Unsupervised Learning II

Roberta De Vito

BROWN
Public Health

Cluster

- groups that are similar
- homogeneous property
- differences among the groups

Cluster analysis in two steps

- Choice of a proximity measure
- Choice of group-building algorithm

Proximity between objects

$$D = \begin{pmatrix} d_{11} & d_{12} & \dots & \dots & \dots & d_{1n} \\ \vdots & d_{22} & & & & \vdots \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & \dots & \dots & d_{nn} \end{pmatrix}$$

Proximity between objects

1. Euclidean distance:

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

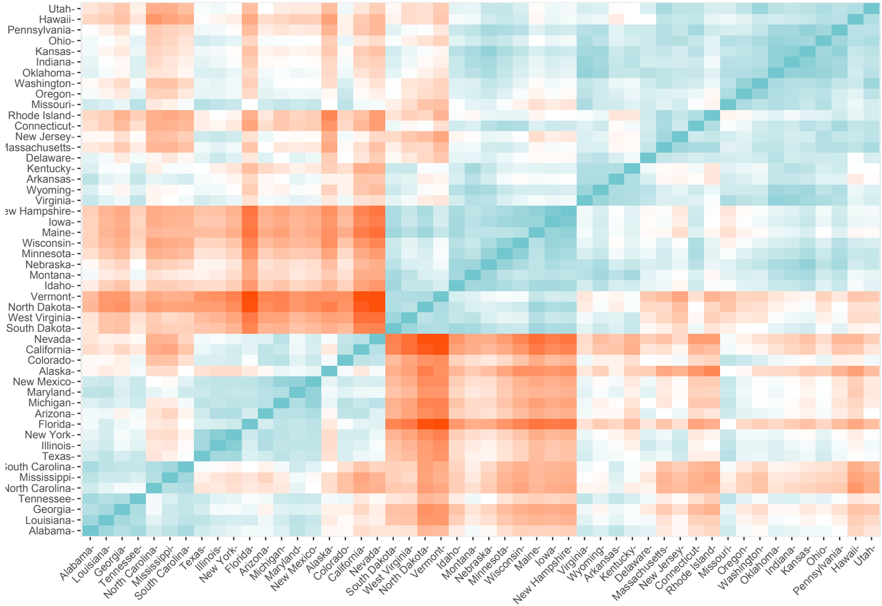
2. Manhattan distance:

$$d_{man}(x, y) = \sum_{i=1}^{n} |(x_i - y_i)|^2$$

The arrest data

```
> head(df)
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
```

The distance matrix: 0 blue, 200 red: Q1 in prismia



K-means Clustering Q2

- High intra-class similarity in the same cluster
- Each cluster is represented by its center (centroids)
- k represents the number of groups
- $C_1 C_2 \cup \cdots \cup C_K = 1, \ldots, n$
- $C_k \cap C_{k'} = \emptyset$ for
- The total within-cluster variation

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

$$min \sum_{k=1}^{K} W(C_k)$$

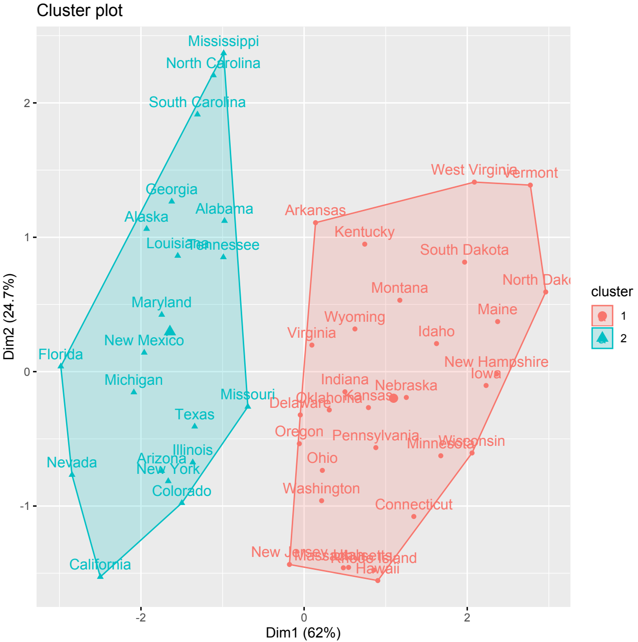
K-means Algorithm

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   2.1 For each of the K clusters, compute the cluster centroid.
   2.2 Assign each observation to the cluster whose centroid is closest (Euclidean distance)

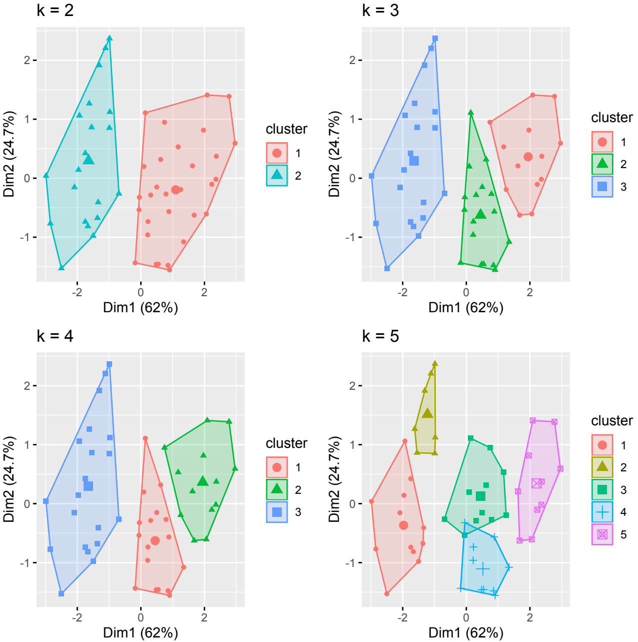
Let's see the output with two clusters

```
> k2 <- kmeans(df, centers = 2, nstart = 25)
> k2
K-means clustering with 2 clusters of sizes 29, 21

Cluster means:
     Murder  Assault UrbanPop     Rape
1  4.841379 109.7586 64.03448 16.24828
2 11.857143 255.0000 67.61905 28.11429

Clustering vector:
       Alabama         Alaska        Arizona       Arkansas     California
             2              2              2              2              2
      Colorado    Connecticut       Delaware        Florida        Georgia
             2              1              2              2              2
        Hawaii          Idaho       Illinois        Indiana           Iowa
             1              1              2              1              1
```

The cluster plot: Q3



How many clusters?



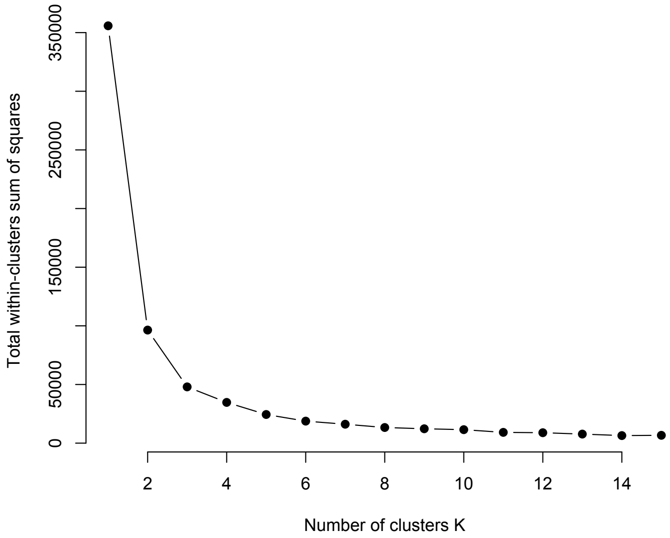
Elbow Method: number of clusters

Minimize

$$\sum_{k=1}^{k} W(C_k)$$

1. Compute clustering algorithm for different values of k.
2. For each k, calculate the total within-cluster sum of square
3. Plot the curve of wss according to the number of clusters k.

How many clusters with the Elbow method?



Gap Method: number of clusters

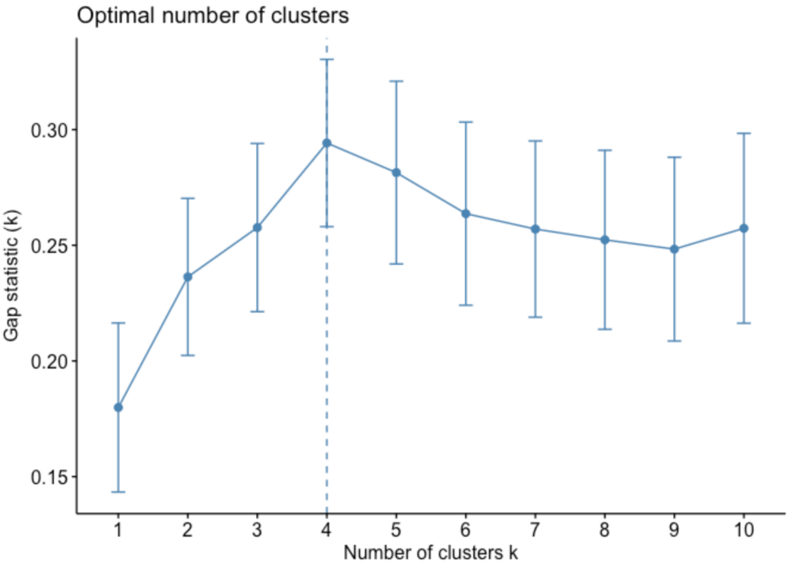
1. Cluster the observed data, varying the number of clusters from $k = 1, \ldots, K$, and compute the corresponding W_k
2. Generate B reference data sets and cluster each of them with varying number of clusters $k = 1, \ldots, k_{max}$. Compute the estimated gap statistics

$$Gap_n(k) = E_n log(W_k) - log(W_k)$$

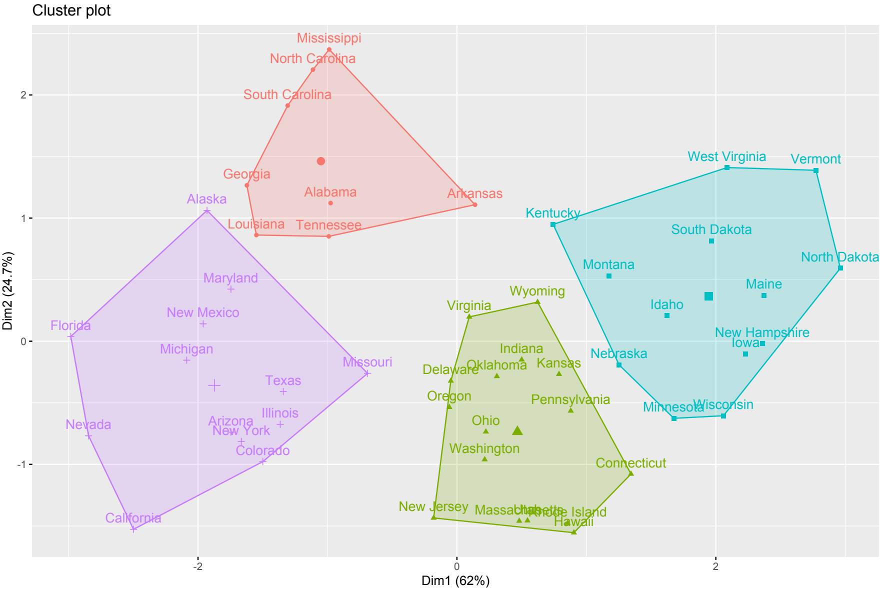
3. E_n is defined via bootstrapping
4. Aim: maximize $Gap_n(k)$
5. Compute the standard deviation s_k
6. Choose the number of clusters as the smallest k such that

$$Gap_k \geq Gap_{k+1} - s_{k+1}$$

How many clusters with the Gap method?



The final output: Q4



The Hierarchical Clustering

- Starting out 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, $n1$ clusters
- Next $n - 2$ clusters
- The algorithm proceeds until all of the observations belong to one single cluster, and the dendrogram is complete

The dendogram