Cluster Analysis

Determining the Number of Clusters

Frank Klawonn

Institute of Applied Informatics, Department of Computer Science
Ostfalia University of Applied Sciences
Wolfenbuettel, Germany
f.klawonn@ostfalia.de

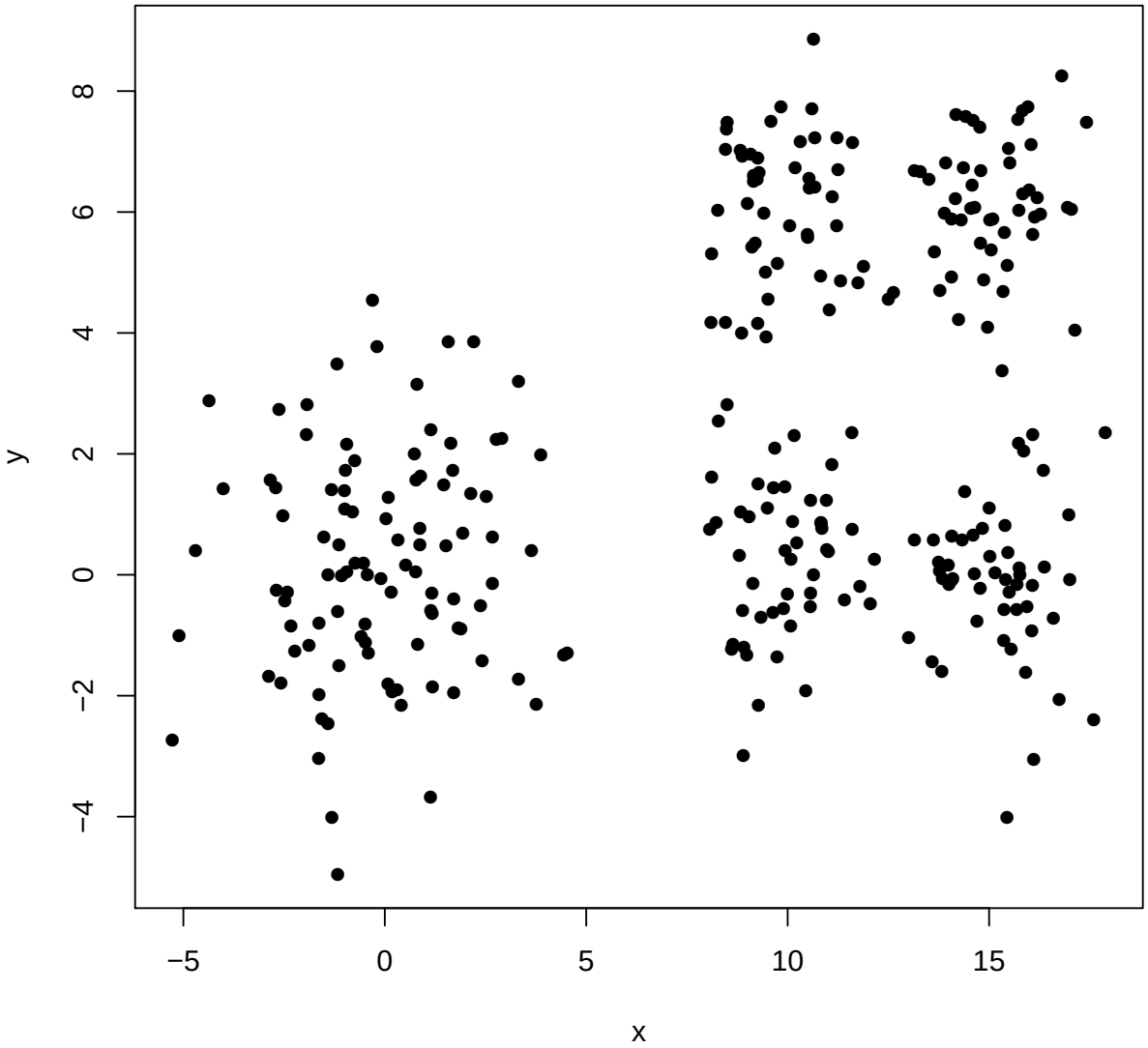
Bioinformatics & Statistics
Helmholtz Centre for Infection Research
Braunschweig, Germany
frank.klawonn@helmholtz-hzi.de

A clustering problem?

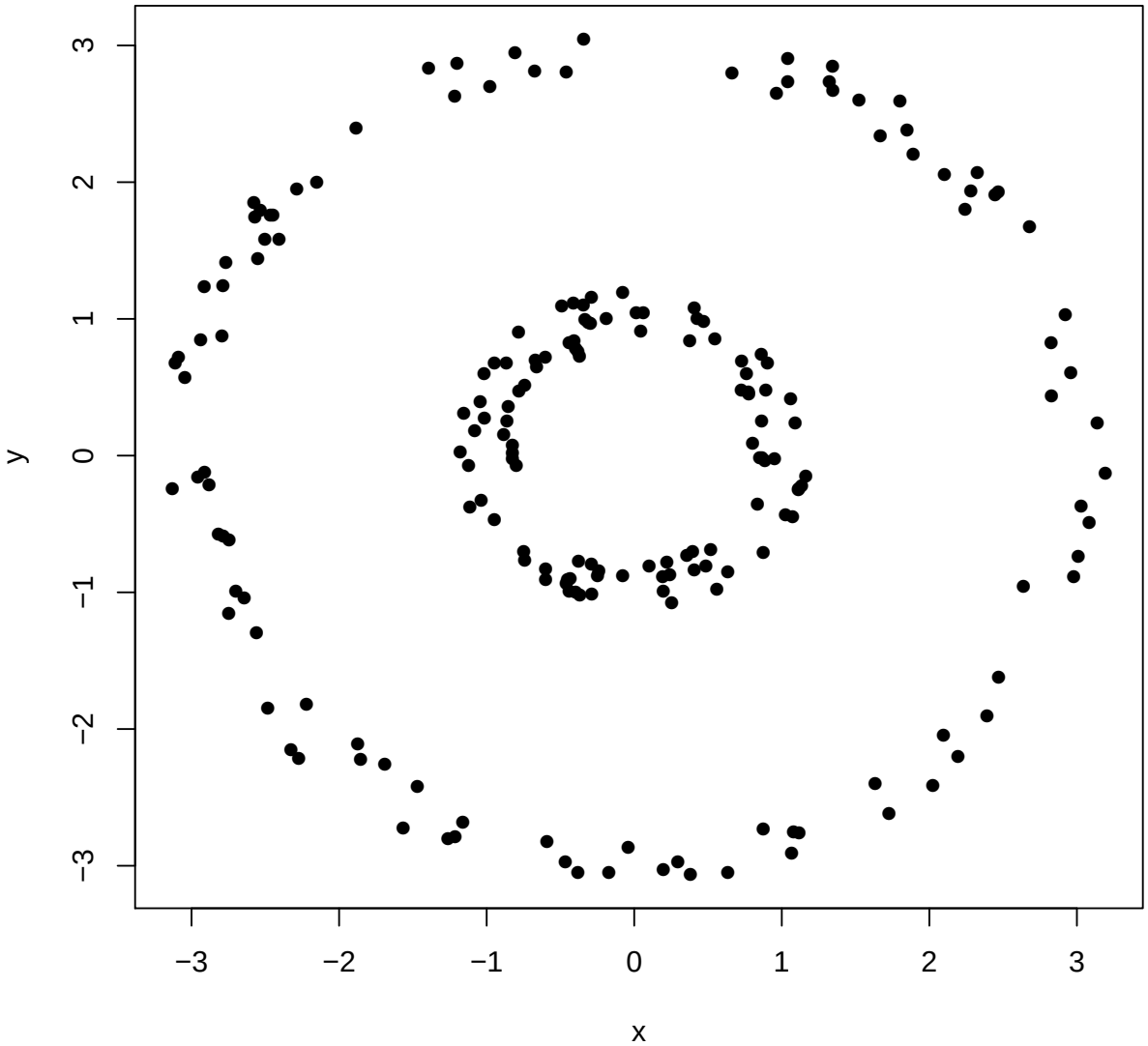
Can you form a small cluster of people you know within this group of people?

Contents

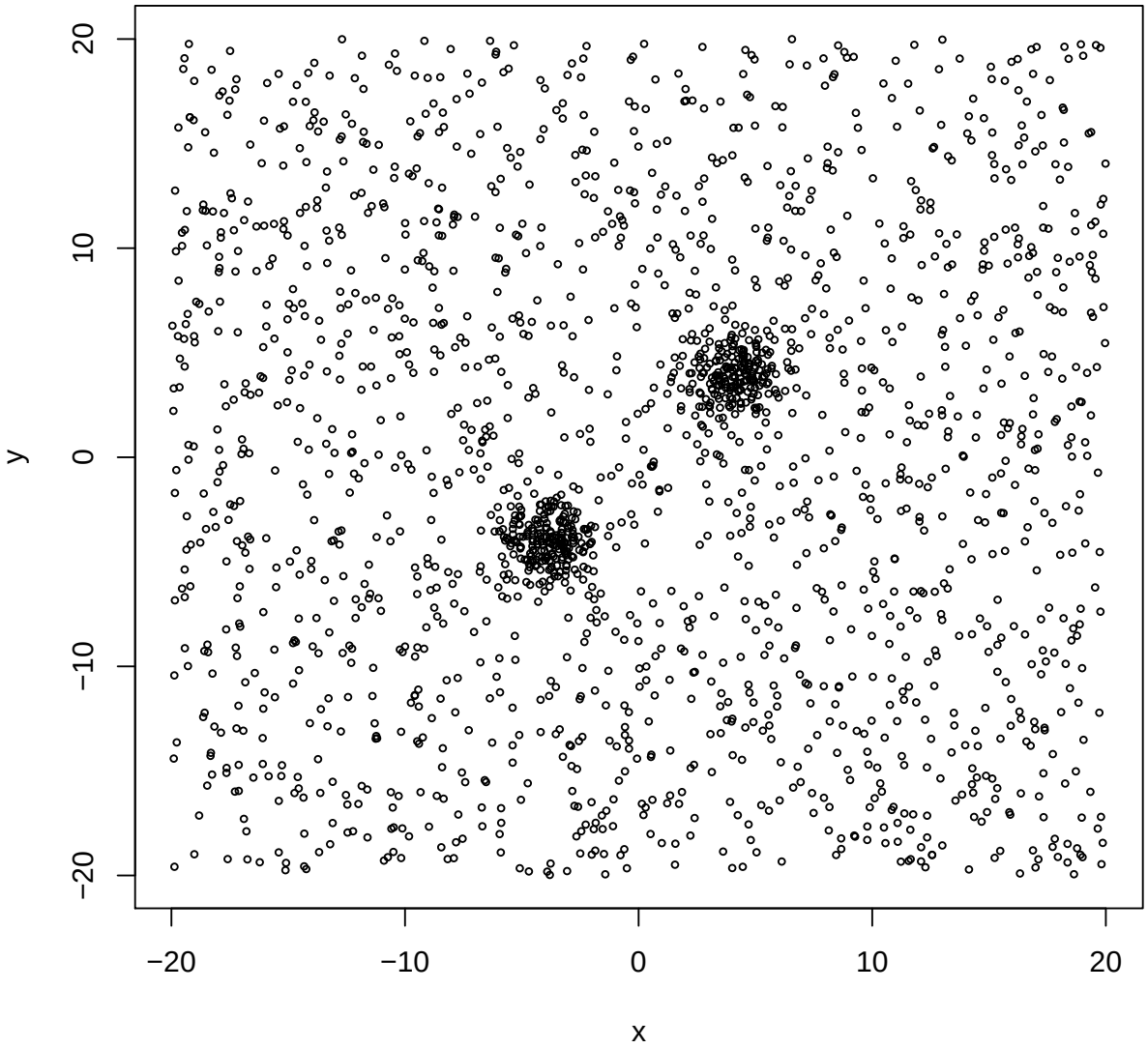
How many clusters are there?



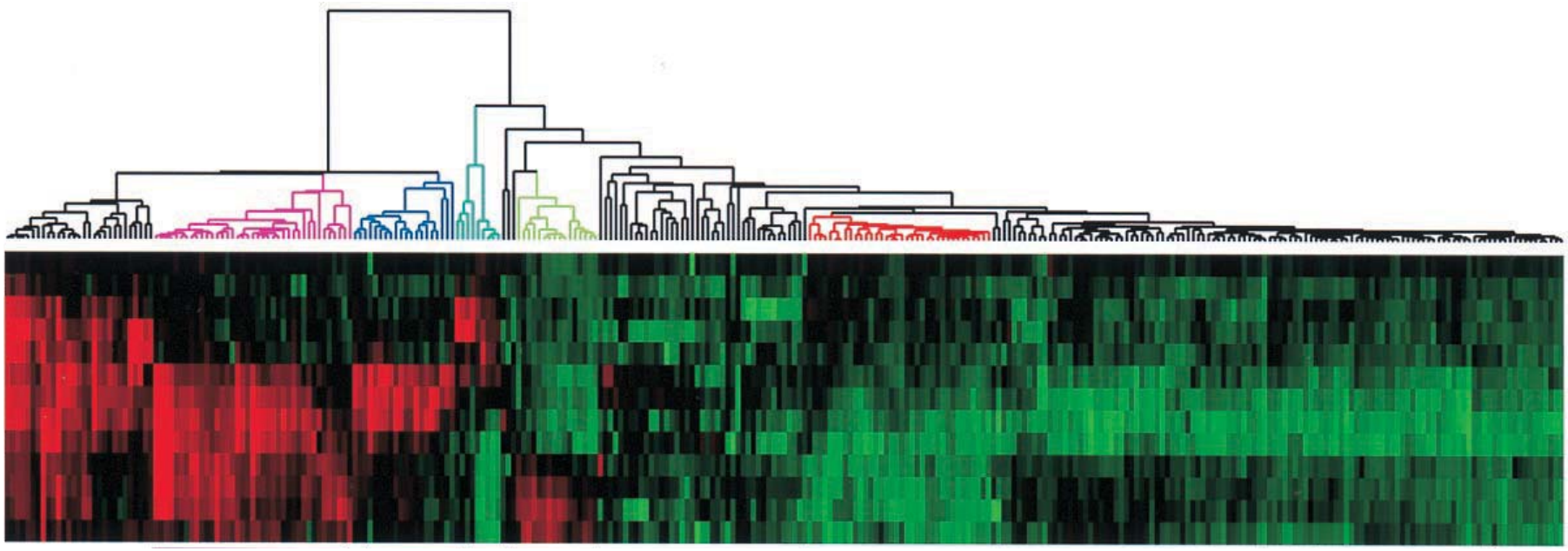
Can clusters look like this?



Is this a clustering problem?



Are all attributes important for the clusters?



What is cluster analysis?

A cluster (in a given data set) is a subset of data, so that the data

- within the cluster are "similar"
- and differ from the data outside the cluster.

Data inside a cluster should be homogeneous, data from different clusters heterogeneous.

Goals of cluster analysis

- Partition a given data set into clusters
- Check whether related objects cluster together
- Classify unknown objects
- Find single "meaningful" clusters (and do not care about the rest of the data)

Goals of cluster analysis

- Partition a given data set into clusters
- Check whether related objects cluster together
- Classify unknown objects
- Find single "meaningful" clusters (and do not care about the rest of the data)

Assignment of data to clusters

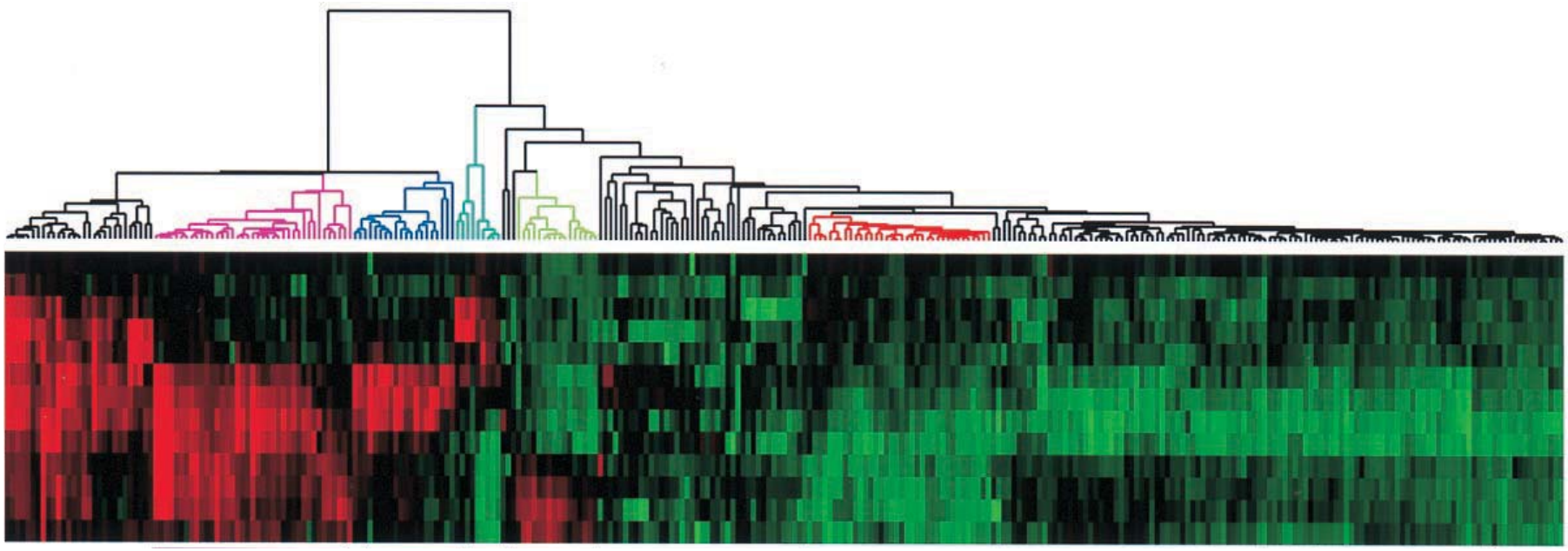
- Crisp clustering: Each data object is assigned to exactly one cluster
- Probabilistic models: For each data object a probability distribution over the clusters is specified.
- Fuzzy "probabilistic" clustering: Each data object is assigned to a cluster with a membership degree. (The membership degrees add up to one, but cannot be interpreted as probabilities.)
- (Fuzzy) possibilistic clustering: Each data object is assigned to a cluster with a membership degree. The sum of the membership degrees can have arbitrary values. (Problems of inconsistency.)

Assignment of data to clusters

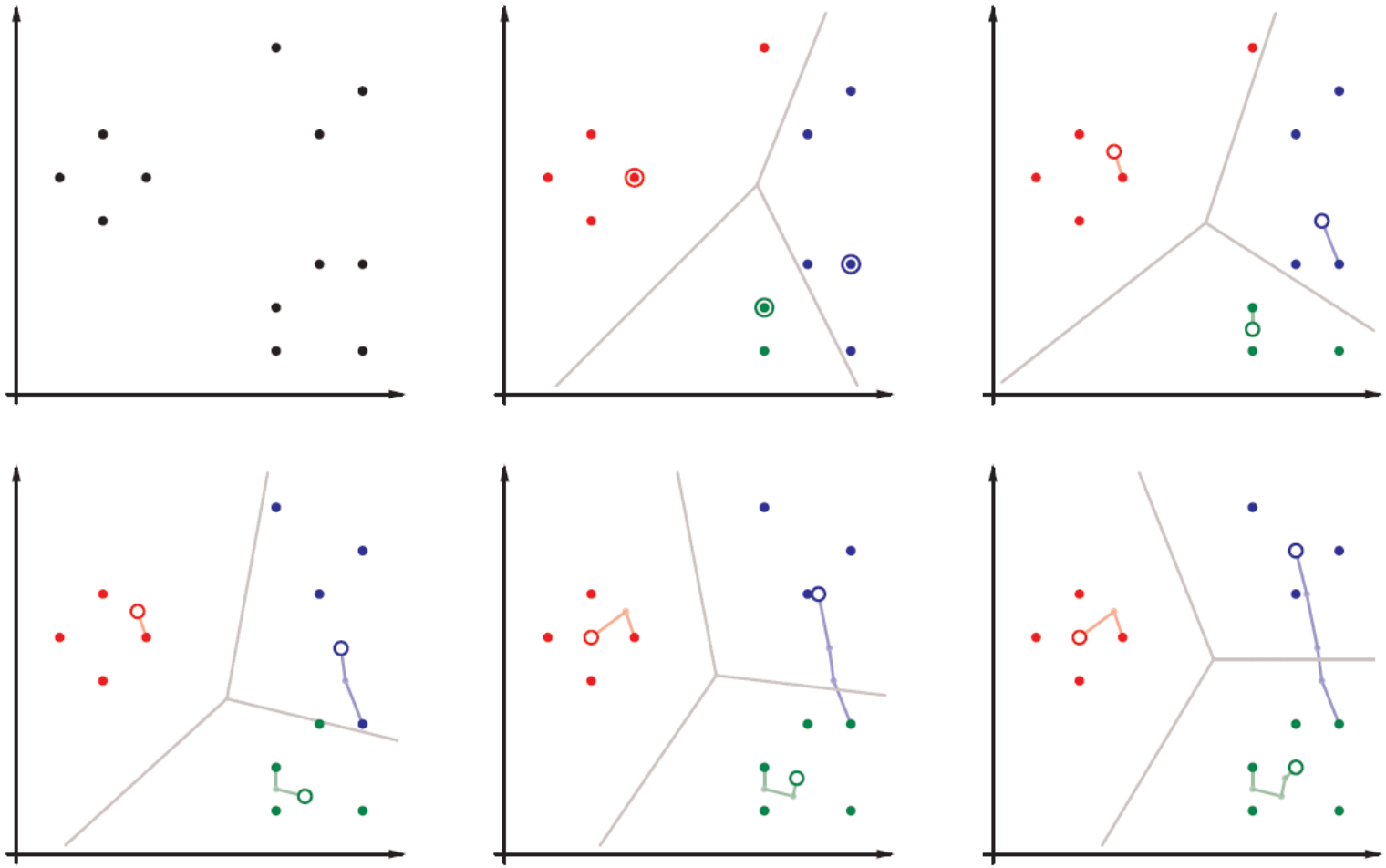
- Crisp clustering: Each data object is assigned to exactly one cluster
- Probabilistic models: For each data object a probability distribution over the clusters is specified.
- Fuzzy "probabilistic" clustering: Each data object is assigned to a cluster with a membership degree. (The membership degrees add up to one, but cannot be interpreted as probabilities.)
- (Fuzzy) possibilistic clustering: Each data object is assigned to a cluster with a membership degree. The sum of the membership degrees can have arbitrary values. (Problems of inconsistency.)

A "noise" cluster might be included, i.e. data might (partially) not be assigned to any cluster.

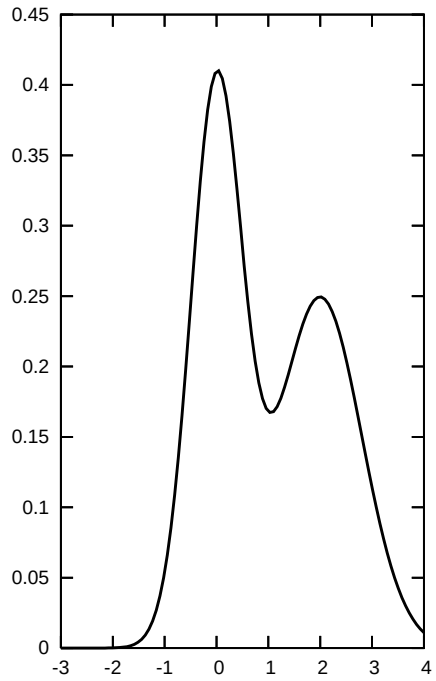
Hierarchical clustering



k-Means clustering

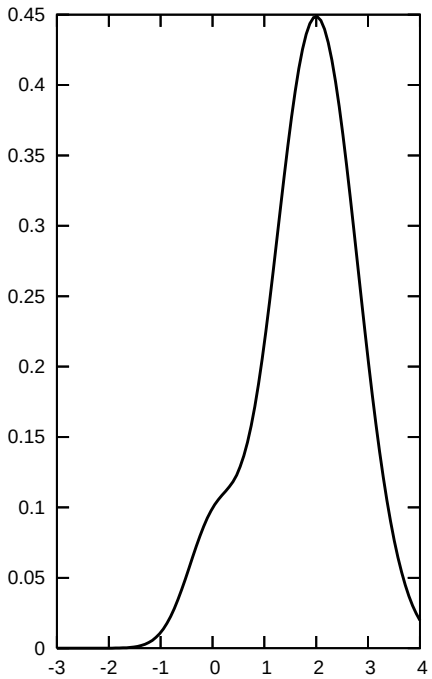


(Gaussian) mixture models



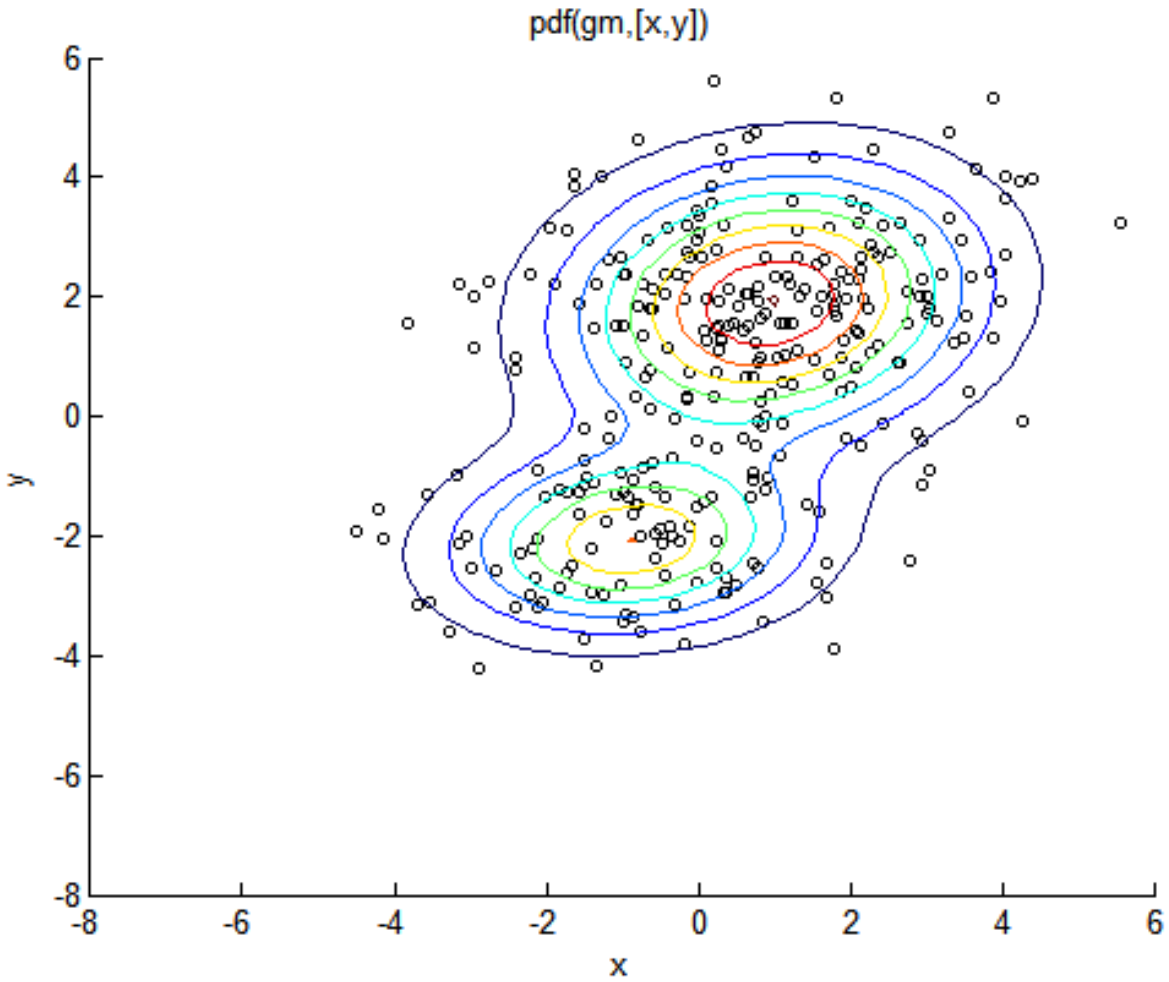
Mixture model (both normal distributions contribute 50%)

(Gaussian) mixture models



Mixture model (one normal distributions contributes 10%, the other 90%)

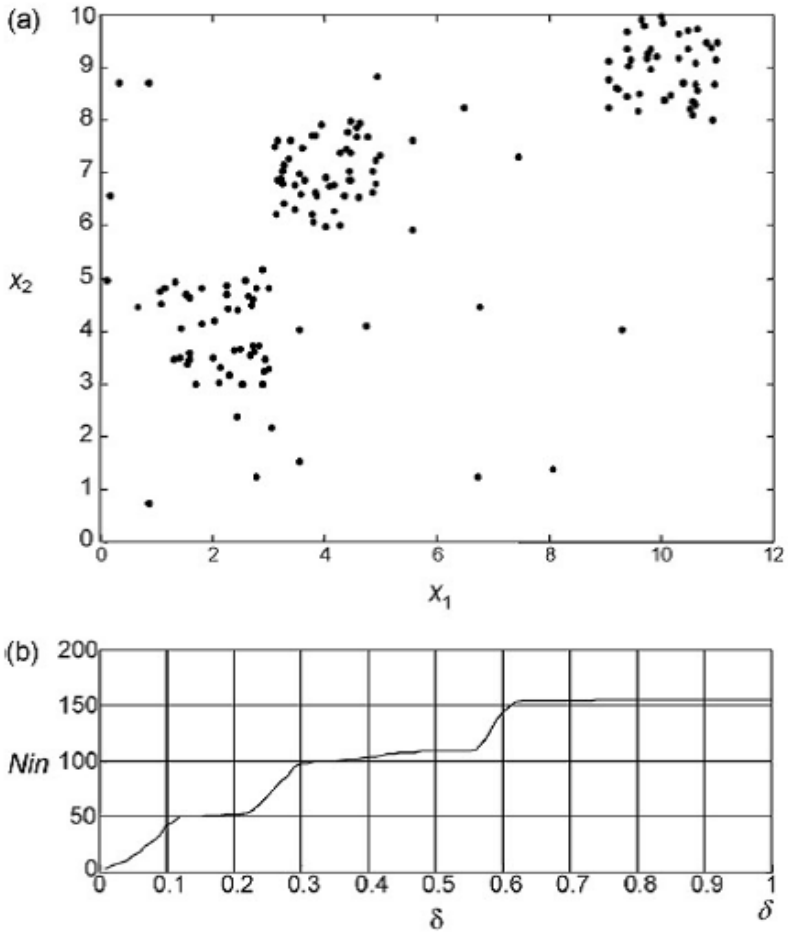
(Gaussian) mixture models

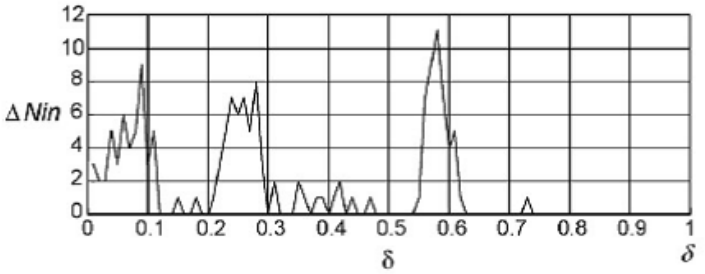


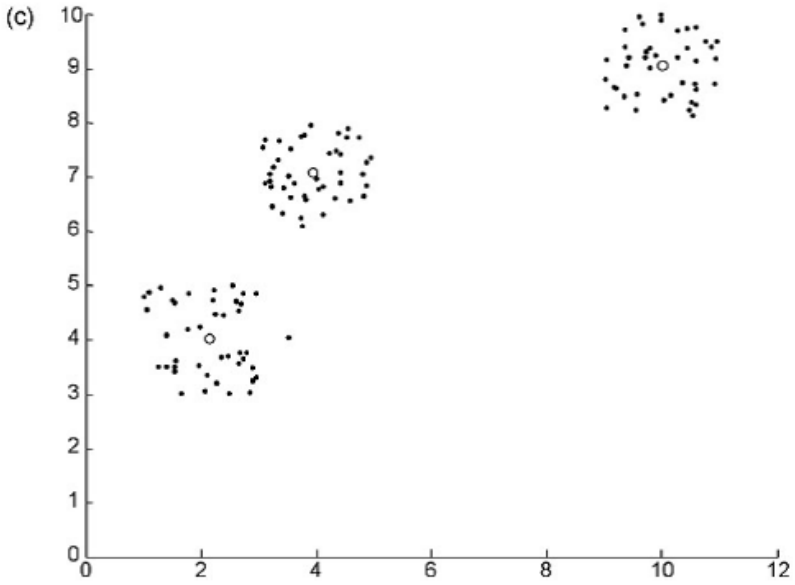
Subtractive clustering

- Identify clusters step by step.
- Find one cluster, remove it from the data set and
- continue this procedure until no data objects are left or no more clusters can be found.

Subtractive clustering





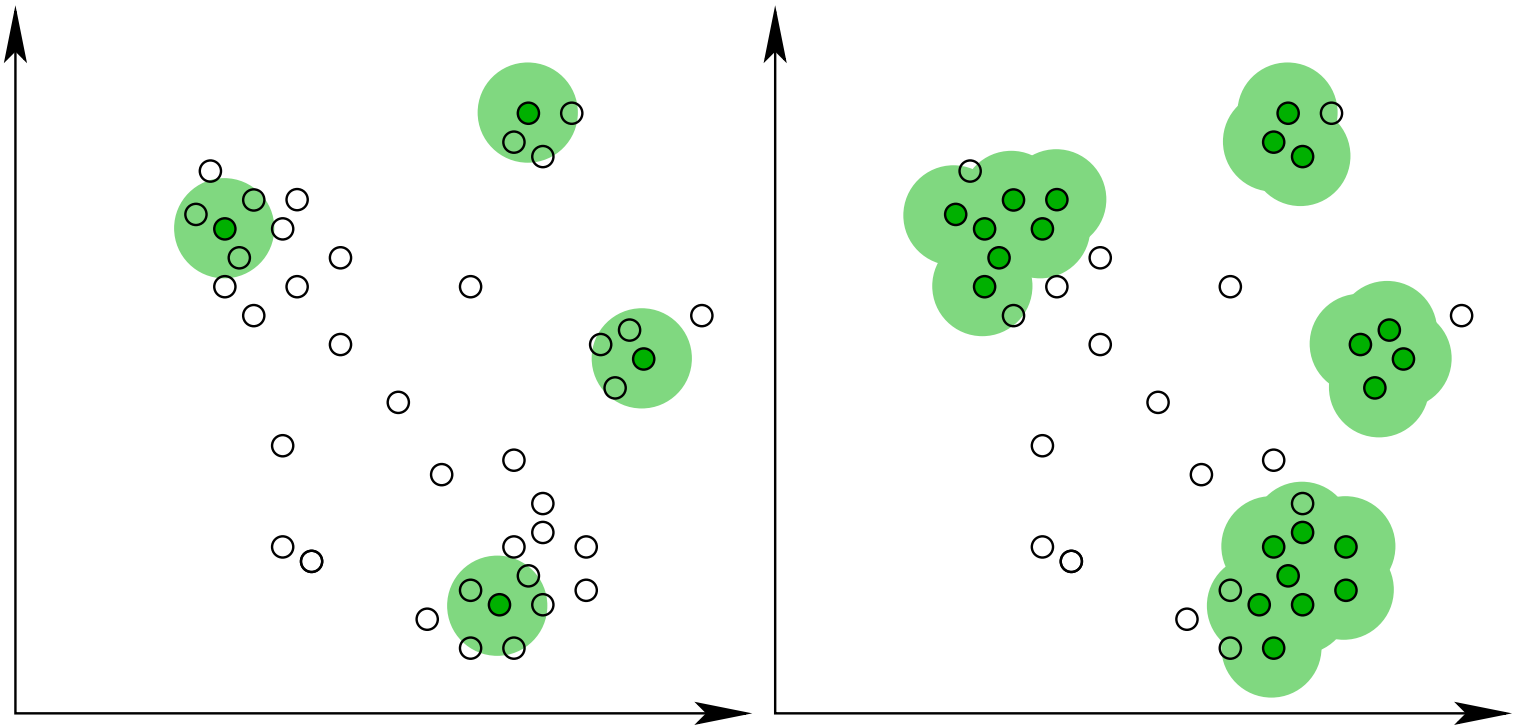


Density-based clustering: DBScan

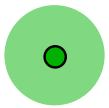
Principle idea of DBScan:

1. Find a data point where the data density is high, i.e. in whose ε-neighbourhood are at least ℓ other points. (ε and ℓ are parameters of the algorithm to be chosen by the user.)
2. All the points in the ε-neighbourhood are considered to belong to one cluster.
3. Expand this ε-neighbourhood (the cluster) as long as the high density criterion is satisfied.
4. Remove the cluster (all data points assigned to the cluster) from the data set and continue with 1. as long as data points with a high data density around them can be found.

Density-based clustering: DBScan



neighbourhood cell with at least 3 hits



neighbourhood cell with at least 3 hits

Are there clusters at all? Hopkins index

- Choose a number $m \ll n$.
- Sample m points $\{y_1, \ldots, y_m\}$ from a uniform distribution over the convex hull of the data.
- Choose m points $\{z_1, \ldots, z_m\}$ randomly from the original data set.
- d_{y_i}: Distance of y_i to the closest point in the data set.
- d_{z_i}: Distance of z_i to its closest neighbour in the data set.

$$\text{Hopkins index} \quad h = \frac{\sum_{i=1}^{m} d_{y_i}}{\sum_{i=1}^{m} d_{y_i} + \sum_{i=1}^{m} d_{z_i}} \qquad \in [0, 1]$$

Are there clusters at all? Hopkins index

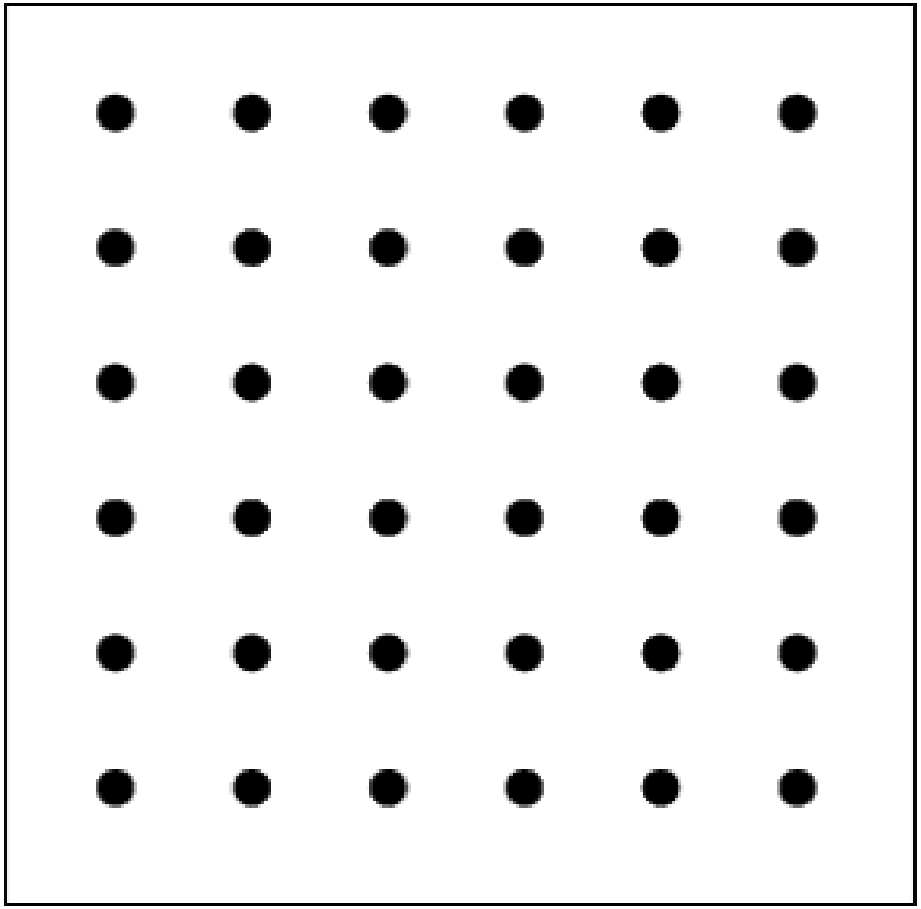
- Choose a number $m \ll n$.
- Sample m points $\{y_1, \ldots, y_m\}$ from a uniform distribution over the convex hull of the data.
- Choose m points $\{z_1, \ldots, z_m\}$ randomly from the original data set.
- d_{y_i}: Distance of y_i to the closest point in the data set.
- d_{z_i}: Distance of z_i to its closest neighbour in the data set.

$$\text{Hopkins index} \quad h = \frac{\sum_{i=1}^{m} d_{y_i}}{\sum_{i=1}^{m} d_{y_i} + \sum_{i=1}^{m} d_{z_i}} \qquad \in [0, 1]$$

h depends very much on the random selection. Repeat the procedure multiple times and compute the mean value of the results.

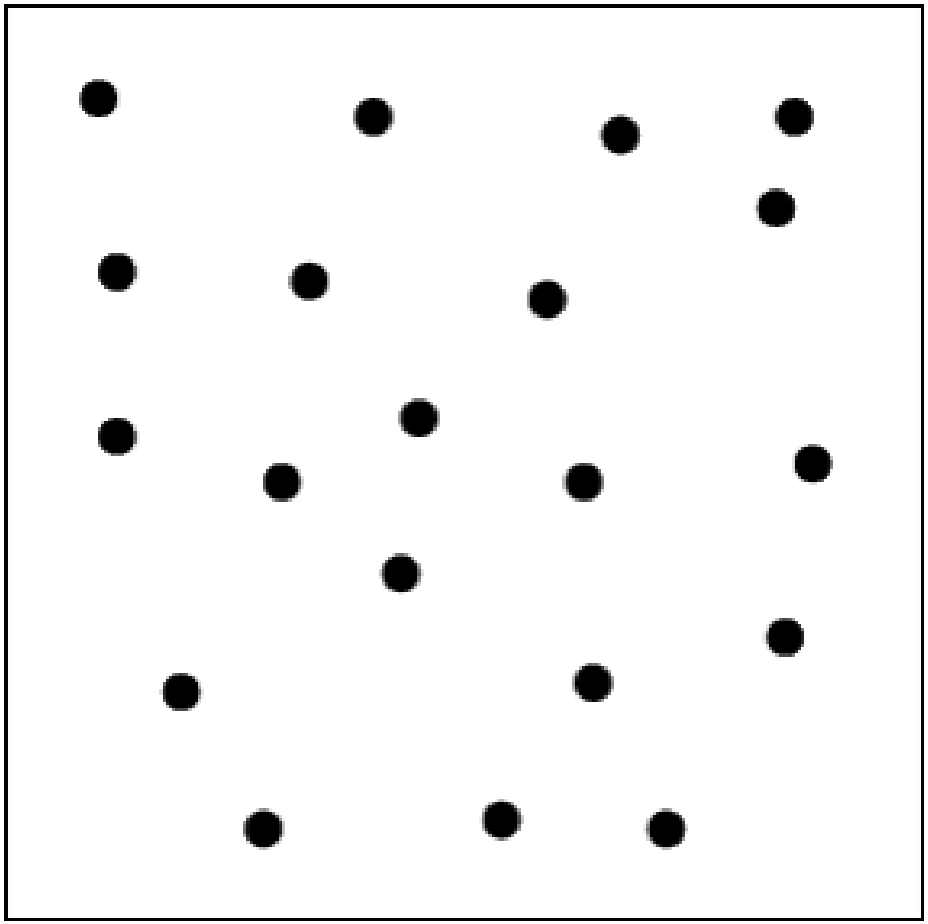
Are there clusters at all? Hopkins index

$h \approx 0$: regular structure, but no clusters



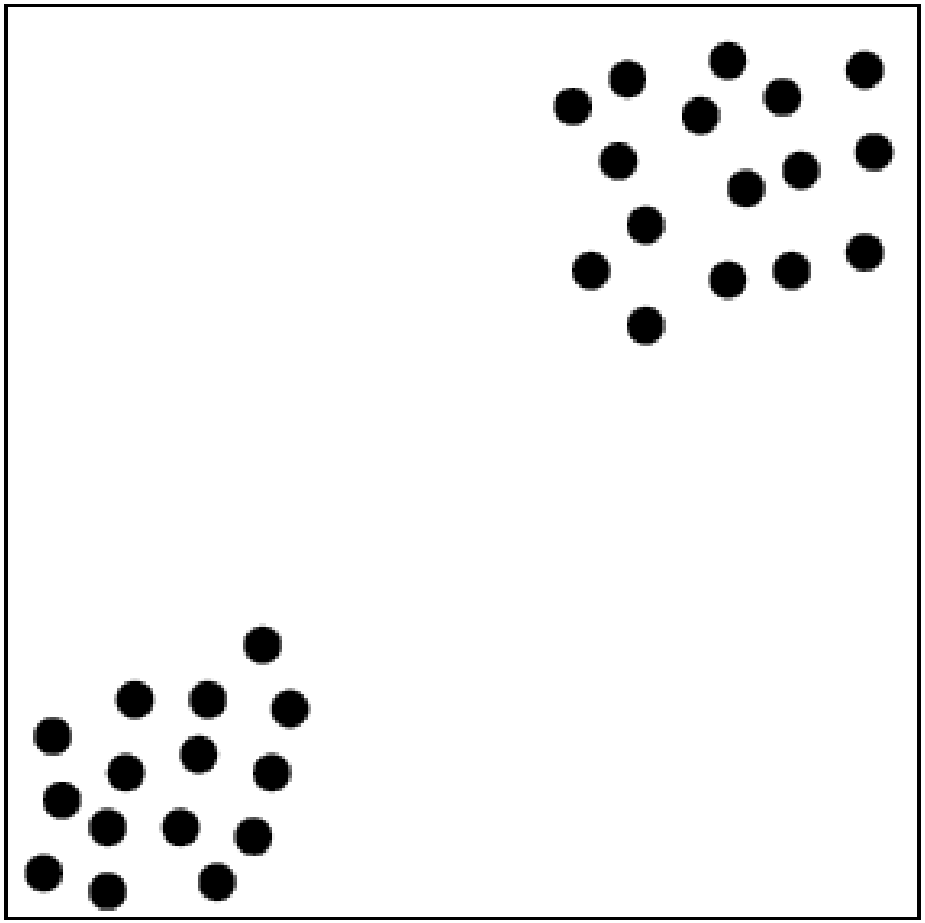
Are there clusters at all? Hopkins index

$h \approx 0.5$: roughly uniform distribution, no clusters



Are there clusters at all? Hopkins index

$h \approx 1$: clusters in the data set



Model selection and clustering

- Model selection refers to choosing the right (statistical) model from a set of candidate models for a given data set.
- The more complex the model, the better it can be fit to the data.
- But the most complex model is usually not the best model.

Model selection and clustering

- Model selection refers to choosing the right (statistical) model from a set of candidate models for a given data set.
- The more complex the model, the better it can be fit to the data.
- But the most complex model is usually not the best model.
- For cluster analysis, the "model" which defines an individual cluster for each data object yields a "perfect fit".
  - Data within a cluster are very similar, they are even equal.
  - Data from different clusters are different.

Model selection and clustering

- Model selection refers to choosing the right (statistical) model from a set of candidate models for a given data set.
- The more complex the model, the better it can be fit to the data.
- But the most complex model is usually not the best model.
- For cluster analysis, the "model" which defines an individual cluster for each data object yields a "perfect fit".
  - Data within a cluster are very similar, they are even equal.
  - Data from different clusters are different.
- Danger of overfitting.

Model selection and clustering

- Common technique to avoid overfitting: Split the data into training data to build the model and test data to validate the model.
- To remove dependence on the specific training and test set: k-fold crossvalidation.

Model selection and clustering

- Common technique to avoid overfitting: Split the data into training data to build the model and test data to validate the model.
- To remove dependence on the specific training and test set: k-fold crossvalidation.
  - Partition the data set into k (usually $k = 10$) random subsets of approximately the same size.
  - Remove one of the subsets and train the model with the remaining data.
  - Validate the model with the test set. (e.g. for regression, compute the mean squared error in the removed subset, for classification, calculate the misclassification rate for the removed subset.)
  - Repeat this for each subset.
  - Take the mean value of the k validation runs as performance measure.

Resampling

Basic idea of resampling:

- Sample subsets of the data set,
- cluster the subsets and

Resampling

Basic idea of resampling:

- Sample subsets of the data set,
- cluster the subsets and
- check whether the clustering results remain "stable".

Resampling

Simple strategy:

- Partition the data set into k (usually $k = 10$) random subsets of approximately the same size.
- Remove one of the subsets and cluster the remaining data.
- Repeat this for each subset.

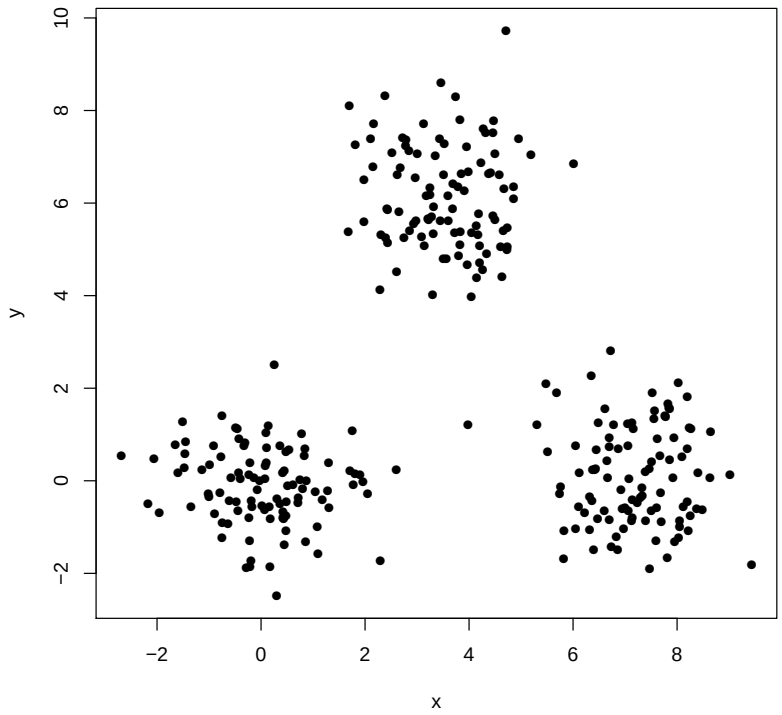
Resampling

Simple strategy:

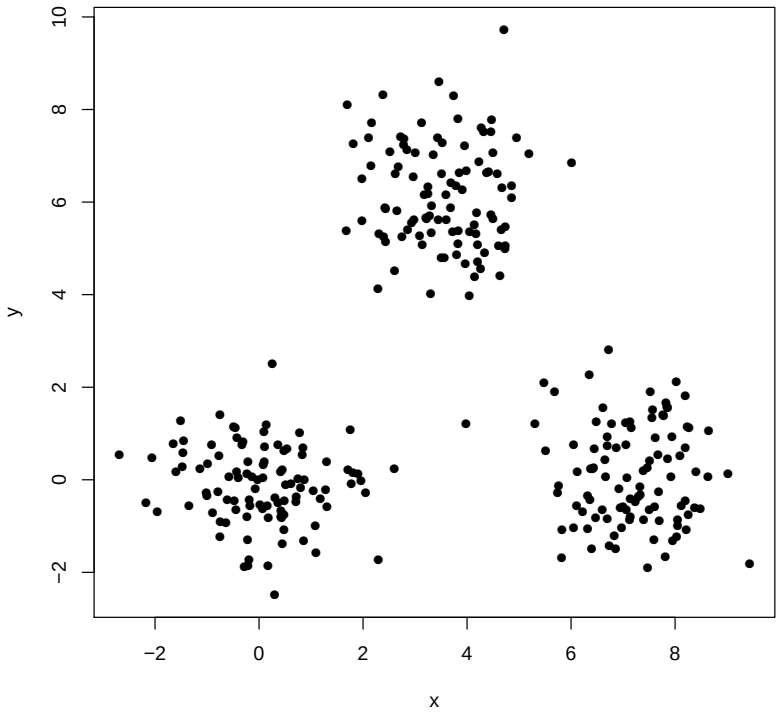
- Partition the data set into k (usually $k = 10$) random subsets of approximately the same size.
- Remove one of the subsets and cluster the remaining data.
- Repeat this for each subset.

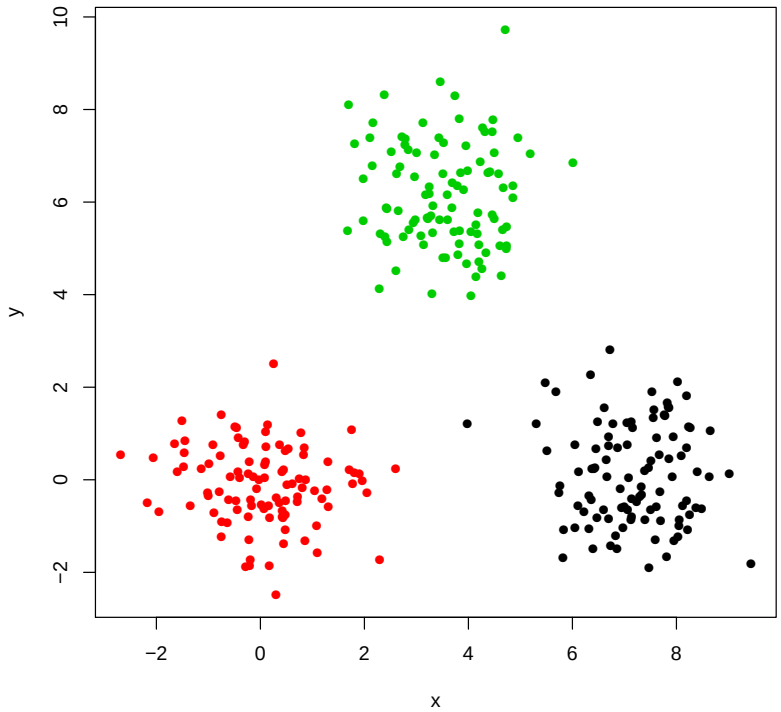
But how can the clustering results be validated?

Resampling: Stability of clusters

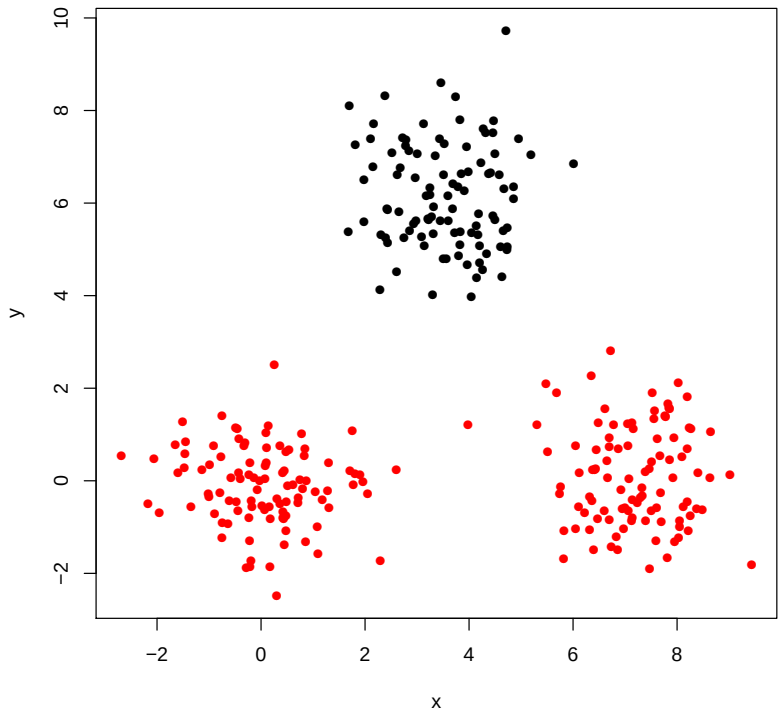


Resampling: Stability of clusters

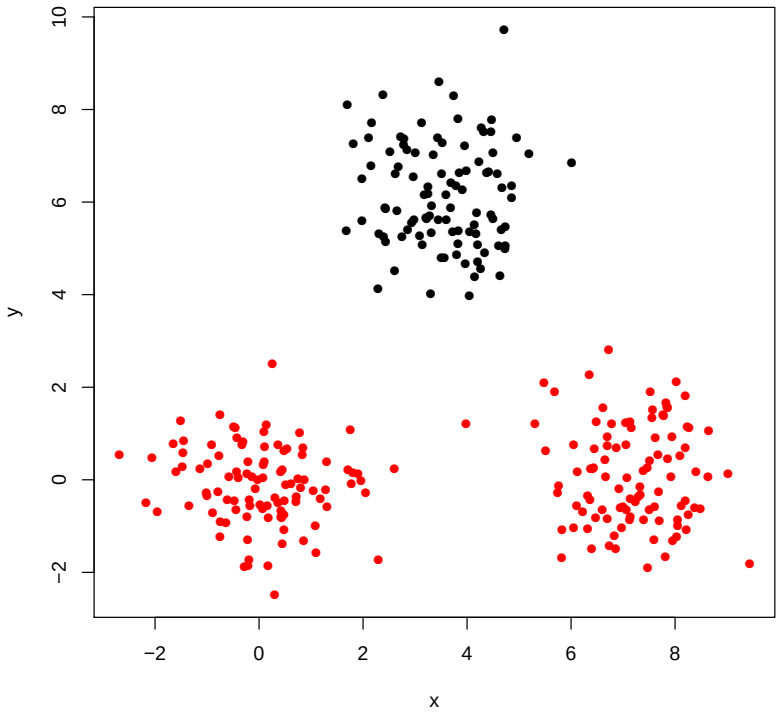


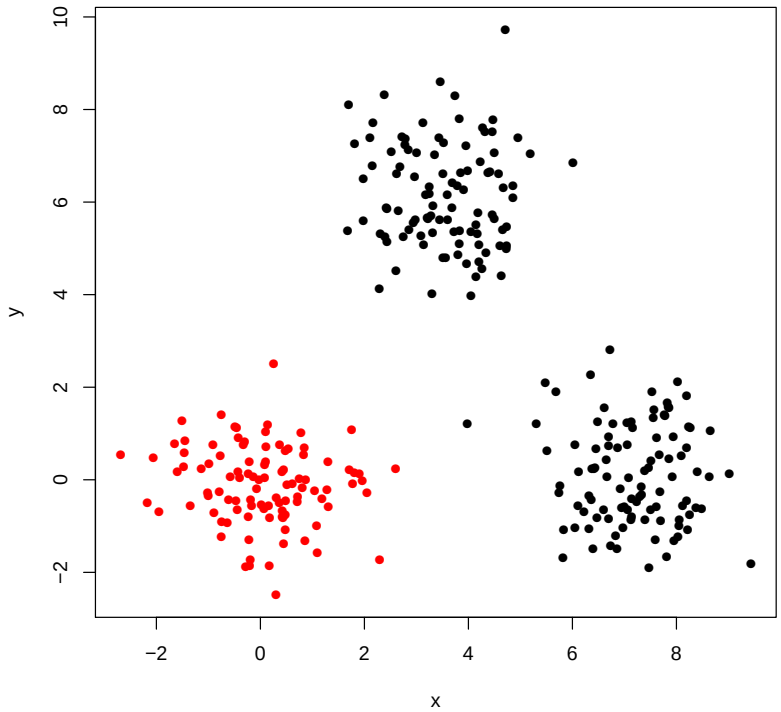


Resampling: Stability of clusters

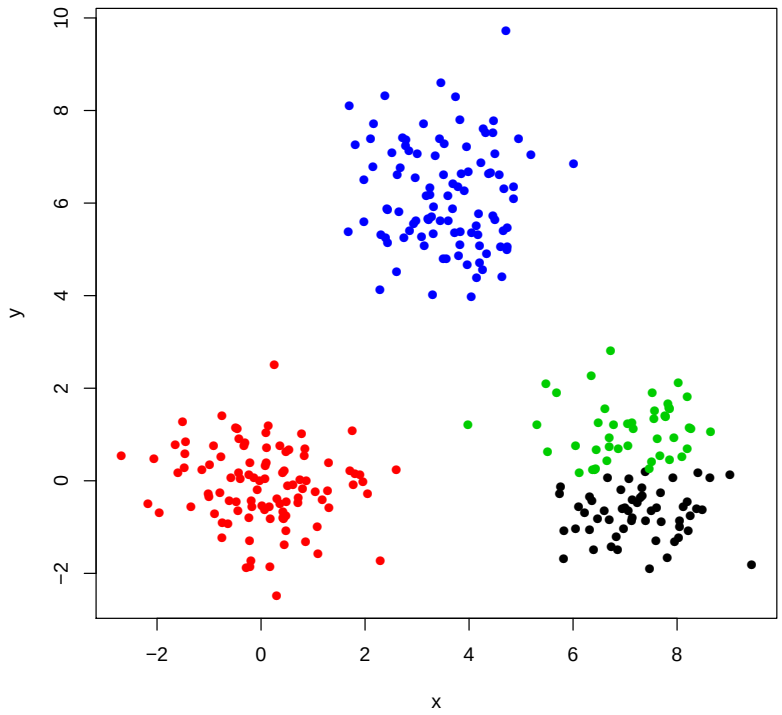


Resampling: Stability of clusters

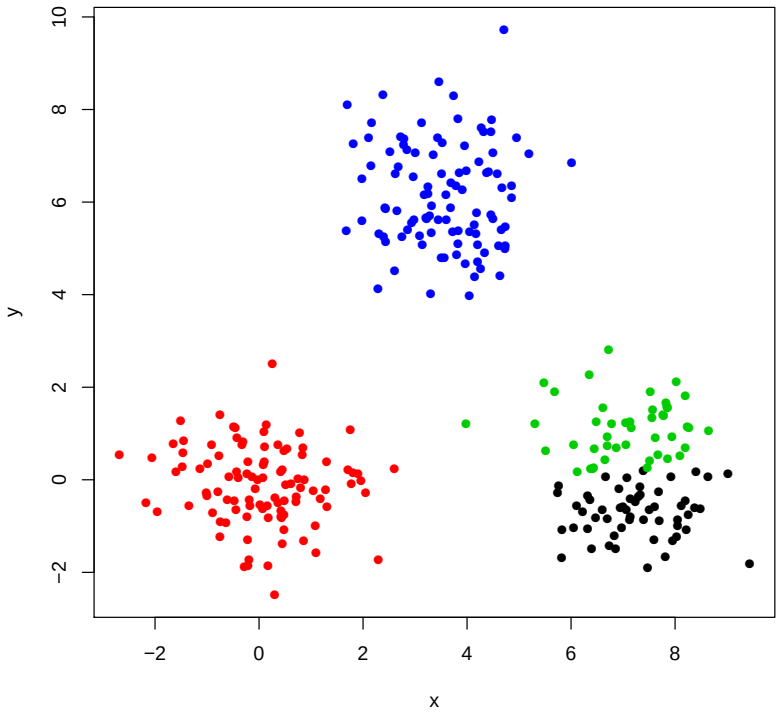


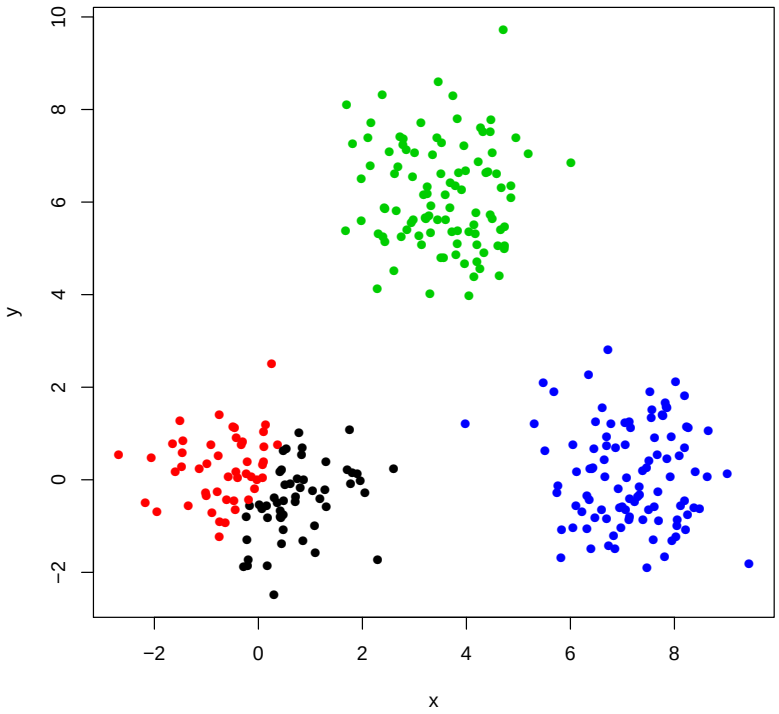


Resampling: Stability of clusters



Resampling: Stability of clusters





Resampling: Simple validation strategy

- Consider two data objects x_1 and x_2 that had both been sampled together for clustering.
  (If the sampling is carried out as in k-fold crossvalidation, they will be selected together at least $(k-2)$ times.)

Resampling: Simple validation strategy

- Consider two data objects x_1 and x_2 that had both been sampled together for clustering.
  (If the sampling is carried out as in k-fold crossvalidation, they will be selected together at least $(k-2)$ times.)
- If the clustering is perfectly stable, in all clustering results where x_1 and x_2 were in the clustering set, they should either
  - always be in the same cluster or
  - always be in different clusters.

Resampling: Simple validation strategy

- Consider two data objects x_1 and x_2 that had both been sampled together for clustering.
  (If the sampling is carried out as in k-fold crossvalidation, they will be selected together at least $(k-2)$ times.)
- If the clustering is perfectly stable, in all clustering results where x_1 and x_2 were in the clustering set, they should either
  - always be in the same cluster or
  - always be in different clusters.

For real data, clustering will seldom be perfectly stable.

As a (basis for a) validation measure, count for each pair of data objects the number of consistent clustering results. (Rand index)

Resampling: More sophisticated cases

- Extension to fuzzy or probabilistic cluster memberships
- Stability/consistency check for clustering algorithms that determine the number of clusters automatically and classify data as noise:
  - How to match (crisp, probabilistic or fuzzy) clustering results with different numbers of clusters?
  - How should the data classified as noise be handled?

Minimum description length principle

The minimum description length principle (MDL) is a model selection techniques based on the following idea:

- A model is understood as a summary or decoding of the data.
- In order to store or transfer the data, the decoding scheme and the encoded (compressed) data are required.
- A model with a perfect fit would make it possible to recover the data without additional information.
  Example: A polynomial regression function

$$y = a_0 + a_1 x + \ldots + a_k x^k$$

  with zero error (i.e. an interpolation function) would make it possible to compute the y_i-values only on the basis of the x_i-values.

Minimum description length principle

- The polynomial (or its coefficients) would be the "decoding scheme", the x_i-values the "compressed" data.
- If a regression instead of an interpolation polynomial is used, the "compressed" data must also contain the residuals e_i in addition to the x_i-values in order to retrieve the y_i-avlues.

Minimum description length principle

- The polynomial (or its coefficients) would be the "decoding scheme", the x_i-values the "compressed" data.
- If a regression instead of an interpolation polynomial is used, the "compressed" data must also contain the residuals e_i in addition to the x_i-values in order to retrieve the y_i-avlues.

- The coding length of the data corresponds to the sum of the coding needed for the decoding scheme (the model) and the required information (compressed data) (the residuals) to recover the original data.
- The "best" model is the one with the shortest coding length.

MDL and clustering

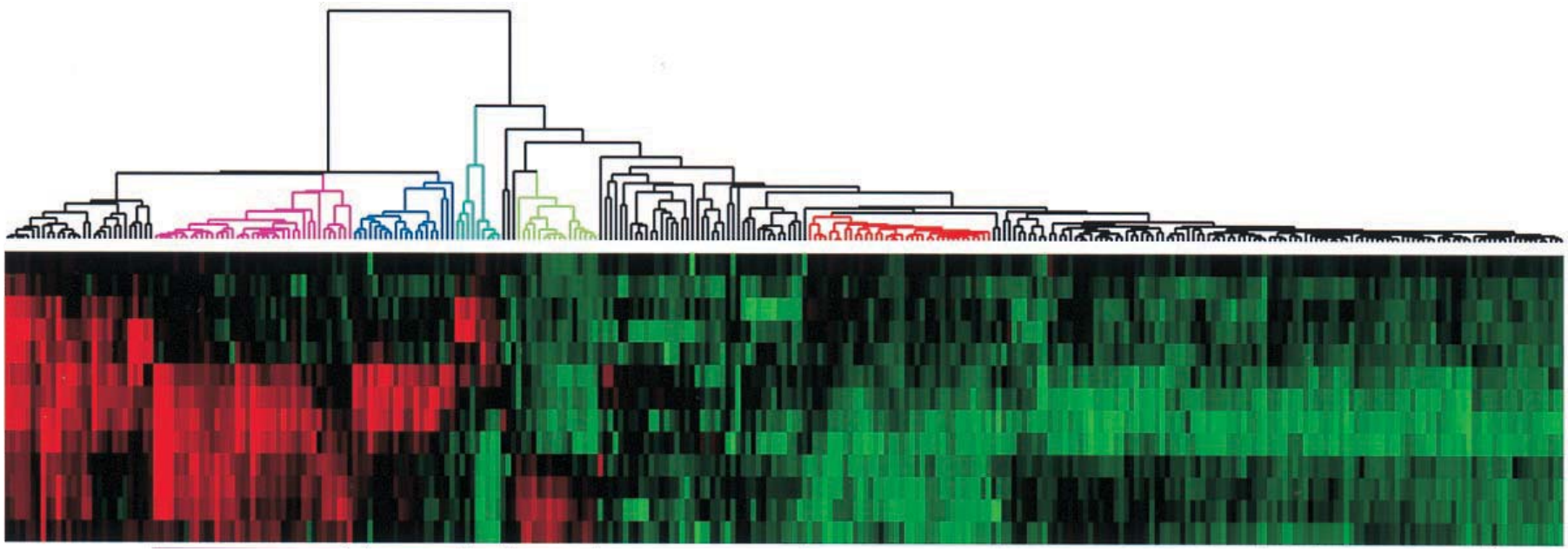
Basic idea (for k-means):

- Each cluster is represented by its clustre centre.
- The cluster centres and the assignments of the data to the clusters correspond to the "model".
- The correction vectors of the data points to the corresponding cluster centres correspond to the compressed data.

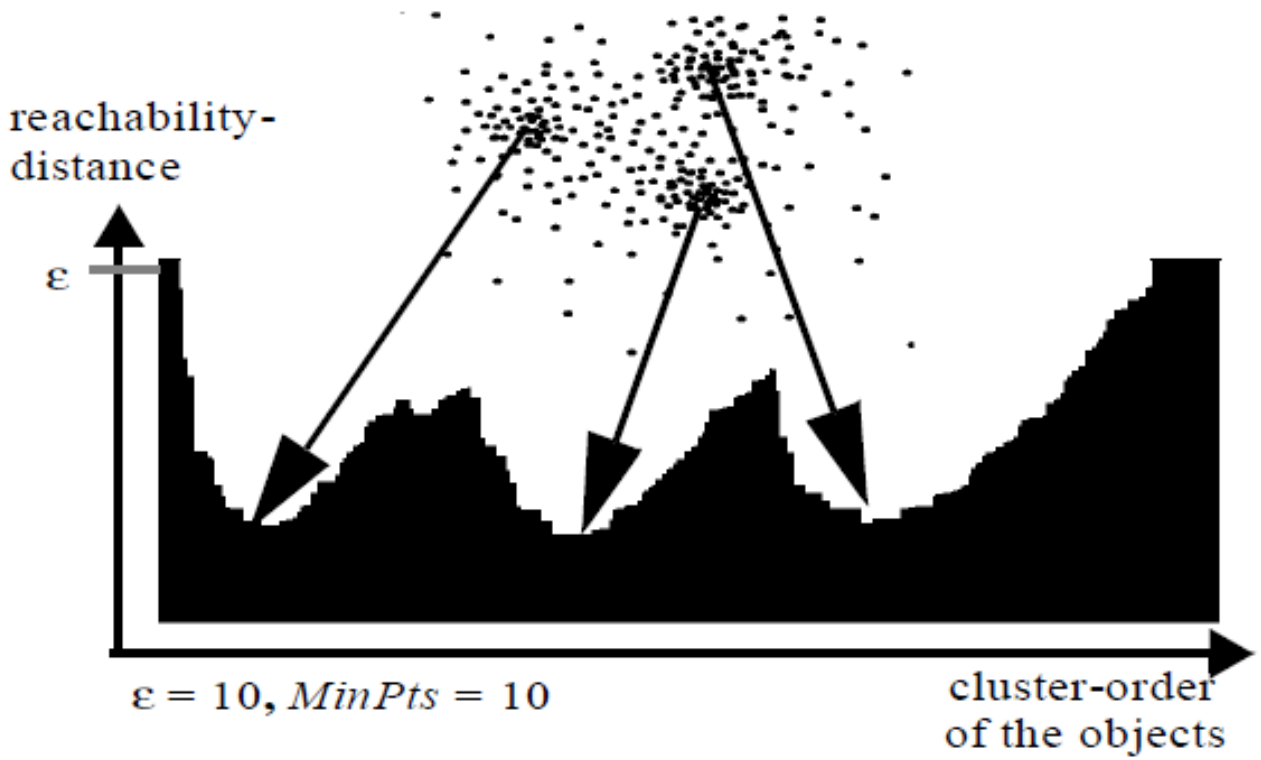
BIC and Gaussian mixture models

For Gaussian mixture models usually the Bayesian information criterion (BIC) is applied to determine the number of clusters.

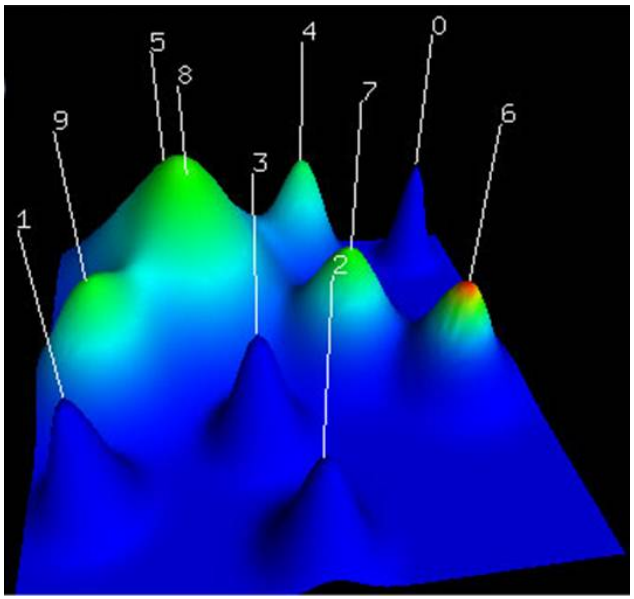
Visualisation techniques

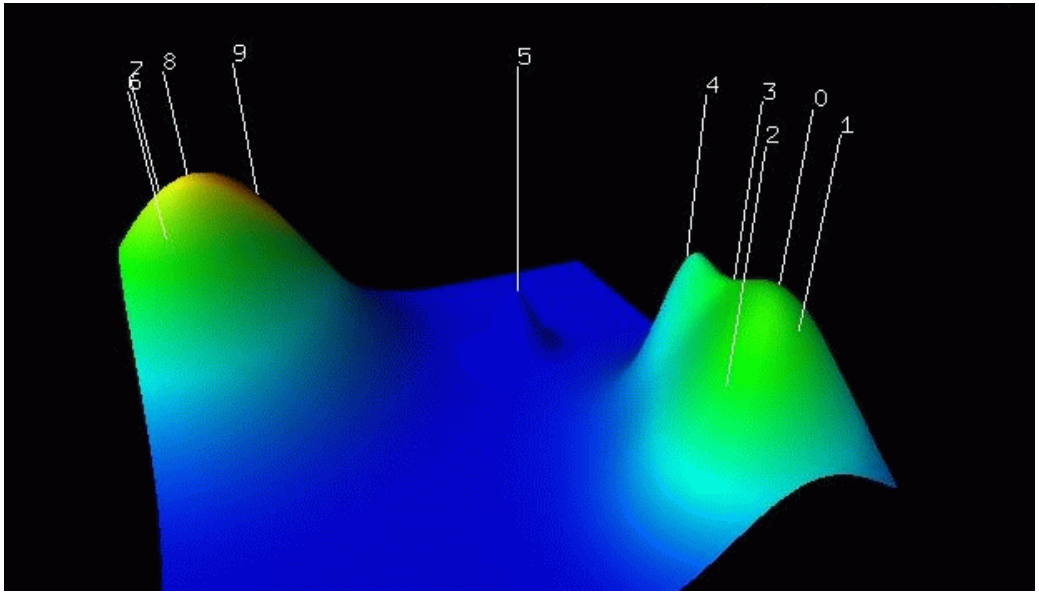


Visualisation: OPTICS and DBSCAN



Visualisation: gCLUTO

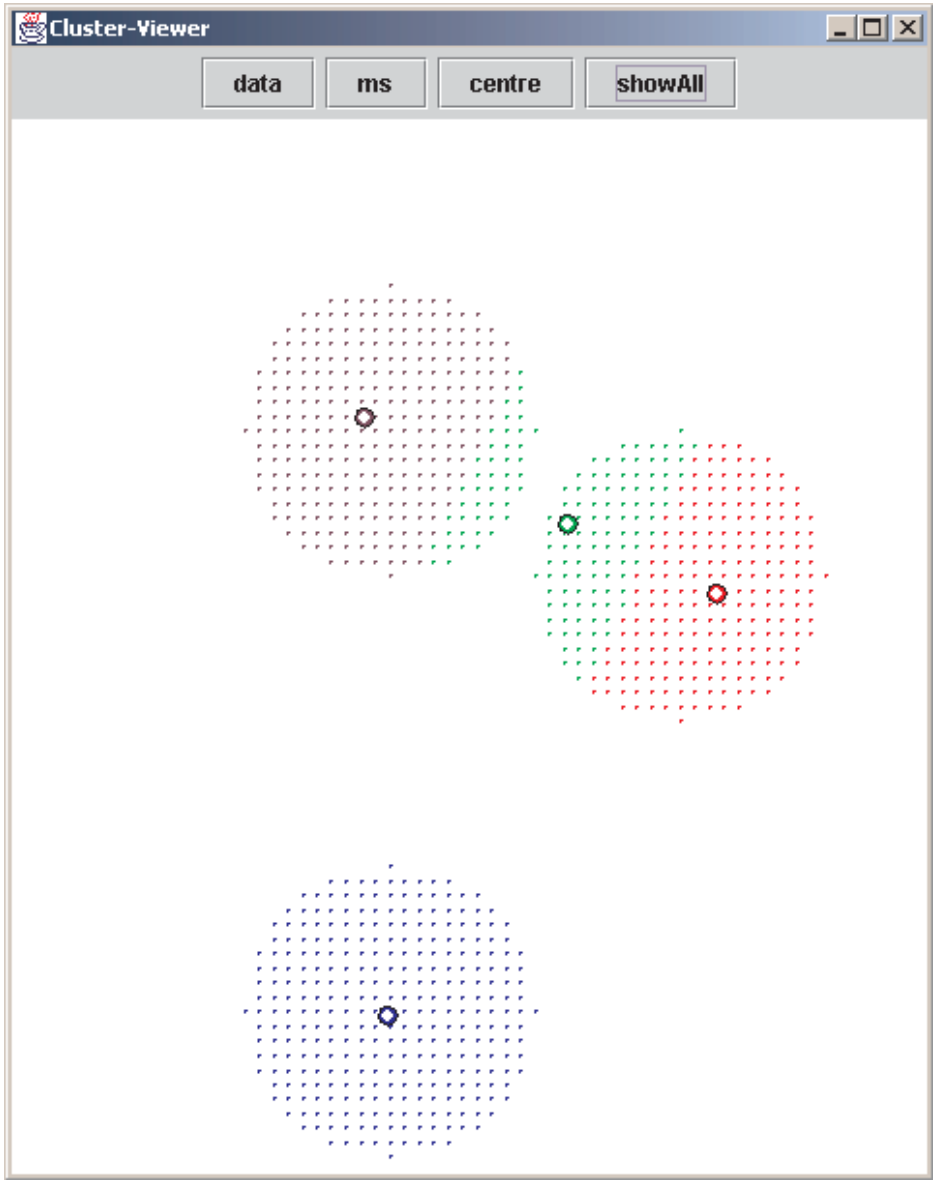




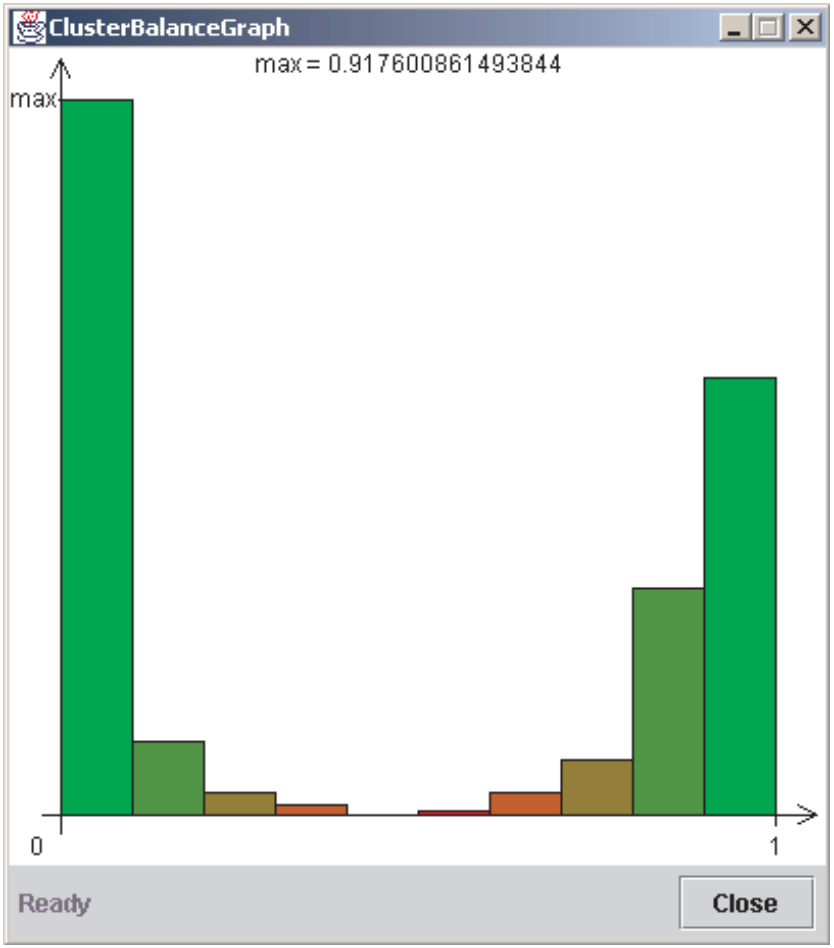
Visualisation: gCLUTO

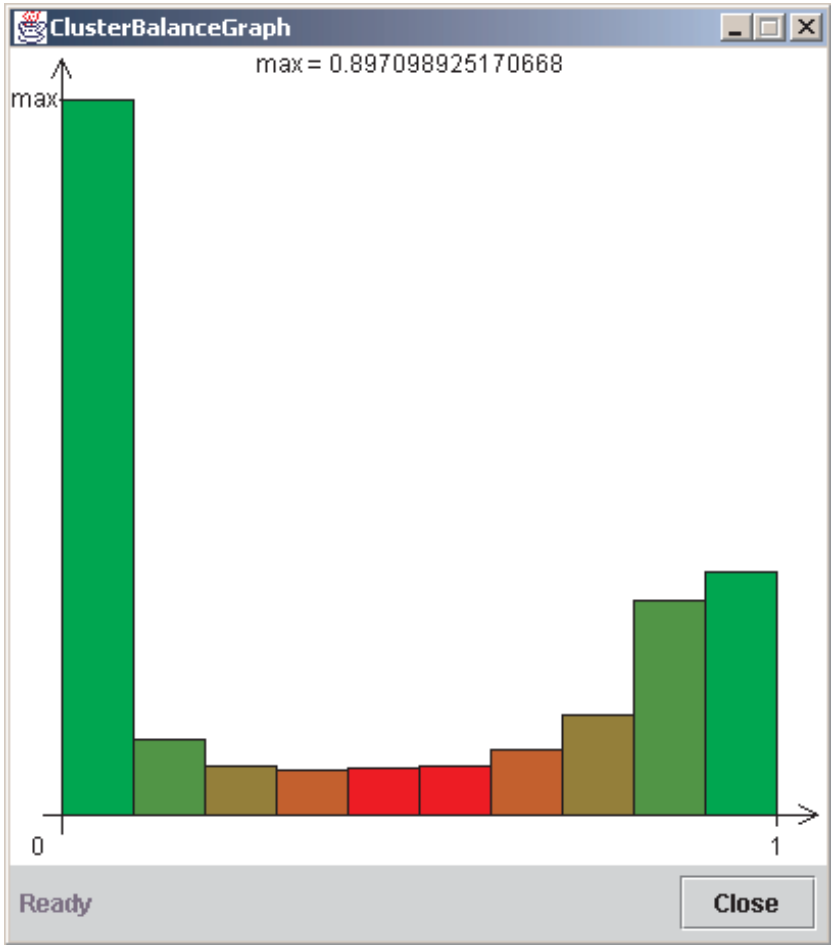
- One peak per cluster
- Distance between peaks corresponds to similarity between clusters
- Height of the peak: Internal similarity of the cluster (average pairwise similarity of the objects in the cluster)
- Volume of the peak proportional to the number of objects in the cluster
- Colour of a peak for internal standard deviation (standard deviation of the pair-wise similarities between the cluster's objects) of the cluster's objects.
  red: low, blue: high

Visualisation techniques

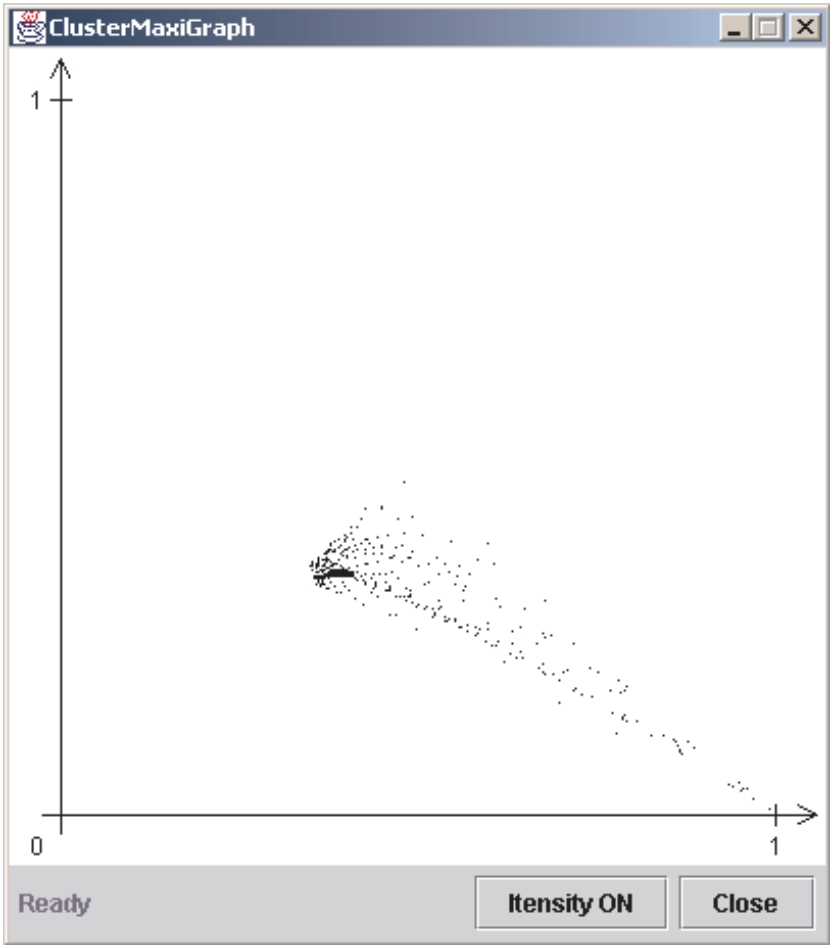


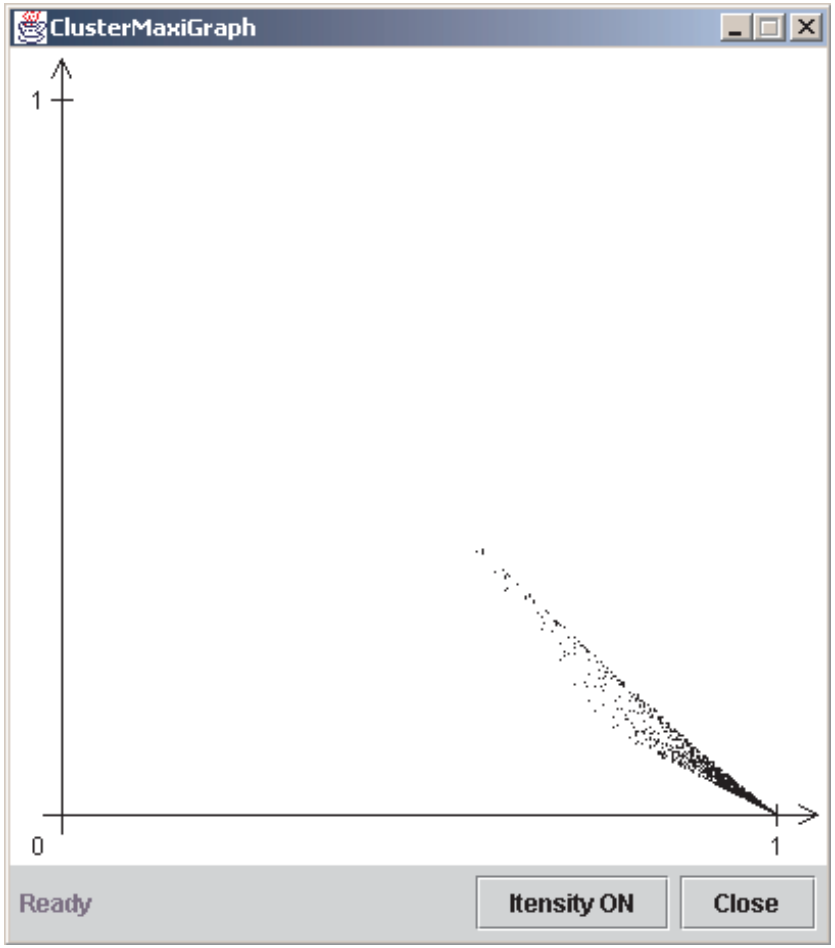
Visualisation techniques



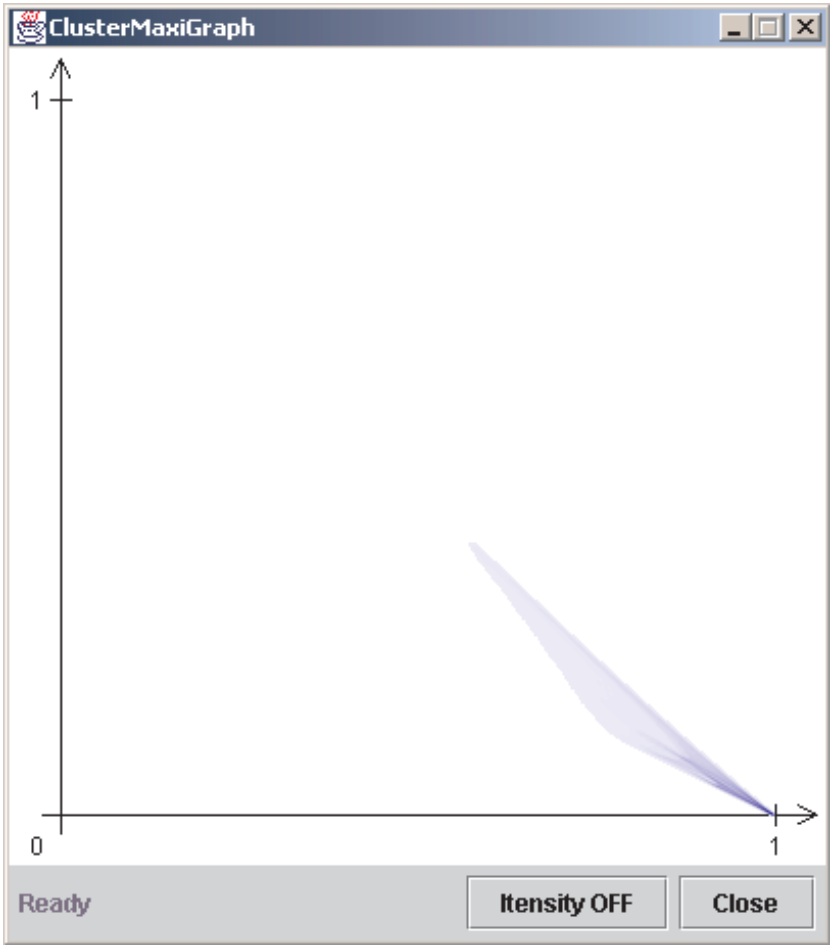


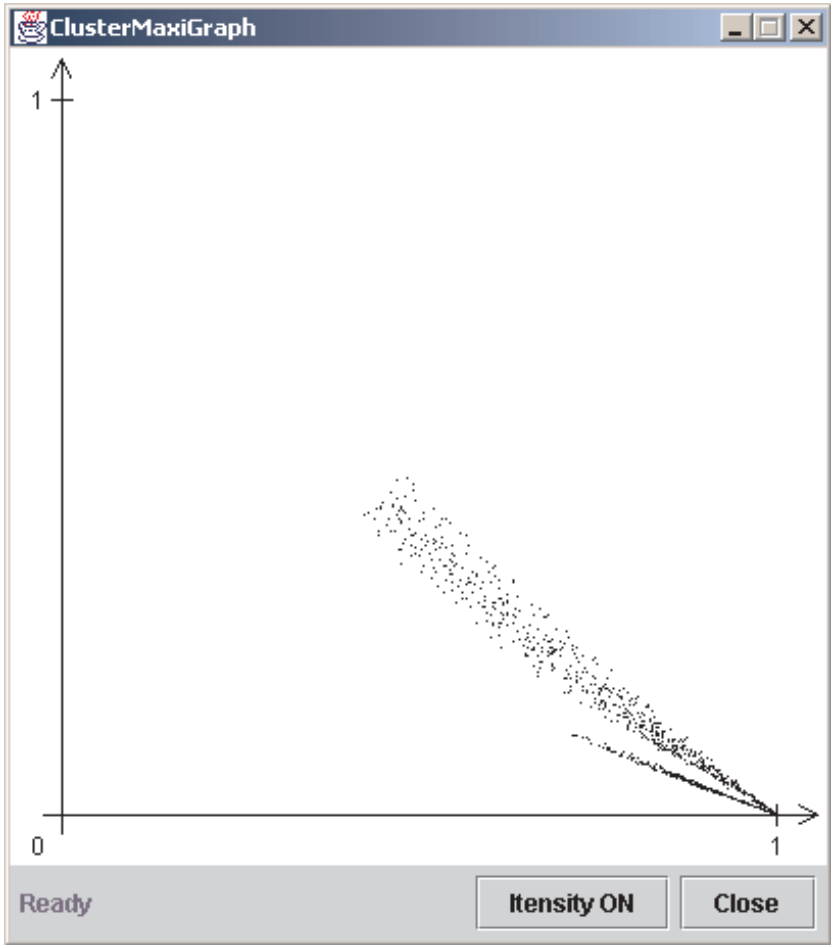
Visualisation techniques



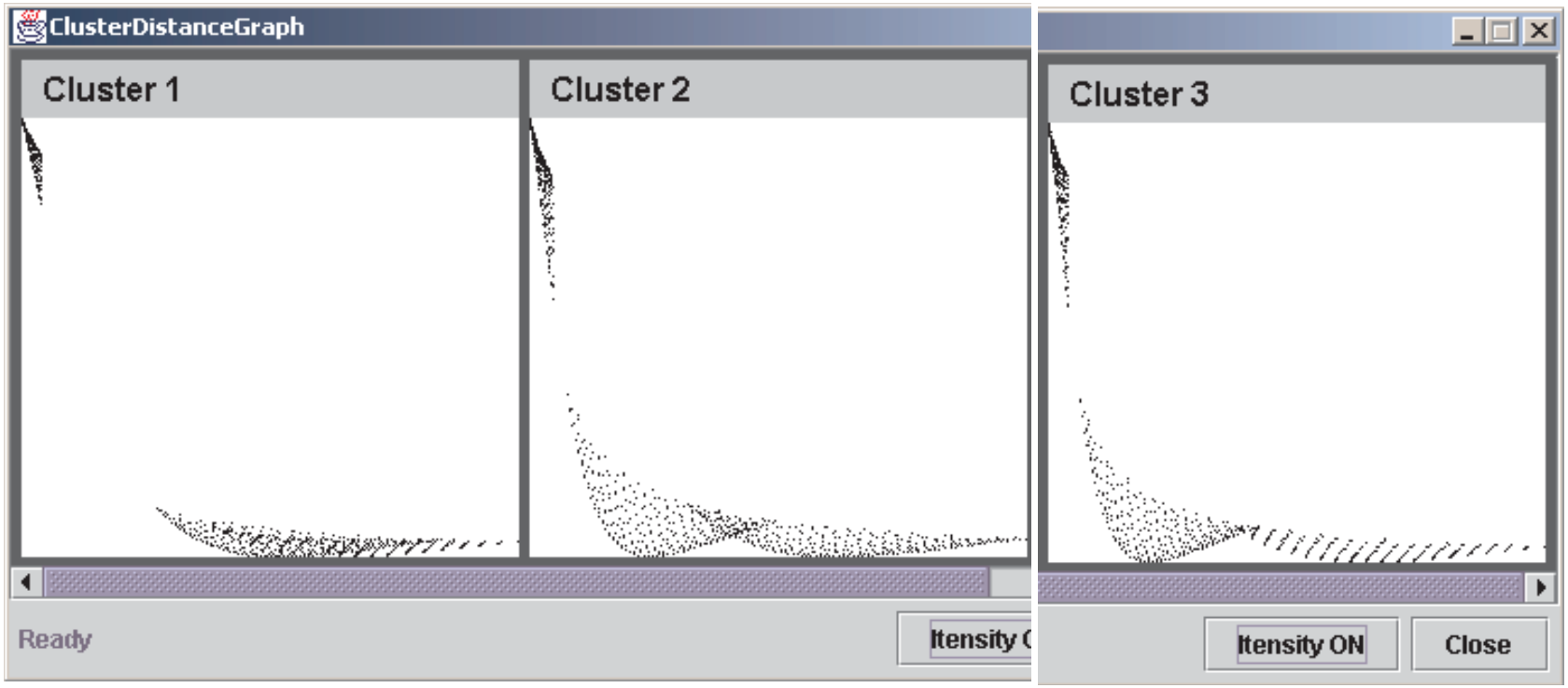


Visualisation techniques

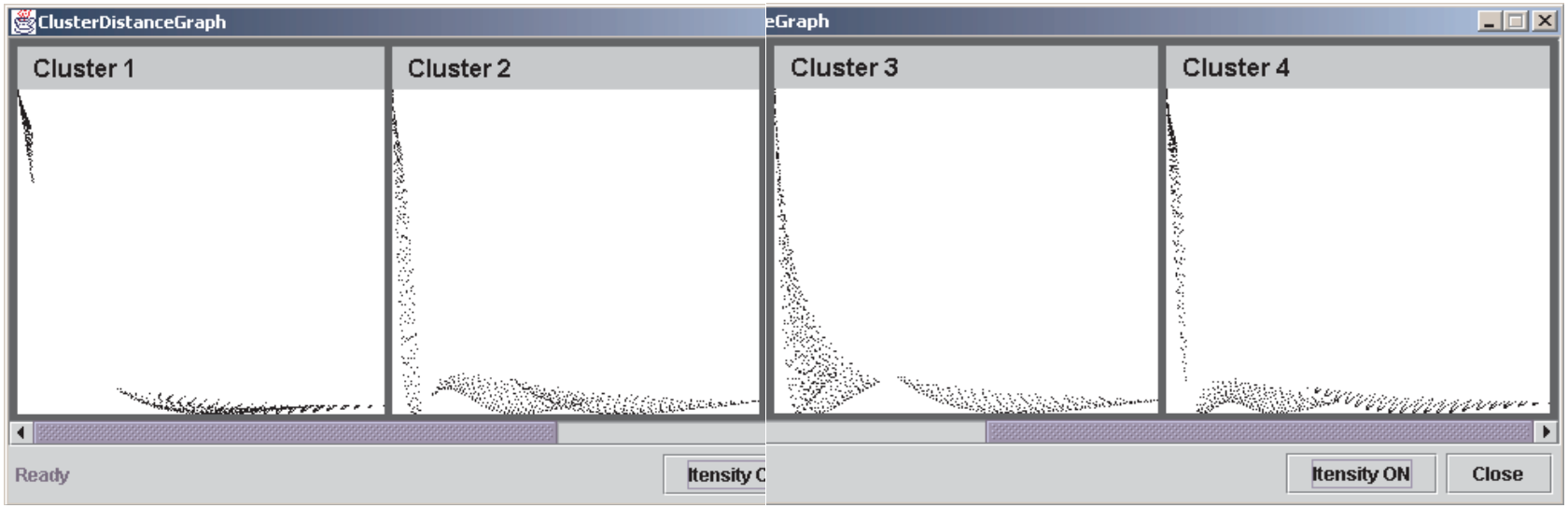




Visualisation techniques



Visualisation techniques



Validity measures

(Global) validity measures provide a numeric value for a clustering result.

The cluster analysis can be repeated with different numbers of clusters and the result with the best value for the validity measure is chosen.

Validity measures exclusively based on membership degrees

Bezdek's partition coefficent

$$I_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \quad \in \left[\frac{1}{c}, 1\right]$$

or its normalised version $1 - \frac{c}{c-1}(1 - I_{PC})$

Validity measures exclusively based on membership degrees

Bezdek's partition coefficent

$$I_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \quad \in \left[\frac{1}{c}, 1\right]$$

or its normalised version $1 - \frac{c}{c-1}(1 - I_{PC})$

Bezdek's (normalised) partition entropy

$$I_{PE} = -\frac{1}{n \cdot \log_2(c)}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log_2(u_{ij})$$

Validity measures based on geometry/topology

Dunn (separation) index

$$I_{Dunn} = \frac{\text{smallest distance between clusters}}{\text{largest distance between objects within the same cluster}}$$

Validity measures based on geometry/topology

Dunn (separation) index

$$I_{Dunn} = \frac{\text{smallest distance between clusters}}{\text{largest distance between objects within the same cluster}}$$

Davies-Bouldin index

$$I_{DB} = \frac{1}{c}\sum_{i=1}^{c}\max_{j\neq i}\frac{\text{dispersion cluster } i + \text{dispersion cluster } j}{\text{distance between clusters } i \text{ and } j}$$

Validity measures based on geometry/topology

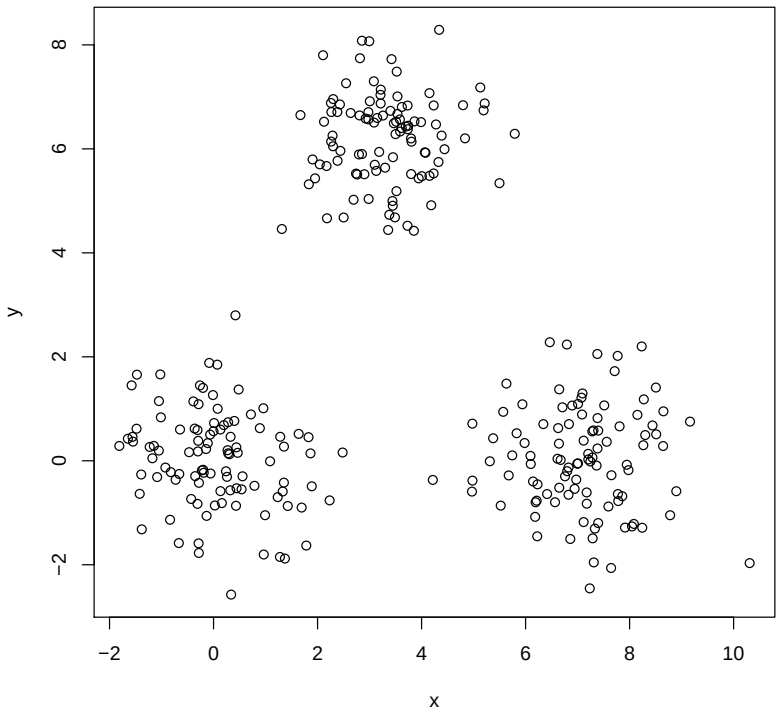
Silhouette index

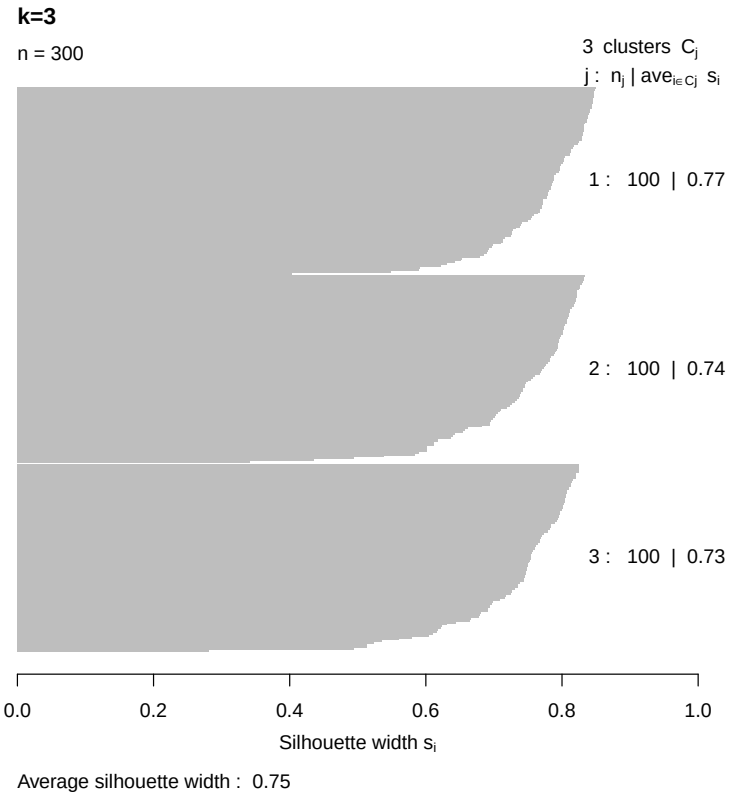
- $a(j)$: Average distance between object j and all other objects in the same cluster
- $b_i(j)$: Average distance between object j and all objects in a different cluster i
- $b(j)$: $\min_i\{b_i(j)\}$
- $s(j) = \frac{b(j)-a(j)}{\max\{a(j),b(j)\}}$

$$I_S = \frac{1}{n}\sum_{i=j}^{n} s(j)$$

Validity measures based on geometry/topology

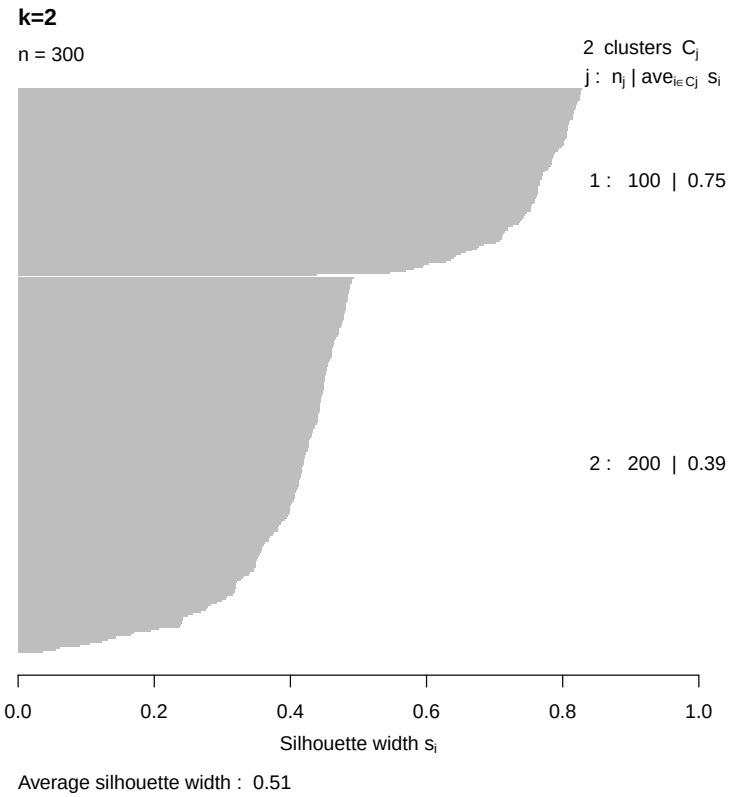
Silhouette index

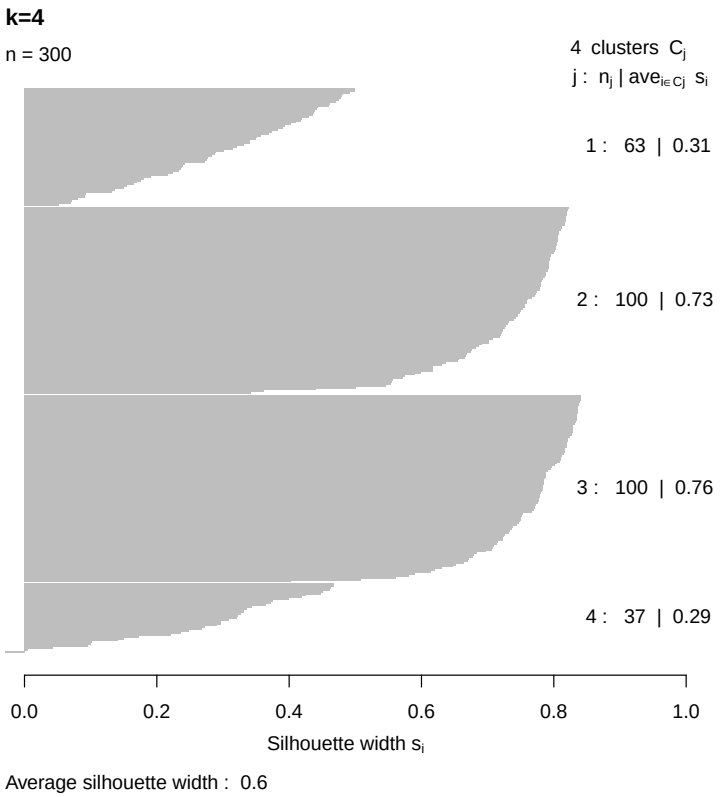




Validity measures based on geometry/topology

Silhouette index





Validity measures based on geometry/topology

Huber's Γ statistic

$$\Gamma = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} d_{jk} \cdot d(v(j), v(k))$$

where

- d_{jk} is the distance between objects j and k and
- $d(v(j), v(k))$ is the distance between the corresponding clusters (cluster centres)

Validity measures based on geometry/topology

Xie-Bieni index

$$I_{XB} = \frac{\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} d_{ij}}{\min_{1\leq i<k\leq c}\{\text{distance between cluster } i \text{ and } k\}}$$

Fukuyama-Sugeno index

$$\begin{aligned} I_{FS} = & \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} d_{ij} \\ & - \sum_{i=1}^{c} (\text{distance between cluster } i \text{ and the centre of the data}) \\ & \cdot \sum_{j=1}^{n} u_{ij}^{m} \end{aligned}$$

Validity measures based on geometry/topology

Average partition density

The (square root of the) determinant of the covariance matrix S_i of a cluster is a measure for the cluster (hyper-)volume.

$$I_{APD} = \frac{1}{c} \sum_{i=1}^{c} \frac{\sum_{j \in I_i} u_{ij}}{\sqrt{\det(S_i)}}$$

where I_i is the set of objects "close" to the cluster centre.

Compatible cluster merging

- Start with a large number of clusters (e.g. k-means, FCM, Gaussian mixture).
- A local cluster index for isolated clusters with a small number of objects is needed.
- A measure to identify similar clusters is needed.
- Discard the small isolated clusters.
- Merge similar clusters together and recalculate the cluster prototypes.
- Cluster again with the recalculated prototypes and the smaller number of clusters.
- Repeat this procedure until no clusters can be merged or removed.

Concluding remarks

- Resampling is a very reliable technique, but with high computational costs.
- It is always assumed that the clustering algorithm has found the best clustering result (best fit for the data) for a given the number of clusters.
  This is not necessarily the case for objective function-based clustering.
- The use of geometrical/topological properties of the data for evaluating the clustering result can be misleading for high-dimensional data.