Clustering algorithms

Konstantinos Koutroumbas

Unit 13

– Cluster validity
– Clustering tendency

koutroum@noa.gr

Cluster validity

- Consider a data set X. Then
  (a) application of an inappropriate algorithm **or**
  (b) application of an appropriate algorithm with inappropriate values of its (hyper)parameters
  may lead to poor results. Hence the need for further evaluation of clustering results is apparent.

- Cluster validity: a task that **evaluates** quantitatively the results of a clustering algorithm.

- A clustering structure $\boldsymbol{C}$, resulting from an algorithm may be either
  - A hierarchy of clusterings or
  - A single clustering.

Cluster validity

Cluster validity may be approached through three possible directions:

- $\boldsymbol{C}$ is **evaluated** in terms of an independently drawn structure, imposed on X ***a priori***. The criteria used in this case are called external criteria.
- $\boldsymbol{C}$ is **evaluated** in terms of quantities that involve the vectors of X themselves (e.g., proximity matrix). The criteria used in this case are called internal criteria.
- $\boldsymbol{C}$ is **evaluated** by comparing it with other clustering structures, resulting from the application of the same clustering algorithm but with different parameter values, or other clustering algorithms, on X. Criteria of this kind are called relative criteria.

Cluster validity

➢ *Cluster validity for the cases of external and internal criteria*

- Hypothesis testing is employed.
- The null hypothesis H_0, which is a statement of randomness concerning the structure of X, is defined.
- The **generation** of a reference data population of size r under the random hypothesis takes place.
- An appropriate statistic, q, whose values are indicative of the structure of a data set, is defined. The value of q that results for our data set X, q^*, is **compared** against the r values of q, $q_1, \ldots, q_r$, associated with the r members of the reference (random) population.
- If q^* is (a) greater than $(1-\rho)\cdot r$, (b) less than $\rho \cdot r$, (c) less than $\frac{\rho}{2}\cdot r$ OR greater than $\left(1-\frac{\rho}{2}\right)\cdot r$, the **null hypothesis** is **rejected**(*).

Ways for generating reference populations under the null hypothesis (each one used in different situations) (**):

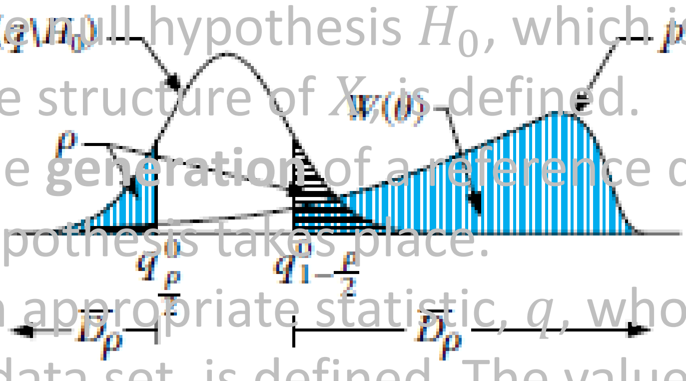
- Random position hypothesis.
- Random graph hypothesis.
- Random label hypothesis.

(*)Actually, we approximate the $p(q|H_0)$, via Monte Carlo simulations.
(**)The three cases are related to the kind of the adopted statistic q (see next slide).

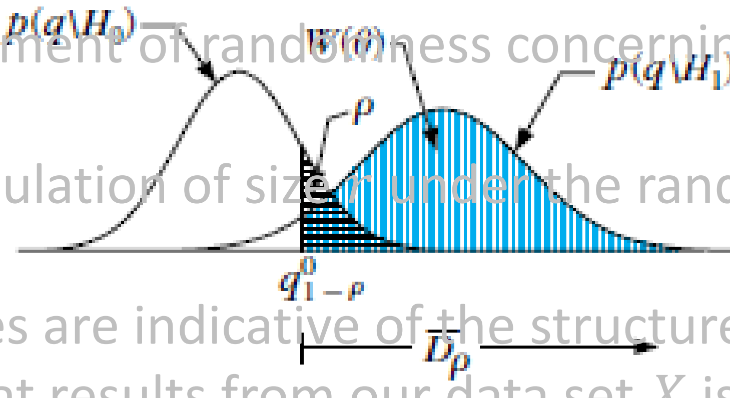
Cluster validity

➢ <u>*Cluster validity for the cases of external and internal criteria*</u>

- Hypothesis testing is employed.
- The null hypothesis H_0, which is a statement of randomness concerning the structure of X, is defined.
- The **generation** of a reference data population of size r under the random hypothesis takes place.
- An appropriate statistic, q, whose values are indicative of the structure of a data set, is defined. The value of q that results from our data set X is **compared** against the r values of q associated with the members of the reference (random) population.



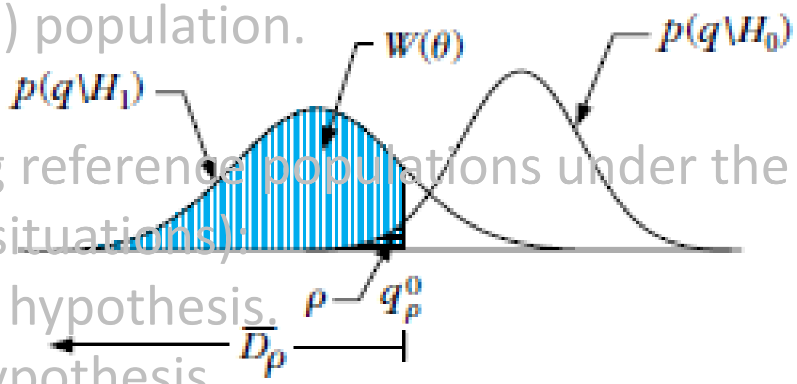
(a)



(b)

Ways for generating reference populations under the null hypothesis (each one used in different situations):

- Random position hypothesis.
- Random graph hypothesis.
- Random label hypothesis.



(c)

Cluster validity

- *Cluster validity for the cases of external and internal criteria*

• Random position hypothesis.

It **requires** that "*all the arrangements of the N vectors in a specific region of the l-dimensional data space are equally likely to occur*".

It can be used with respect to both external and internal criteria.

Cluster validity

➢ *Cluster validity for the cases of external and internal criteria*

Statistics suitable for external criteria

- For the comparison of $\boldsymbol{C}$ with an independently drawn partition $\boldsymbol{P}$ of X
  - Rand statistic
  - Jaccard statistic
  - Fowlkes-Mallows index
  - Hubert's Γ statistic
  - Normalized Γ statistic
- For assessing the agreement between $\boldsymbol{P}$ and the proximity matrix P.
  - Γ statistic.

Statistics suitable for internal criteria

- Validation of hierarchy of clusterings
  - Cophenetic correlation coefficient ($CPCC$)
  - γ statistic
  - Kudall's τ statistic.
- Validation of individual clusterings
  - Γ statistic
  - Normalized Γ statistic

Cluster validity

➢ ***Cluster validity for the cases of external and internal criteria***

Statistics suitable for external criteria

- For the comparison of $\boldsymbol{C}$ with an independently drawn partition $\boldsymbol{P}$ of X
  - Rand statistic

Let $\boldsymbol{P}$ be an external partition of X into groups and $\boldsymbol{C}$ a clustering

A pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is denoted as

SS if $\boldsymbol{x}_i, \boldsymbol{x}_j$ belong to the same cluster in $\boldsymbol{C}$ and to the same group in $\boldsymbol{P}$.

SD if $\boldsymbol{x}_i, \boldsymbol{x}_j$ belong to the same cluster in $\boldsymbol{C}$ and to different groups in $\boldsymbol{P}$.

DS if $\boldsymbol{x}_i, \boldsymbol{x}_j$ belong to different clusters in $\boldsymbol{C}$ and to the same group in $\boldsymbol{P}$.

DD if $\boldsymbol{x}_i, \boldsymbol{x}_j$ belong to different clusters in $\boldsymbol{C}$ and to different groups in $\boldsymbol{P}$.

Let a= number of SS, b= number of SD, c=number of DS, d= number of DD

M=total number of pairs of points ($= a + b + c + d$)

Rand statistic $R = (a + d)/M$

The greater the value of R the greater the degree of agreement between $\boldsymbol{P}$ and $\boldsymbol{C}$.

Cluster validity

➢ ***Cluster validity for the cases of external and internal criteria***

Statistics suitable for external criteria

Example: Consider a data set $X = \{\boldsymbol{x}_i \in H_l \equiv [0,1]^l, i = 1, \dots, 100\}$ so that the first 25 ($\boldsymbol{x}_1 - \boldsymbol{x}_{25}$) stem from $N(\boldsymbol{\mu}_1, 0.2 \cdot I)$, the next 25 ($\boldsymbol{x}_{26} - \boldsymbol{x}_{50}$) from $N(\boldsymbol{\mu}_2, 0.2 \cdot I)$, the next 25 ($\boldsymbol{x}_{51} - \boldsymbol{x}_{75}$) from $N(\boldsymbol{\mu}_3, 0.2 \cdot I)$ and the final 25 ($\boldsymbol{x}_{76} - \boldsymbol{x}_{100}$) from $N(\boldsymbol{\mu}_4, 0.2 \cdot I)$, where $\boldsymbol{\mu}_1 = [0.2,0.2,0.2]^T$, $\boldsymbol{\mu}_2 = [0.5,0.2,0.8]^T$, $\boldsymbol{\mu}_3 = [0.5,0.8,0.2]^T$, $\boldsymbol{\mu}_4 = [0.8,0.8,0.8]^T$ and I is the 3×3 identity matrix.

External information: The points form the following four different groups $P_1 = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_{25}\}$, $P_2 = \{\boldsymbol{x}_{26}, \dots, \boldsymbol{x}_{50}\}$, $P_3 = \{\boldsymbol{x}_{51}, \dots, \boldsymbol{x}_{75}\}$, $P_4 = \{\boldsymbol{x}_{76}, \dots, \boldsymbol{x}_{100}\}$. Thus, we have the partition $\boldsymbol{P} = \{P_1, P_2, P_3, P_4\}$.

We run the k-means algorithm for $m = 4$ and let $\boldsymbol{C} = \{C_1, C_2, C_3, C_4\}$ be the resulting clustering.

Question: Are $\boldsymbol{C}$ and $\boldsymbol{P}$ in **good agreement** with each other?

Cluster validity

➢ *Cluster validity for the cases of external and internal criteria*

Statistics suitable for external criteria

Example (cont.):

- **Compute** the $Rand(\boldsymbol{C}, \boldsymbol{P})$ $(= 0.91)$.
- **For** $i = 1$ to r $(= 100)$
  - ➢ **Generate** a data set X^i of 100 vectors in H_3, so the vectors are uniformly distributed in it.
  - ➢ **Assign** each vector $\boldsymbol{y}_j{}^i \in X^i$ to the group where the respective $\boldsymbol{x}_j \in X$ belongs according to P.
  - ➢ **Run** the k-means algorithm for X^i and let $\boldsymbol{C}^i$ be the resulting clustering
  - ➢ **Compute** $Rand(C^i, P)$
- **End for**
- Set the significance level ρ to 0.05.

It turns out that $Rand(\boldsymbol{C}, \boldsymbol{P})$ is greater than $(1-\rho)\cdot r = 95$ values $Rand\left(\boldsymbol{C}^i, \boldsymbol{P}\right), i = 1, \ldots, r$ (actually, it is greater than all 100 values).

Thus, the null hypothesis that $\boldsymbol{C}$ is in **agreement** with $\boldsymbol{P}$ by chance is **rejected** at significance level 0.05.

Exercise: What would be the case if the clusters variances where $0.8 \cdot I$?

Cluster validity

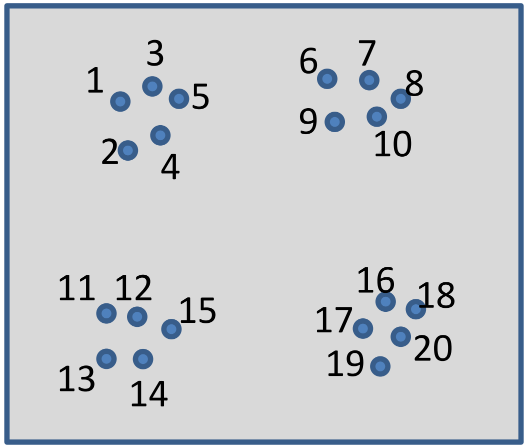
➢ *Cluster validity for the cases of external and internal criteria*

Statistics suitable for external criteria

Example (cont.): External information:

$$\boldsymbol{P} = \{P_1, P_2, P_3, P_4\} = \{\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_5\}, \{\boldsymbol{x}_6, \dots, \boldsymbol{x}_{10}\}, \{\boldsymbol{x}_{11}, \dots, \boldsymbol{x}_{15}\}, \{\boldsymbol{x}_{16}, \dots, \boldsymbol{x}_{20}\}\}$$

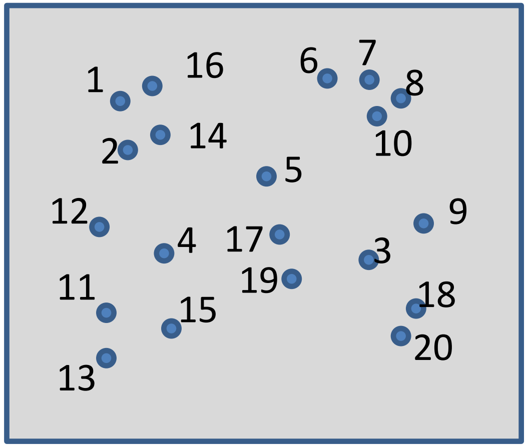
Data set under study



Clustering result

$$\boldsymbol{C} = \{C_1, C_2, C_3, C_4\}$$
$$= \left\{ \begin{matrix} \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_5\}, \{\boldsymbol{x}_6, \dots, \boldsymbol{x}_{10}\}, \\ \{\boldsymbol{x}_{11}, \dots, \boldsymbol{x}_{15}\}, \{\boldsymbol{x}_{16}, \dots, \boldsymbol{x}_{20}\} \end{matrix} \right\}$$

Randomly generated data set



Clustering result

$$\boldsymbol{C} = \{C_1, C_2, C_3, C_4\}$$
$$= \left\{ \begin{matrix} \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_{14}, \boldsymbol{x}_{16}\}, \\ \{\boldsymbol{x}_{12}, \boldsymbol{x}_4, \boldsymbol{x}_{11}, \boldsymbol{x}_{13}, \boldsymbol{x}_{15}, \boldsymbol{x}_{17}, \boldsymbol{x}_{19}\}, \\ \{\boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_{10}, \boldsymbol{x}_5\}, \{\boldsymbol{x}_3, \boldsymbol{x}_9, \boldsymbol{x}_{18}, \boldsymbol{x}_{20}\} \end{matrix} \right\}$$

Cluster validity

➢ *Cluster validity for the cases of external and internal criteria*

Statistics suitable for internal criteria

- Validation of individual clusterings
  - Γ statistic

Consider two $N \times N$ matrices $X = [x_{ij}]$ and $Y = [y_{ij}]$, drawn independently from each other. Then

$$\Gamma(X,Y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}y_{ij}$$

or, for symmetric matrices,

$$\Gamma(X,Y) = \frac{1}{M}\sum_{i=1}^{N}\sum_{j=i+1}^{N} x_{ij}y_{ij}$$

$$M = \frac{N(N-1)}{2}$$

Cluster validity

➢ *Cluster validity for the cases of external and internal criteria*

Statistics suitable for internal criteria

• Validation of individual clusterings

Example: Consider a data set $X = \{x_i \in H_l \equiv [0,1]^l, i = 1, \dots, 100\}$ so that the first 25 ($\boldsymbol{x}_1 - \boldsymbol{x}_{25}$) stem from $N(\boldsymbol{\mu}_1, 0.1 \cdot I)$, the next 25 ($\boldsymbol{x}_{26} - \boldsymbol{x}_{50}$) from $N(\boldsymbol{\mu}_2, 0.1 \cdot I)$, the next 25 ($\boldsymbol{x}_{51} - \boldsymbol{x}_{75}$) from $N(\boldsymbol{\mu}_3, 0.1 \cdot I)$ and the final 25 ($\boldsymbol{x}_{76} - \boldsymbol{x}_{100}$) from $N(\boldsymbol{\mu}_4, 0.1 \cdot I)$, where $\boldsymbol{\mu}_1 = [0.2, 0.2]^T$, $\boldsymbol{\mu}_2 = [0.8, 0.2]^T$, $\boldsymbol{\mu}_3 = [0.2, 0.8]^T$, $\boldsymbol{\mu}_4 = [0.8, 0.8]^T$ and I is the 2×2 identity matrix.

Run the k-means algorithm and let $C = \{C_1, C_2, C_3, C_4\}$ be the resulting clustering.

Question: Does the clustering **agrees** with the “internal structure“ of the data by chance (H_0 hypothesis) or not (alternative hypothesis)?

Let the internal structure of X be **reflected** in the dissimilarity matrix $P_{N \times N}$, based on the squared Euclidean distance.

Cluster validity

➢ *Cluster validity for the cases of external and internal criteria*

Statistics suitable for internal criteria

• Validation of individual clusterings

Example (cont.):

Define the matrix $Y_{N\times N} = [y_{ij}]$ as follows

$$y_{ij} = \begin{cases} 1, & if\ \boldsymbol{x}_i\ and\ \boldsymbol{x}_j\ belong\ to\ different\ clusters \\ 0, & otherwise \end{cases}$$

Compute $\Gamma(Y, P)(= 0.57)$.

- **For** $i = 1$ to r $(= 100)$
  - ➢ **Generate** a data set X^i of 100 vectors uniformly distributed in H_2.
  - ➢ **Compute** the associated P^i dissimilarity matrix.
  - ➢ **Run** the k-means algorithm for X^i and let $\boldsymbol{C}^i$ be the resulting clustering
  - ➢ **Form** the $N \times N$ matrix Y^i as above and **compute** $\Gamma(Y^i, P^i)$
- **End for**
- Set the significance level ρ to 0.05.

It turns out that $\Gamma(Y, P)$ is greater than $(1-\rho)\cdot r = 95$ values $\Gamma(Y^i, P^i), i = 1, \ldots, r$ (actually, it is greater than 99 values).

Thus, the null hypothesis that $\boldsymbol{C}$ is in **agreement** with $\boldsymbol{P}$ by chance is **rejected** at significance level 0.05.

Cluster validity

➢ *Cluster validity for the cases of external and internal criteria*

Statistics suitable for internal criteria

- Validation of individual clusterings

Exercise 1: What would be the case if the clusters variances where $0.2 \cdot I$?

Exercise 2: What would change in the above procedure of y_{ij}'s were defined as

$$y_{ij} = \begin{cases} 1, & if\ \boldsymbol{x}_i\ and\ \boldsymbol{x}_j\ belong\ to\ the\ same\ cluster \\ 0, & otherwise \end{cases}$$

Cluster validity

- ***Cluster validity for the cases of relative criteria***

Let $\boldsymbol{A}$ denote the set of parameters of a clustering algorithm.

Statement of the problem

- *"Among the clusterings produced by a specific clustering algorithm, for different values of the parameters in $\boldsymbol{A}$, choose the one that best fits the data set X".*

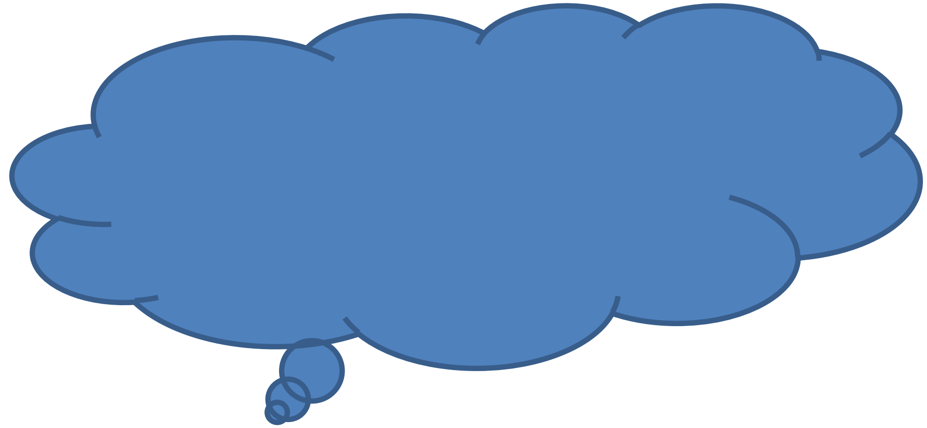
We consider two cases

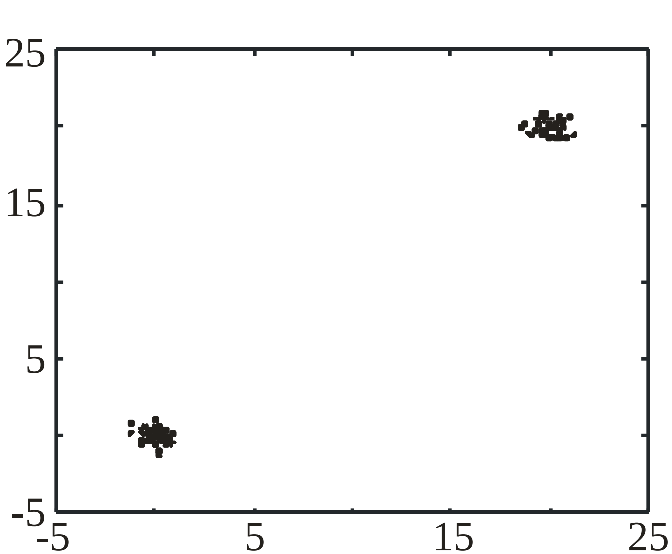
(a) $\boldsymbol{A}$ does not contain the number of clusters m.

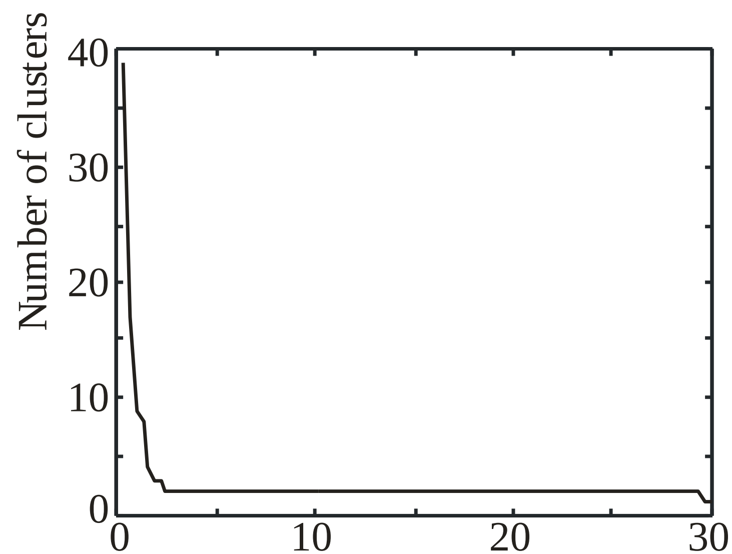
The estimation of the best set of parameter values is carried out as follows:

- **Run** the algorithm for a wide range of values of its parameters.
- **Plot** the number of clusters, m, **versus** the parameters of $\boldsymbol{A}$.
- **Choose** the widest range for which m remains constant.
- **Adopt** the clustering that corresponds to the values of the parameters in $\boldsymbol{A}$ that lie in the **middle** of this range.

Example:







Cluster validity

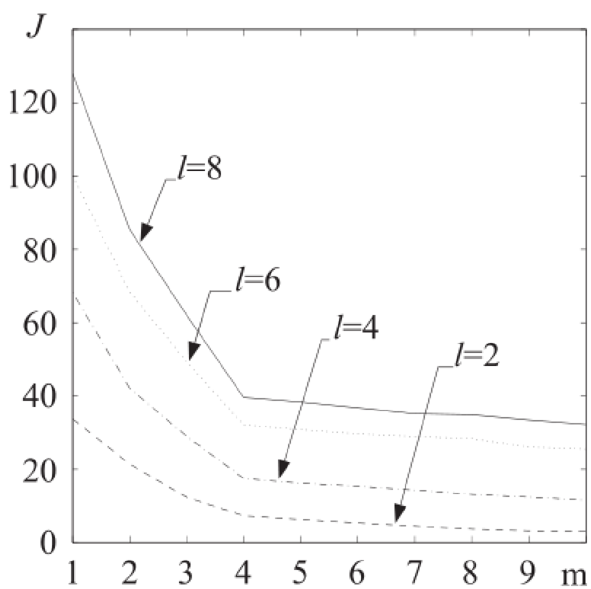
➢ *Cluster validity for the cases of relative criteria*

(b) $\boldsymbol{A}$ *does contain the number of clusters m.*

The estimation of the best set of parameter values is carried out as follows:

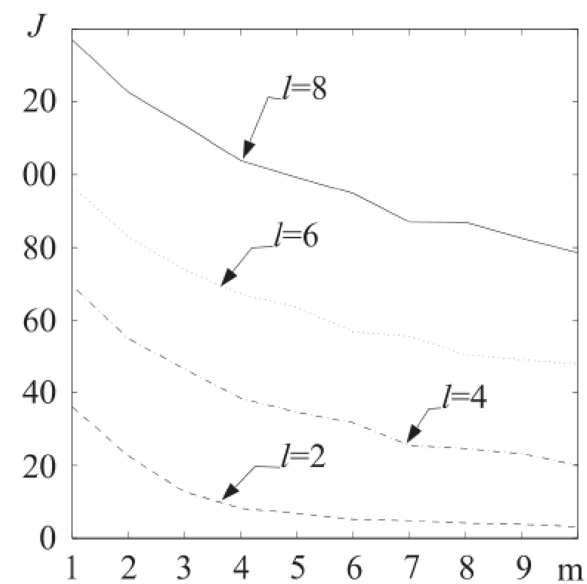
- **Select** a suitable performance index q (the best clustering is identified in terms of q).
- **For** $m = m_{min}$ to m_{max}
  - **Run** the algorithm r times using different sets of values for the other parameters of $\boldsymbol{A}$ and each time compute q.
  - **Choose** the clustering that corresponds to the best q.
- End for
- **Plot** the best values of q for each m versus m.
- The **presence** of a significant knee **indicates** the number of clusters underlying X. Adopt the clustering that corresponds to that knee.
- The absence of such a knee **indicates** that X possesses **no** clear clustering structure.

Example:



Clustered data

Non-clustered data



Cluster validity

- ***Cluster validity for the cases of relative criteria***
- *Statistics suitable for relative criteria*
  - Hard clustering
    - Modified Hubert Γ statistic
    - Dunn and Dunn-like indices

$$D_m = min_{i=1,\dots,m}\left\{min_{j=i+1,\dots,m}\left(\frac{d(C_i,C_j)}{max_{k=1,\dots,m}diam(C_k)}\right)\right\}$$

$$d(C_i,C_j) = min_{\boldsymbol{x}\in C_i,\boldsymbol{y}\in C_j} d(\boldsymbol{x},\boldsymbol{y})$$

$$diam(C) = max_{\boldsymbol{x},\boldsymbol{y}\in C} d(\boldsymbol{x},\boldsymbol{y})$$

    - Davies-Bouldin (DB) and DB-like indices
    - The silhouette index
  - Fuzzy clustering
    - Indices for clusters with point representatives
      - Partition coefficient (PC)
      - Partition entropy coefficient (PE)
      - Xie-Beni (XB) index
      - Fukuyama-Sugeno index
      - Total fuzzy hypervolume
      - Average partition density
      - Partition density

Cluster validity

- ***Cluster validity for the cases of relative criteria***
- *Statistics suitable for relative criteria*
  - Fuzzy clustering (cont.)
    - Indices for shell-shaped clusters
      - o Fuzzy shell density
      - o Average partition shell density
      - o Shell partition density
      - o Total fuzzy average shell thickness

Cluster validity

➢ *Cluster validity for the cases of relative criteria*

➢ *Statistics suitable for relative criteria*

- Hard clustering - The silhouette index

C_{c_i}: The cluster where $\boldsymbol{x}_i$ belongs.

a_i: The **average distance** of $\boldsymbol{x}_i$ from all $\boldsymbol{x}_j \in C_{c_i}$.

b_i: The **average distance** of $\boldsymbol{x}_i$ from its closest cluster C_q.

$s_i = \frac{b_i - a_i}{\max(b_i, a_i)}$: **Silhouette width** of $\boldsymbol{x}_i$ ($s_i \in [-1,1]$).

Values of s_i **close** to

$+1$ indicate that $\boldsymbol{x}_i$ is **well clustered**,

0 indicate that $\boldsymbol{x}_i$ is at the **border** of two clusters

-1 indicate that $\boldsymbol{x}_i$ is **poorly clustered**.

Silhouette index of a **cluster**: $S_j = \frac{1}{n_j} \sum_{i: \boldsymbol{x}_i \in C_j} s_i, j = 1, \ldots, m$ ($S_j \in [-1,1]$)

Global silhouette index: $\mathcal{S}_m = \frac{1}{m} \sum_{j=1}^{m} S_j$ ($\mathcal{S}_m \in [-1,1]$)

Note: The higher the value of $\mathcal{S}_m$ the better the clustering.

Usage: Plot $\mathcal{S}_m$ versus m. The **position** of the **maximum** indicates the **true number** of clusters.

Clustering tendency

Facts

- Most clustering algorithms impose a clustering structure to the data set X at hand.

- However, X may not possess a clustering structure.

- Before we apply any clustering algorithm on X, it must first be verified that X possesses a clustering structure. This is known as the clustering tendency procedure.

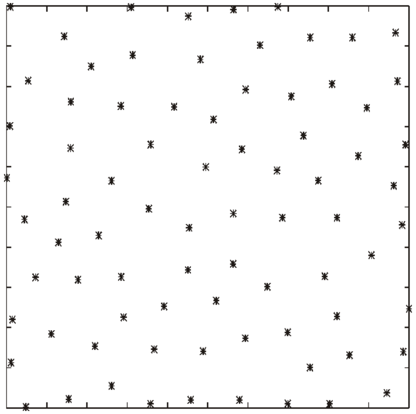
- Clustering tendency is heavily based on hypothesis testing. Specifically, it is based on testing the randomness (null) hypothesis (H_0) against the regularity (H_1) hypothesis and the clustering (H_2) hypothesis .
  - Randomness hypothesis (H_0): "The vectors of X are randomly distributed, according to the uniform distribution in the sampling window (the compact convex support set for the underlying distribution of the vectors of the data set X) of X".
  - Regularity hypothesis (H_1): "The vectors of X are regularly spaced (that is they are not too close to each other) in the sampling window".
  - Clustering hypothesis (H_2): "The vectors of X form clusters".

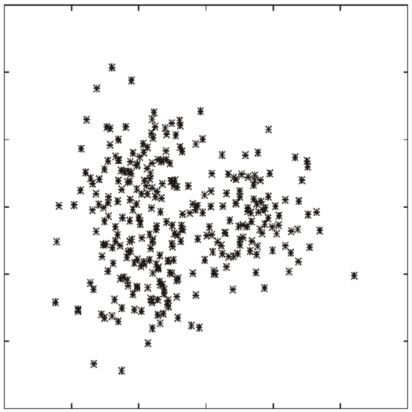
Clustering tendency

• $p(q|H_0), p(q|H_1)$, $p(q|H_2)$ are estimated via Monte Carlo simulations

Some tests for spatial randomness, when the input space dimensionality greater than or equal to 2 are:

- Tests based on structural graphs
  - *Test that utilizes the idea of the minimum spanning tree (MST)*
- Tests based on nearest neighbor distances
  - *The Hopkins test*
  - *The Cox-Lewis test*
- A method based on sparse decomposition.





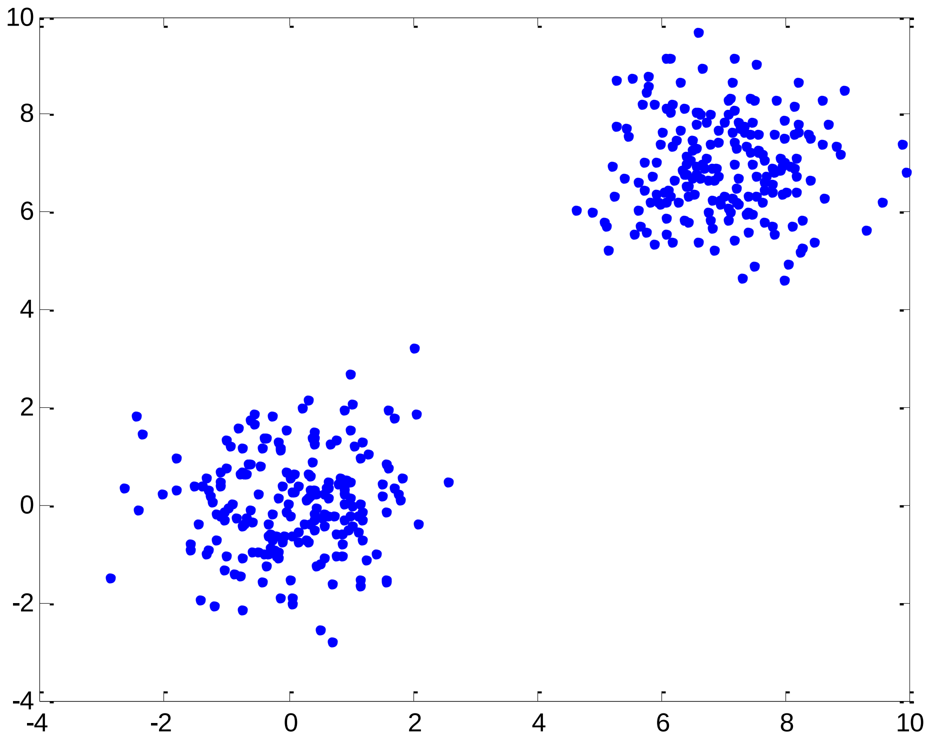
Clustering tendency

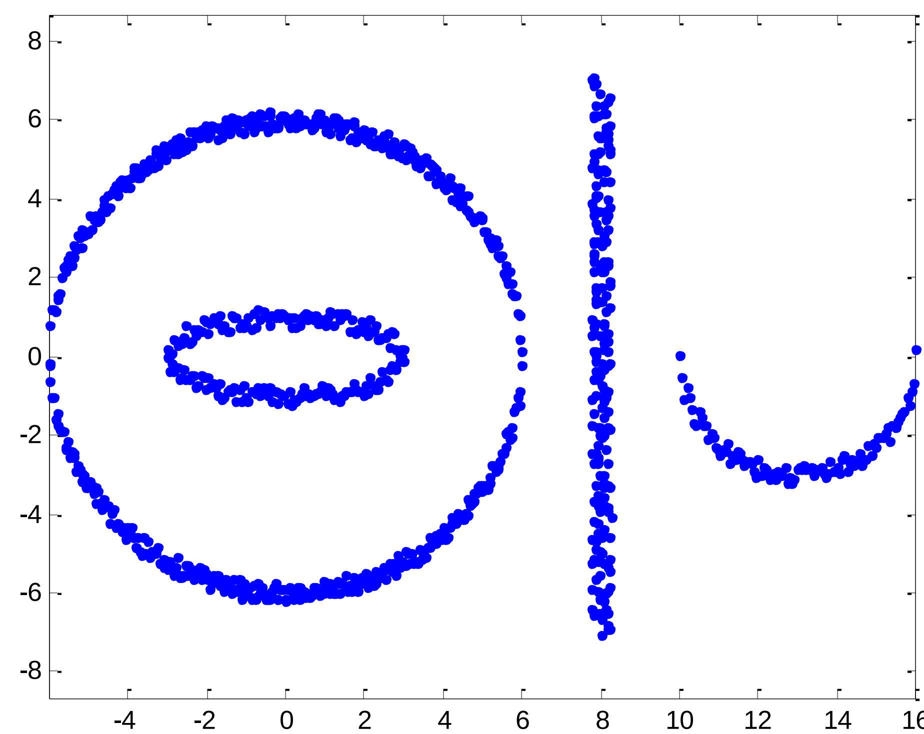
- Important notes:
  - Clustering algorithms should be applied on X, only if the randomness and the regularity hypotheses are rejected. Otherwise, methods different than clustering must be used to describe the structure of X.
  - Most studies in clustering tendency focus on the detection of compact clusters.

- The **basic steps** of the clustering tendency philosophy are:
  - Definition of a test statistic q suitable for the detection of clustering tendency.
  - Estimation of the pdf of q under the null (H_0) hypothesis, $p(q|H_0)$.
  - Estimation of $p(q|H_1)$ and $p(q|H_2)$ (they are necessary for measuring the power of q (the probability of making a correct decision when H_0 is rejected) against the regularity and the clustering tendency hypotheses).
  - Evaluation of q for the data set at hand, X, and examination whether it lies in the critical interval of $p(q|H_0)$, which corresponds to a predetermined significance level ρ.

Clustering Algorithms: case study 1 (optional)

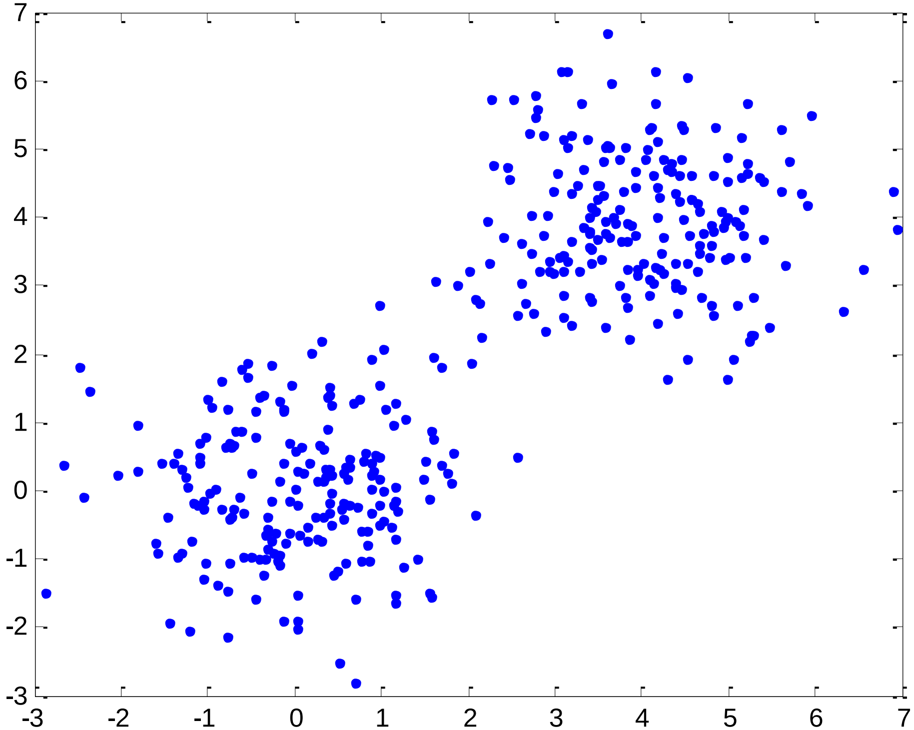
The problem: Propose a method/methodology in order to have an indication of whether the data s set under study possesses a clustering structure or not.

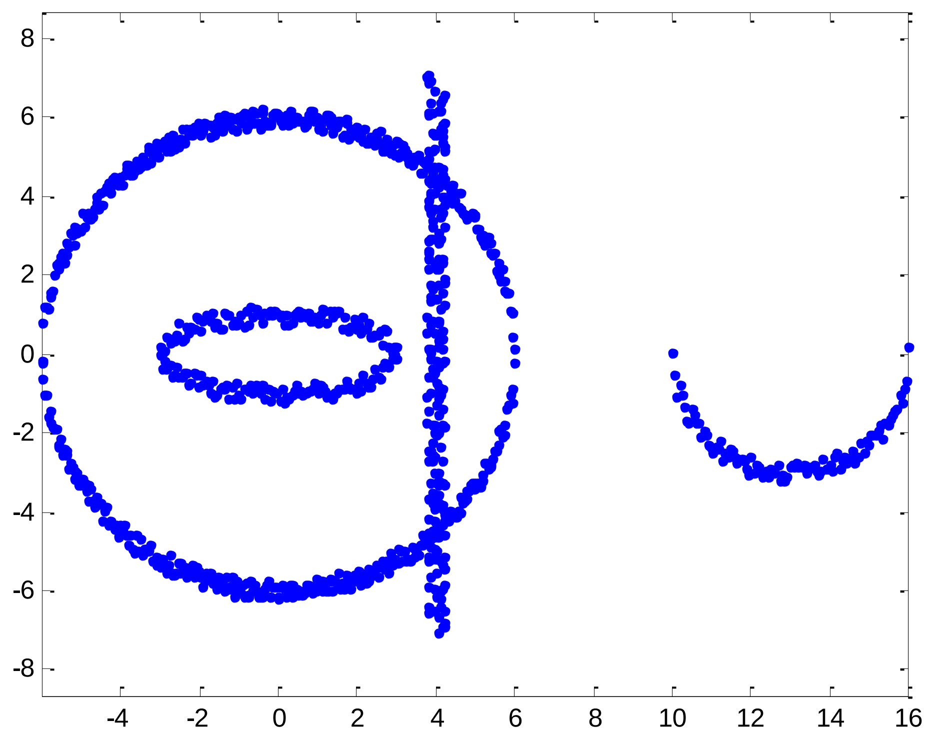




Clustering Algorithms: case study 1 (optional)

The problem: Propose a method/methodology in order to have an indication of whether the data s set under study possesses a clustering structure or not.





Clustering Algorithms: case study 1 (optional)

The problem: Propose a method/methodology in order to have an indication of whether the data s set under study possesses a clustering structure or not.

A possible solution:

- Consider the associated graph where the edges are weighted by the distance of the corresponding data points.
- **Determine** the Minimum Spanning Tree (MST) of the graph.
- Check whether its largest edge is "several standard deviations" away from the mean of the weights of the edges of the MST.
- Alternatively, one can use the statistical hypothesis testing path. That is, to generate a set of N uniformly randomly distributed data in the space where the data live and to check the distance of the largest MST edge weight from the mean of the MST edge weights.

Limitations:

- **Overlapping clusters**. **A possible solution:** If we know the "shape" of the clusters that are expected to be formed by the data, we can run e.g., k-means (for compact clusters) or algorithms like *Fuzzy C Ellipsoidal Shells* (FCES) for the case of ellipsoidally-shaped clusters, or Gustafson-Kessel for linearly-shaped clusters, we can run the algorithm for a range of the number of clusters m and to search for a significant "knee" in the graph of the cost function vs m.

Clustering Algorithms: case study 2 (optional)

The problem: Consider a pool of N **manuscripts** concerning a specific space mission. Given **(a)** a set of characteristic keywords provided by the experts of the application, **(b)** the main thematic categories to which the manuscripts for this mission can be assigned and **(c)** the category(ies) to which a manuscript is assigned, perform clustering on the data to see to what degree the resulting clustering agrees with the labeling provided by the experts.

The possible success of this experiment may open the gate for the automatic annotation of a large number of manuscripts.

TF-IDF computation:

if

- i is the i-th keyword
- j is the j-th document
- app(i,j) is the number of appearances of the i-th keyword to the j-th document
- occ(i) is the number of documents in which the i-th keyword appears
- p is the number of the available documents

then

TF-IDF(i,j)=app(i,j)*$\log_2$(p/occ(i))

***TF-IDF** is*:

(1)**higher** when the **keyword occurs many times** within a **small number** of **documents** (thus lending high discriminating power to those documents),

(2)**lower** when the **keyword occurs fewer times** in a **document OR occurs** in **many documents** (thus offering a less pronounced relevance signal),

(3)**lower** when the **keyword occurs** in **virtually all documents**.

THE END