Lecture 13

Cluster Analysis

Readings: Zelterman, 2015, Chapters 11

DSA 8070 Multivariate Analysis
November 8- November 12, 2021

Whitney Huang
Clemson University

Agenda

1. Overview
2. k-Means Clustering
3. Hierarchical Clustering
4. Model-Based Clustering

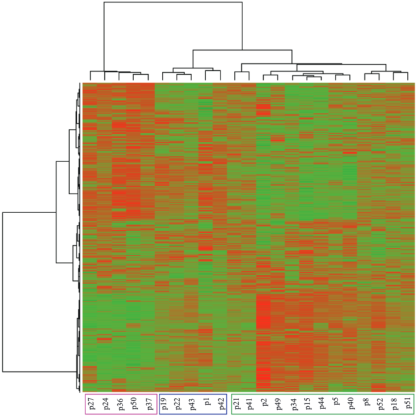
What is Cluster Analysis?

- Cluster: a collection of data objects
  - “Similar” to one another within the same cluster
  - “Dissimilar” to the objects in other clusters
- Cluster analysis: Grouping a set of data objects into clusters
- Clustering is unsupervised classification, unlike classification, there is no predefined classes, and the number of clusters is usually unknown

Some Examples of Clustering Applications

- Market Segmentation: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs

- Clustering Gene Expression Data:



Source: Fig. 1 of M. Garncarz et al, 2016

What Is Good Clustering?

- A good clustering method will produce clusters with
  - high within-class similarity
  - low between-class similarity

  For example, one can use the Euclidean distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{k=1}^{p} [x_{i,k} - x_{j,k}]^2}$ to quantify the similarity

- The quality of a clustering result depends on both the similarity measure used and its implementation

- The performance of a clustering method is measured by its ability to discover the hidden patterns

Major Clustering Approaches

- **Partitioning algorithm:** partition the observations into a pre-specified number of clusters, for example, k-means clustering
- **Hierarchy algorithm:** Construct a hierarchical decomposition of the observations to build a hierarchy of clusters, for example, hierarchical agglomerative clustering
- **Model-based Clustering:** A model is hypothesized for each of the clusters, for example, Gaussian mixture models

Partitioning Algorithm

Let $C_1, \cdots, C_K$ denote sets containing the indices of the observations $\{\boldsymbol{x}_i\}_{i=1}^n$ in each cluster. These sets satisfy two properties:

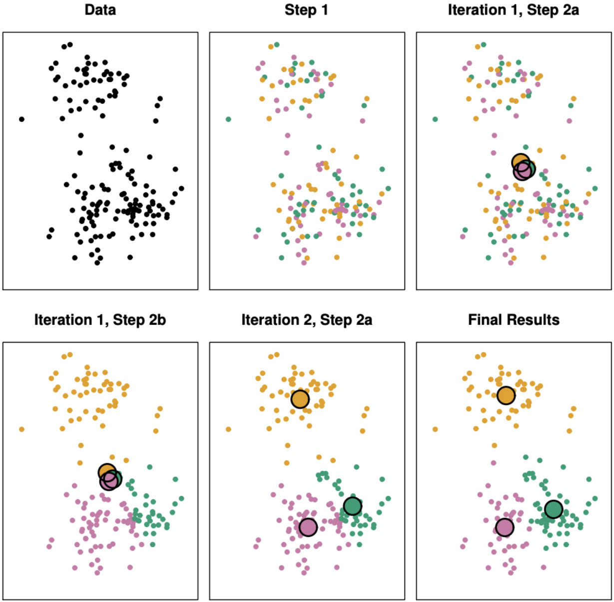
- $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \cdots, n\} \Rightarrow$ each observation belongs to at least one of the K clusters
- $C_k \cap C_{k'} = \varnothing \, \forall k \neq k' \Rightarrow$ no observation belongs to more than one cluster

For instance, if the i_{th} observation (i.e. $\boldsymbol{x}_i$) is in the k_{th} cluster, then $i \in C_k$

The k-Means Algorithm

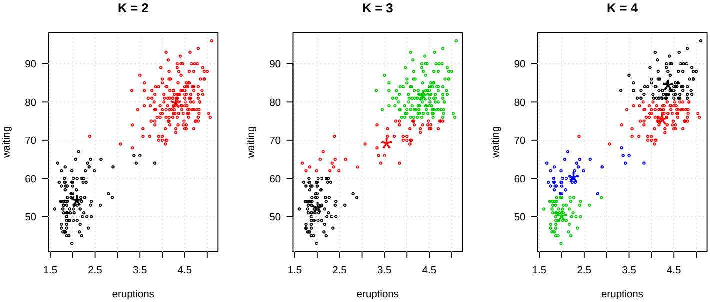
- **Step 0:** Choose the number of clusters K
- **Step 1:** Randomly assign a cluster (from 1 to K), to each of the observations. These serve as the initial cluster assignmemts
- **Step 2:** Iterate until the cluster assignment stop changing
  - For each of the K cluster, compute the cluster centroid. The k_{th} cluster centroid is the mean vector of the observations in the k_{th} cluster
  - Assign each observations to the cluster whose centroid is closest in terms of Euclidean distance

k-Means Clustering Illustration



K-Means Clustering in R

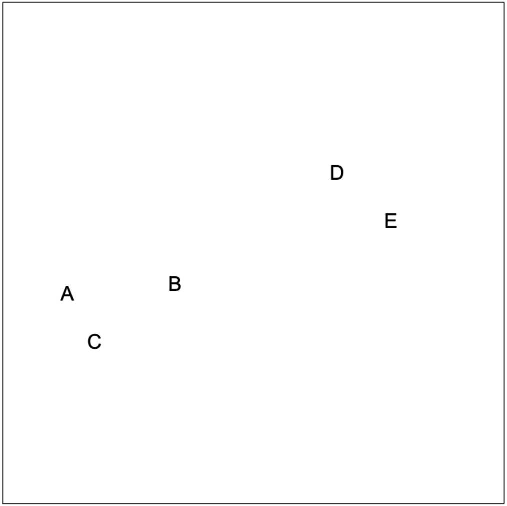
```r
kmean3.faithful <- kmeans(x = faithful, centers = 3)
```



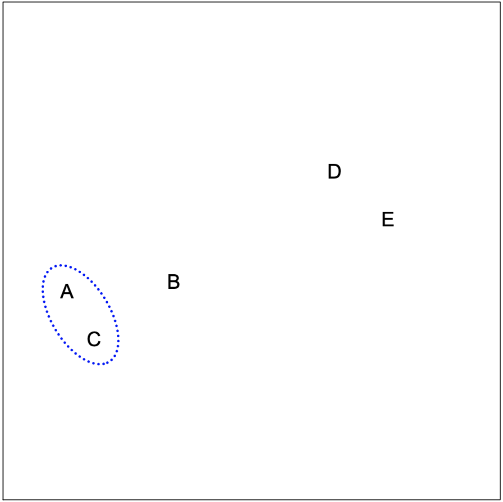
Hierarchical Clustering

- k-means clustering requires us to pre-specify the number of clusters K
- Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K
- Agglomerative clustering: This is a “bottom-up” approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy

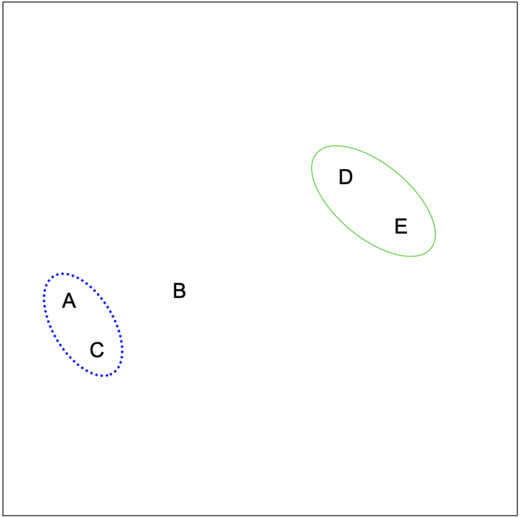
Hierarchical Agglomerative Clustering Illustration



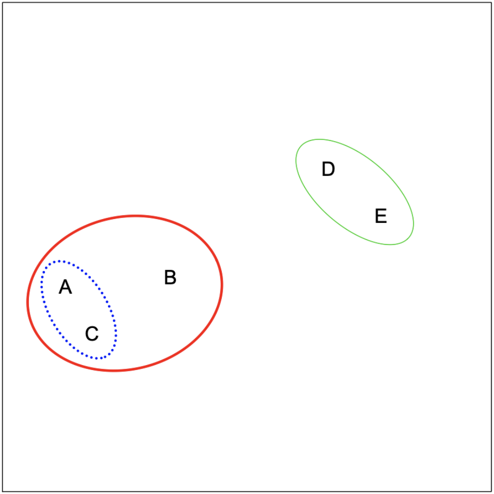
Hierarchical Agglomerative Clustering Illustration



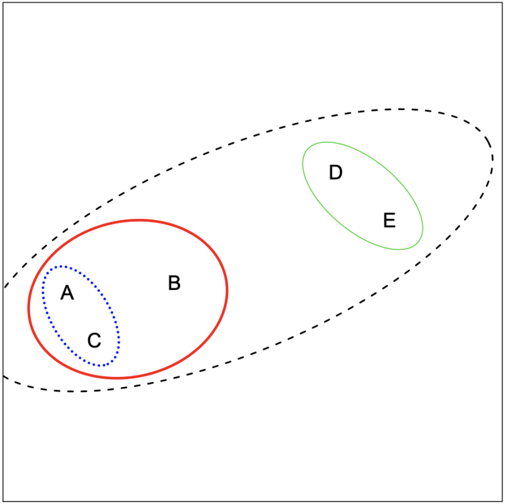
Hierarchical Agglomerative Clustering Illustration



Hierarchical Agglomerative Clustering Illustration

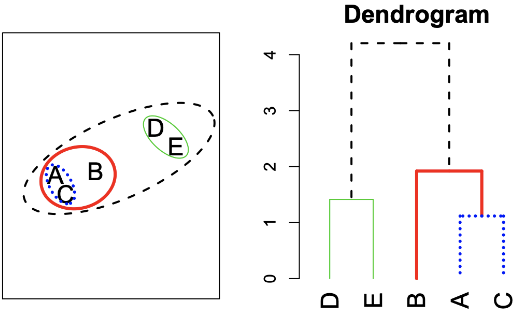


Hierarchical Agglomerative Clustering Illustration



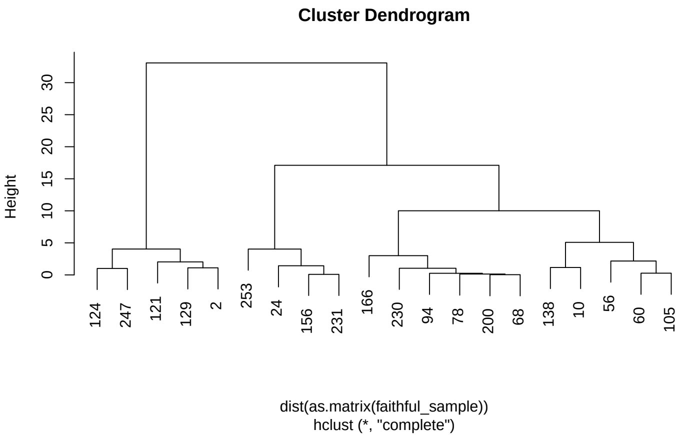
Hierarchical Agglomerative Clustering Algorithm

1. Start with each observation in its own cluster
2. Identify the closest two clusters and merge them
3. Repeat
4. Ends when all observations are in a single cluster



Hierarchical Agglomerative Clustering in R

```
hc.faithful <- hclust(dist(faithful_sample))
plot(hc.faithful)
```



Model-based clustering

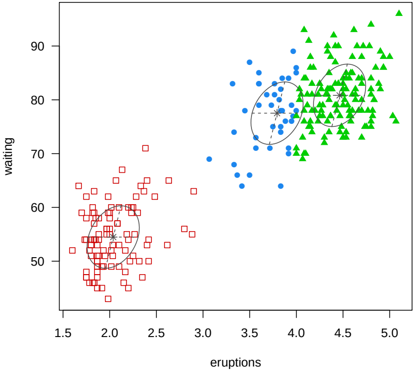
- One disadvantage of k-means is that they are largely heuristic and not based on formal statistical models. Formal inference is not possible
- Model-based clustering is an alternative:
  - Sample observations arise from a mixture distribution of two or more components
  - Each component (cluster) is described by a probability distribution and has an associated probability in the mixture.
  - In Gaussian mixture models, we assume each cluster follows a multivariate normal distribution
  - Therefore, in Gaussian mixture models, the model for clustering is a mixture of multivariate normal distributions

Fitting a Gaussian Mixture Model in R

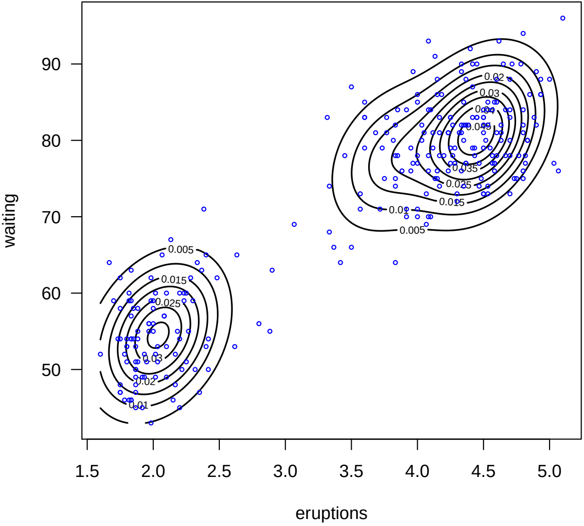
```r
library(mclust)
```

```
## Package 'mclust' version 5.4.5
## Type 'citation("mclust")' for citing this R package in publications.
```

```r
BIC <- mclustBIC(faithful)
model1 <- Mclust(faithful, x = BIC)
```



Fitting a Gaussian Mixture Model in R Cond't



Model-Based Clustering Analysis for Iris Data