Spectral Clustering : principle and perspectives

Sandrine Mouysset

ENSEEIHT-IRIT : APO Team

Gene around the world at CERFACS

Spectral Clustering

- One of the most important method in unsupervised classification
- Applications in a large variety of fields :
  - Biology
  - Information retrieval
  - Image segmentation...

Spectral Clustering

1. Introduction
2. Principle
   - The 'ideal' case
   - Algorithm
   - Clustering Challenges
3. A "good" affinity matrix : rule of σ
4. Quality measures
   - Ratio of Frobenius norms
   - Percentage of misclustered points
5. Perspectives

Introduction

- Objectives :

  Partition a $m \times n$ data set in K clusters in order to have lower within-cluster distances and larger between-clusters distances

- Difficulties :
  - Determine the number of clusters
  - Define the clusters

Introduction

- Data matrix X : $X = [x_{ij}] \in \mathcal{M}_{m,n}(\mathbb{R})$
- Affinity matrix : $A = [a_{ij}] \in \mathcal{M}_{m,m}(\mathbb{R})$:

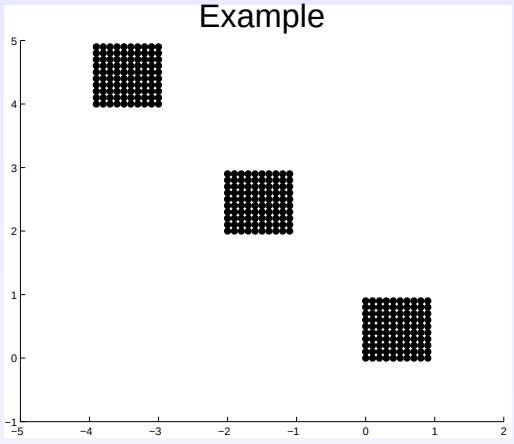
$$\begin{cases} \forall\ (i,j) & a_{ij} = a_{ji} \\ \forall\ (i,j) & 0 \leq a_{ij} \leq 1 \\ \forall\ (i,j) & a_{ij} = 0 \Leftrightarrow i = j \end{cases}$$

- Given a number of clusters : k

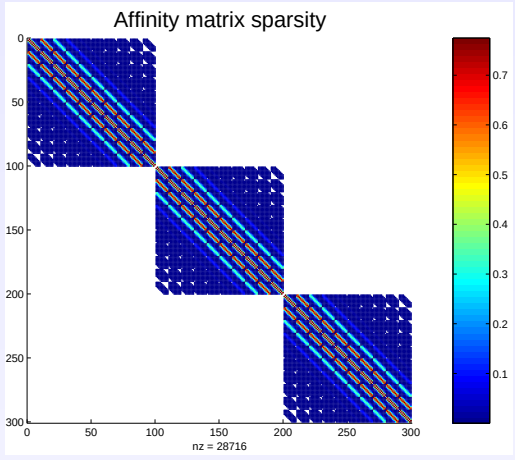
The 'ideal' case

- 3 well-separated clusters : X_1, X_2 and X_3 with respective sizes n_1, n_2, n_3.
- Near block-diagonal affinity matrix : $\hat{A} \approx \begin{bmatrix} A^{(11)} & 0 & 0 \\ 0 & A^{(22)} & 0 \\ 0 & 0 & A^{(33)} \end{bmatrix}$
- $\hat{L}$ normalized matrix $\hat{A}$: $\hat{L}$ row-stochastic block-diagonal.
- $\lambda = 1$ eigenvalue of $\hat{L}$ with multiplicity 3.
  Let u_1, u_2 and u_3 the associated eigenvectors.
- $\hat{Y}$ normalized matrix of $U = \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix}$
- By orthogonal transformation, $\hat{Y} = \begin{bmatrix} \overbrace{\vec{1}}^{1/\sqrt{n_1}} & \overbrace{\vec{0}}^{1/\sqrt{n_2}} & \overbrace{\vec{0}}^{1/\sqrt{n_3}} \\ \vec{0} & \vec{1} & \vec{0} \\ \vec{0} & \vec{0} & \vec{1} \end{bmatrix}$

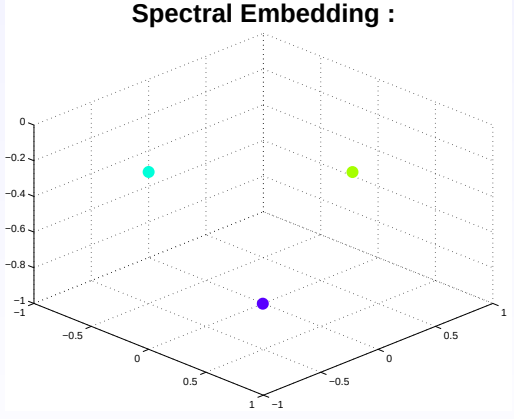
The 'ideal' case : an example



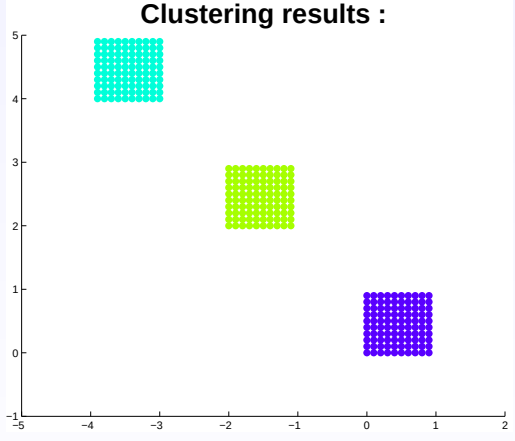
m=300



Near block-diagonal affinity matrix



$\hat{Y}$'s rows



3 well-separated clusters

Algorithm *Ng, Jordan and Weiss*

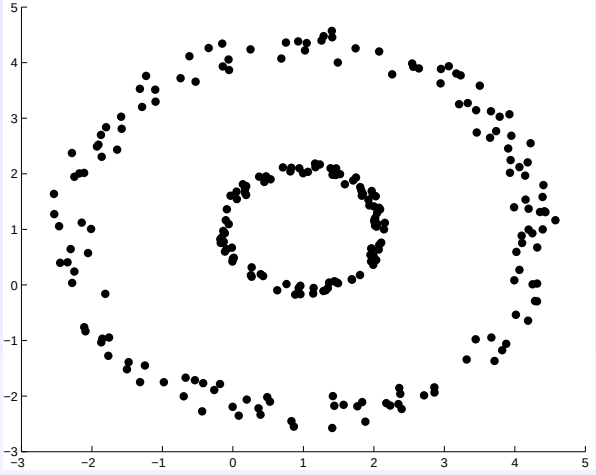
- Form the affinity matrix $A \in \mathbb{R}^{m \times m}$ defined by :

$$A_{ij} = \begin{cases} exp(-\|x_i - x_j\|^2 / 2\sigma^2) \text{ if } i \neq j, \\ 0 \text{ otherwise} \end{cases}$$

- Construct the normalized matrix : $L = D^{-1/2} A D^{-1/2}$ with $D_{i,i} = \sum_{j=1}^{m} A_{ij}$
- Construct the matrix $X = [x_1 x_2 .. x_k] \in \mathbb{R}^{m \times k}$ by stacking the k "largest" eigenvectors of L.
- Form the matrix Y by normalizing each of the X's rows
- Treat each row of Y as a point in $\mathbb{R}^k$ and cluster them in k clusters via *K-means* method
- Assign the original point x_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

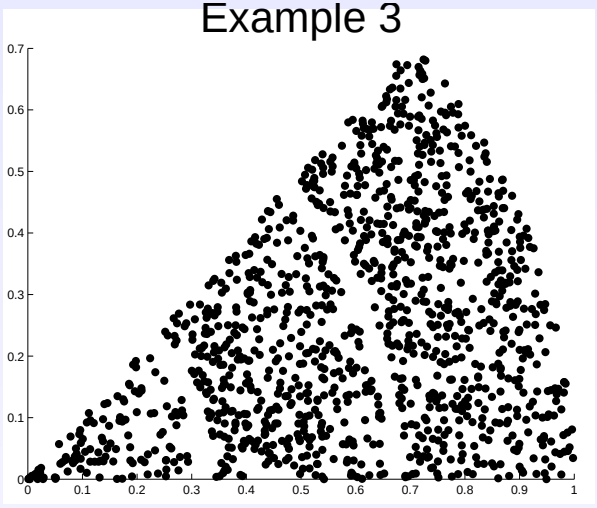
Clustering Challenges

Example 1



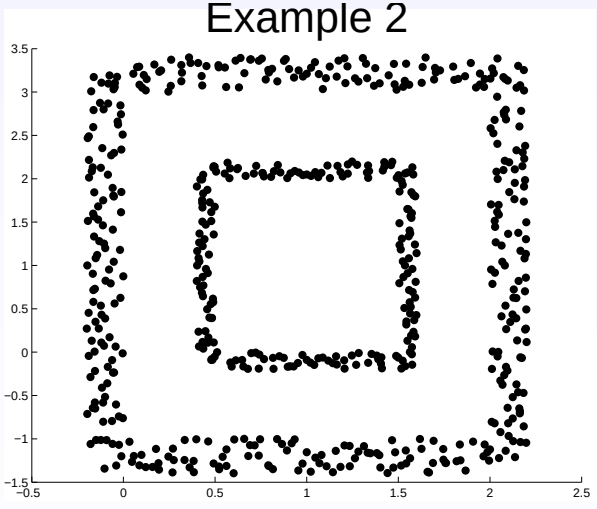
$m = 250$

Example 3



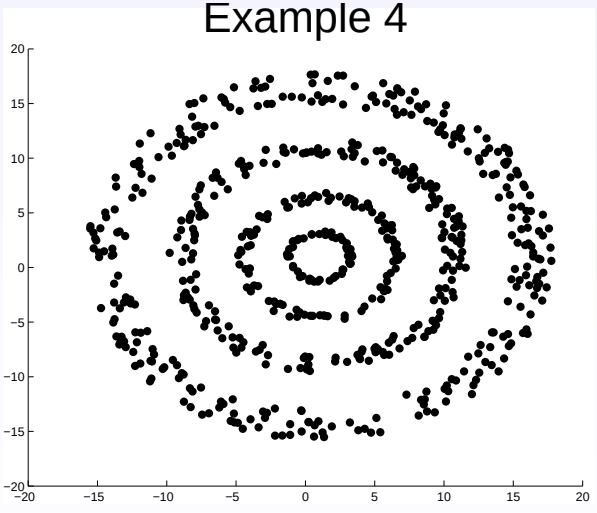
$m = 1200$

Example 2



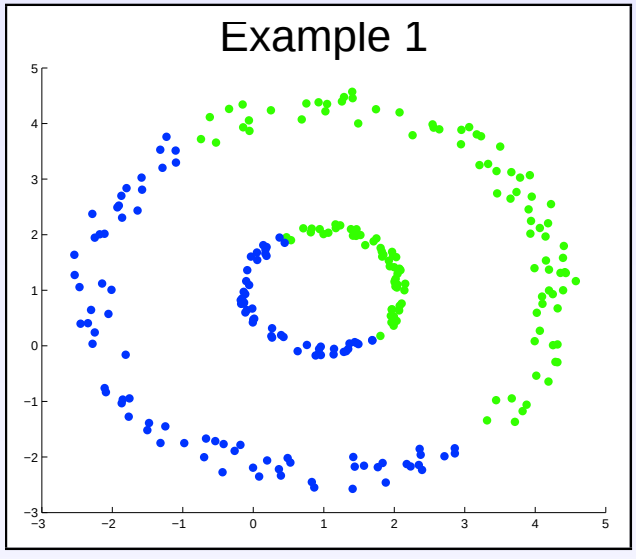
$m = 600$

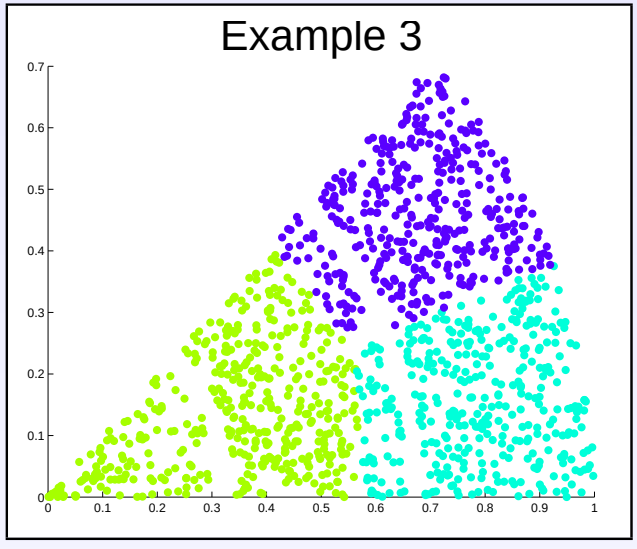
Example 4

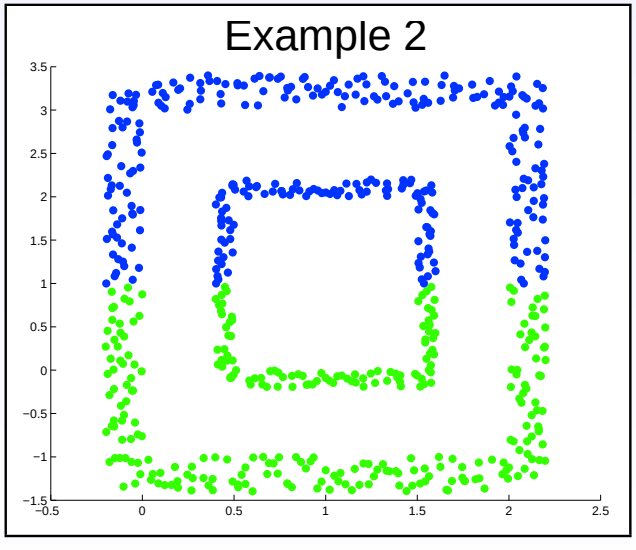


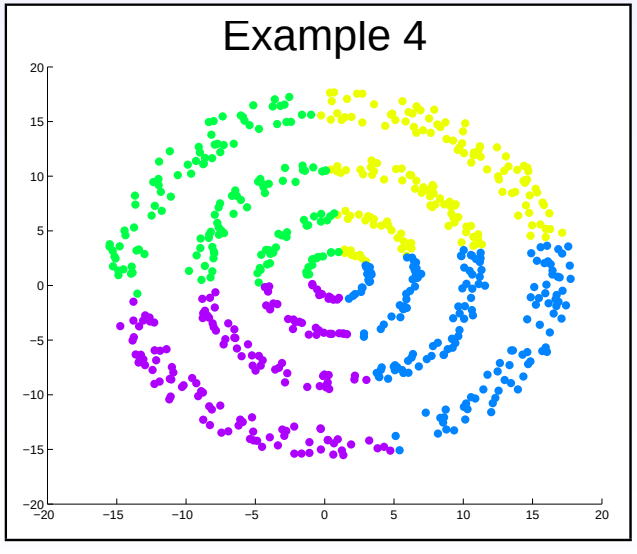
$m = 650$

Clustering Challenges : *K-means* methods applied directly to the original data set



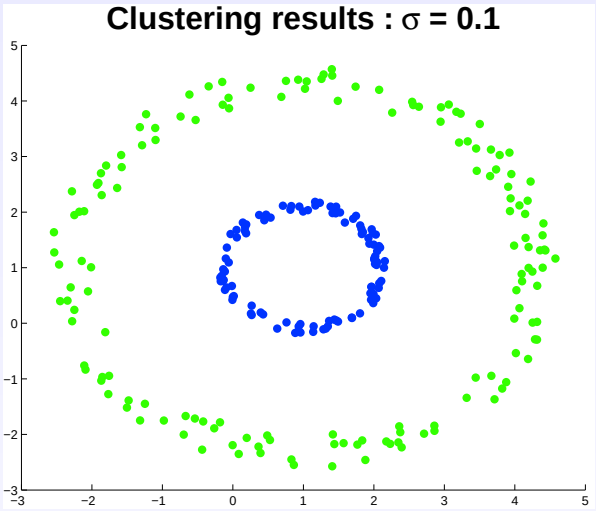


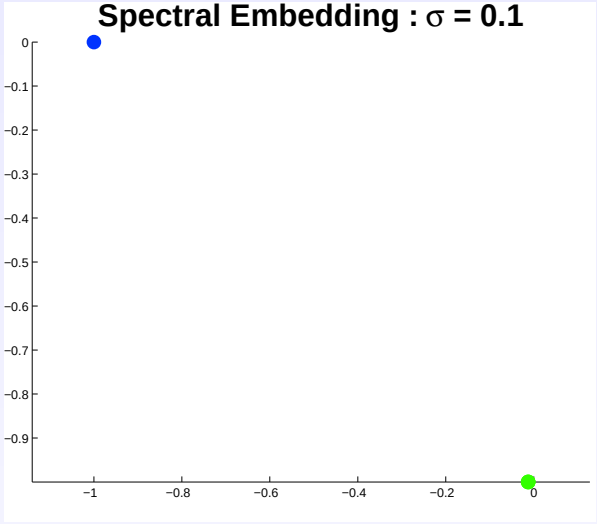


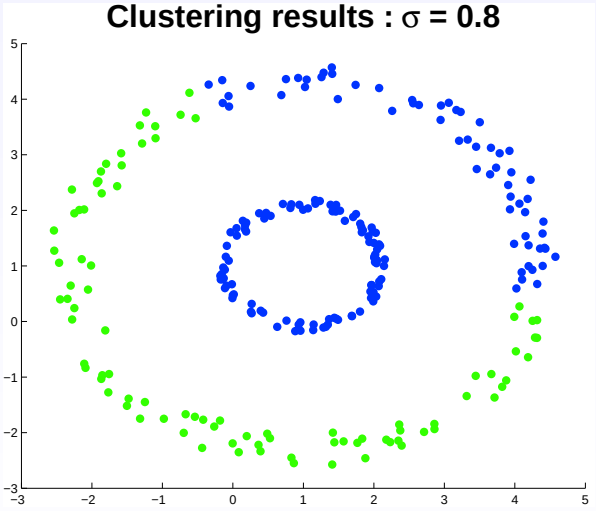


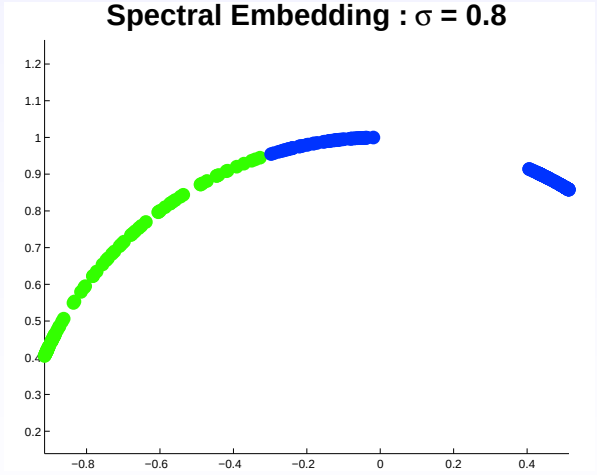
A "good" affinity matrix : rule of σ

Search over σ and pick the value that, after clustering Y's rows, gives the tightest clusters on the surface of the k-sphere. (*Ng, Jordan and Weiss*)



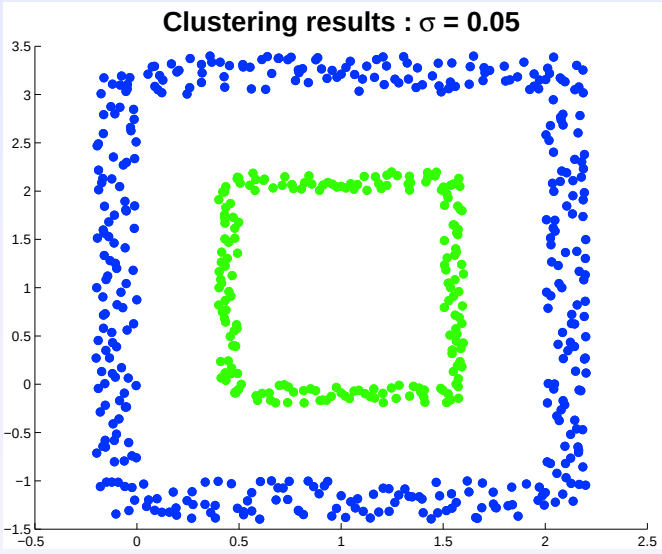


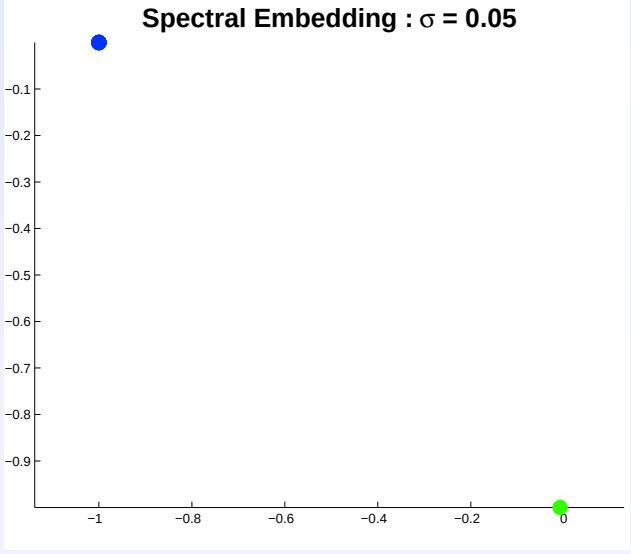


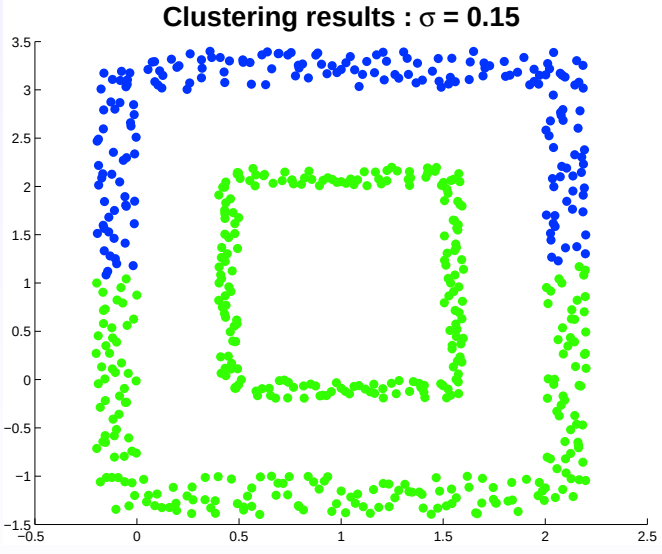


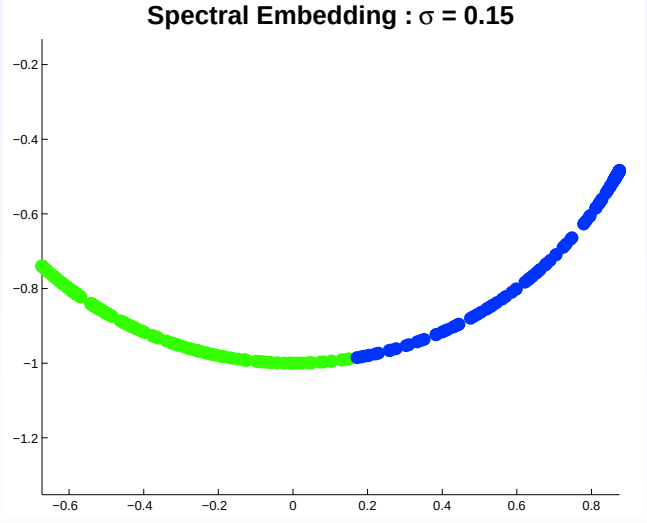
Interval of 'good' σ values : $\sigma \in]0.1, 0.8[$

A "good" affinity matrix : rule of σ









Interval of 'good' σ values : $\sigma \in]0.03, 0.15[$

Definitions for σ :

Brand

A global parameter σ so that :
σ is the mean of the distances between each point and its closest neighbor.

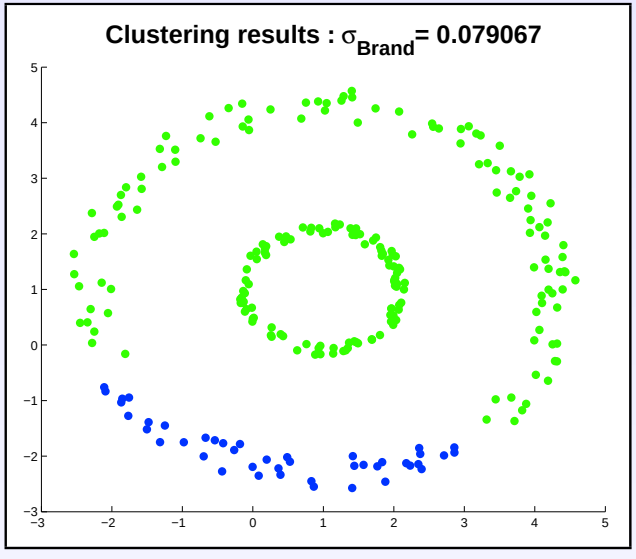
Perona, Zelnik-Manor

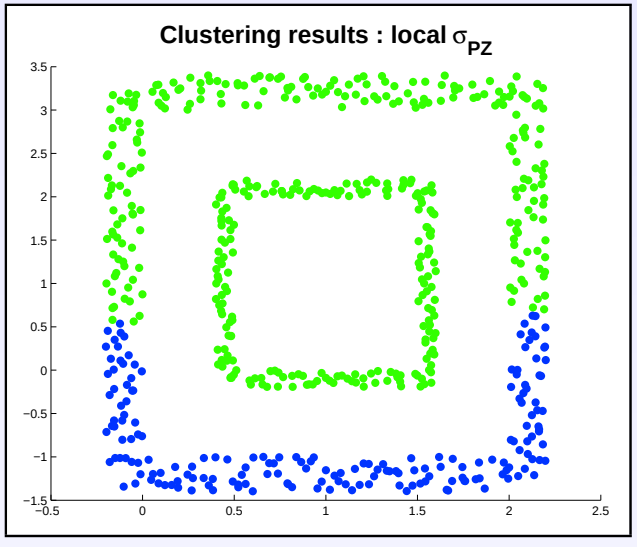
A local parameter σ defined by :

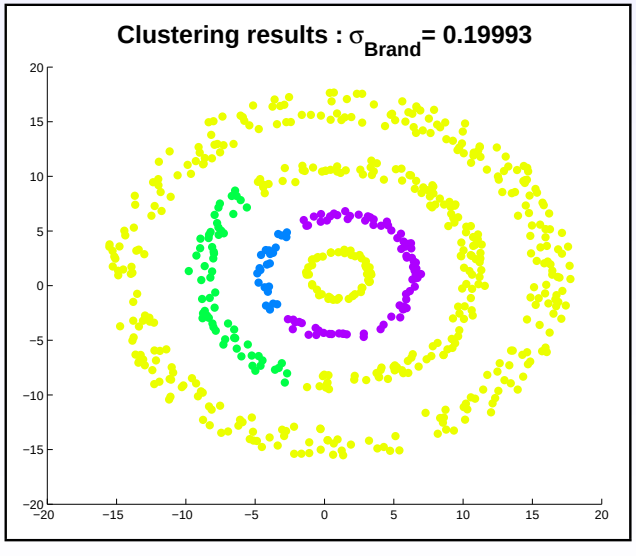
$$\sigma_i = d(x_i, x_K)$$

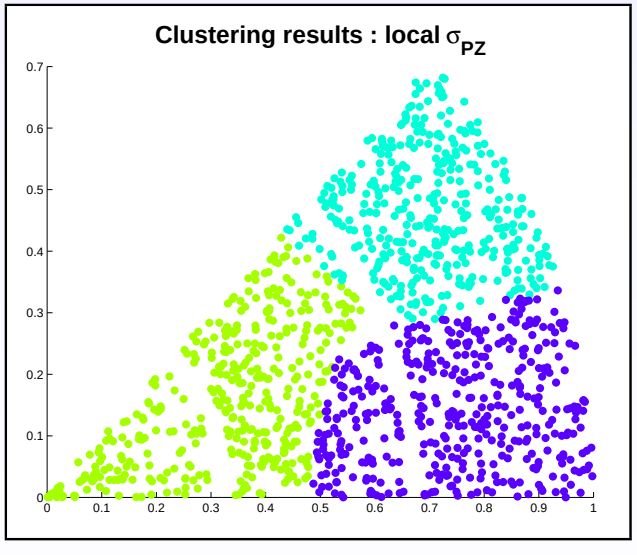
with x_k the Kth neighbor of the point x_i $(K = 7)$, $\forall i \in 1, .., m$.

Some results :









An heuristic global parameter σ_h :

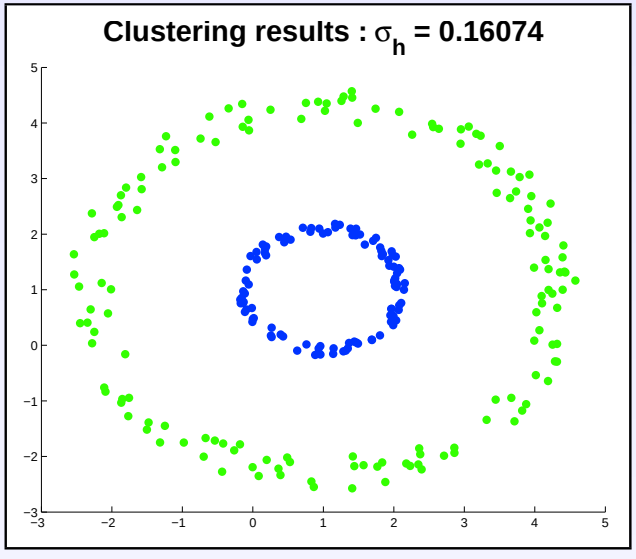
Global heuristic parameter

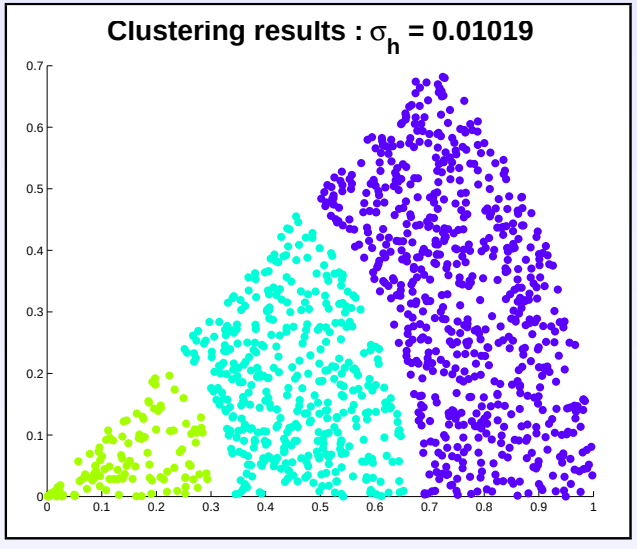
A global parameter σ including density information :

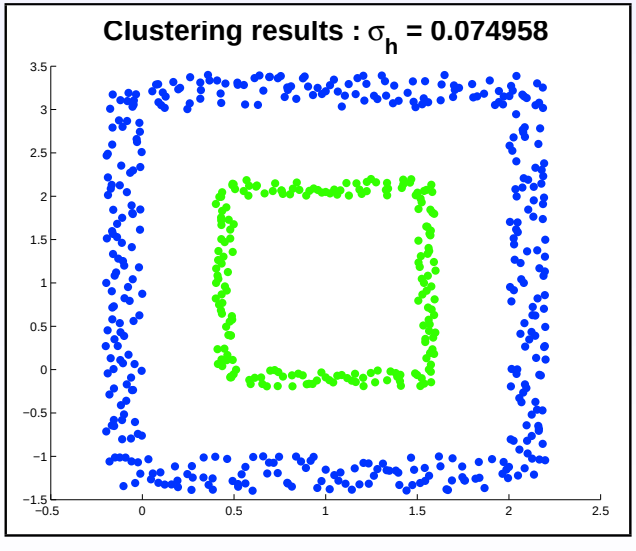
$$\sigma_h = \frac{\max_{i<j}(\|x_i - x_j\|)}{\sqrt{8m}}$$

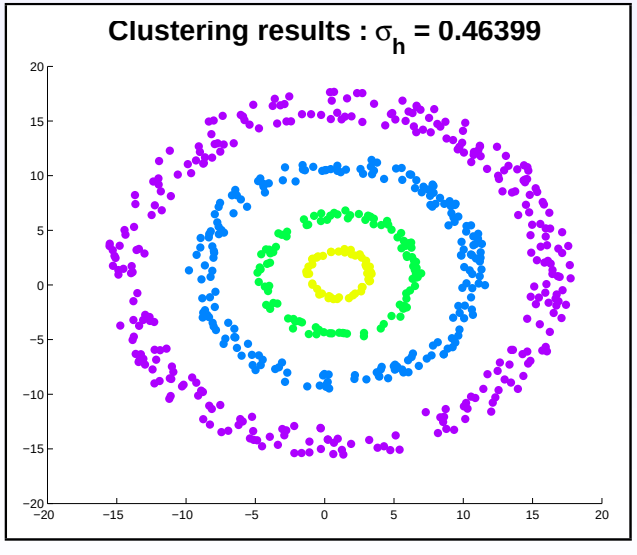
where m is the number of data points.

Some results with σ_h :

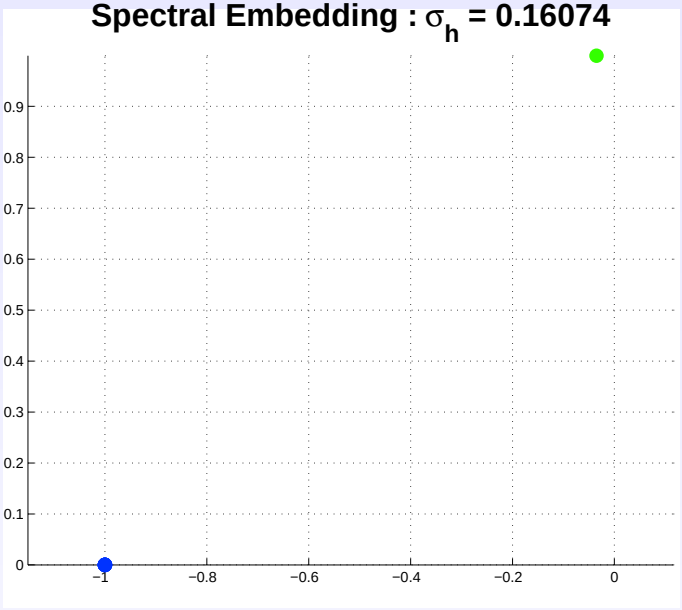


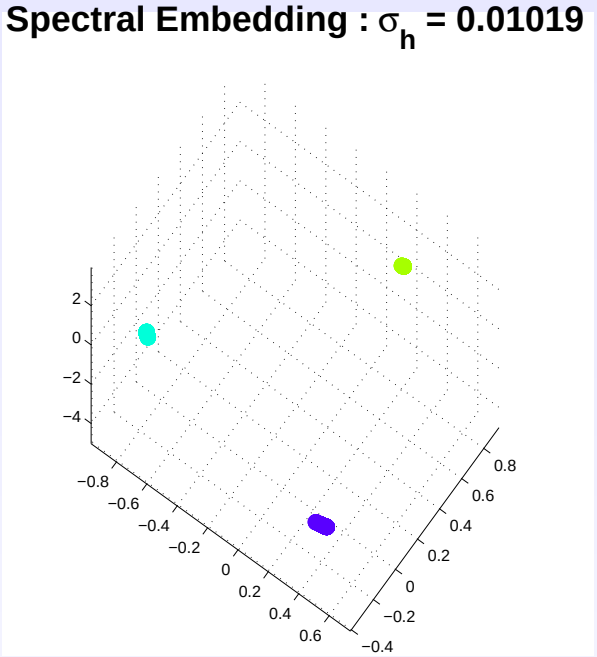


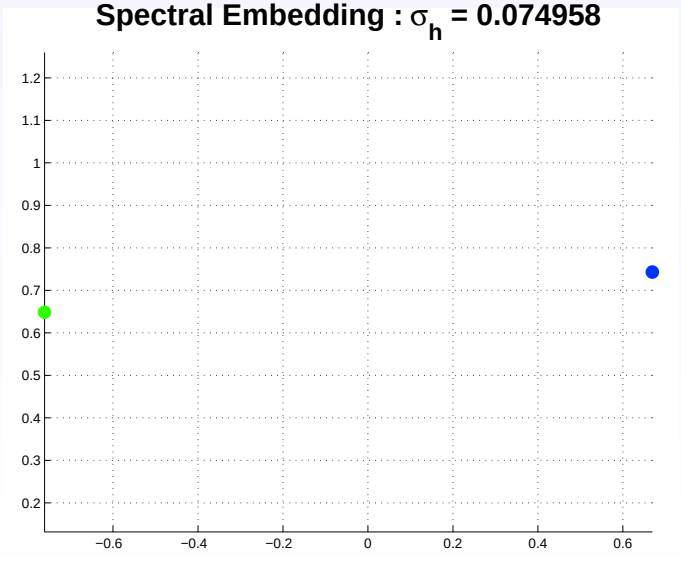


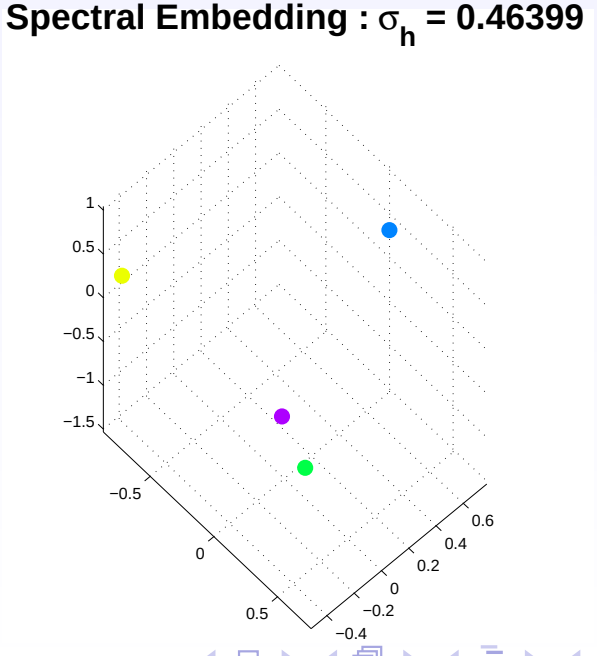


Some results with σ_h :









Quality measures : ratio of Frobenius norms

In general cases, $\hat{A}$'s off-diagonal blocks are non-zero so, with $k = 3$, :

$$\hat{L} = \begin{bmatrix} L^{(11)} & L^{(12)} & L^{(13)} \\ L^{(21)} & L^{(22)} & L^{(23)} \\ L^{(31)} & L^{(32)} & L^{(33)} \end{bmatrix}$$

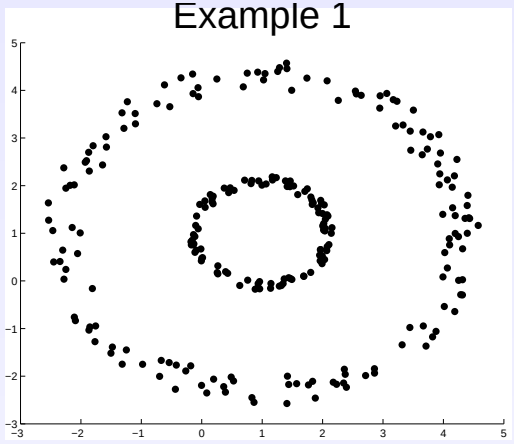
Evaluate the ratio between off-diagonal-blocks Frobenius norm and diagonal-blocks one

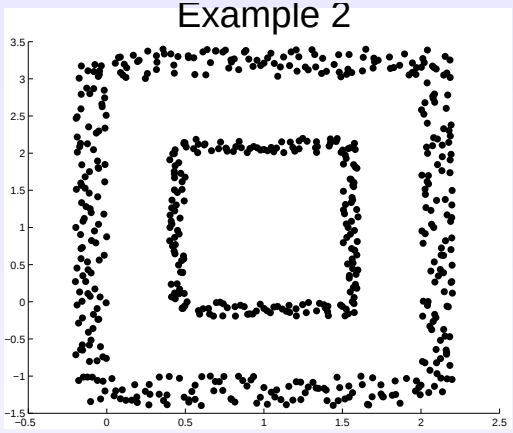
$$r = \frac{\|L^{(ij)}\|_F}{\|L^{(ii)}\|_F}$$

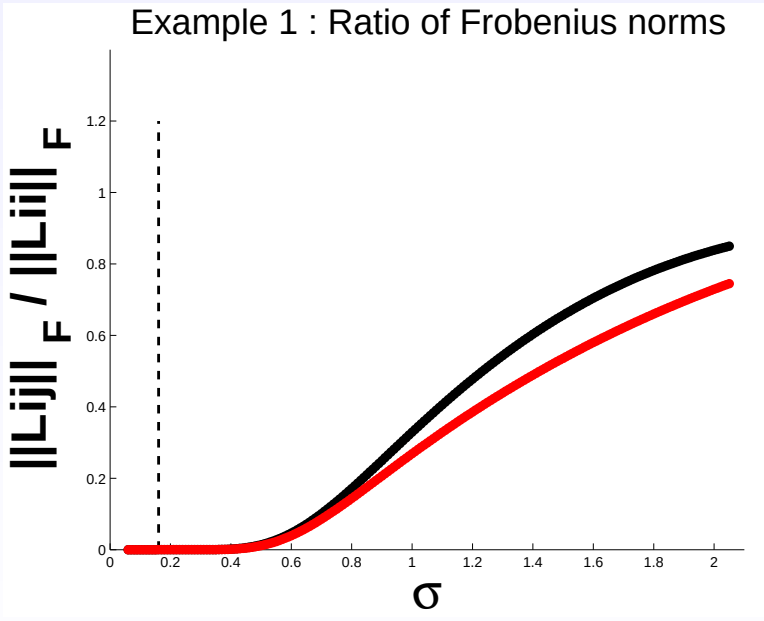
with $i \neq j$ and $i, j \in 1, .., k$

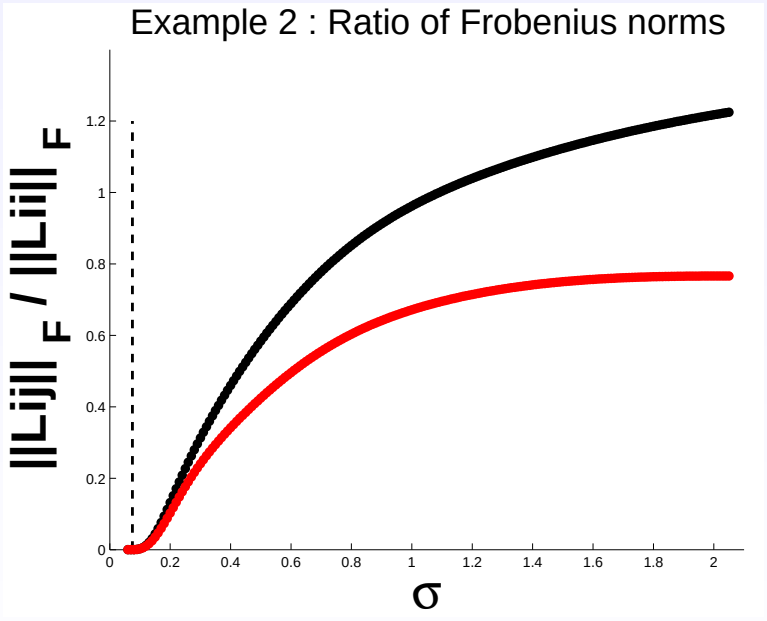
- if $r \approx 0$, near block diagonal structure

Quality measures : examples









Legend : $\cdot$ $\frac{\|L_{12}\|_F}{\|L_{11}\|_F}$ $\cdot$ $\frac{\|L_{12}\|_F}{\|L_{22}\|_F}$ $- - \sigma_h$

Quality measures : confusion matrix

Let $W \in \mathcal{M}_{k,k}(\mathbb{R})$ be the confusion matrix :

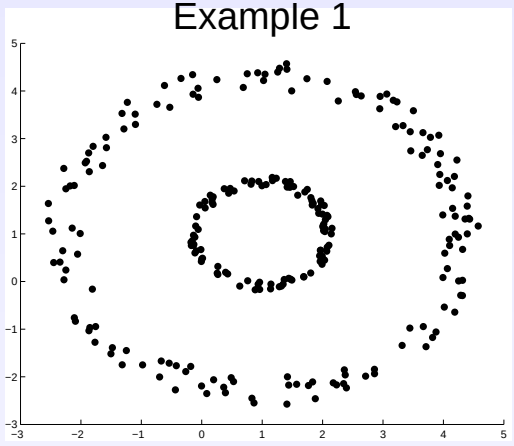
$$W = \begin{bmatrix} W^{(11)} & W^{(12)} & W^{(13)} \\ W^{(21)} & W^{(22)} & W^{(23)} \\ W^{(31)} & W^{(32)} & W^{(33)} \end{bmatrix}$$

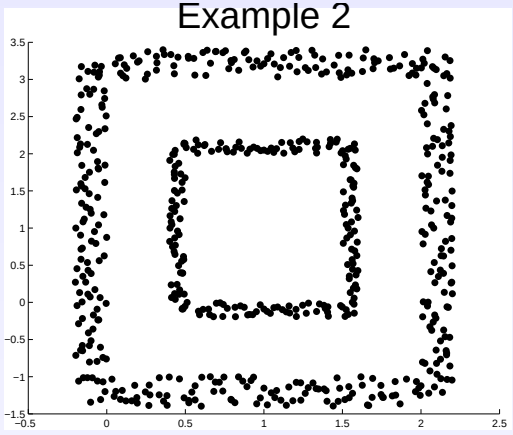
where $W^{(ij)}$ is the number of points that were assigned in cluster j instead of cluster i.

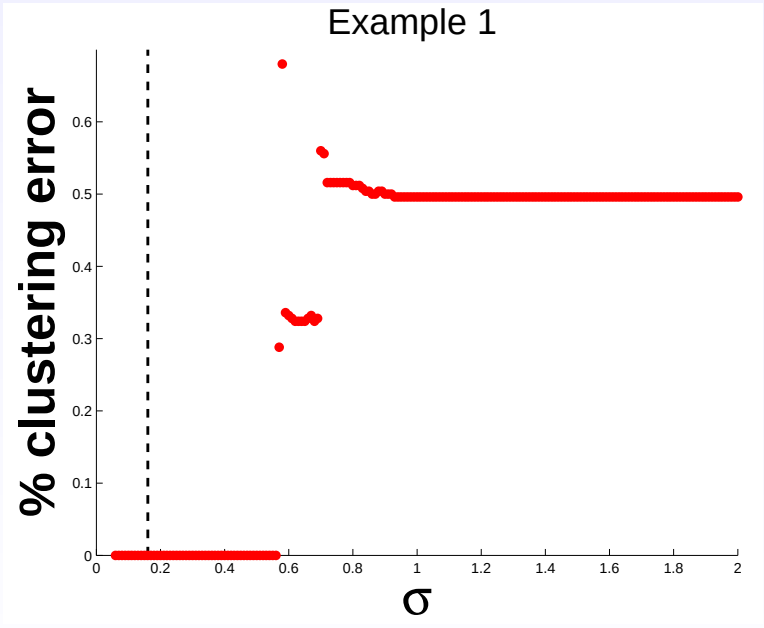
We define the *percentage of misclustered points* by :

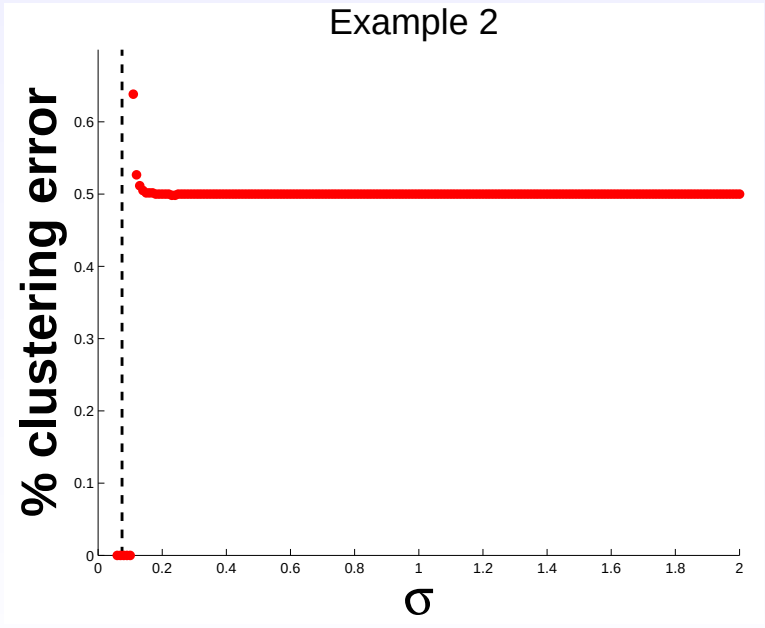
$$p = \frac{\sum_{i \neq j}^{k} W^{(ij)}}{m}$$

Quality measures : examples









Legend : – – σ_h

Perspectives

- Spectral clustering generalization : various data dimensions and increase the number of clusters.
- Spectral analysis : find methods in linear algebra and/or parallel calculus which permit treating important data set.
- Cluster definition : study the robustness of the global parameter σ_h with different densities in the data set.
- Data from Biology : how to incorporate time dependency as a parameter ?