session 3.4: Hierarchical Clustering

Reto Wüest

Department of Political Science and International Relations  
University of Geneva



# Clustering

---

# Clustering

---

## Hierarchical Clustering

# Hierarchical Clustering

- A potential disadvantage of  $K$ -means clustering is that it requires us to **pre-specify** the number of clusters  $K$ .
- **Hierarchical clustering** is an alternative approach that does **not** require us to do that.
- Hierarchical clustering results in a tree-based representation of the observations, called a **dendrogram**.
- We focus on **bottom-up** or **agglomerative** clustering, which is the most common type of hierarchical clustering.

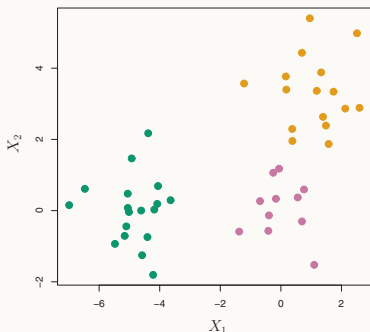
# Clustering

---

## Interpreting a Dendrogram

# Interpreting a Dendrogram

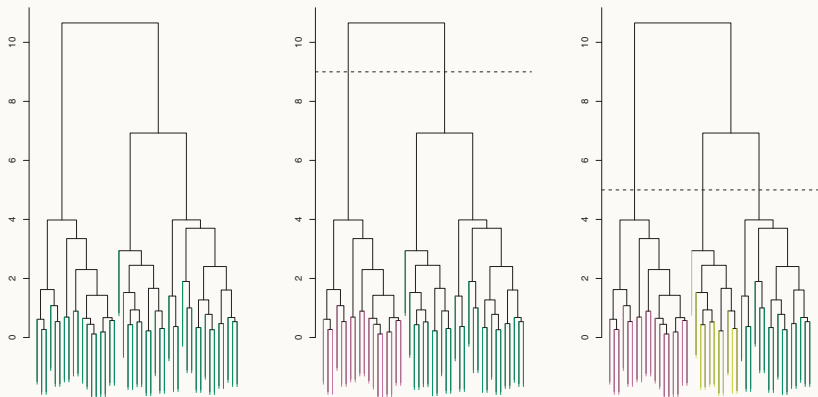
- We have (simulated) data consisting of 45 observations in two-dimensional space.
- The data were generated from a three-class model.
- However, suppose that the data were observed without the class labels and we want to perform hierarchical clustering.



(Source: James et al. 2013, 391)

# Interpreting a Dendrogram

Results obtained from hierarchical clustering (with complete linkage)



(Source: James et al. 2013, 392)

# Interpreting a Dendrogram

- Each **leaf** of the dendrogram represents an observation.
- As we move up the tree, leaves **fuse** into branches and branches into other branches.
- Observations that fuse at the **bottom** of the tree are **similar** to each other, whereas observations that fuse close to the **top** are **different**.
- We compare the **similarity** of two observations based on the location on the **vertical axis** where the branches containing the observations are first fused.
- We **cannot** compare the similarity of two observations based on their proximity along the **horizontal axis**.



## Interpreting a Dendrogram

- How do we **identify clusters** on the basis of a dendrogram?
- To do this, we make a **horizontal cut** across the dendrogram (see center and right panels above).
- The sets of observations **beneath the cut** can be interpreted as clusters.
- One single dendrogram can be used to obtain **any number** of clusters.
- The **height** of the cut to the dendrogram serves the same role as the  $K$  in  $K$ -means clustering: it controls the **number of clusters** obtained.

## Hierarchical Clustering vs. $K$ -Means Clustering

- Hierarchical clustering is called **hierarchical** because clusters obtained by a cut at a given height are **nested** within clusters obtained by cuts at any greater height.
- However, this assumption of hierarchical structure might be **unrealistic** for a given data set.
- Suppose that we have a group of people with a 50-50 split of males and females, evenly split among three countries of origin.

## Hierarchical Clustering vs. $K$ -Means Clustering

- Suppose further that the best division into two groups splits these people by gender, and the best division into three groups splits them by country.
- In this case, the clusters are **not nested**.
- Hierarchical clustering might yield **worse** (less accurate) results than  $K$ -means clustering.

# Clustering

---

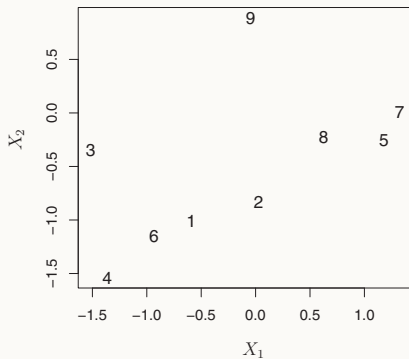
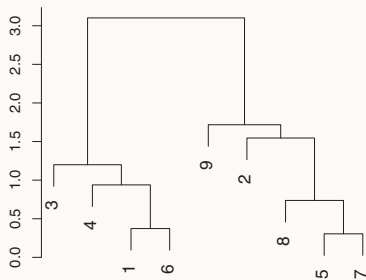
## The Hierarchical Clustering Algorithm

# The Hierarchical Clustering Algorithm

- The hierarchical clustering dendrogram is obtained via the following algorithm.
- We first define a **dissimilarity measure** between each pair of observations (most often, Euclidean distance is used).
- Starting at the **bottom** of the dendrogram, each of the  $n$  observations is treated as its **own cluster**.
- The two clusters that are **most similar** to each other are then **fused** so that there are now  $n - 1$  clusters.
- Next the two clusters that are **most similar** to each other are **fused** again, leaving us with  $n - 2$  clusters.
- The algorithm proceeds until all observations belong to one **single cluster**.

# The Hierarchical Clustering Algorithm – Example

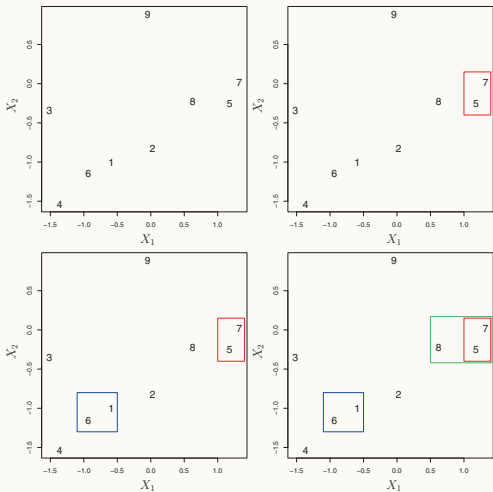
Hierarchical clustering dendrogram and initial data



(Source: James et al. 2013, 393)

# The Hierarchical Clustering Algorithm – Example

First few steps of the hierarchical clustering algorithm



(Source: James et al. 2013, 396)

# The Hierarchical Clustering Algorithm

- In the figure above, how did we determine that the cluster  $\{5, 7\}$  should be fused with the cluster  $\{8\}$ ?
- We have a concept of the dissimilarity between **pairs of observations**, but how do we define the dissimilarity between **two clusters** if they contain **multiple observations**?
- We need to **extend** the concept of dissimilarity between a pair of observations to a **pair of groups of observations**.
- The **linkage** defines the dissimilarity between two groups of observations.



# The Hierarchical Clustering Algorithm

Summary of the four most common types of linkage

<i>Linkage</i>	<i>Description</i>
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the <i>largest</i> of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the <i>smallest</i> of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the <i>average</i> 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 <i>inversions</i> .

# Clustering

---

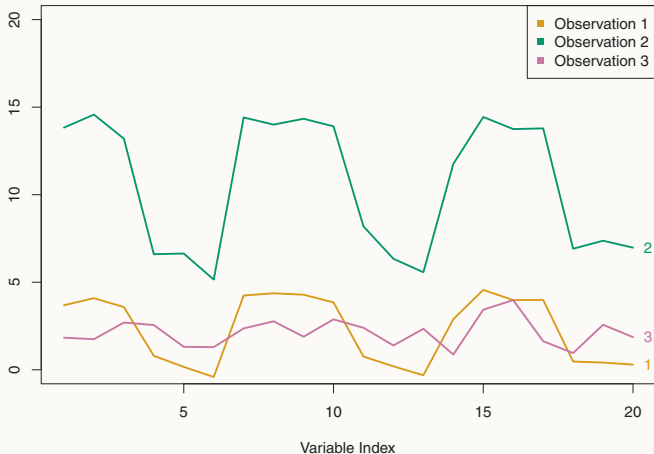
Choice of Dissimilarity Measure

## Choice of Dissimilarity Measure

- So far, we have used **Euclidean distance** as the **dissimilarity measure**.
- Sometimes other dissimilarity measures might be preferred.
- An alternative is **correlation-based distance** which considers two observations to be similar if their **features are highly correlated**.
- Correlation-based distance focuses on the **shapes** of observation profiles rather than their **magnitudes**.

# Choice of Dissimilarity Measure

Three observations with measurements on 20 variables



(Source: James et al. 2013, 398)

# Practical Issues in Clustering

---

# Practical Issues in Clustering

In order to perform clustering, some **decisions** must be made.

- Should the observations or features first be **standardized**?
- In the 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 the case of  **$K$ -means clustering**, how **many clusters** should we look for in the data?