Statistical Models in Cluster Stability Problem

Z. Volkovich

Software Engineering Department, ORT Braude College of Engineering, Karmiel 21982, Israel

Cluster Analysis

Clustering problems arise in various areas of machine learning, pattern recognition, optimization and statistics. Roughly speaking, clustering is the categorization of items into dissimilar groups, or more specifically, the partitioning of a data into subsets (clusters), with the intention that the data in each subset (ideally) possess some shared characteristic which usually is proximity consistent with certain distance measure. In various cases, cluster analysis is the crucial technique of data analysis, apart from the explicit questions of interest. Fundamental to any analysis of data is the scientific question of interest.

Background

- *Clustering* is the process of identifying natural groupings in the data
- *Unsupervised learning* technique

  No predefined class labels
- *Several implementations*

  R, SPSS, S-PLUS, SAS, MATLAB

Applications

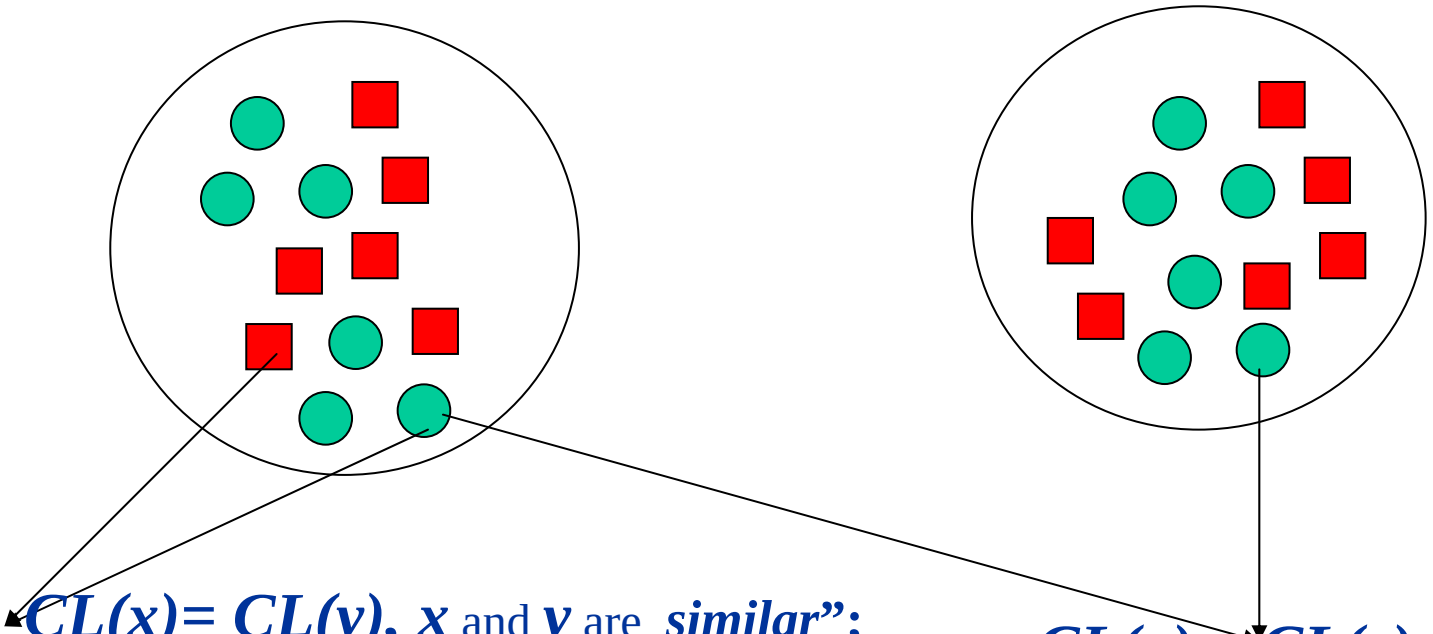
- Pattern Recognition, e.g. handwritten characters
- Speech Recognition
- Image compression (using code-book vectors)
- Texture maps

  Classification of cloud pattern (cumulus etc.)

A similarity measures is a crucial component in cluster problems. It must be well defined according to the actual task.

Clustering Purpose

Partitioning of a set by means of a clustering algorithm ***CL***



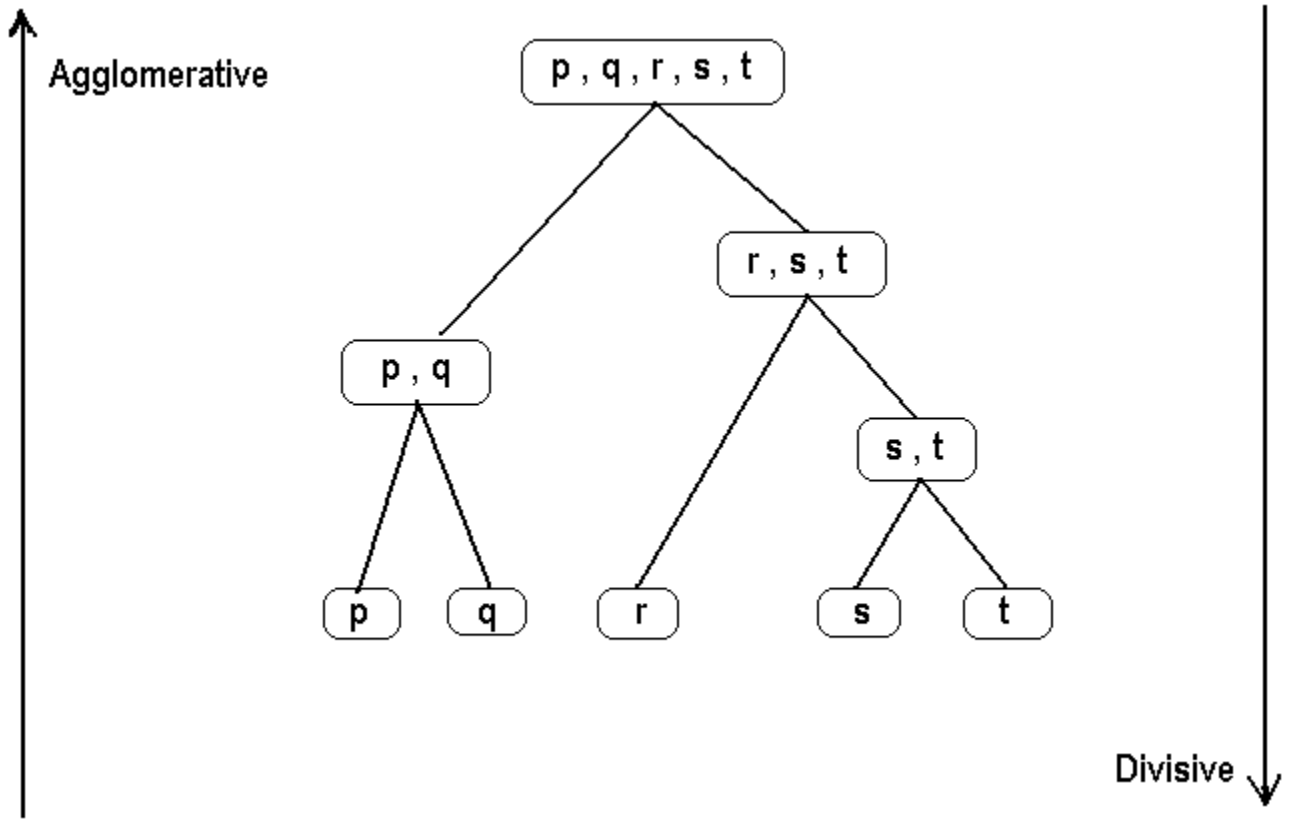
CL(x)= CL(y), x and ***y*** are ***similar"***;

CL(x) ≠CL(y),
x and ***y*** are ***"dissimilar".***

Hierarchical Clustering

Typically, clustering algorithms are either hierarchical or partitional. Hierarchical ones based on the dissimilarity matrix of all pairwise distances seek consecutive clusters by means of formerly constructed clusters. Hierarchical clustering methods fall into two classes: agglomerative nesting methods ("bottom-up") and divisive analysis methods ("top-down"). Agglomerative algorithms start with n singleton clusters (n is the number of items), and merge the closest pair of clusters leaving $(n - 1)$ singleton clusters and one cluster with two jointed objects; and so on until one cluster consisting of all objects remains. Divisive algorithms start opposite with the entire set and continue to split it into successively smaller clusters until n singleton clusters are obtained.

An Example of Hierarchical Clustering



Hierarchical Clustering

The main disadvantage of hierarchical clustering approaches is that a cluster membership of a data element cannot change once it has been put in a cluster. On the contrary, partition methods can recalculate cluster assignments at every iterations.

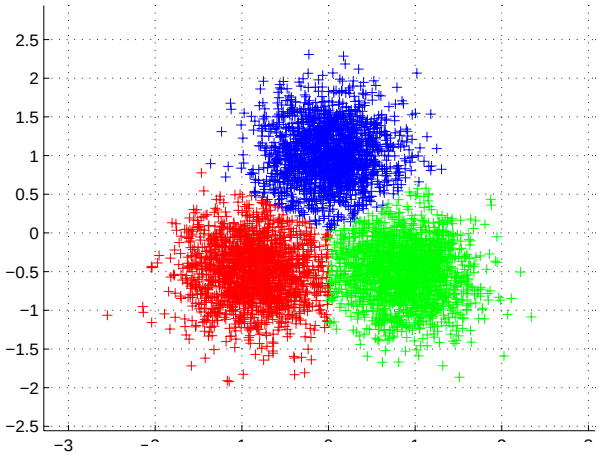
Partition Clustering

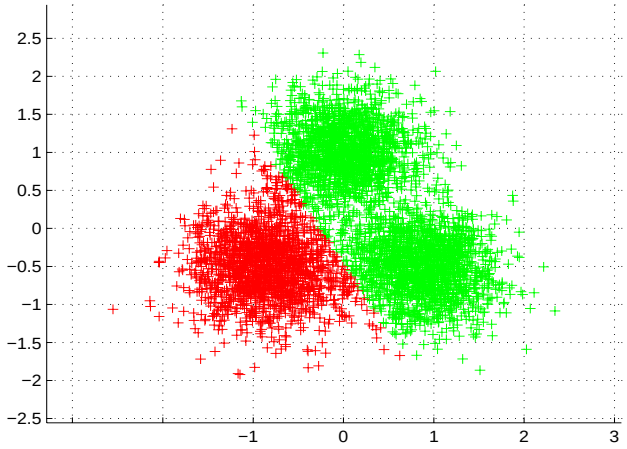
A partitioning iterative clustering process is usually carried out in two phases: a partitioning phase and a quality assessment phase. The partitioning phase is expressed by assigning a label to each element. This label represents the cluster membership. The quality assessment phase measures the partition quality. The outcome of the clustering process is a partition that receives the highest quality score.

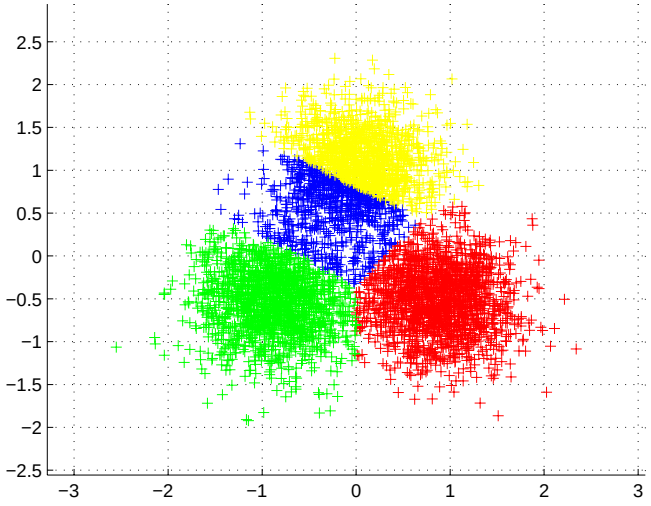
k-means algorithm

The most widespread iterative algorithm - the *k*-means algorithm. It supposes that the clustered objects are drawn form a vector space. The number of clusters *k* is assumed to be much smaller of the items number. This algorithm allocates each item to the cluster having the nearest center (also called centroid). The centroid is the mean of all the cluster points. Indeed, the algorithm attempts to minimizes the total intra-cluster variance. It must be noted that the algorithm does not always achieves a global optimum. The final solution is strongly depending on the initial partition, and can, actually, yield a far from optimal partition. Additionally, the algorithm has the number of clusters *k* is an input parameter. A badly chosen *k* may bear poor outcomes.

Example: a tree-cluster set partitioned by the k-means algorithm into 2 and 4 clusters







The Problem

The problem of an appropriate choice of the number of clusters arise in many attempts to employ the clustering approaches. It is important to note that these procedures require that k be specified in advance. The problem of choosing the number of clusters k is not a trivial one, and in fact there have been many proposed solutions and is known as ”ill posed” connected, for instance, the scale in which the data is measured.

Many approaches have been suggested to handle this problem. So far, none of them has been accepted as superior to the others due to the frequent complexities of clusters’ configurations.

Concept-1

We consider a finite subset $X = \{x_1,\ldots,x_n\}$ in an Euclidian space as a sample drawn from a population having an underlying distribution $f_{X.}$. Consider a partition $\Pi_k = \{\pi_1,\ldots,\pi_k\}$ of the set, i.e.:

$$\bigcup_{i=1}^{k} \pi_i = X;\ \pi_i \cap \pi_j = \varnothing,\ \ i \neq j.$$

and corresponding representation

$$f(x) = \sum_{i=1}^{k} p_i f_i$$

where f_i is the inner distribution of the cluster π_i and p_i is the cluster probability. Our concept suggests to measure the partition quality via the steadiness of the partition expressed by the resistance of the corresponding distributions f_i, $i=1,..,k$ to a rerun of a clustering algorithm CL.

Concept-2

Following the customary statistical scheme we consider a sequence of samples $S_1, S_2, \ldots, S_T$ having size N drawn without replacement from X and autonomously cluster them by the algorithm Cl.

According to our concept, the occurrences of the samples in the clusters

$$X_{ij} = \pi_i \cap S_j \ , \ i = 1, \ldots, k, \ j = 1, .., N$$

appear to be for each $i=1, \ldots, k$ as if they were a sample chosen independently from the same population.

Concept-3

The most appropriated statistical tool applicable to this issue is two sample test statistic. Recall, that two sample tests are intended to test the null hypothesis which suggests that the elements of two considered samples have been drawn from the same distribution. Actually, we can compare here pairs of samples in the purpose to test if the samples have been drawn from the same distribution. However, several problems arise in this straightforward approach.

Applicable two sample tests

No prior knowledge of the data distribution is available. Thus, an applied two sample test has to be distribution-free two-sample test. The Kolmogorov-Smirnov test, the Cramer-von Mises test, the Friedman's nonparametric ANOVA test and the Wald-Wolfowitz test must be reminded as the classical uni-variate procedures for this purpose. The following multi-variate tests appear to be most appropriated to be used in the cluster context.

- The Friedman-Rafsky test, 1979. (Minimal spanning tree based);
- The Nearest Neighbors test by Henze, 1988;
- The Baringhaus- Franz test 2004;
- The energy test, Zech and Aslan 2005;
- The N-distances test by Klebanov, 1989, 2005.

The Baringhaus- Franz test and the energy tests are, indeed, variants of the Klebanov test.

Difficulties

Outliers in the samples and the limitations of clustering algorithms heavily contribute to the noise level. To overcome this difficulty, we have to extract our conclusions based on a sufficiently large amount of information. In other words, we repeat this procedure many times. Another difficulty results from the well known fact that the concentration of the distance distribution increases as a function of the examined number of clusters. To neutralize this effect it is common to normalize the obtained distance values.

Implementation

Thus, the process can roughly be described as the generation of an empirical distance's distribution, including normalization, and then the testing the distribution concentration at zero. This test can be provided by means of several known statistics like the sample mean and the size of the sample first quartile.

Friedman-Rafsky's MST two sample test statistic

We measure the dissimilarity by means of the *Friedman-Rafsky's* Minimal Spanning Tree (*MST)* two sample test statistic. Recall that, two sample tests are intended to test the null hypothesis which suggests that the elements of two considered samples have been drawn from the same distribution. Obviously, this is not the case here. However, we presume that, inside stable well-defined clusters, the elements of the two samples are mingled as if they were selected from close distributions.

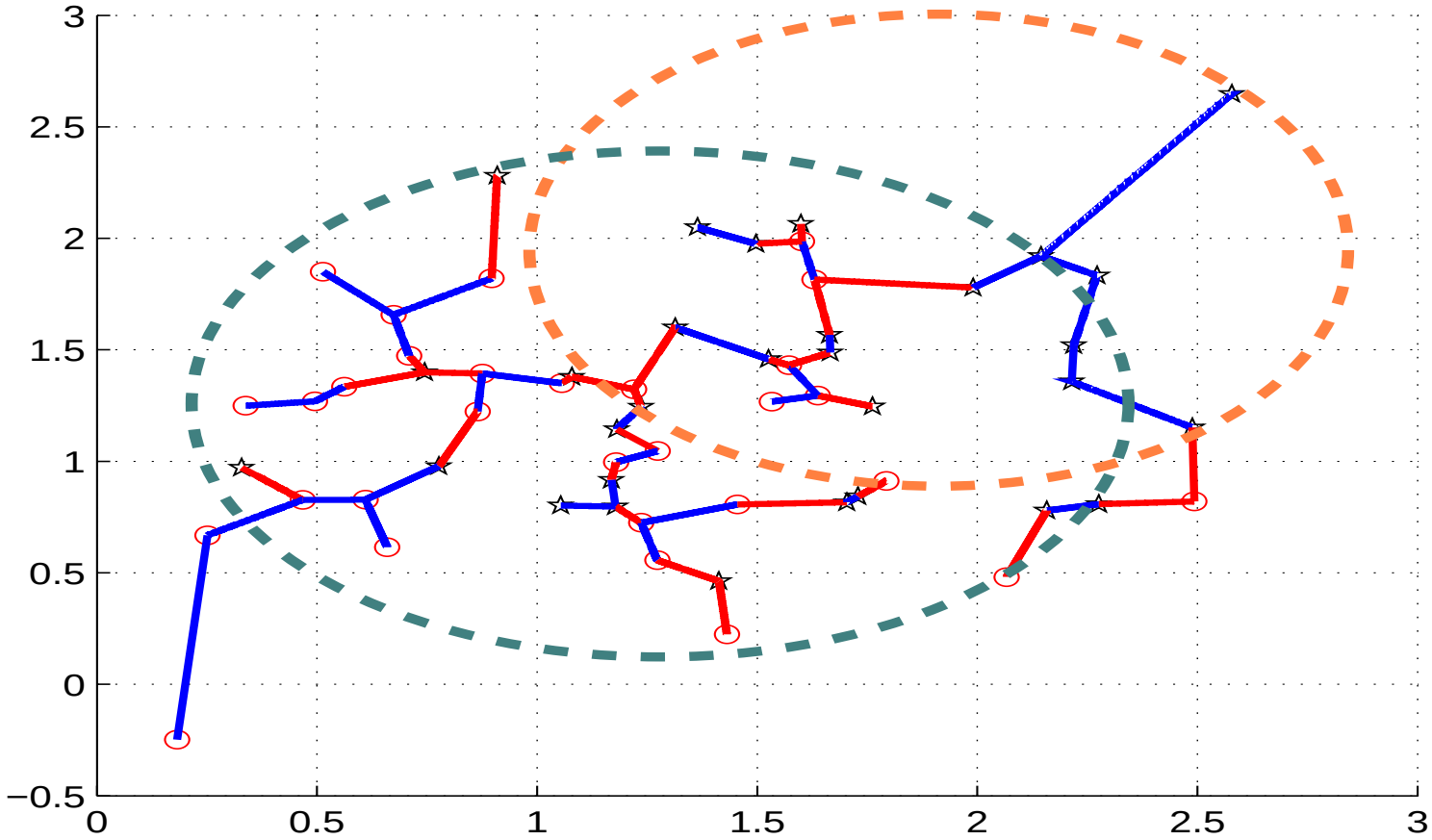
Friedman-Rafsky's MST two sample test statistic (cont.)

For given samples S and T, the *Friedman-Rafsky's MST test* considers an *MST* created for the pooled set $V=S \cup T$. If all distances between elements of V are distinct, the set is called nice. In this case, there is only one *MST* that connects all points of V, such that sum of the lengths of the edges is minimal.

Friedman-Rafsky's MST two sample test statistic

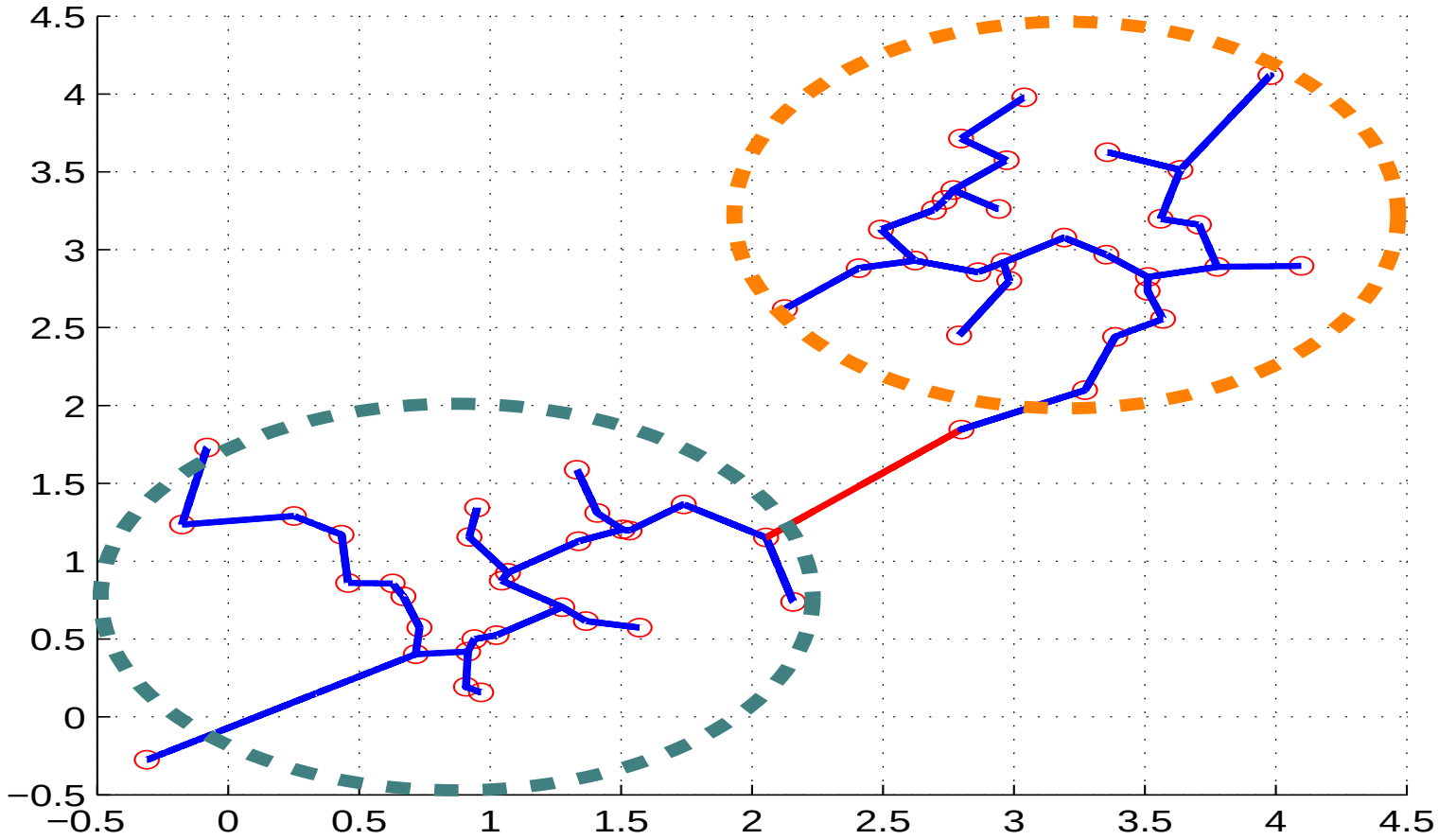
An *MST* can be built in $O(|V|^2)$ time, including distance calculations, using the well known Prim's, Kruskal's, Boruvka's or Dijkstra's algorithms. The Friedman and Rafsky's test statistic is defined as the number of edges of the MST which connects elements from different samples. Under the null hypothesis, this statistic is normally distributed. This result coincides with our intuition since if two sets are closed then there are many edges which unit points from different samples. In the spirit of *The Central Limit Theorem,* this quantity is expected to be normally distributed.

A graphical illustration. Stable



The edges connecting points from different samples are marked by red.

A graphical illustration. Non-stable



The edges connecting points from different samples are marked by red.

Route

To implement the method, for each possible number of clusters $k=2,\ldots,k^*$, we draw many pairs of disjoint samples

$$S_1^{(m)},\ S_2^{(m)},\ m = 1,\ldots,M\ .$$

having the same sufficiently big size (n) according to the cluster core Densities and the underlying density and calculate

$$S_{1l}^{(m)} = S_1^{(m)} \cap \pi_l^{(m)}(S^{(m)}),\quad S_{2l}^{(m)} = S_2^{(m)} \cap \pi_l^{(m)}(S^{(m)})$$

as subsets of the samples respectively belonging to each cluster.

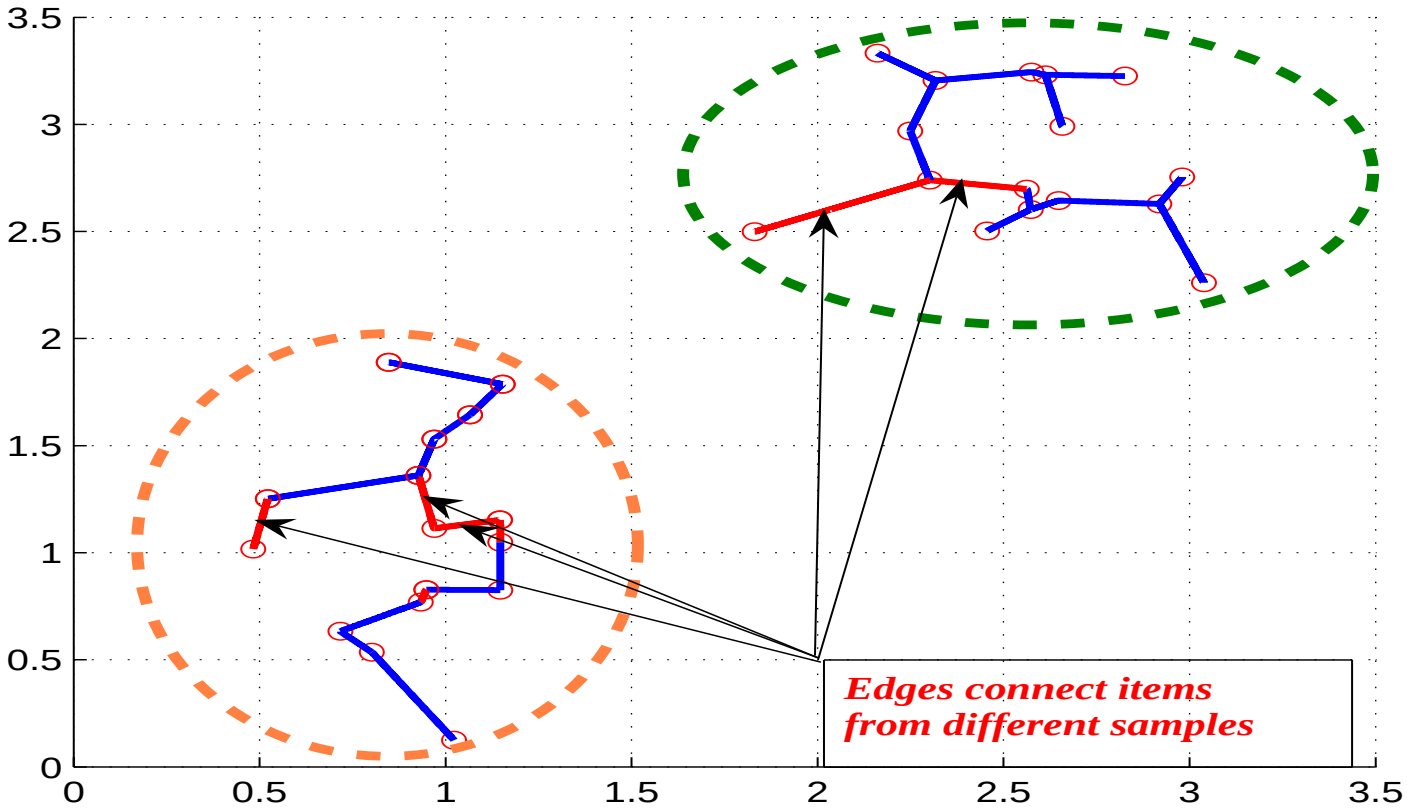
The test statistic

For each cluster $l=1,\ldots,k$ we compute a value of the two-sample MST-test statistics $r_l^{(m)}$ and

$$R^{(m)} = \min\{r_l^{(m)} \mid l = 1,2,\ldots,k\}$$

We would like to note that this value presents the minimal number of edges connected the samples within the clusters. Using of such a characteristic appears to be very natural in the light of the considered early examples. Small values of $R^{(m)}$ can witness about samples separation inside a cluster.

Example: Calculation of



$$R=min(3,2)=2;\ n=10,\ k=2$$

Route

Thus, under the mentioned assumptions, the random variable $R^{(m)}$ is distributed as the minimal value of k normal i.i.d variables. In order to estimate the mean μ and the variance σ^2 of this distribution we calculate the average of $r_l^{(m)}$ $(l=1,\ldots,k)$:

$$T^{(m)}=mean(r_l^{(m)},\ l=1,\ldots,k).$$

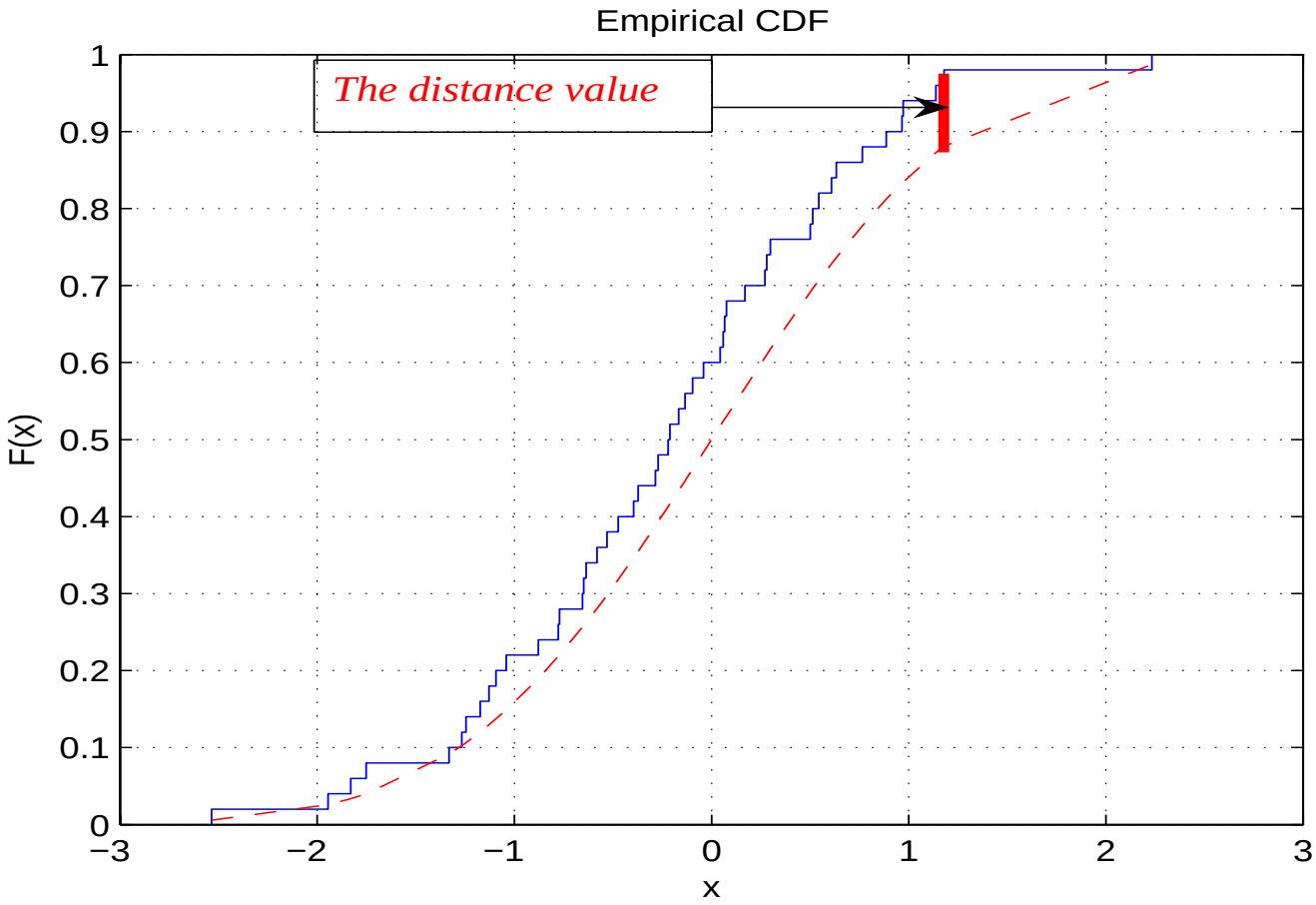
From the M averages, $T^{(m)}$, $m=1,\ldots,M$ their mean and their variance are obtained. For this end we substitute:

$$\mu=mean(T^{(m)},\ m=1,\ldots,M\ ),$$
$$\sigma^2=var(T^{(m)},\ m=1,\ldots,M\ ).$$

Routine

Then, we apply a one-dimensional two sample test statistic to assess the distance between empirical distributions of $\{ R^{(m)} \}$ and $\{ Y^{(m)} \}$. This can be done, for example, by means of the famous Kolmogorov–Smirnov distance, often called the *K-S* distance.

The Kolmogorov-Smirnov Distance



Remarks

1. We indirectly assume that the sets $S_{1l}^{(m)}$ and $S_{2l}^{(m)}$ are permanently nice. This suggestion holds for almost all real datasets.
2. The presence of clusters having the same geometrical structure and size could be rarely guarantied for real datasets. However, we are still interested to judge from this point of view. Apparently, we can consider the proposed distribution model as a theoretical etalon characterizing a stable partition situation.

Remarks

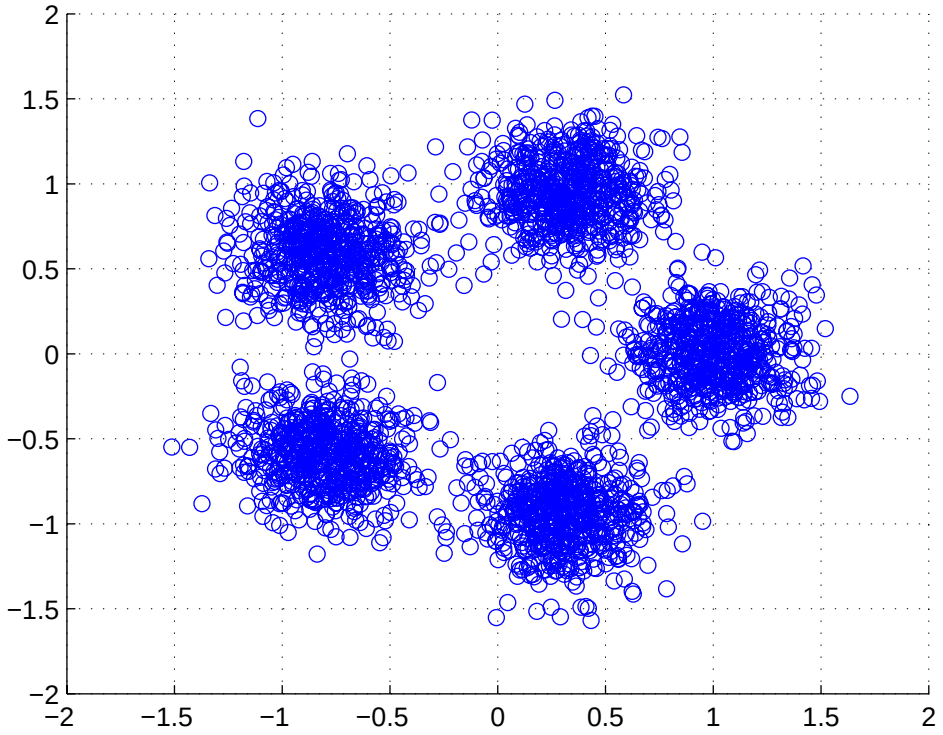
3. The idea to explore differences between the high and low density regions by the *Friedman-Rafsky's MST* is presented in (Smith et al., 1984) and (Jain et al., 2002). The goal was to point out an "inconsistent" edge whose length is significantly larger than the average length of the nearby edges. We apply the MST two sample test statistic in the "departure from normality" manner that allows to attain more detailed results about the cluster structure.

Remarks

4. The high and low density regions have been characterized in the papers by (Smith and el., 1984) and (Jain et al., 2002) via the MST which appears to be computational costly for big datasets. We use a variant of the k-nearest neighbors approach for the underlying density estimation.

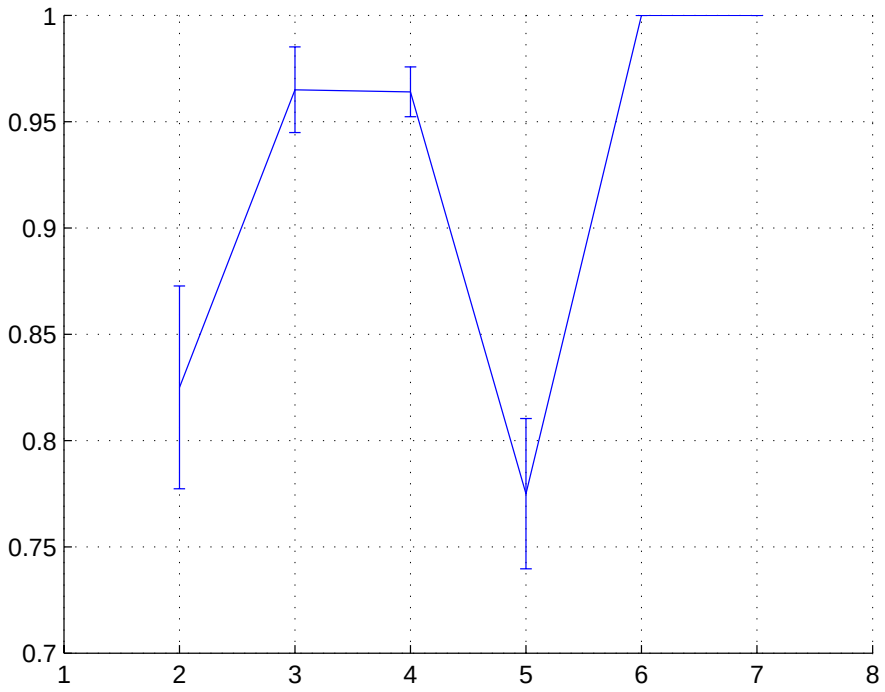
Example 1: 5 components Gaussian dataset

This set of 3000 items is composed of 5 Gaussian components centered at points located in the unit circle with $\sigma=0.2$.



5 components Gaussian dataset (cont.)

Error-bar plot of the K-S distance of the Gaussian dataset among *10* trials. Sample size is *300*. Number of samples pairs: *M=100* .



Example 2: A three text collection set

These datasets is chosen from

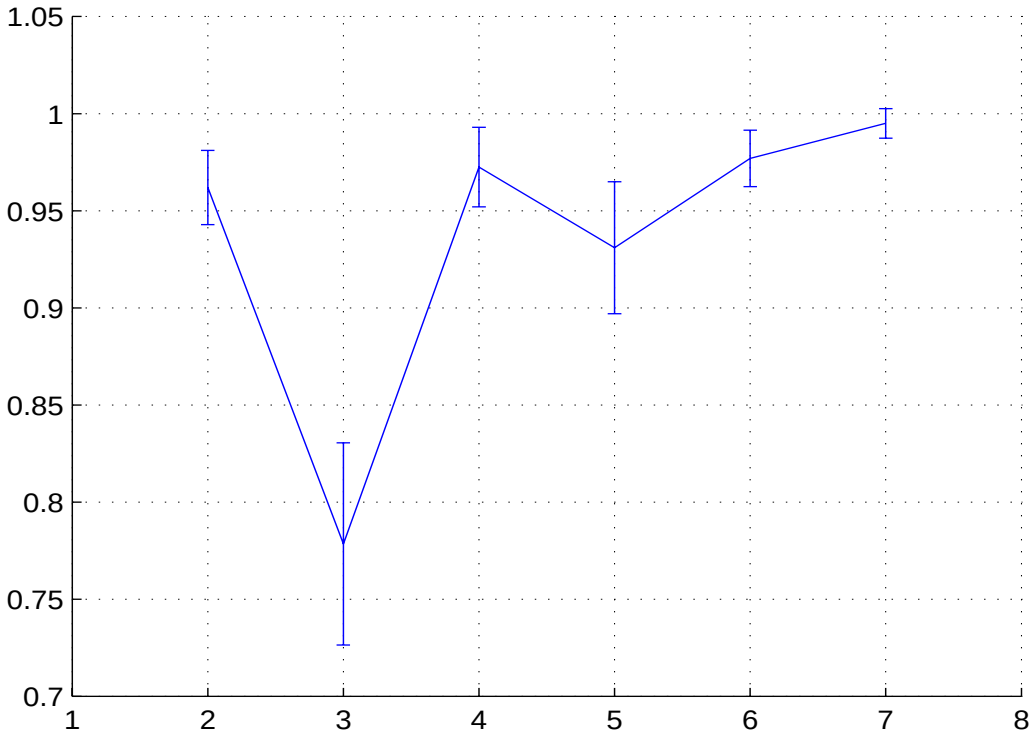
http : *//www.dcs.gla.ac.uk/idom/ir resources/test collections/*

and consists of the following three text ollections

- DC0–Medlars Collection (1033 medical abstracts).
- DC1–CISI Collection (1460 information science abstracts).
- DC2–Cranfield Collection (1400 aerodynamics abstracts).

A three text collection set-600 terms

Error-bar plot of the K-S distance for the three text collection dataset among 10 trials.



Example 3: The Iris Flower Dataset

The well known Iris Flower Dataset is available at *http ://fmwww.bc.edu/ec − p/data/micro/iris.dta*.

This dataset contains features for three different classes of flowers:

0 - Iris Setosa;

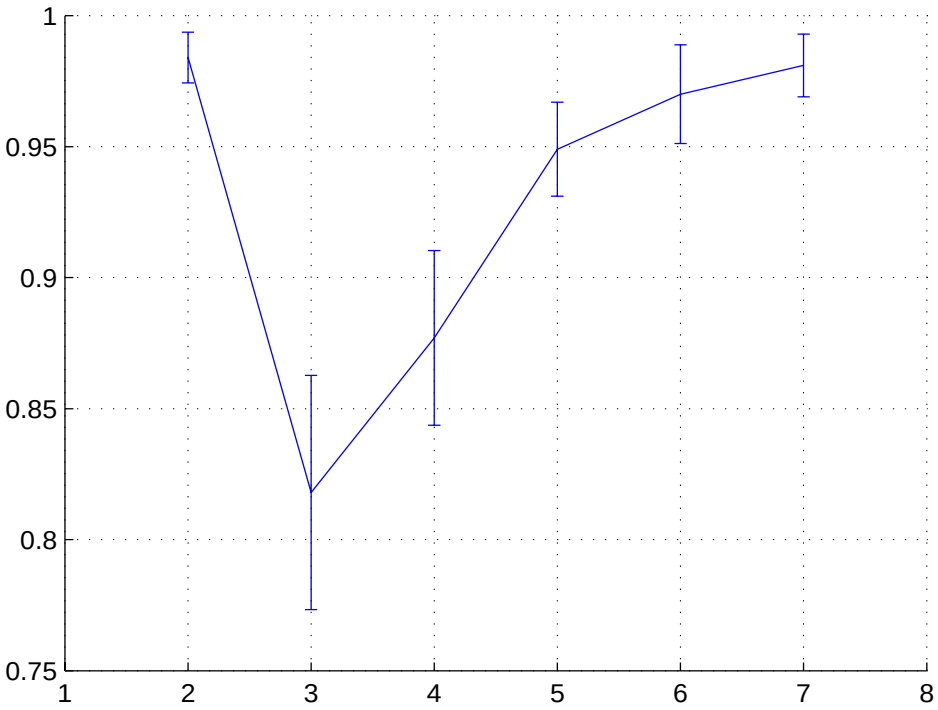
1 - Iris Versicolour;

2 - Iris Virginica.

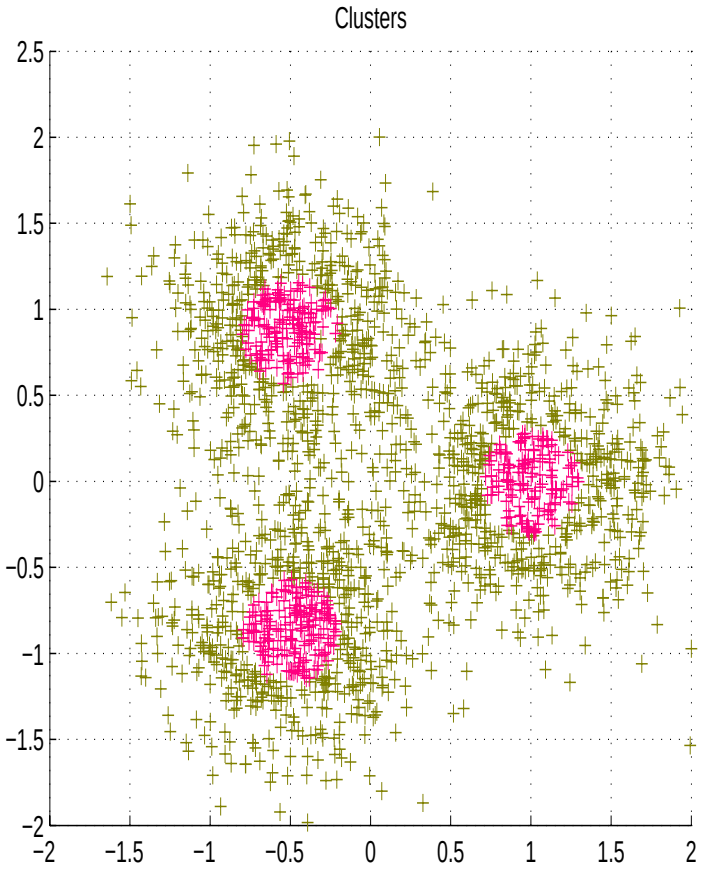
There are 50 examples for each class in the dataset. Each example has a four dimensional feature vector. It is well known that one data set is linearly separable from the others while the other two are not.

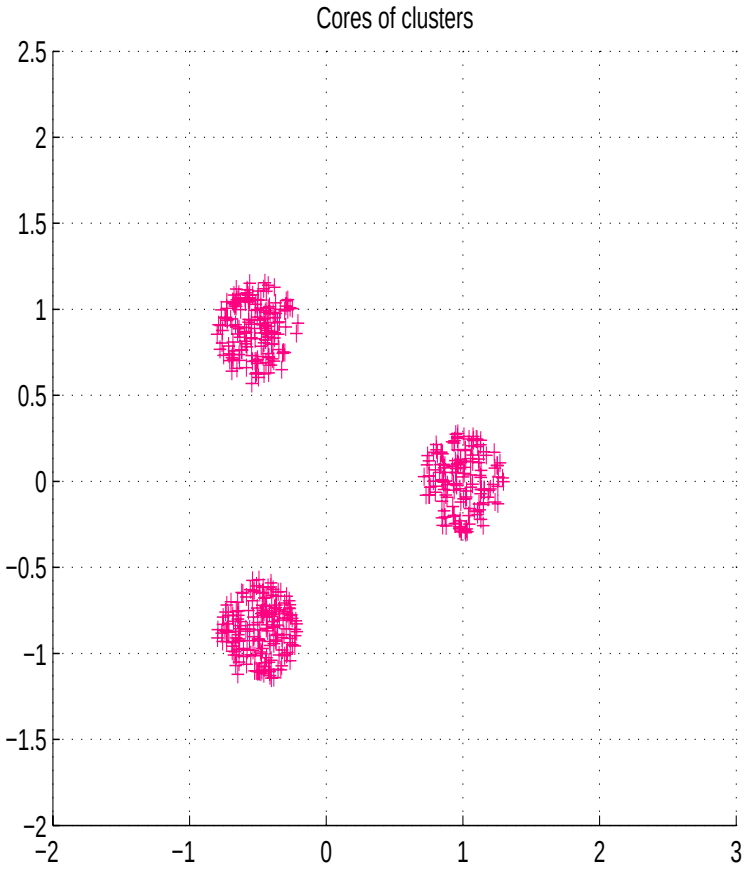
The Iris Flower Dataset (cont.)

Error-bar plot of the K-S distance the Iris dataset among 10 trials. Sample size: *70*. Number of samples pairs: *M=100*.



Clusters and their cores





Clusters and their noisy versions for different level of noise