Workshop on "Enhancing your Computer Coding Skills"

Exploratory data analysis and clustering I

Thierry Denœux

tdenoeux@utc.fr
https://www.hds.utc.fr/~tdenoeux

Université de technologie de Compiègne

August 2022

Exploratory data analysis

- Exploratory Data Analysis (EDA): techniques for summarizing the main characteristics of data using statistical graphics and data visualization techniques.
- Basic methods (1D or 2D plots):
  - histograms
  - boxplots
  - scatter plots
  - ...
- More advanced techniques (visualizing high-dimensional data):
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling (MDS)
  - ...

Clustering

- Finding groups in data, such that observations within each group are similar, and observations from different groups are dissimilar.
- Applications
  - EDA (finding groups of companies, countries, etc., with similar characteristics for further analysis)
  - In marketing: finding groups of customers with similar characteristics and/or purchasing behavior (customer segmentation)
  - ...

What we will see

- Today:
  - How to draw some basic plots
  - Clustering algorithms for cross-sectional numerical data:
    - Partitional clustering: c-means algorithm
    - Clustering evaluation criteria
    - Fuzzy clustering: fuzzy c-means (FCM) algorithm
    - (Hierarchical clustering)
- Next class (Aug 20, 2022):
  - More advances data visualization techniques (PCA, MDS)
  - Clustering qualitative and hybrid (numerical/qualitative) data
  - Time series clustering

Running example: KOF globalization data

- The KOF Swiss Economic Institute publishes each year globalisation indices measures the economic, social and political dimensions of globalisation.
- These indices aggregate various economic, social and politcal data (see next slides)
- We will consider the 2014 cross-section of 157 countries to illustrate various data visualization and clustering techniques

Economic globalization index

A.	**Economic Globalization**		
	i) Data on actual Flows		
	Trade (percent of GDP)	World Bank (2014)	Trade is the sum of exports and imports of goods and services measured as a share of gross domestic product. Data are in percent of GDP.
	Foreign Direct Investment, stocks (percent of GDP)	UNCTAD (2013)	Sum of inward and outward FDI stock as a percentage of GDP.
	Portfolio Investment (percent of GDP)	IMF (2014)	Portfolio investment is the sum of portfolio investment assets stocks and portfolio investment liabilities stocks. Data are in percent of GDP.
	Income Payments to Foreign Nationals (percent of GDP)	World Bank (2014)	Income payments refer to employee compensation paid to nonresident workers and investment income (payments on direct investment, portfolio investment, other investments). Income derived from the use of intangible assets is excluded. Data are in percent of GDP.
	ii) Data on restrictions		
	Hidden Import Barriers	Gwartney et al. (2013)	The index is based on the Global Competitiveness Report's survey question: "In your country, tariff and non-tariff barriers significantly reduce the ability of imported goods to compete in the domestic market." The question's wording has varied slightly over the years.
	Mean Tariff Rate	Gwartney et al. (2013)	As the mean tariff rate increases, countries are assigned lower ratings. The rating declines toward zero as the mean tariff rate approaches 50%.
	Taxes on International Trade (percent of current revenue)	World Bank (2014)	Taxes on international trade include import duties, export duties, profits of export or import monopolies, exchange profits, and exchange taxes. Current revenue includes all revenue from taxes and nonrepayable receipts (other than grants) from the sale of land, intangible assets, government stocks, or fixed capital assets, or from capital transfers from nongovernmental sources. It also includes fines, fees, recoveries, inheritance taxes, and nonrecurrent levies on capital. Data are for central government and in percent of all current revenue.
	Capital Account Restrictions	Gwartney et al. (2013)	Index based on two components: (i) Beginning with the year 2002, this sub-component is based on the question: "Foreign ownership of companies in your country is (1) rare, limited to minority stakes, and often prohibited in key sectors or (2) prevalent and encouraged". For earlier years, this sub-component was based on two questions about "Access of citizens to foreign capital markets and foreign access to domestic capital markets". (ii) Index based on the IMF's Annual Report on Exchange Arrangements and Exchange Restrictions, including 13 different types of capital controls. It is constructed by subtracting the number of restrictions from 13 and multiplying the result by 10.

Social globalization index

B. Social Globalization		
i) Data on Personal Contact		
Telephone Traffic	International Telecommunication Union (2013)	International voice traffic is the sum of international incoming and outgoing fixed telephone traffic (in minutes per person).
Transfers (percent of GDP)	World Bank (2014)	Sum of gross inflows and gross outflows of goods, services, income, or financial items without a quid pro quo. Data are in percent of GDP.
International Tourism	World Bank (2014)	Sum of arrivals and departures of international tourists as a share of population.
Foreign Population (percent of total population)	World Bank (2014)	Foreign population is the number of foreign or foreign-born residents in a country. Data are in percent of total population.
International letters (per capita)	Universal Postal Union, Postal Statistics database	Number of international letters sent and recieved per capita.
ii) Data on Information Flows		
Internet Users (per 1000 people)	World Bank (2014)	Internet users are people with access to the worldwide internet network.
Television (per 1000 people)	World Bank (2007), International Telecommunication Union (2013)	Share of households with a television set.
Trade in Newspapers (percent of GDP)	United Nations Commodity Trade Statistics Database (2013)	The sum of exports and imports in newspapers and periodicals in percent of GDP. Data are provided by the Statistical Division of the United Nations and correspond to those published in the U.N. World Trade Annual.
iii) Data on Cultural Proximity		
Number of McDonald's Restaurants (per capita)	various sources	Number of McDonald's Restaurants (per capita).
Number of Ikea (per capita)	Ikea	Number of Ikea (per capita).
Trade in books (percent of GDP)	UNESCO (various years), United Nations Commodity Trade Statistics Database (2013)	The sum of exports and imports in books and pamphlets in percent of GDP. Data are provided by the Statistical Division of the United Nations and correspond to those published in the U.N. World Trade Annual.

Political globalization index

C. Political Globalization

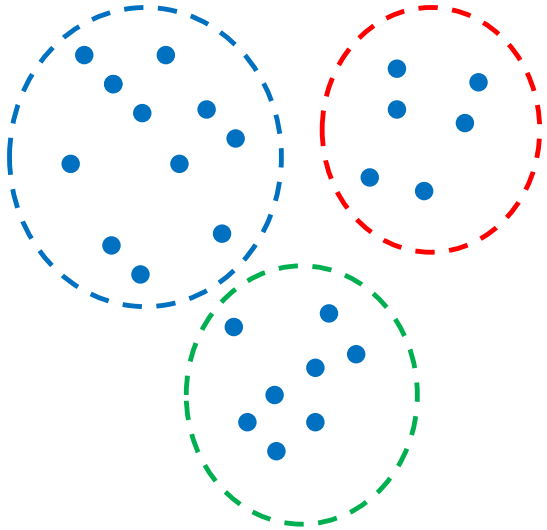
Embassies in Country	Europa World Yearbook (various years)	Absolute number of embassies in a country.
Membership in International Organizations	CIA World Factbook (various years)	Absolute number of international inter-governmental organizations.
Participation in U.N. Security Council Missions	Department of Peacekeeping Operations, UN	Personnel contributed to U.N. Security Council Missions per capita.
International Treaties	United Nations Treaties Collection	Any document signed between two or more states and ratified by the highest legislative body of each country since 1945. Not ratified treaties, or subsequent actions, and annexes are not included. Treaties signed and ratified must be deposited in the Office of Secretary General of the United Nations to be included.

Overview

Questions

1. Are different groups of countries, with similar globalization characteristics?
2. How many groups are there?
3. How to assign observations to each group?

Partition



Definition

A partition of a set $\mathcal{X}$ is a collection of subsets $\mathcal{X}_1, \ldots, \mathcal{X}_c$ such that $\mathcal{X}_1 \cup \ldots \cup \mathcal{X}_c = \mathcal{X}$, and for any $i \neq j$, $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$. Each subset $\mathcal{X}_k$ is called a class, a group or a cluster.

Partitional clustering aims at finding a partition of n observations (objects) in a dataset.

Representation of a partition

- Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a set of n objects (p-dimensional attribute vectors). A partition in c groups can be represented in several ways:
  1. As a vector $\mathbf{y} = (y_1, \ldots, y_n)$ where $y_i = k$ if $x_i \in \mathcal{X}_k$
  2. As an $n \times c$ matrix $U = (u_{ik})$, where $u_{ik} = 1$ if $y_i = k$, $u_{ik} = 0$ otherwise
  3. As an $n \times n$ matrix $R = (r_{ij})$ such that $r_{ij} = 1$ if $y_i = y_j$, $r_{ij} = 0$ otherwise
- We will mainly use the 1st and 2nd representations.
- Matrix U verifies:

$$\sum_{k=1}^{c} u_{ik} = 1, \quad i = 1, \ldots, n.$$

- The number of observations in cluster k is

$$n_k = \sum_{i=1}^{n} u_{ik}, \quad k = 1, \ldots, c.$$

Partitional clustering algorithms

- There exist a lot of partitional clustering algorithms.
- The (hard) c-means (HCM) algorithm was introduced in the 1960's but it is still the most widely used today, because of its simplicity and speed.
- It is applicable to data with numerical attributes.

Overview

(Hard) c-Means Algorithm

1. Fix the number c of clusters
2. Initialize cluster centers (prototypes) $v_1, \ldots, v_c$ randomly.
3. Compute distances $d_{ik} = \|x_i - v_k\|$ between each observation x_i and each prototype v_k, and assign each x_i to the cluster of its nearest prototype:

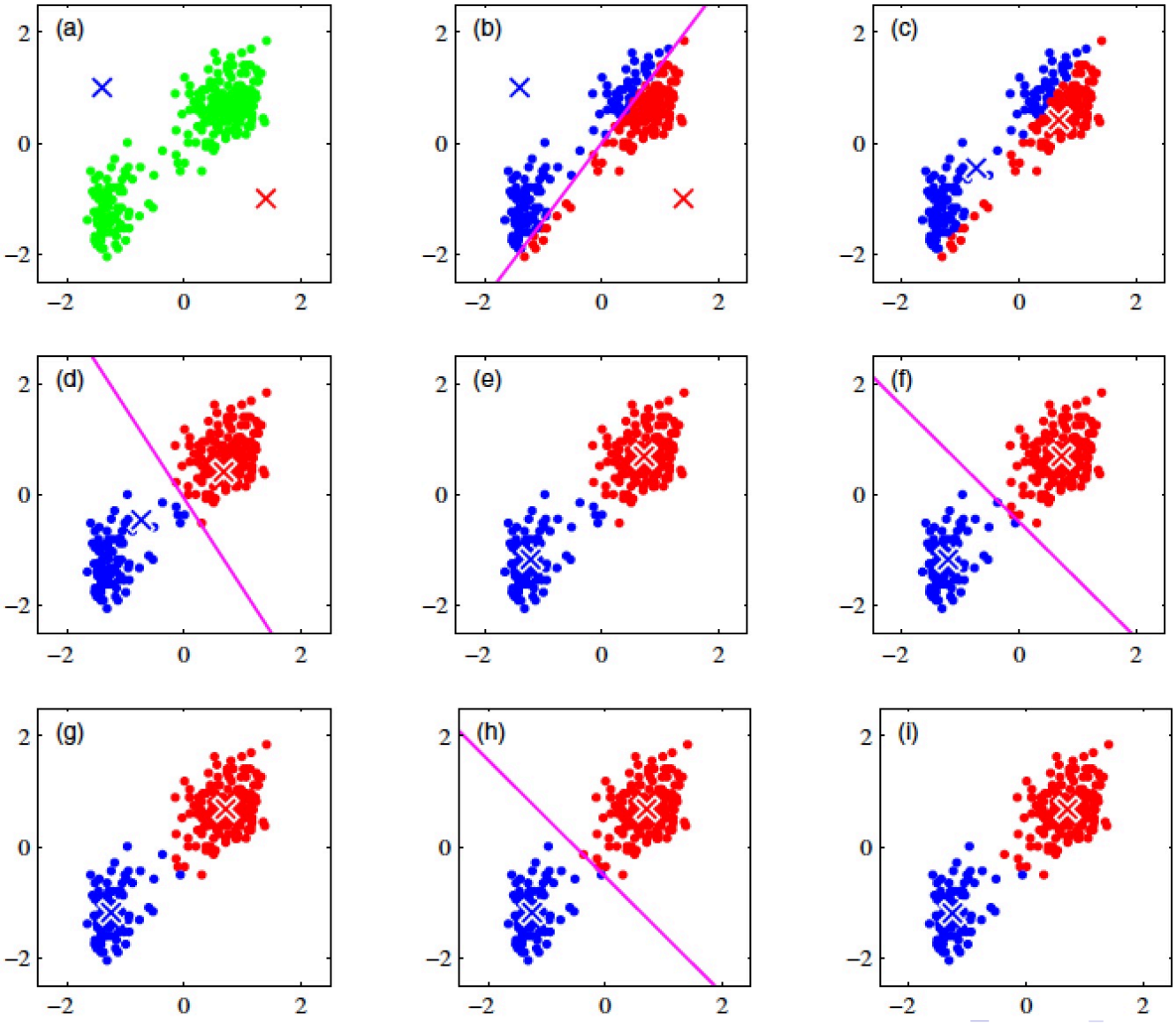
$$u_{ik} := \begin{cases} 1 & \text{if } d_{ik} = \min_\ell d_{i\ell} \\ 0 & \text{otherwise} \end{cases}$$

4. Recompute each prototype v_k as the center of mass of cluster k:

$$v_k := \frac{\sum_{i=1}^n u_{ik} x_i}{\sum_{i=1}^n u_{ik}} = \frac{1}{n_k} \sum_{\{i : u_{ik}=1\}} x_i, \quad k = 1, \ldots, c$$

5. If the prototypes have not changed in the last iteration, stop. Otherwise, return to Step 3.

Illustration of the HCM algorithm



Why does it work?

- Let $U = (u_{ik})$ and $V = (v_1, \ldots, v_c)$. Consider the following cost function:

$$\boxed{J_{\text{HCM}}(U, V) = \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik} d_{ik}^2}$$

  where $d_{ik} = \|x_i - v_k\|$ is the Euclidean distance between x_i and v_k.
- It can be shown that $J_{\text{HCM}}(U, V)$ decreases at each step of HCM.
- As a consequence, the algorithm converges to a local minimum of $J_{\text{HCM}}(U, V)$.

Proof that $J_{\text{HCM}}(U, V)$ decreases at each iteration

- The HCM algorithm alternates 2 steps:
  1. Update of U with V fixed
  2. Update of V with U fixed
- Step 1: the cost function can be written as

$$J_{\text{HCM}}(U, V) = \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{c} u_{ik} d_{ik}^2}_{d_{i,k(i)}^2} = \sum_{i=1}^{n} d_{i,k(i)}^2$$

When updating U for fixed V, each $k(i)$ is chosen to minimize $d_{i,k(i)}^2$, i.e., $d_{i,k(i)} = \min_\ell d_{i\ell}$, and hence $J_{\text{HCM}}(U, V)$.

Proof that $J_{\text{HCM}}(U, V)$ decreases at each iteration (cont.)

- Step 2: the cost function can alternatively be written as

$$J_{\text{HCM}}(U, V) = \sum_{k=1}^{c} \underbrace{\sum_{i=1}^{n} u_{ik} d_{ik}^2}_{\mathcal{I}(v_k)}$$

where

$$\mathcal{I}(v_k) = \sum_{i=1}^{n} u_{ik} (x_i - v_k)^T (x_i - v_k)$$

- We have

$$\frac{\partial \mathcal{I}(v_k)}{\partial v_k} = 0 \Leftrightarrow v_k = \frac{1}{n_k} \sum_{\{i:u_{ik}=1\}} x_i.$$

so updating V in such a way that v_k is the center of cluster k minimizes $J_{\text{HCM}}(U, V)$. Proof

Remarks

1. The prototypes may be initialized with randomly selected observations.
2. The final solution usually depends on the initial prototypes. It is recommended to run the algorithm several times with different random initializations, and keep the best solution according to J_{HCM}.
3. The choice of the number c of clusters is a difficult problem in clustering. This problem is addressed in the next section.

Overview

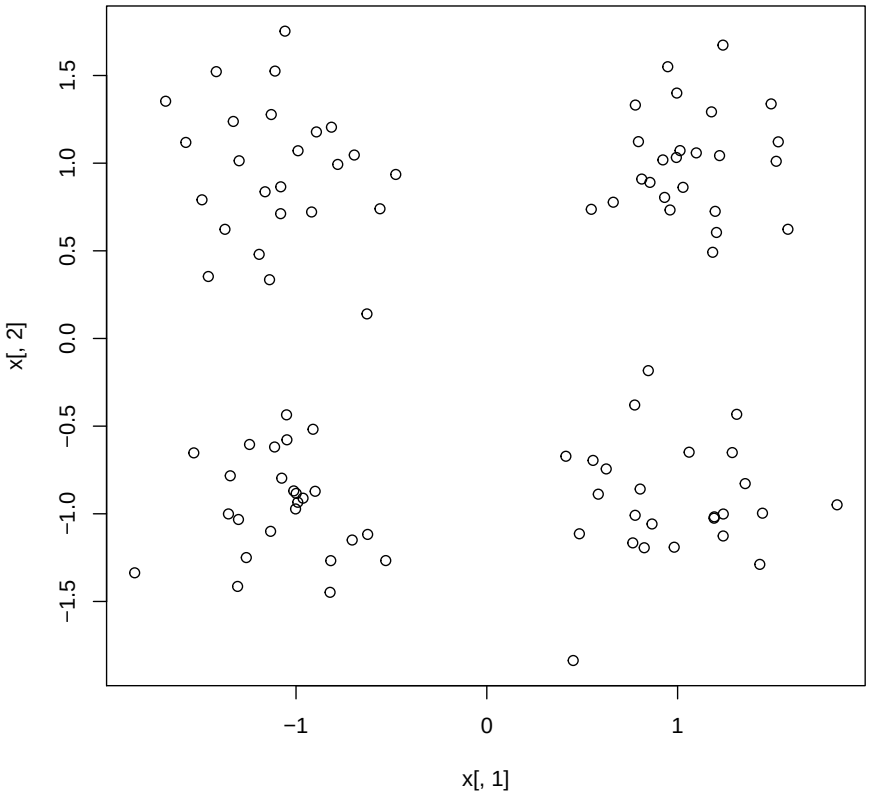
Validity of a partition

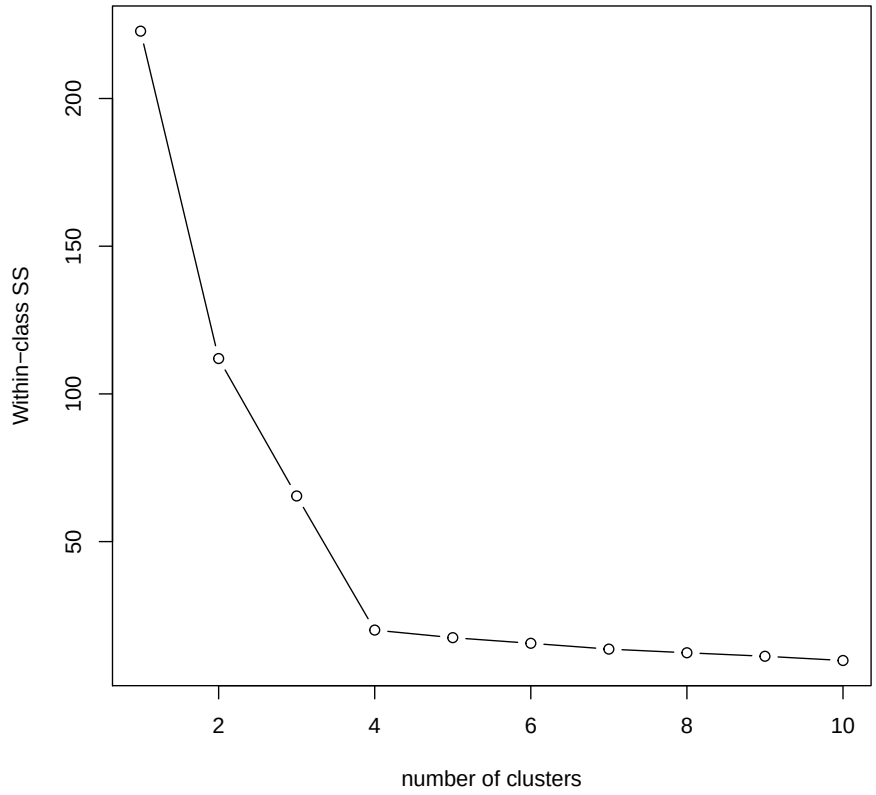
- After we have generated a partition using a clustering algorithm, we need ways to evaluate the validity/quality of this partition.
- If several partitions are generated (e.g., with different numbers of clusters), we need ways to compare them.
- The main approaches include:
  - Graphical representations
  - Internal indices

Graphical representations

- If p is small (say, $p \leq 5$), we can visually inspect the data using 2D or 3D plots. If p is large, we need more sophisticated methods.
- A simple method to determine the number of clusters is to plot J_{HCM} as a function of c and see if the curve decreases more slowly beyond some value of c ("knee"). This method works only of the clusters are well separated (see next slides).
- The silhouette plot is a more advanced graphical representation that provides visual information about the quality of the partition.

The knee method for data with well-separated clusters





Silhouette plot

- The silhouette plot is a graphical representation of data that can be used to visually evaluate the validity of a partition into clusters.
- For each object i in cluster $y_i = k(i)$, let a_i be the mean distance to the other objects in the same cluster

$$a_i = \frac{1}{n_{k(i)} - 1} \sum_{j \neq i, y_j = k(i)} d(i,j)$$

where $d(i,j)$ is the distance between objects i and j, y_j is the cluster of object j. (We assume $n_{k(i)} > 1$).

- Let b_i be the smallest mean distance of object i to all objects in any other cluster, to which i does not belong:

$$b_i = \min_{k \neq k(i)} \frac{1}{n_k} \sum_{j : y_j = k} d(i,j)$$

Silhouette (continued)

- We define the silhouette value of object i as

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- We can see that

$$s_i = \begin{cases} 1 - a_i/b_i & \text{if } a_i < b_i \\ 0 & \text{if } a_i = b_i \\ b_i/a_i - 1 & \text{if } a_i > b_i \end{cases}$$

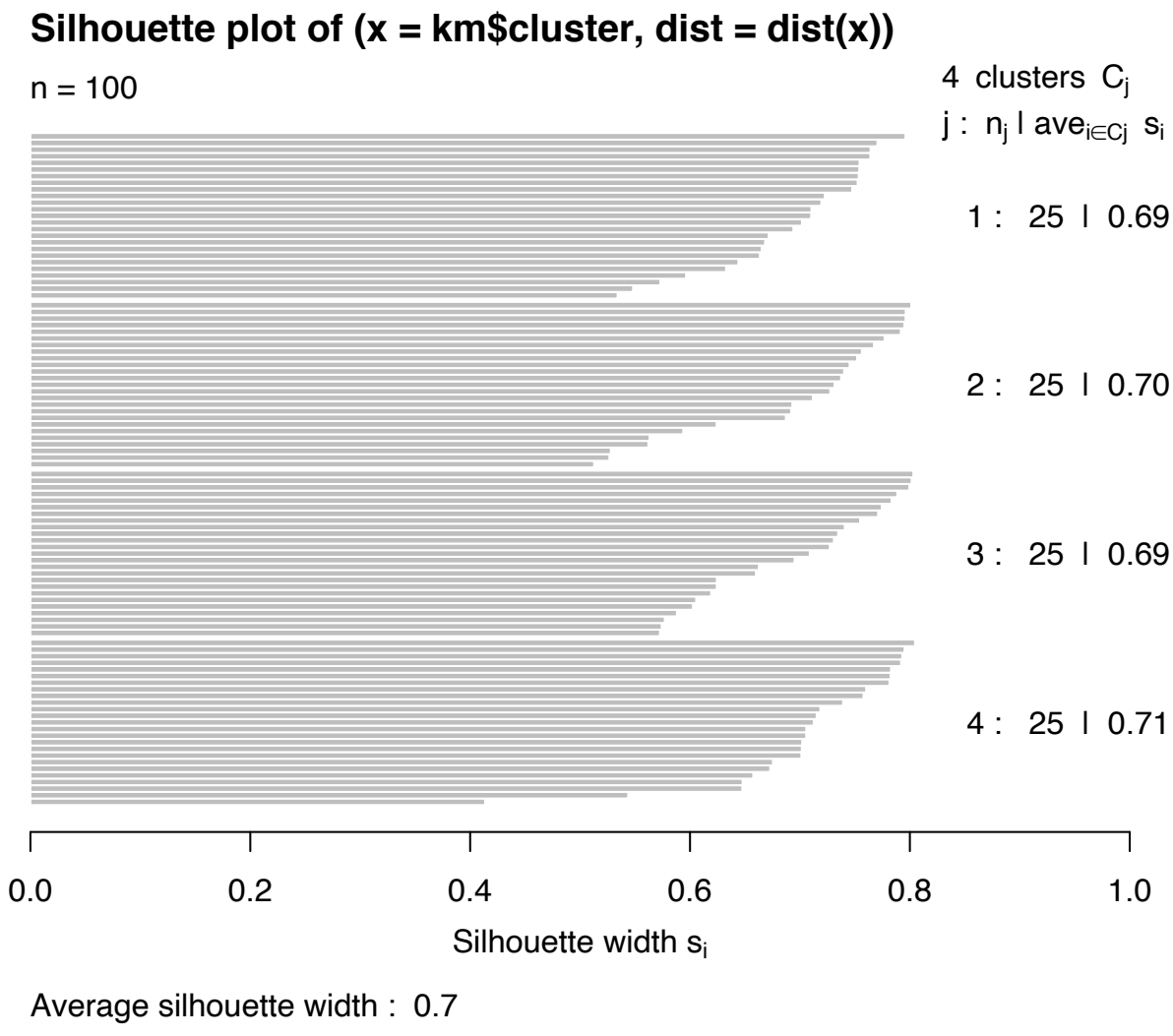
  and $-1 \leq s_i \leq 1$.

- Observations with a large s_i (close to 1) are well clustered, a small s_i (around 0) means that the observation lies between two clusters, and observations with $s_i < 0$ are probably placed in the wrong cluster.

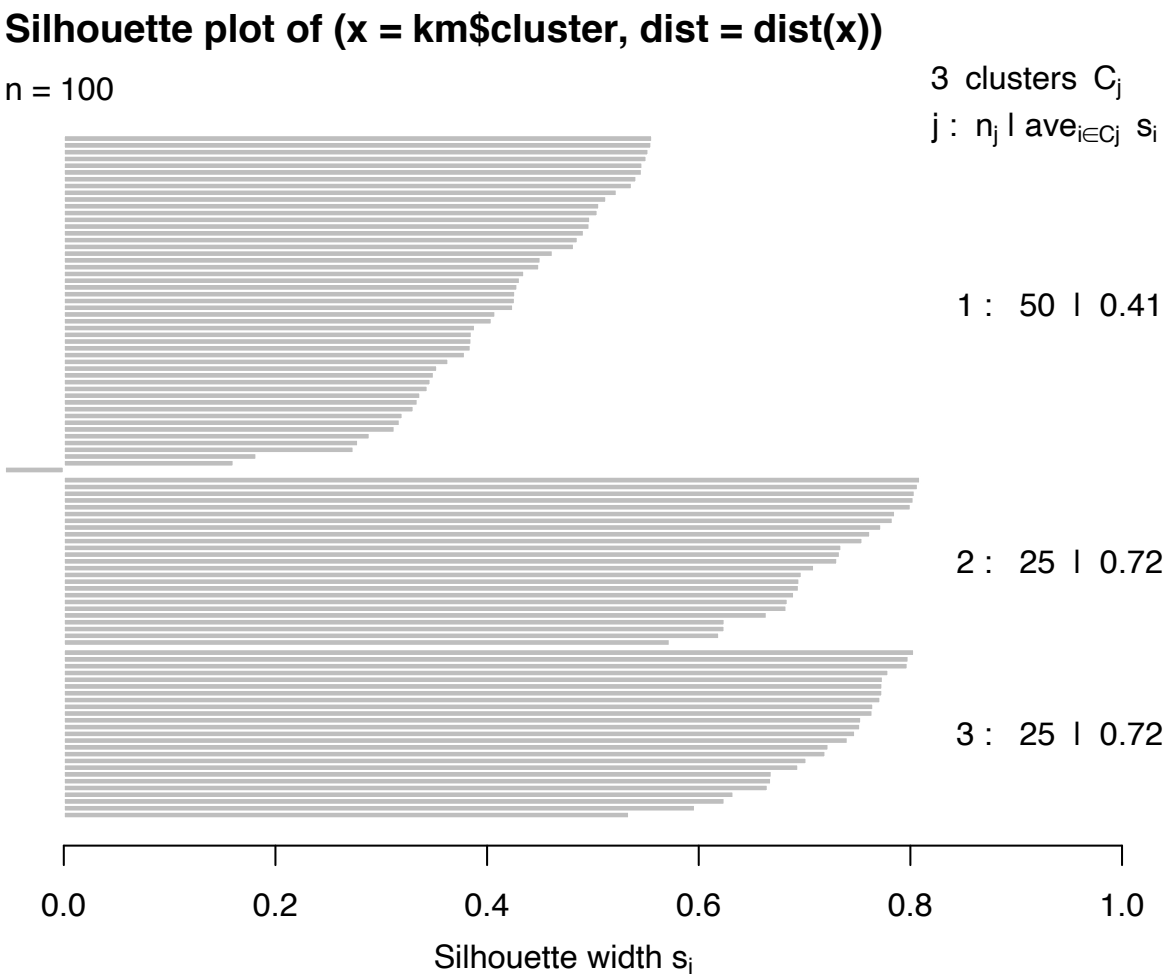
Silhouette plot in R

```
library(cluster)

km<-kmeans(x,centers=3,nstart=10)
D<- dist(x)
sil<-silhouette(km$cluster,D)
plot(sil)
```

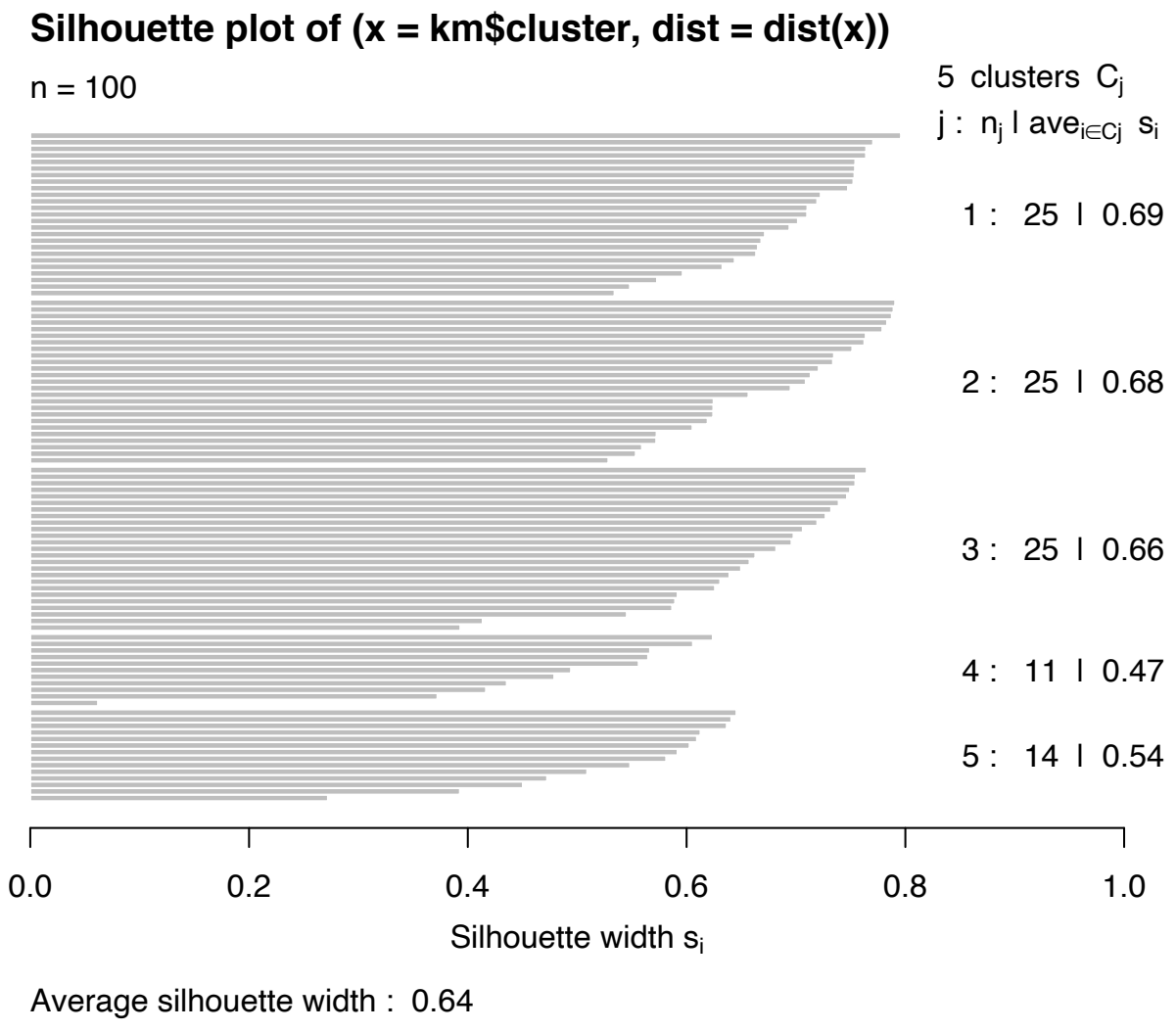
Silhouette plot of the 4-cluster synthetic data with $c = 4$



Silhouette plot of the 4-cluster synthetic data with $c = 3$



Silhouette plot of the 4-cluster synthetic data with $c = 5$



Internal indices

- Internal indices measure the "intrinsic" quality of a partition (how well-separated the clusters are).
- We have seen that the mean silhouette value

$$\overline{s} = \frac{1}{n} \sum_{i=1}^{n} s_i$$

can be used as an internal index. (The larger $\overline{s}$, the better clustering).

- There exist many other internal indices. Two of the most widely used are the Davies-Bouldin and the Dunn indices.

Davis-Bouldin index

- Let $\overline{d}_k$ be the mean distance between an object of cluster k and the center v_k of that cluster:

$$\overline{d}_k = \frac{1}{n_k} \sum_{i=1}^{n} u_{ik} d_{ik}$$

  It is a measure of the scatter/spread of cluster k.

- A measure of within-to-between spread between clusters k and ℓ is

$$R_{k\ell} = \frac{\overline{d}_k + \overline{d}_\ell}{d(v_k, v_\ell)}$$

  $R_{k\ell}$ is small if clusters k and ℓ are well separated.

Davis-Bouldin index (continued)

- The index of cluster k is

$$R_k = \max_{\ell \neq k} R_{k\ell}$$

  R_k is small if cluster k is well separated from all other clusters.
- The Davis-Bouldin (DB) index is defined as

$$DB = \frac{1}{c} \sum_{k=1}^{c} R_k$$

- The smaller DB, the better clustering.

Dunn index

- Let $\delta_{k\ell}$ denote the smallest distance between a vector of cluster k and a vector of cluster ℓ:

$$\delta_{k\ell} = \min_{\{(i,j):u_{ik}=1,u_{j\ell}=1\}} d(i,j)$$

- Let δ_{min} denote the smallest such distance:

$$\delta_{\text{min}} = \min_{k\neq\ell} \delta_{k\ell}$$

Dunn index (continued)

- Let Δ_k denote the diameter of cluster k, defined as the largest distance separating two distinct points in that cluster:

$$\Delta_k = \max_{\{(i,j):u_{ik}=u_{jk}=1\}} d(i,j)$$

- Let $\Delta_{\max}$ denote the maximum diameter:

$$\Delta_{\max} = \max_{1 \leq k \leq c} \Delta_k$$

- The Dunn index is defined as the quotient of $\delta_{\min}$ and $\Delta_{\max}$:

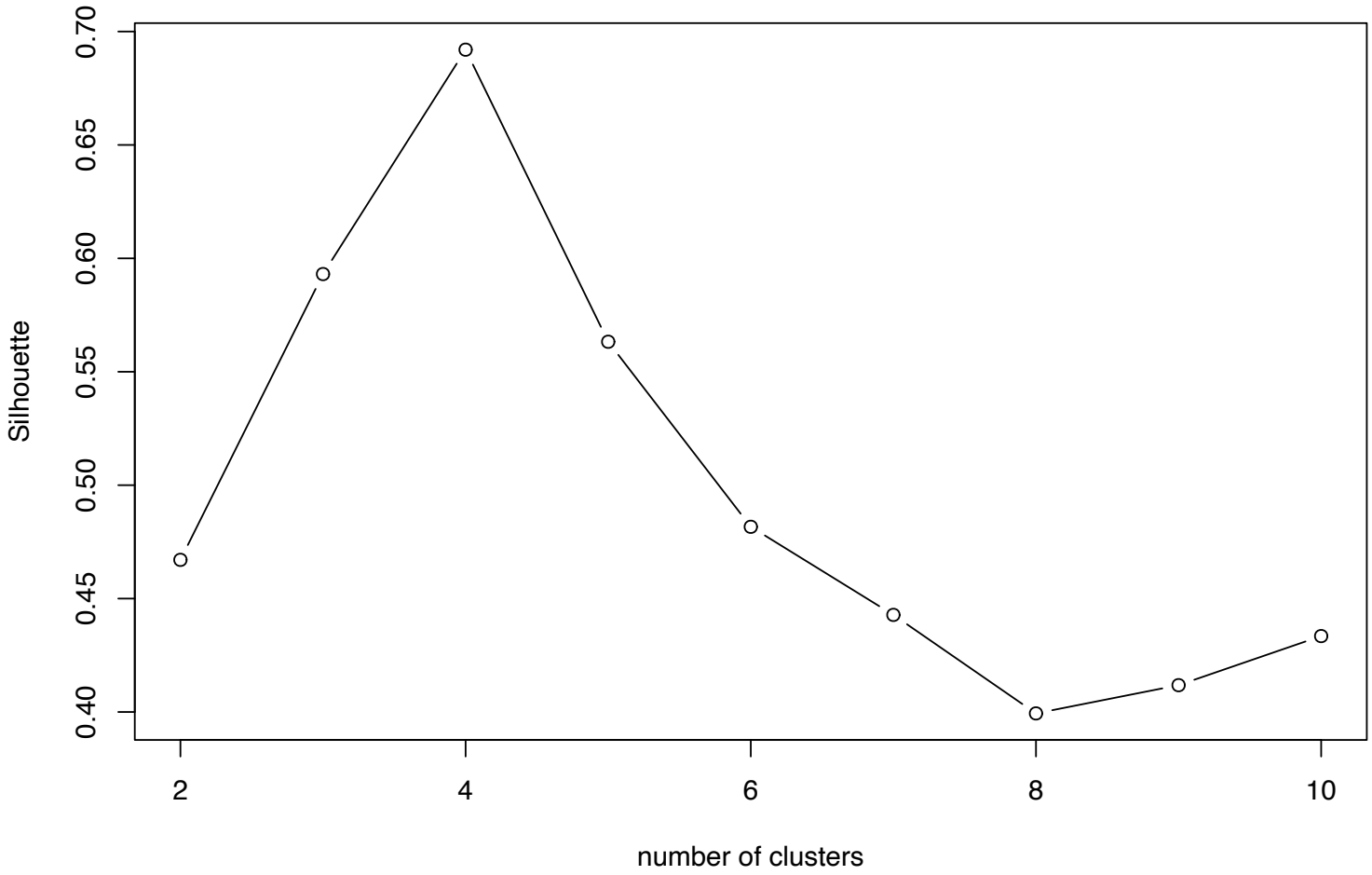
$$\text{Dunn} = \frac{\delta_{\min}}{\Delta_{\max}}$$

- The higher Dunn index, the better.

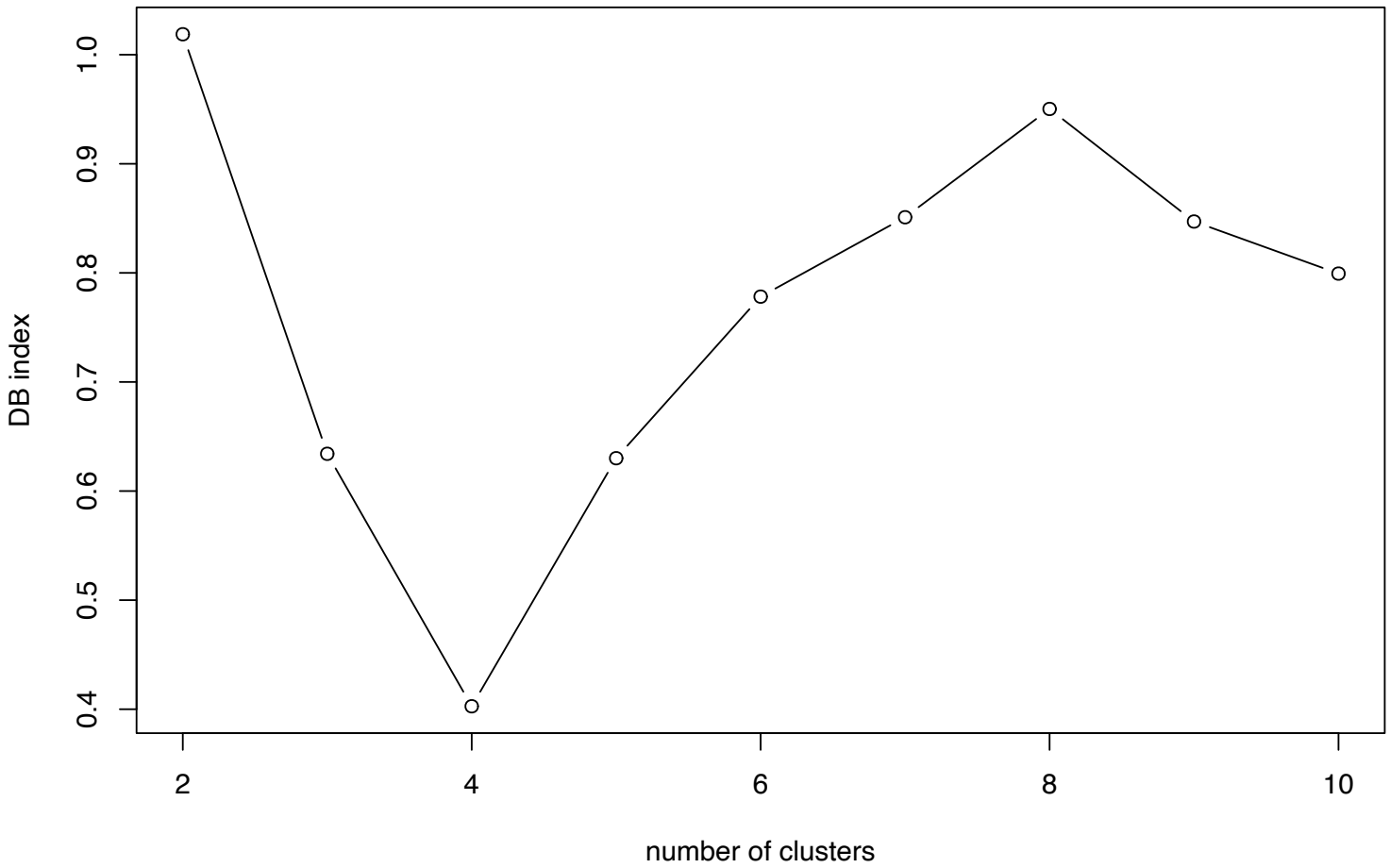
Davis-Bouldin index in R

```
library(clusterCrit)

C<- 2:10
N<-length(C)
DB<-rep(0,N)
Du<-rep(0,N)
Si<-rep(0,N)
for(i in 1:N){
km<-kmeans(x,centers=C[i],nstart=10)
DB[i]<-intCriteria(as.matrix(x), km$cluster, crit="Davies_Bouldin")
Du[i]<-intCriteria(as.matrix(x), km$cluster, crit="Dunn")
Si[i]<-intCriteria(as.matrix(x), km$cluster, crit="Silhouette")
}
plot(C,DB,type="b",xlab="number of clusters",ylab="DB index")
plot(C,Du,type="b",xlab="number of clusters",ylab="Dunn index")
plot(C,Si,type="b",xlab="number of clusters",ylab="Silhouette")
```

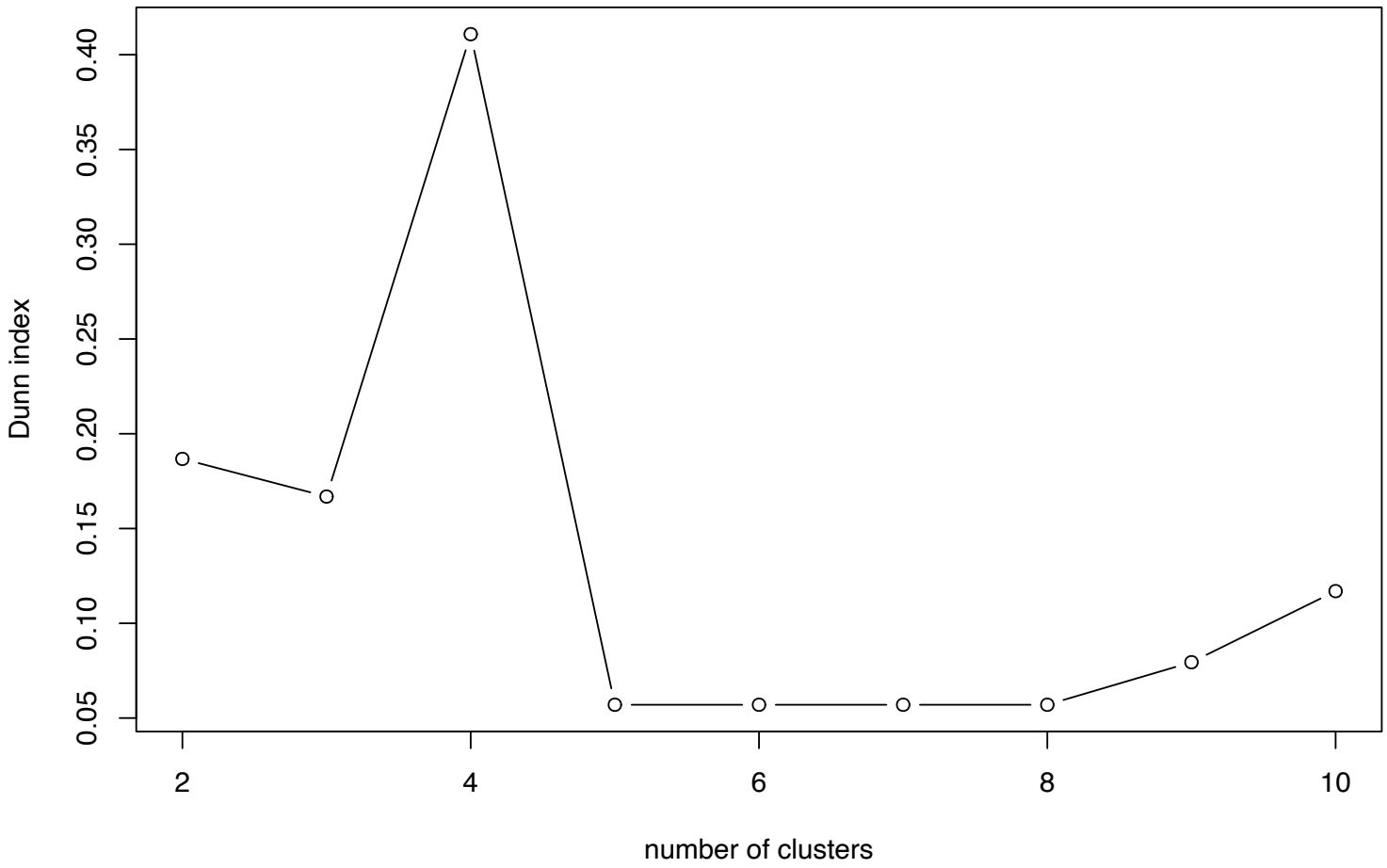
Sihouette index for the 4-cluster synthetic data



DB index for the 4-cluster synthetic data



Dunn index for the 4-cluster synthetic data

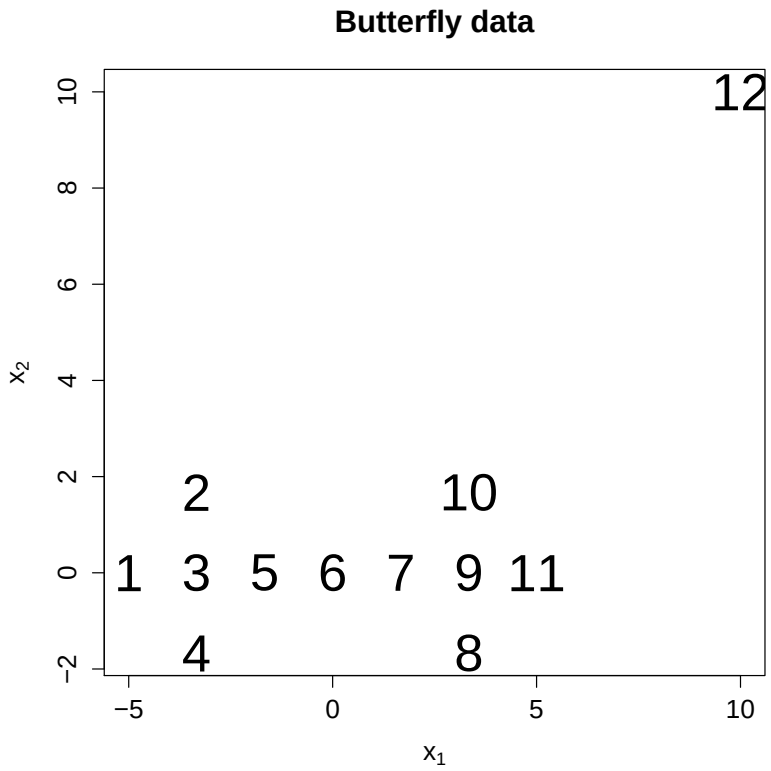


Overview

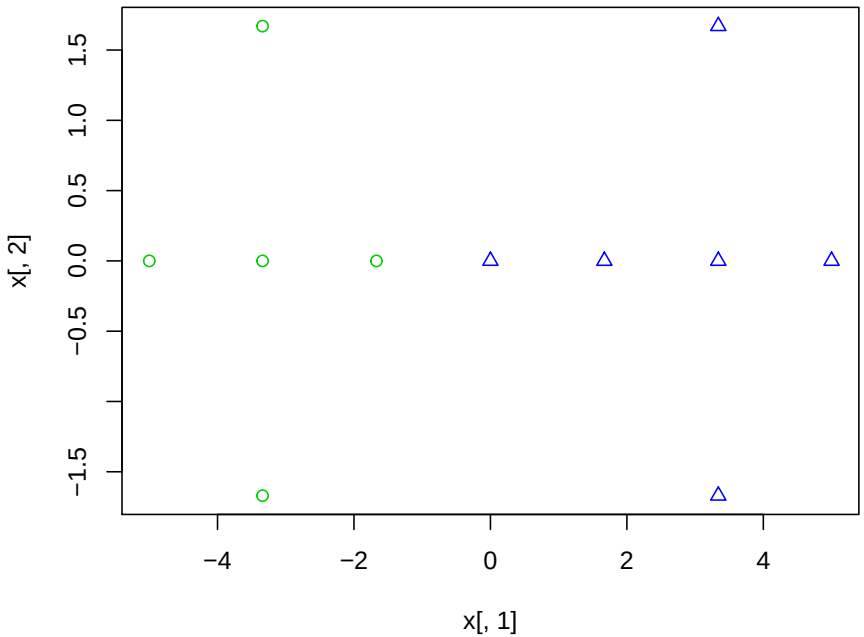
Overview

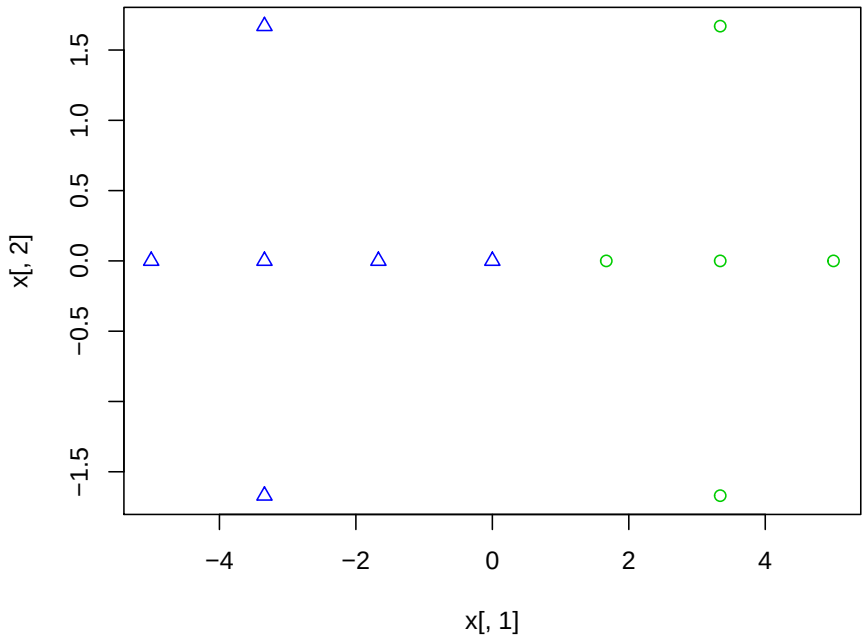
Limitation of partitional clustering

- In partitional clustering, an observation is assigned unambiguously to one and only one cluster.
- This may be arbitrary when the observation lies at the boundary between two or more clusters.
- Example ("butterfly" data):



Butterfly data: 2 solutions with HCM





Fuzzy partition

- A fuzzy partition is described by an $n \times c$ matrix $U = (u_{ik})$, where $u_{ik} \in [0, 1]$ is the degree of membership of observation i to cluster k.
- $u_{ik} = 1$ means full membership, $u_{ik} = 0$ means no membership at all, and $0 < u_{ik} < 1$ means partial membership.
- We still impose the n equality constraints

$$\sum_{k=1}^{c} u_{ik} = 1, \quad i = 1, \ldots, n.$$

- Each cluster becomes a fuzzy set of observations.
- How to generate a fuzzy partition?

Overview

Fuzzy c-means (FCM)

- We consider the following optimization problem:

$$\text{minimize } J_{\text{FCM}}(U, V) = \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik}^{\beta} d_{ik}^2$$

with $d_{ik} = \|x_i - v_k\|$, subject to the constraints

$$\sum_{k=1}^{c} u_{ik} = 1, \quad i = 1, \ldots, n$$

$$u_{ik} \geq 0, \quad i = 1, \ldots, n \text{ and } k = 1, \ldots, c$$

- With $\beta = 1$, the solution is the same as that of HCM.
- To obtain a fuzzy partition, we need to set $\beta > 1$ (default: $\beta = 2$).

Solution of the optimization problem

As with HCM, we start with randomly selected prototypes $v_1, \ldots, v_c$ and we use a grouped coordinate descent strategy by alternating 2 steps

1. Minimize $J_{\text{FCM}}(U, V)$ with respect to U for fixed V
2. Minimize $J_{\text{FCM}}(U, V)$ with respect to V for fixed U

until some stopping criterion is met, for instance

$$\max |U^{(t+1)} - U^{(t)}| < \epsilon$$

or

$$\max |V^{(t+1)} - V^{(t)}| < \epsilon$$

Minimization of $J_{\text{FCM}}(U, V)$ w.r.t. U for fixed V

- We can write the cost function as

$$J_{\text{FCM}}(U, V) = \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{c} u_{ik}^{\beta} d_{ik}^2}_{J_i(u_{i\cdot})} = \sum_{i=1}^{n} J_i(u_{i\cdot}),$$

  with $u_{i\cdot} = (u_{i1}, \ldots, u_{ic})$.
- We can minimize each function J_i independently, subject to $\sum_k u_{ik} = 1$ (and $u_{ik} \geq 0$ but we can ignore these positivity constraints).
- To solve these constrained optimization problems, we use the method of Lagrange multipliers.

Lagrange multipliers

- The method of Lagrange multipliers is a general method for solving constrained optimization problems of the form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) = 0 \end{array} \tag{1}$$

- We consider the Lagrange function

$$\mathcal{L}(x, \lambda) = f(x) - \lambda g(x)$$

and we solve the equations

$$\frac{\partial \mathcal{L}}{\partial x}(x, \lambda) = 0, \quad g(x) = 0 \tag{2}$$

- Under some technical conditions, the solution of (2) gives us the solution of the optimization problem (1).

Minimization of $J_{\text{FCM}}(U, V)$ w.r.t. U for fixed V (cont.)

- We minimize

$$J_i = \sum_{k=1}^{c} u_{ik}^{\beta} d_{ik}^2 \quad \text{subject to} \quad \sum_{k} u_{ik} = 1$$

- The Lagrange function is

$$\mathcal{L}(u_{i\cdot}, \lambda) = \sum_{k=1}^{c} u_{ik}^{\beta} d_{ik}^2 - \lambda \left(\sum_{k=1}^{c} u_{ik} - 1 \right)$$

Minimization of $J_{\text{FCM}}(U, V)$ w.r.t. U for fixed V (cont.)

- The solution must verify

$$\frac{\partial \mathcal{L}}{\partial u_{ik}} = \beta u_{ik}^{\beta-1} d_{ik}^2 - \lambda = 0, \quad k = 1, \ldots, c$$

$$\sum_{k=1}^{c} u_{ik} = 1$$

- After some manipulation we get

$$\boxed{u_{ik} = \frac{d_{ik}^{-2/(\beta-1)}}{\sum_{\ell=1}^{c} d_{i\ell}^{-2/(\beta-1)}}, \quad k = 1, \ldots, c}$$

Proof

Minimization of $J_{\text{FCM}}(U, V)$ w.r.t. V for fixed U

- We write the cost function as

$$J_{\text{FCM}}(U, V) = \sum_{k=1}^{c} \underbrace{\sum_{i=1}^{n} u_{ik}^{\beta} (x_i - v_k)^T (x_i - v_k)}_{\mathcal{I}(v_k)} = \sum_{k=1}^{c} \mathcal{I}(v_k)$$

- We can minimize each $\mathcal{I}(v_k)$ w.r.t. v_k independently. Solving

$$\frac{\partial \mathcal{I}(v_k)}{\partial v_k} = 0$$

we get

$$\boxed{v_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta} x_i}{\sum_{i=1}^{n} u_{ik}^{\beta}}, \quad k = 1, \ldots, c}$$

Proof

FCM algorithm

1. Initialize prototypes $V = (v_1, \ldots, v_c)$ randomly.
2. Update U for fixed V:

$$u_{ik} = \frac{d_{ik}^{-2/(\beta-1)}}{\sum_{\ell=1}^{c} d_{i\ell}^{-2/(\beta-1)}} \quad \text{for all } i, k$$

3. Update V for fixed U:

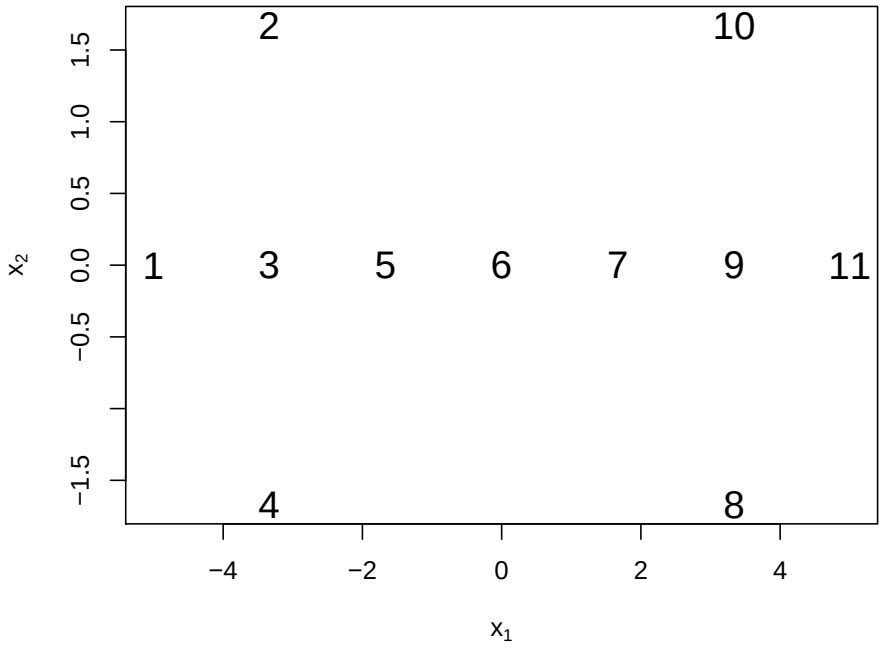
$$v_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta} x_i}{\sum_{i=1}^{n} u_{ik}^{\beta}} \quad \text{for all } k$$

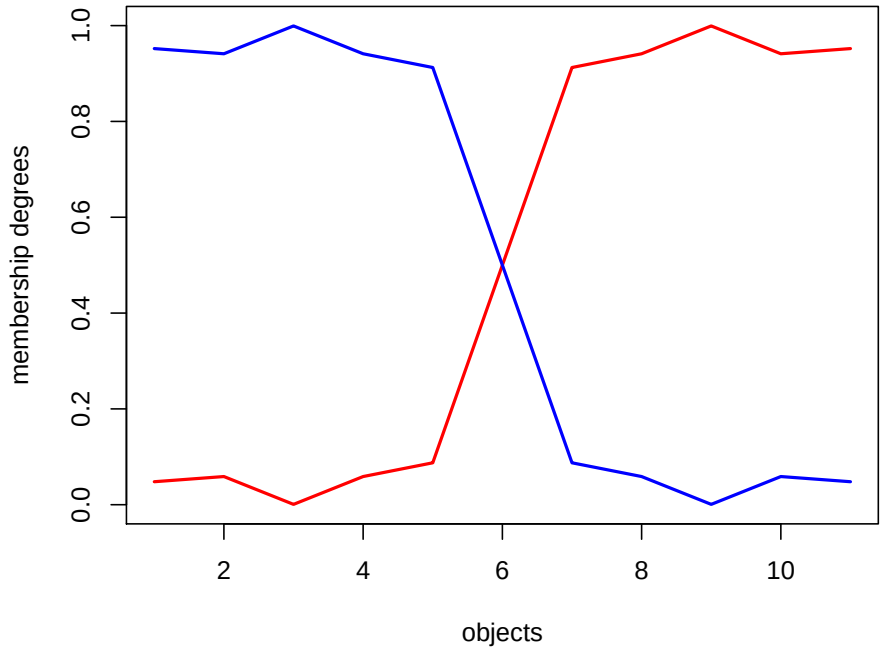
4. Return to Step 2 while the change in V or U is greater than some threshold.

Example in R: butterfly data

```
library(fclust)

fm<-FKM(x,2,RS=5)

plot(1:11,fm$U[,1],type="l",ylim=c(0,1),xlab="objects",
ylab="membership degrees",col="red",lwd=2)

lines(1:11,fm$U[,2],lty=1,col="blue",lwd=2)
```

Result


2
10
1
3
5
6
7
9
11
4
8
−4
−2
0
2
4
−1.5
−0.5
0.0
0.5
1.0
1.5
x1
x2


membership degrees
objects
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10

Fuzzy silhouette index

- Cluster validity indices have been defined for fuzzy clustering as well.
- One such index is the fuzzy silhouette index, defined as

$$\overline{s}_f = \frac{\sum_{i=1}^{n} \left(u_{ik(i)} - u_{ik'(i)}\right)^{\alpha} s_i}{\sum_{i=1}^{n} \left(u_{ik(i)} - u_{ik'(i)}\right)^{\alpha}}$$

  where $k(i)$ and $k'(i)$ are the fuzzy clusters to which object i has, respectively, the highest and second highest membership degrees, and $\alpha \geq 0$ is a weighting coefficient (by default, $\alpha = 1$).
- An object in the near vicinity of a cluster prototype is given more importance than another object located in an overlapping area (where the membership degrees of the objects to two or more fuzzy clusters are similar).

Derivation of the HCM algorithm

- We have

$$\mathcal{I}(v_k) = \sum_{i=1}^{n} u_{ik}(x_i - v_k)^T(x_i - v_k)$$
$$= -2\left(\sum_{i=1}^{n} u_{ik}x_i\right)^T v_k + v_k^T v_k \underbrace{\sum_{i=1}^{n} u_{ik}}_{n_k} + C$$

- Consequently,

$$\frac{\partial \mathcal{I}(v_k)}{\partial v_k} = -2\sum_{i=1}^{n} u_{ik}x_i + 2n_k v_k = 0 \Leftrightarrow v_k = \frac{1}{n_k}\sum_{i=1}^{n} u_{ik}x_i$$

Back

Derivation of the FCM algorithm (1/2)

- From $\beta u_{ik}^{\beta-1} d_{ik}^2 - \lambda = 0$, we get

$$u_{ik} = \left(\frac{\lambda}{\beta d_{ik}^2} \right)^{1/(\beta-1)} = \left(\frac{\lambda}{\beta} \right)^{1/(\beta-1)} d_{ik}^{-2/(\beta-1)}$$

- From $\sum_\ell u_{i\ell} = 1$,

$$\left(\frac{\lambda}{\beta} \right)^{1/(\beta-1)} \sum_\ell d_{i\ell}^{-2/(\beta-1)} = 1 \Rightarrow \left(\frac{\lambda}{\beta} \right)^{1/(\beta-1)} = \frac{1}{\sum_\ell d_{i\ell}^{-2/(\beta-1)}}$$

- Finally,

$$u_{ik} = \frac{d_{ik}^{-2/(\beta-1)}}{\sum_\ell d_{i\ell}^{-2/(\beta-1)}}$$

Back

Derivation of the FCM algorithm (2/2)

- We have

$$\mathcal{I}(v_k) = \sum_{i=1}^{n} u_{ik}^{\beta} (x_i - v_k)^T (x_i - v_k)$$

$$= -2 \left(\sum_i u_{ik}^{\beta} x_i \right)^T v_k + \left(\sum_i u_{ik}^{\beta} \right) v_k^T v_k + C$$

- Consequently,

$$\frac{\partial \mathcal{I}(v_k)}{\partial v_k} = -2 \left(\sum_i u_{ik}^{\beta} x_i \right) + 2 \left(\sum_i u_{ik}^{\beta} \right) v_k$$

and

$$\frac{\partial \mathcal{I}(v_k)}{\partial v_k} = 0 \Leftrightarrow v_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta} x_i}{\sum_{i=1}^{n} u_{ik}^{\beta}}$$

Back