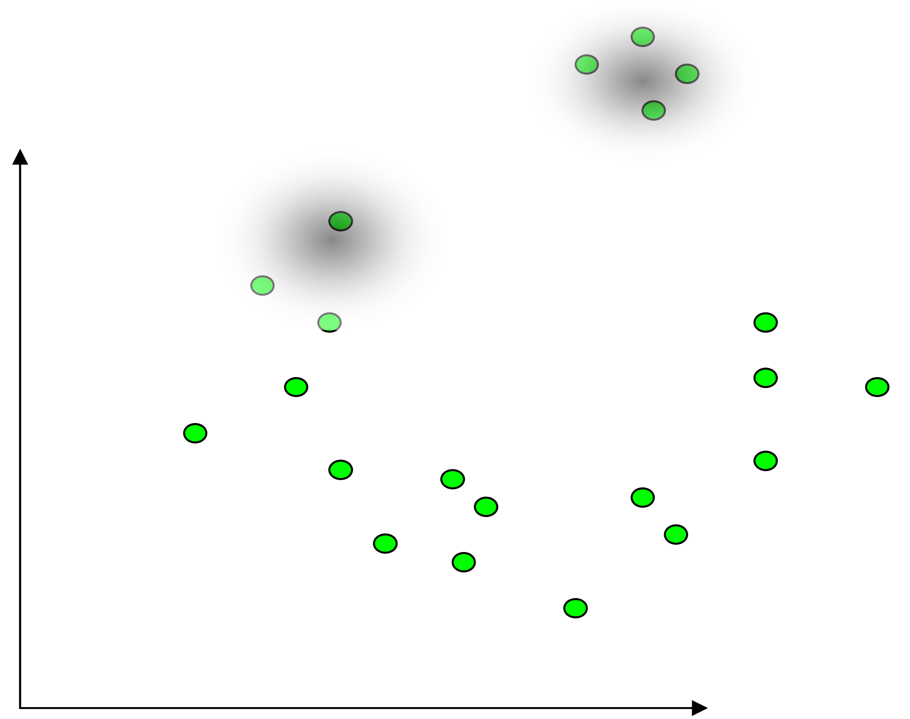
Learning the topology of a data set

Michaël Aupetit – Researcher Engineer (CEA)
Pierre Gaillard – Ph.D. Student
Gerard Govaert – Professor (University of Technology of Compiegne)

Introduction

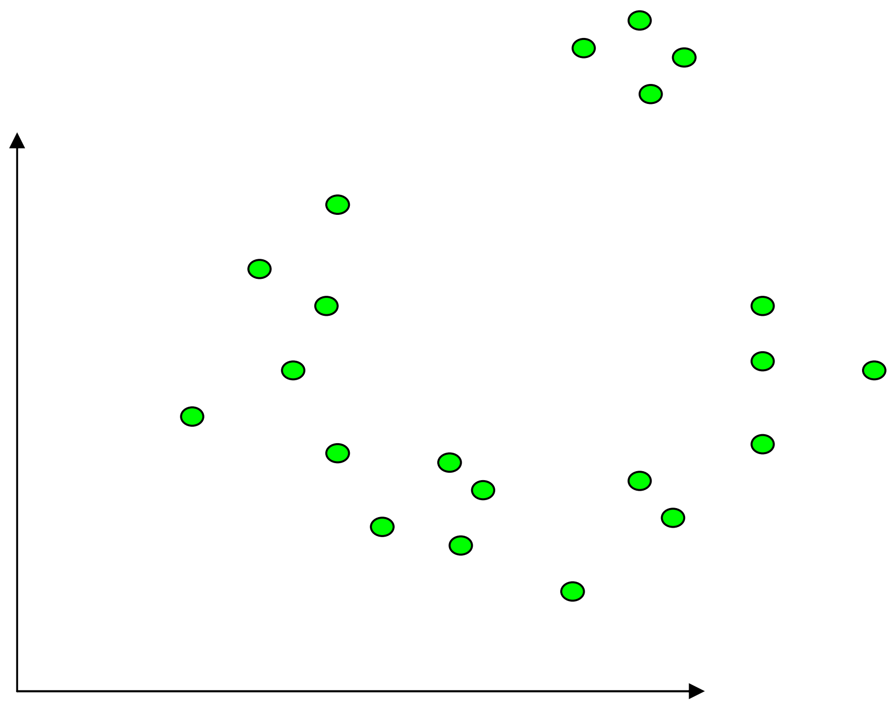
Given a set of M data in R^D, the estimation of the density allow solving various problems :

classification, clustering, regression



A question without answer…

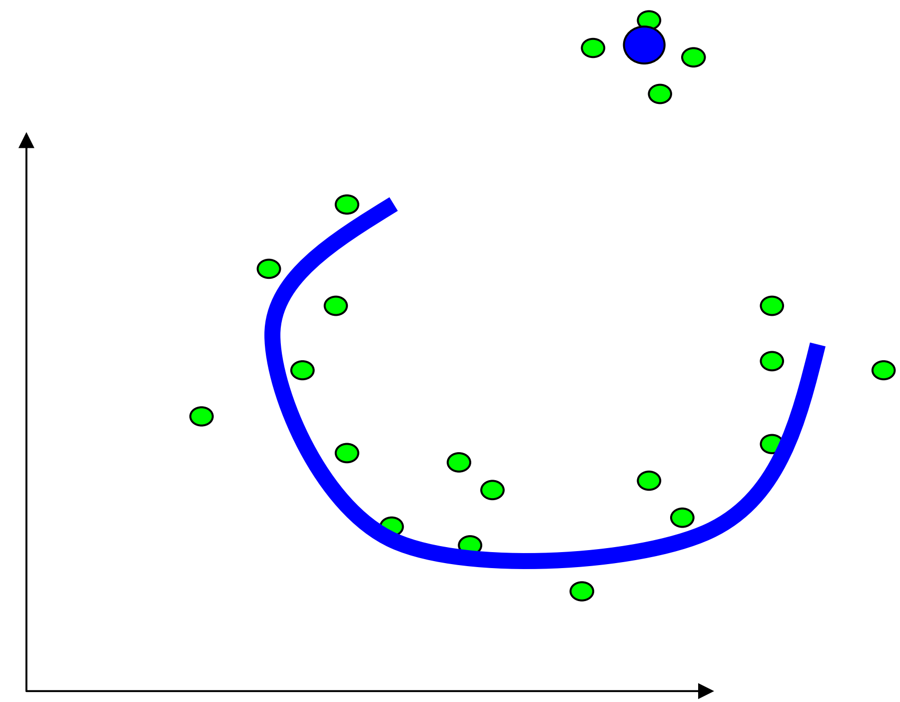
The generative models cannot aswer this question:
Which is the « shape » of this data set ?



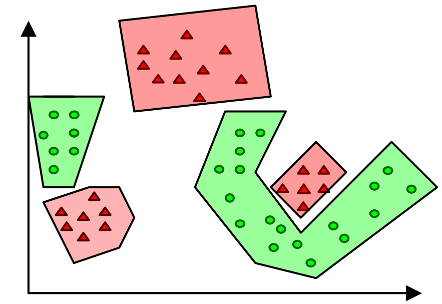
An subjective answer

The expected answer is :

1 point and 1 curve not connected to each other



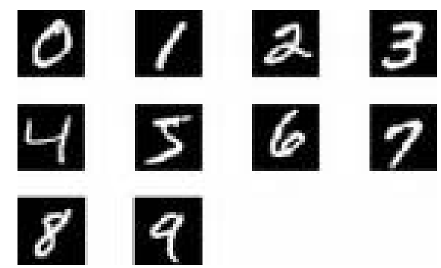
The problem : what is the topology of the principal manifolds



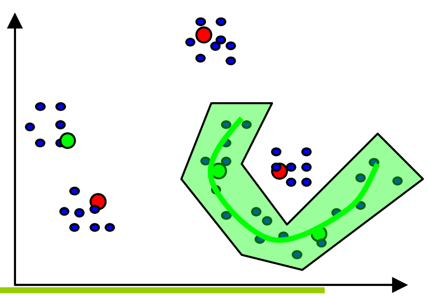
Estimate the complexity of the classification task
[Lallich02, Aupetit05Neurocomputing]

Add a topological *a priori* to design a classifier
[Belkin05Nips]

$$f^* = \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2$$



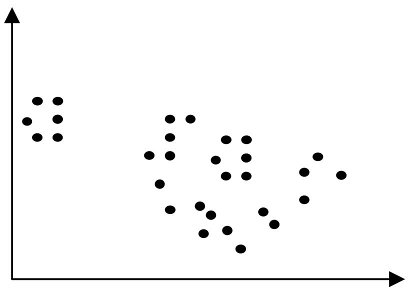
Add topological features to statistical features



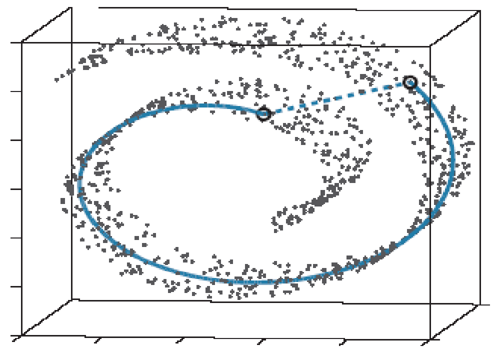
Classify through the connected components or the intrinsic dimension.
[Belkin]

Why learning topology : unsupervised applications

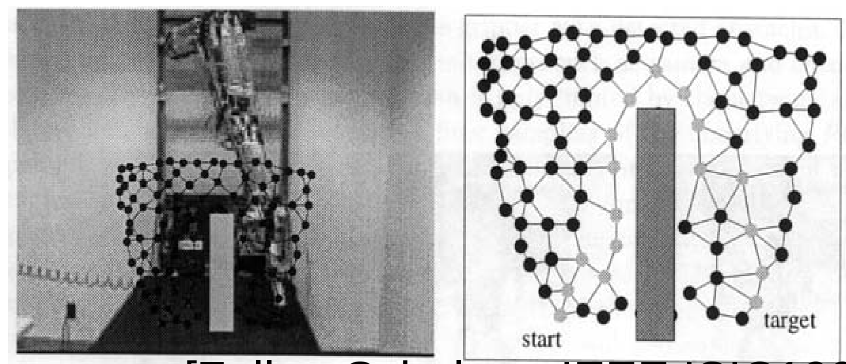
Clusters defined by the connected components



Data exploration (e.g. shortest path)



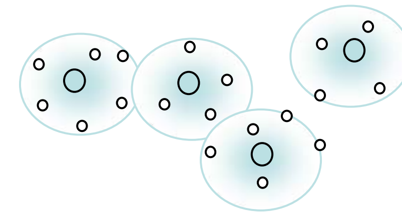
Robotic (Optimal path, inverse kinematic)



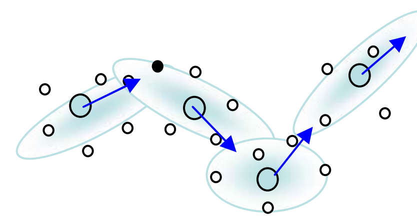
[Zeller, Schulten -IEEE ISIC1996]

Generative manifold learning

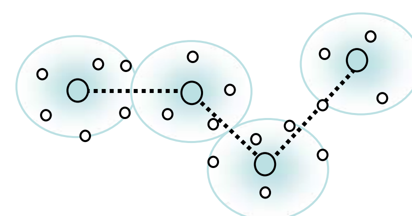
Gaussian Mixture



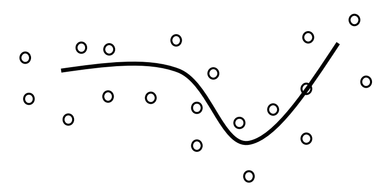
MPPCA [Bishop]



GTM [Bishop]



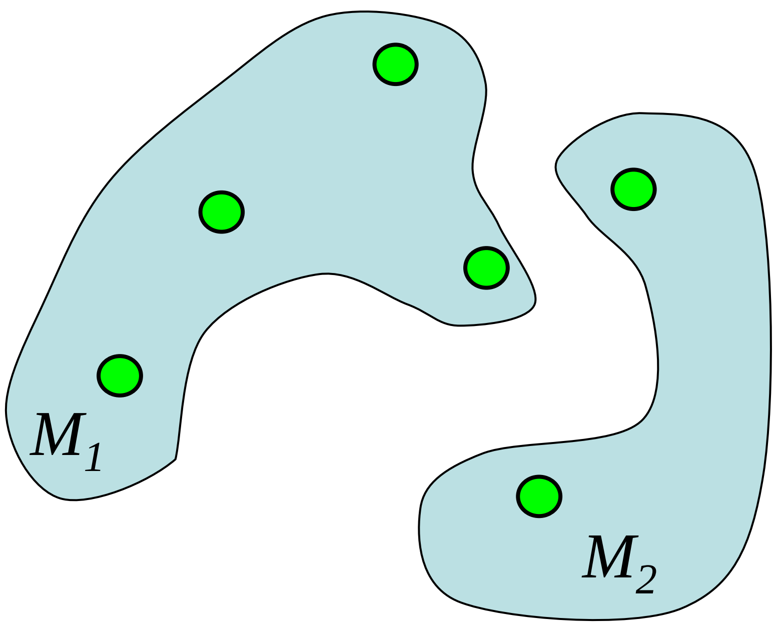
Revisited Principal Curves [Hastie,Stuetzle]



Problems: fixed or incomplete topology

Computational Topology

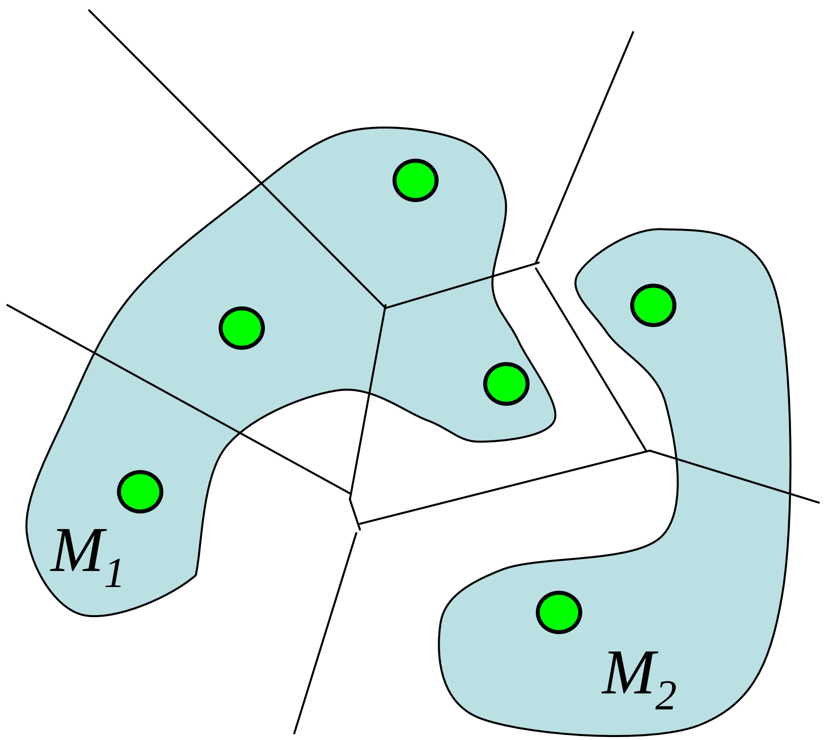
All the previous work about topology learning has been grounded on the result of Edelsbrunner and Shah (1997) which proved that given a manifold and a set of N prototypes nearby M, it exists a subgraph* of the Delaunay graph of the prototypes which has the same topology as M



* more exactly a subcomplex of the Delaunay complex

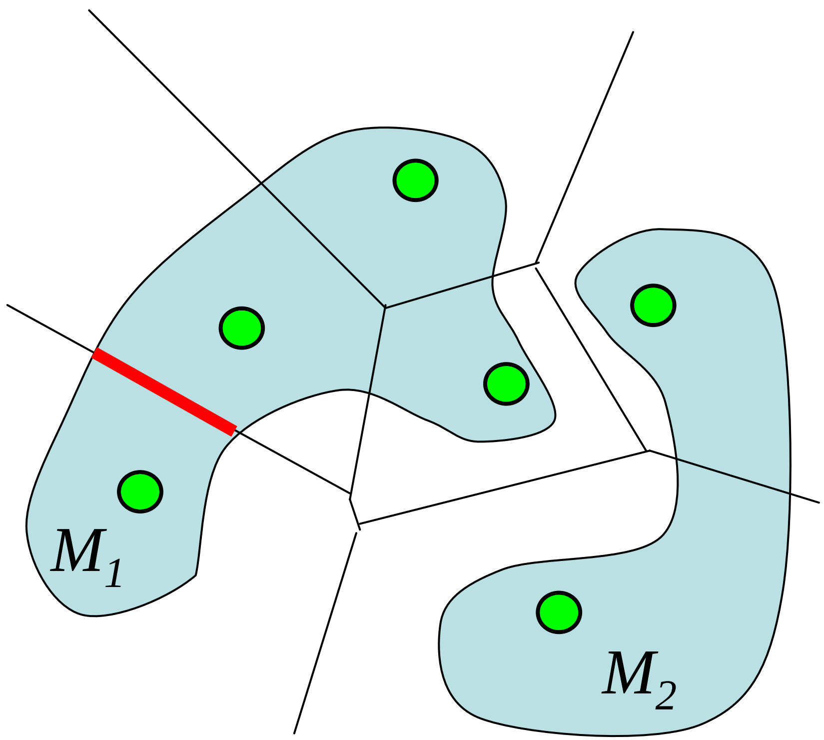
Computational Topology

All the previous work about topology learning has been grounded on the result of Edelsbrunner and Shah (1997) which proved that given a manifold and a set of N prototypes nearby M, it exists a subgraph of the Delaunay graph of the prototypes which has the same topology as M



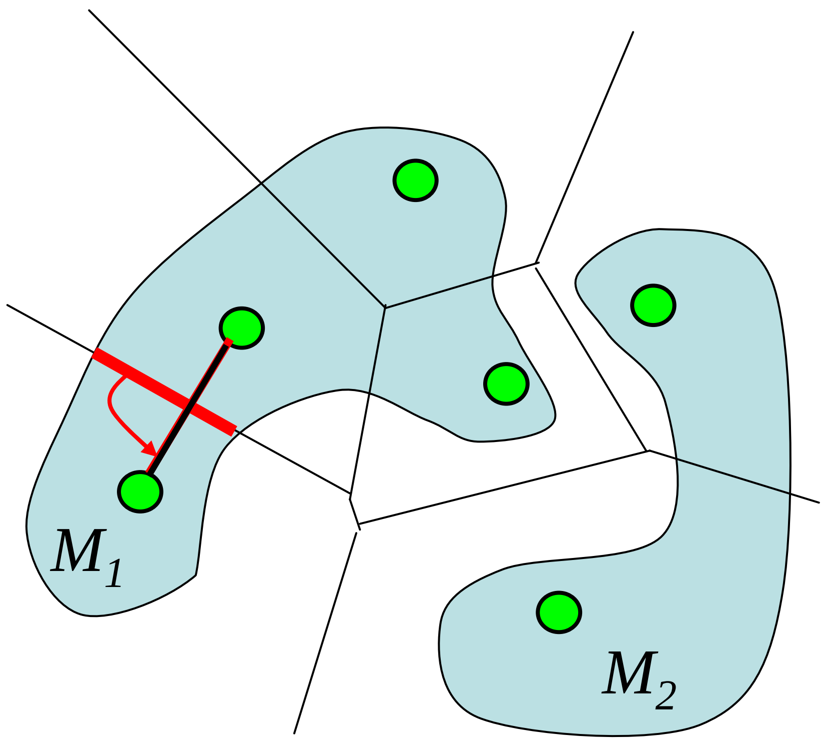
Computational Topology

All the previous work about topology learning has been grounded on the result of Edelsbrunner and Shah (1997) which proved that given a manifold and a set of N prototypes nearby M, it exists a subgraph of the Delaunay graph of the prototypes which has the same topology as M



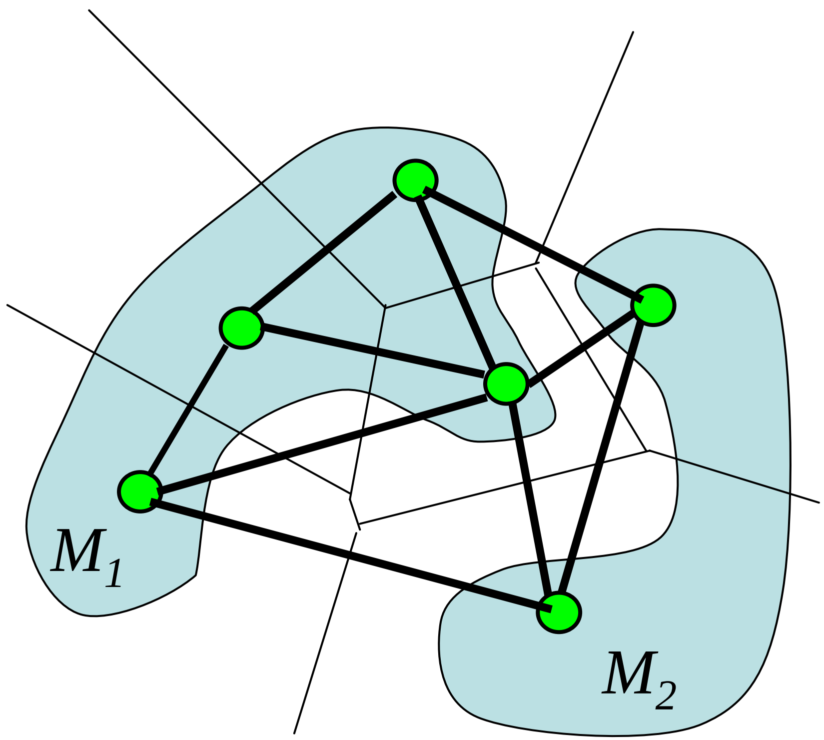
Computational Topology

All the previous work about topology learning has been grounded on the result of Edelsbrunner and Shah (1997) which proved that given a manifold and a set of N prototypes nearby M, it exists a subgraph of the Delaunay graph of the prototypes which has the same topology as M



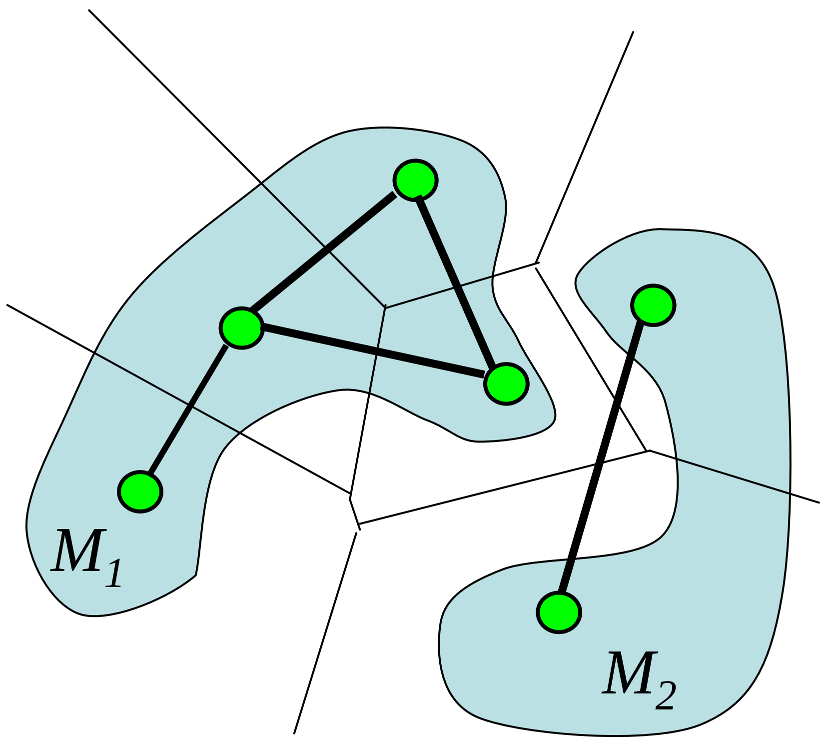
Computational Topology

All the previous work about topology learning has been grounded on the result of Edelsbrunner and Shah (1997) which proved that given a manifold and a set of N prototypes nearby M, it exists a subgraph of the Delaunay graph of the prototypes which has the same topology as M



Computational Topology

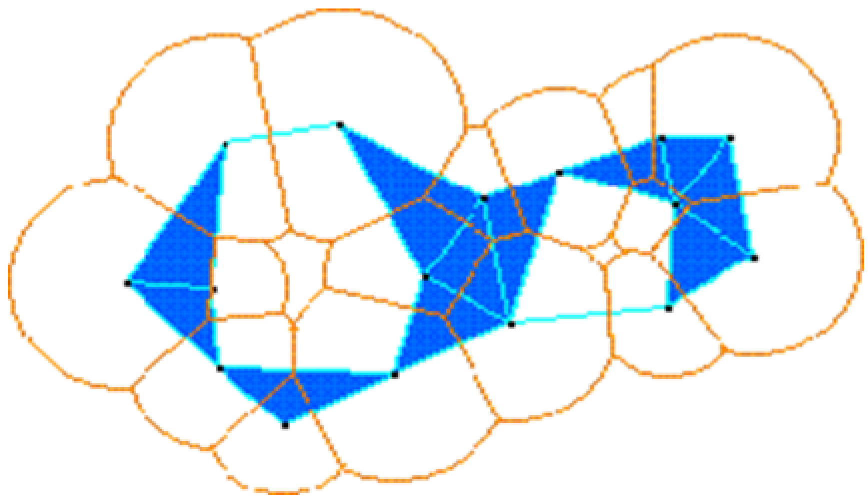
All the previous work about topology learning has been grounded on the result of Edelsbrunner and Shah (1997) which proved that given a manifold and a set of N prototypes nearby M, it exists a subgraph of the Delaunay graph of the prototypes which has the same topology as M



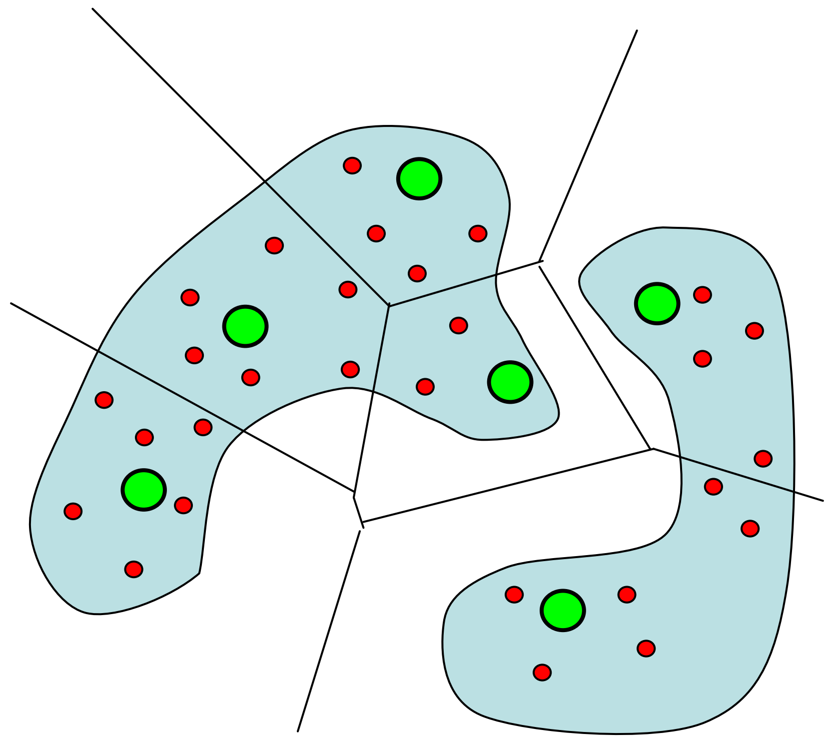
Extractible topology
$O(DN^3)$

Application : known manifold

Topology of molecules
[Edelsbrunner1994]

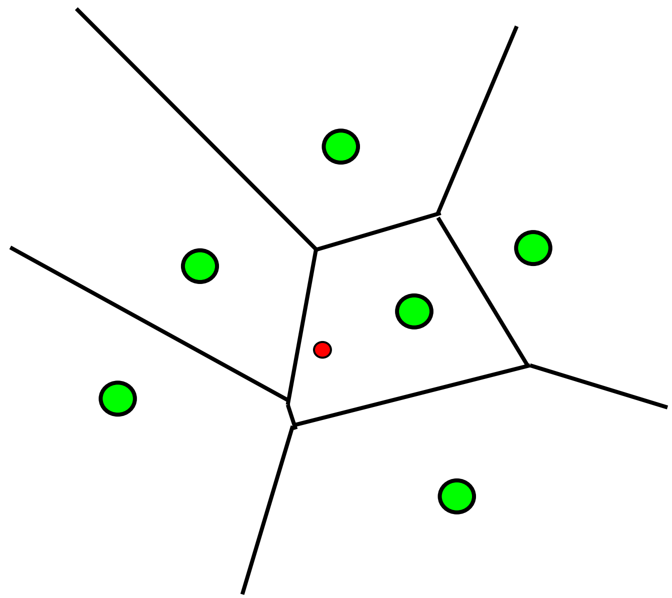


Approximation : manifold known throught a data set



Topology Representing Network

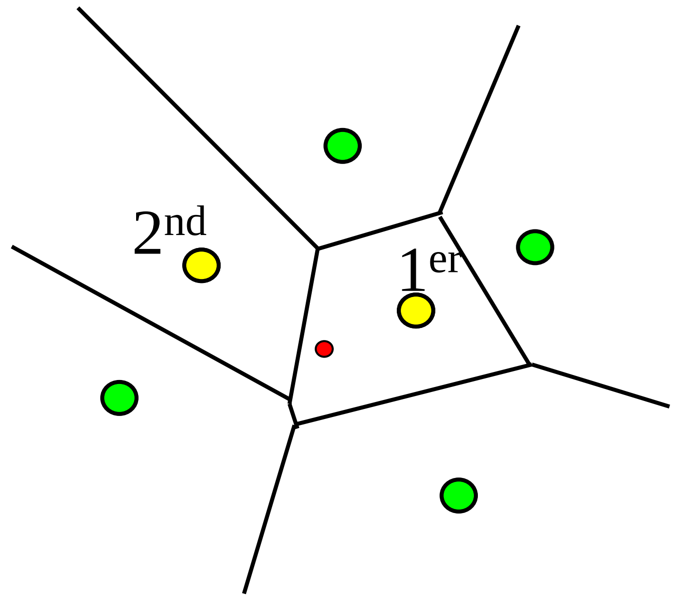
- Topology Representing Network [Martinetz, Schulten 1994]



Connect the 1st and 2nd NN of each data

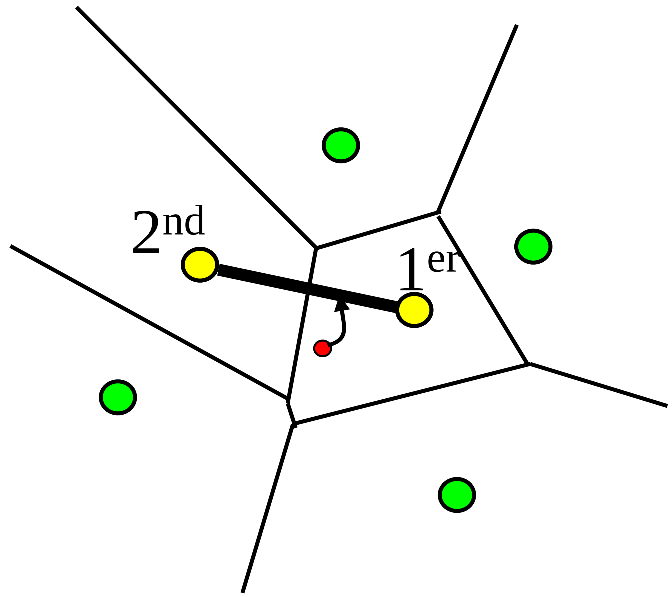
Topology Representing Network

- Topology Representing Network [Martinetz, Schulten 1994]



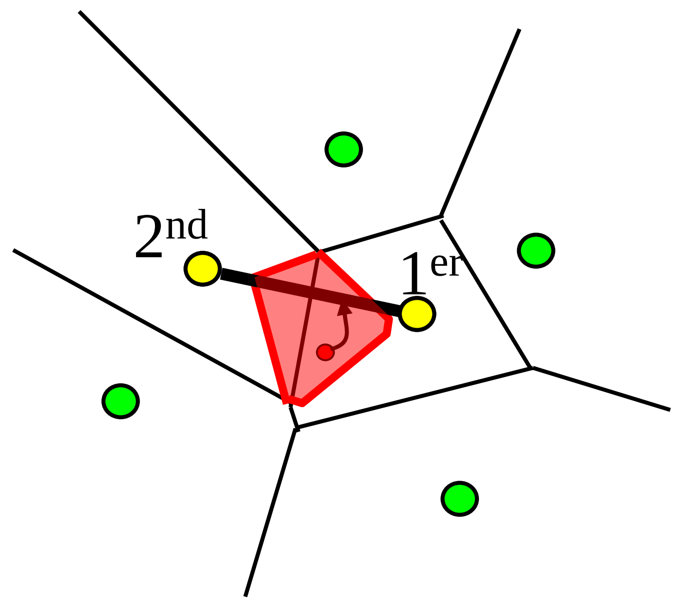
Topology Representing Network

- Topology Representing Network [Martinetz, Schulten 1994]



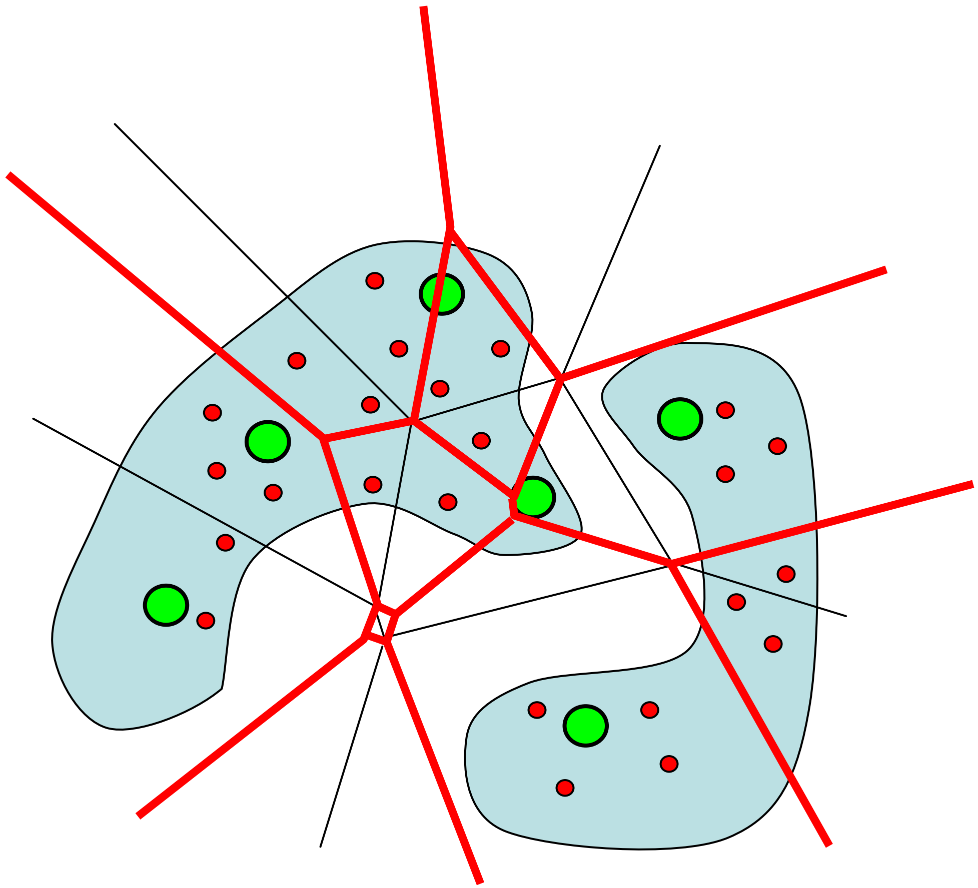
Topology Representing Network

- Topology Representing Network [Martinetz, Schulten 1994]



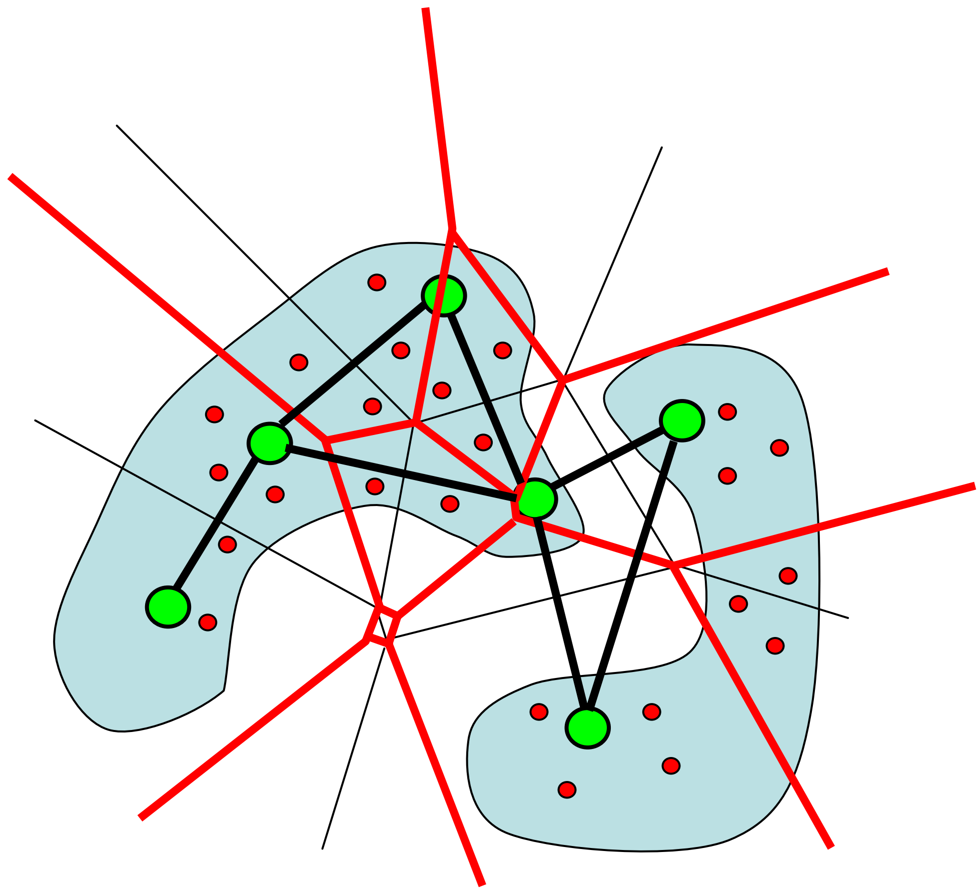
Topology Representing Network

- Topology Representing Network [Martinetz, Schulten 1994]

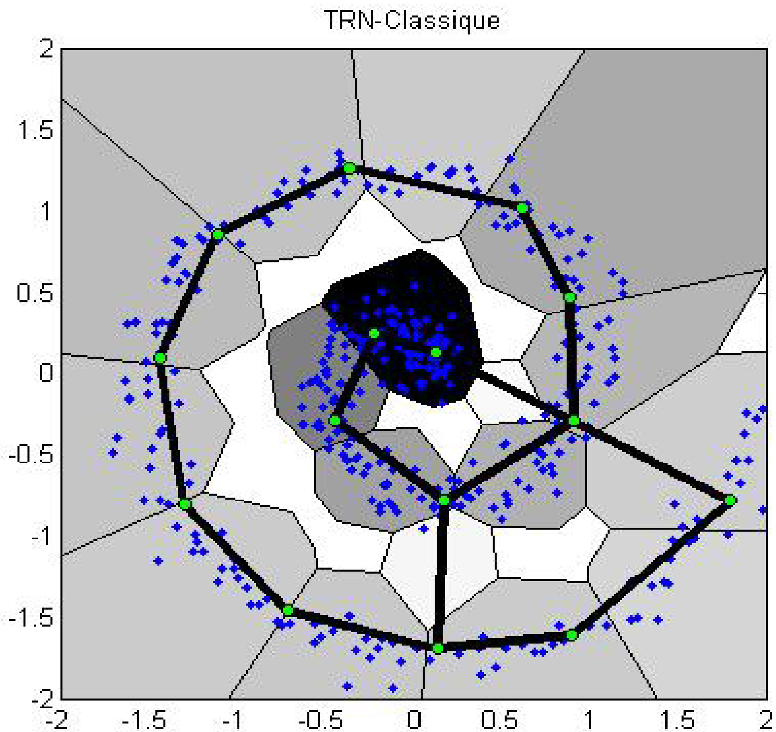


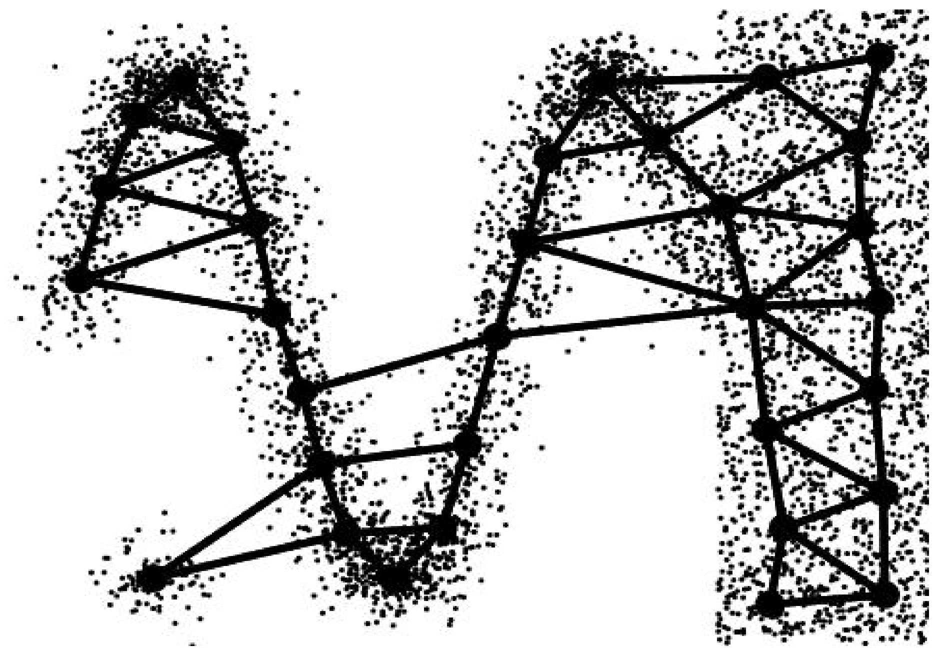
Topology Representing Network

- Topology Representing Network [Martinetz, Schulten 1994]



Topology Representing Network





Good points :

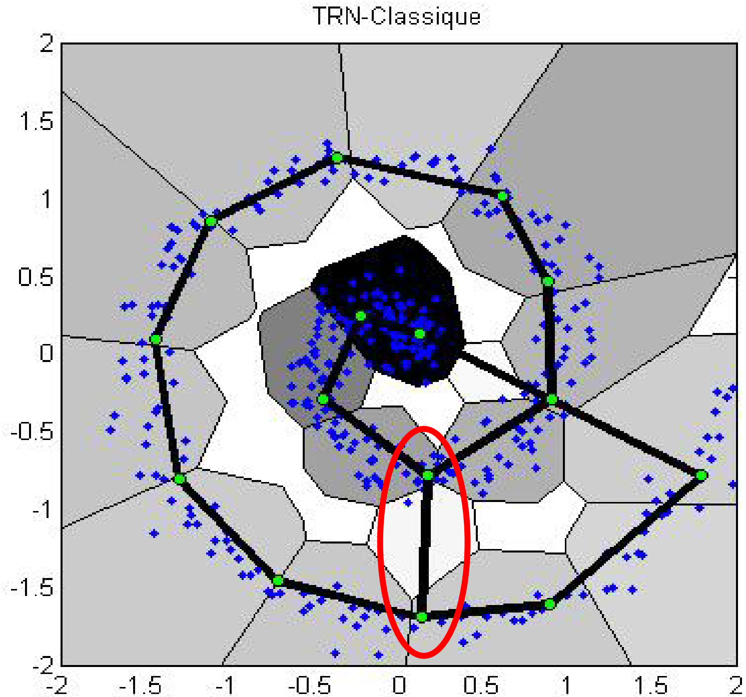
1- O(DNM)

2- If there are enough prototypes and if they are well located then resulting graph is « good » in practice.

Some drawbacks from the machine learning point of view

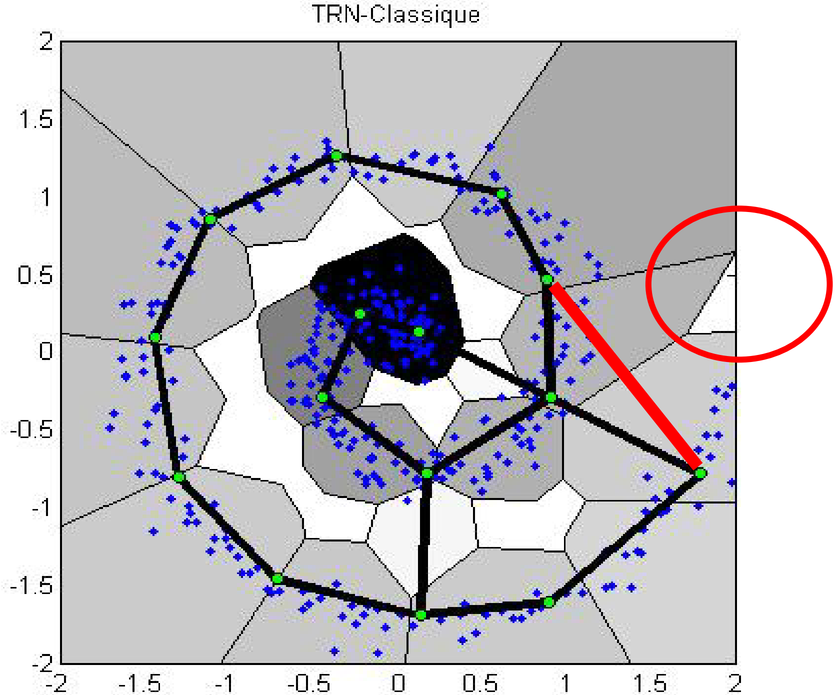
Topology Representing Network : some drawbacks

- **Noise sensitivity**



Topology Representing Network : some drawbacks

- **Not self-consistent [Hastie]**

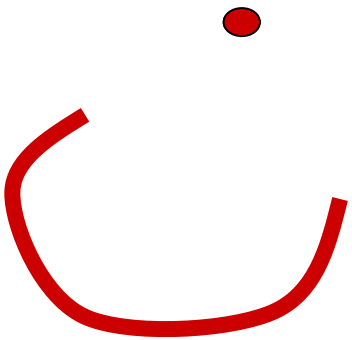


Topology Representing Network : some drawbacks

- **No quality measure**
  - How to measure the quality of the TRN if D >3 ?
  - How to compare two models ?

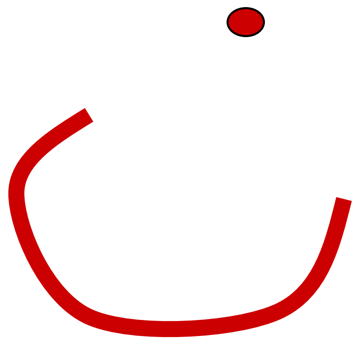
For all these reasons, we propose a generative model

General assumptions on data generation

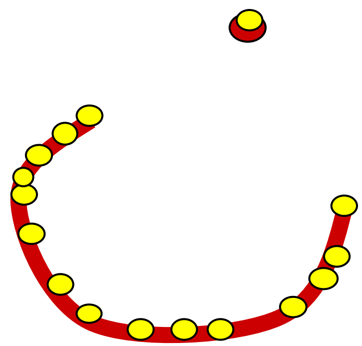


Unknown
principal
manifolds

General assumptions on data generation

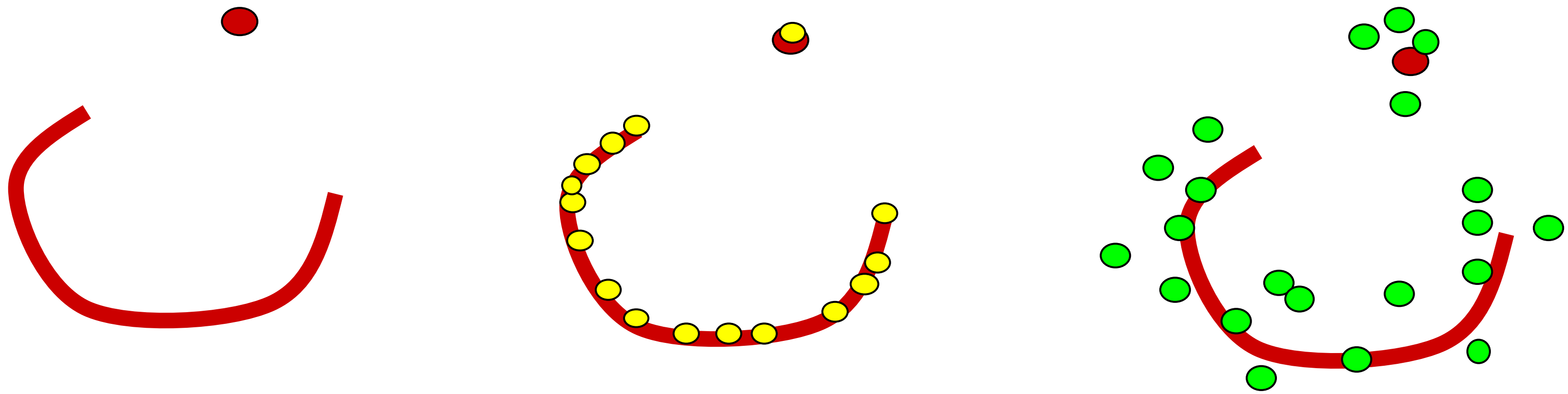


Unknown
principal
manifolds



…from which are drawn
data with a unknown pdf

General assumptions on data generation

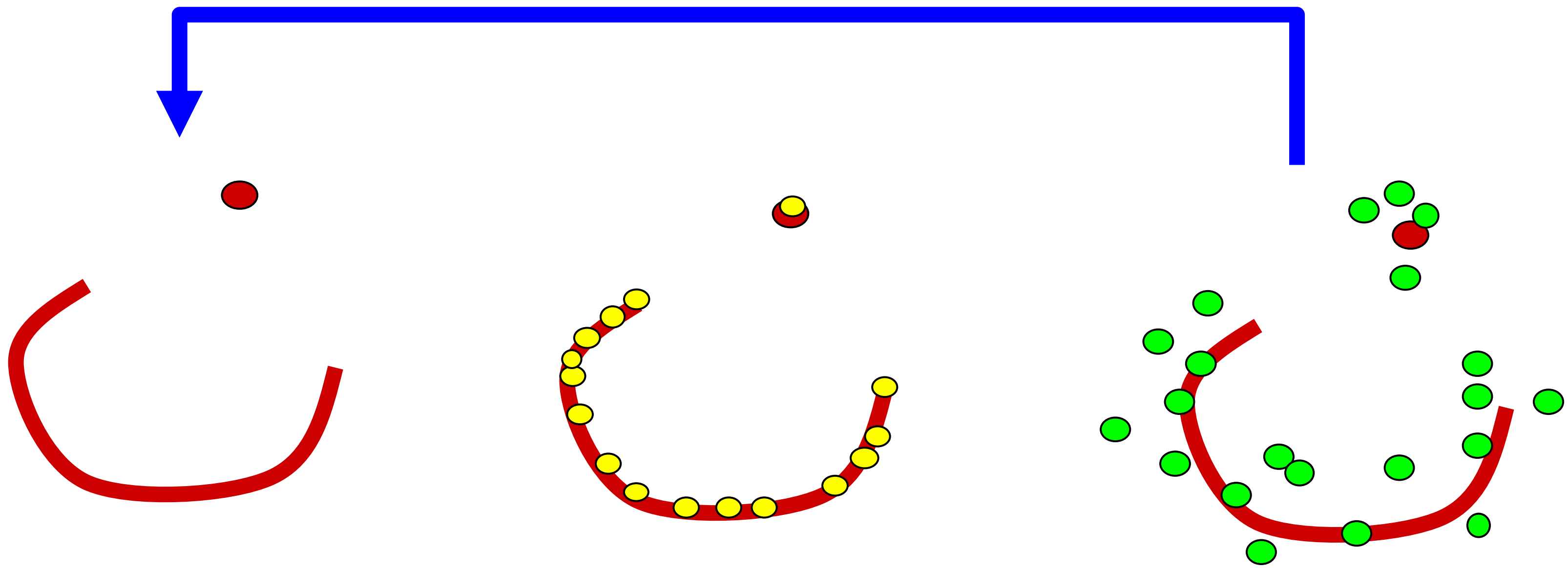


Unknown principal manifolds

…from which are drawn data with a unknown pdf

…corropuPted with some unknown noise leading to the observarion

General assumptions on data generation



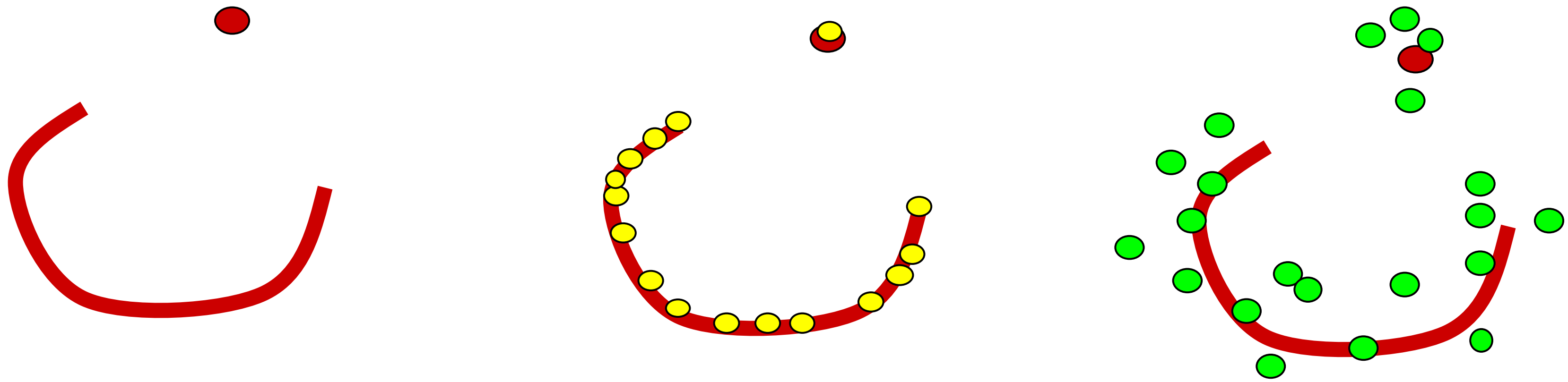
Unknown principal manifolds

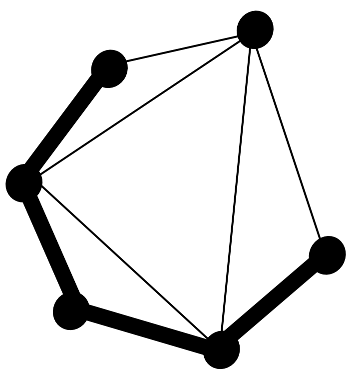
…from which are drawn data with a unknown pdf

…corroputed with some unknown noise leading to the observarion

The goal is to learn from the observed data, the principal manifolds such that their topological features can be extracted

3 assumptions…1 generative model

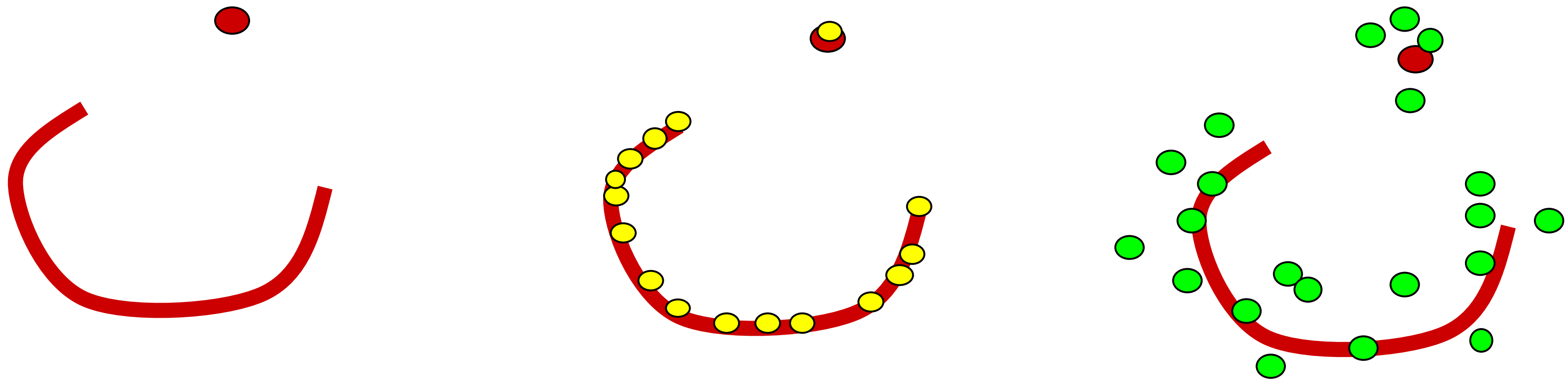


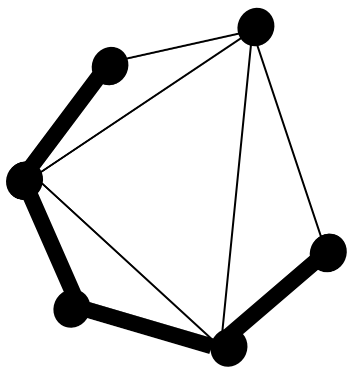


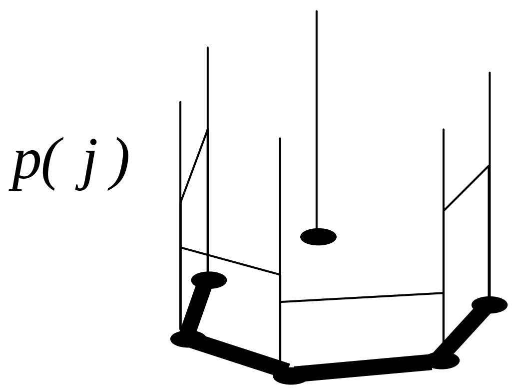
The manifold is close to the DG of some prototypes

$$p(x) = \sum_{j \in J}$$

3 assumptions…1 generative model





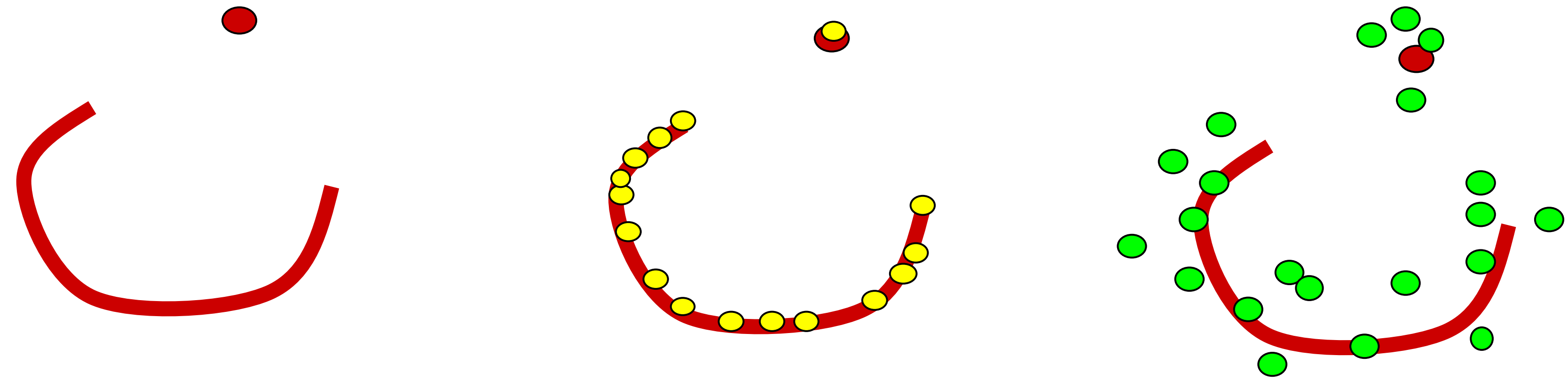


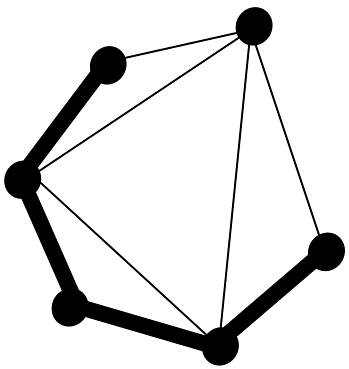
The manifold is close to the DG of some prototypes

we associate to each component a weighted uniform distribution

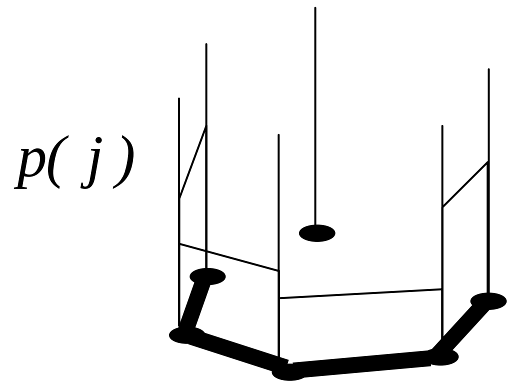
$$p(x) = \sum_{j \in J} p(j)$$

3 assumptions…1 generative model

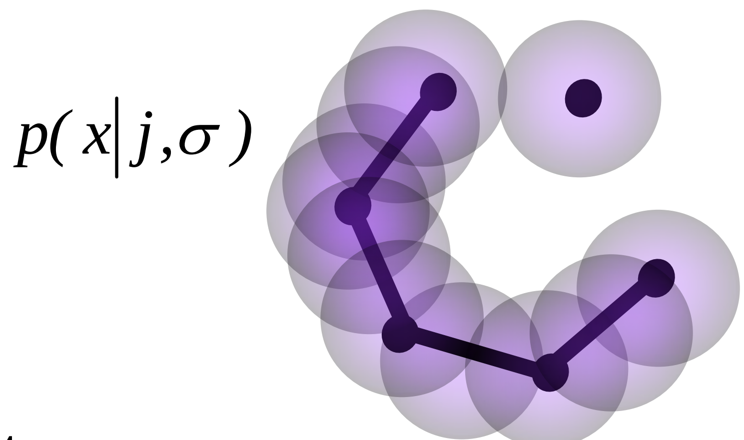




The manifold is close to the DG of some prototypes



we associate to each component a weighted uniform distribution



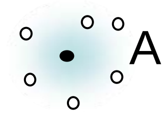
we convolve the components by an isotropic Gaussian noise

$$p(x) = \sum_{j \in J} p(j)\; p(x \mid j,\sigma)$$

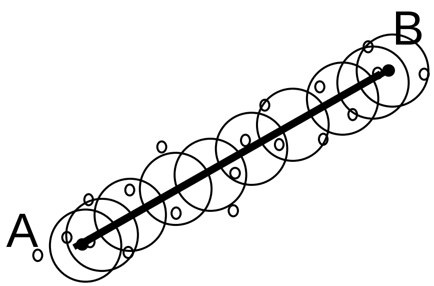
$$\sum_j p(j) = 1 \qquad p(j) \geq 0$$

A Gaussian-point and a Gaussian-segment

How to define a generative model based on points and segments ?

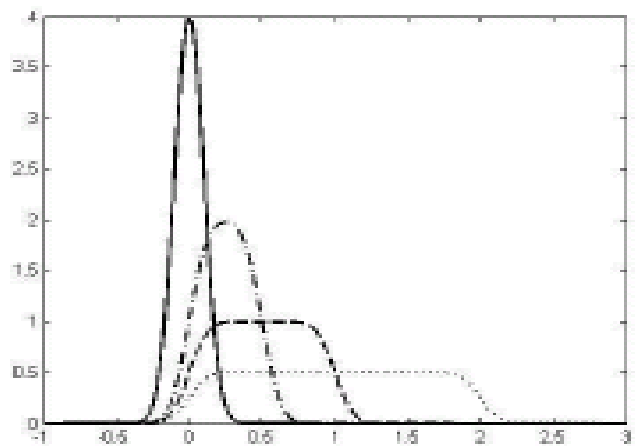


$$p^0(x|A,\sigma) = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{(x-A)^2}{2\sigma^2}\right)$$

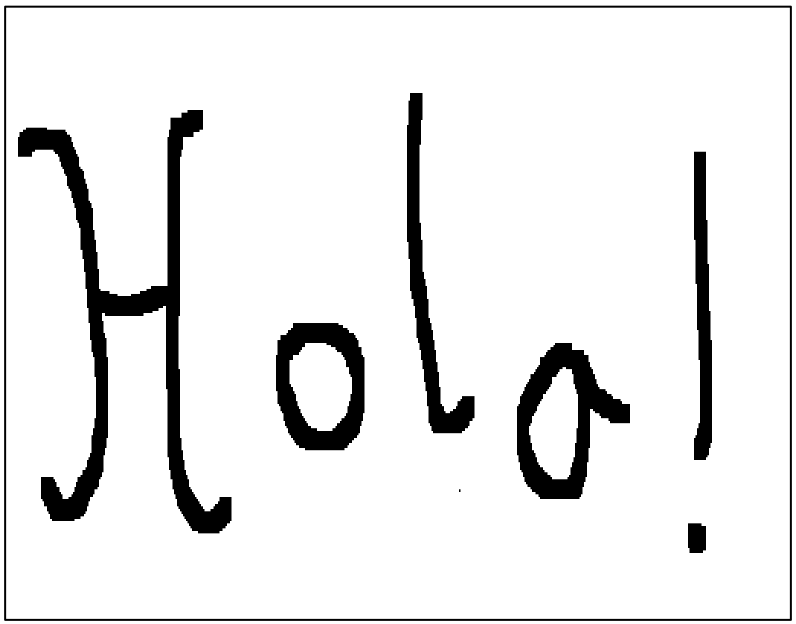


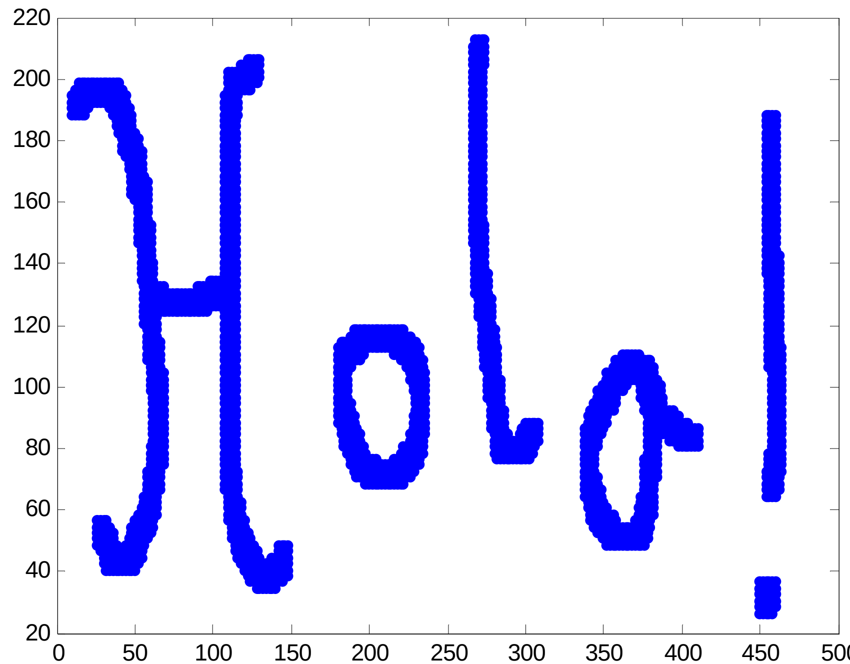
$$p^1(x\|[AB],\sigma) = \int_{[AB]} p(x|v,\sigma)dv$$

can be expressed in terms of « erf »

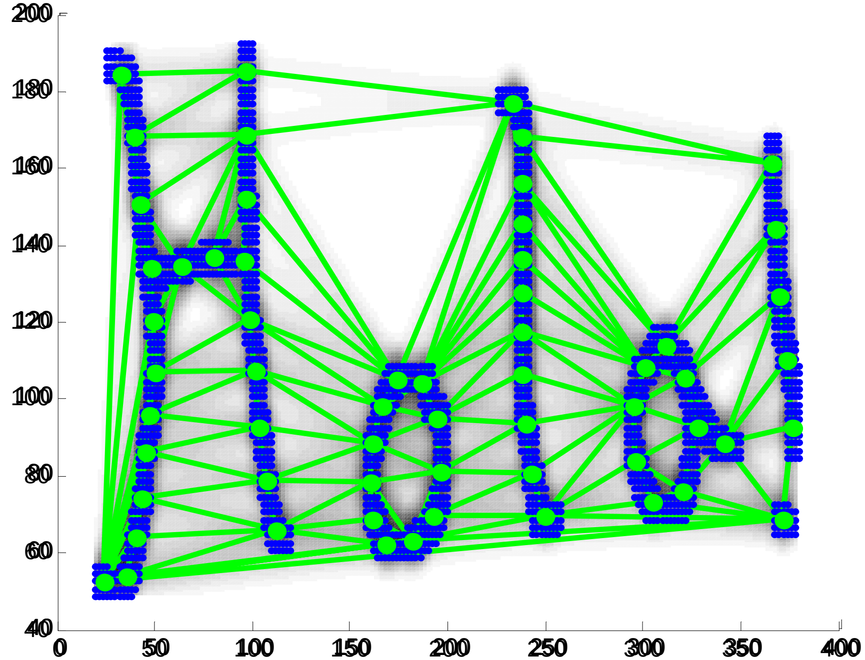


Hola !




220
200
180
160
140
120
100
80
60
40
20
0
50
100
150
200
250
300
350
400
450
500

1. Initialization

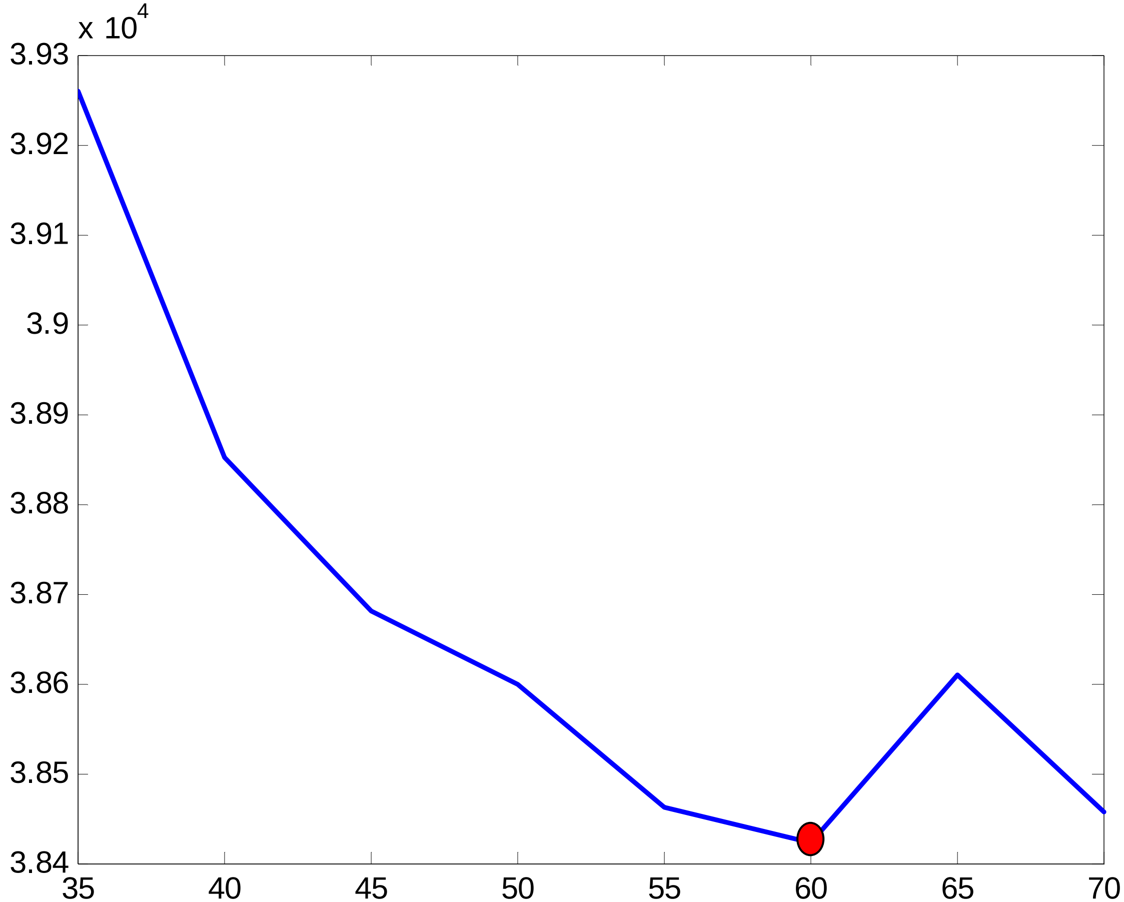


Location of the prototypes with a « classical » isotropic GM

…and then building of the **Delaunay Graph**
Initialize the generative model
(equiprobability of the components)

Number of prototypes

min BIC ~ - Likelihood + Complexity of the model



2. Learning

$$p(x) = \sum_{j \in J} p(j)\; p(x \mid j, \sigma)$$

update
the **variance of the Gaussian noise**,
the **weights of the components,**
and the location of the prototypes

with the **EM algorithm**

in order to maximize the **Likelihood** of the model
w.r.t the *N* observed data :

$$L(\underline{\pi}, \sigma; x, DG) = \prod_{i=1}^{N} p(x_i; DG, \underline{\pi}, \sigma)$$

EM updates

$$p(x, c; \underline{\pi}, \underline{\beta}, \underline{w}, \sigma, DG) = \sum^{1} \sum^{N_d} \pi_j^d \beta_{cj}^d g(x|(d,j);\sigma)$$

$$L(\underline{\pi}, \underline{\beta}, \underline{w}, \sigma, DG) = \prod_{i=1}^{M} p(x_i, c_i; \underline{\pi}, \underline{\beta}, \underline{w}, \sigma, DG)$$

$$\begin{aligned}
\pi_j^{d[\text{new}]} &= \frac{1}{M} \sum_{i=1}^{M} p((d,j)|x_i, c_i) \\
\sigma^{2[\text{new}]} &= \frac{1}{DM} \sum_{i=1}^{M} \Bigg[\sum_{j=1}^{N_0} p((0,j)|x_i, c_i)(x_i - w_j)^2 \\
&\quad + \sum_{j=1}^{N_1} p((1,j)|x_i, c_i) \frac{(2\pi\sigma^2)^{-D/2} \exp(-\frac{(x_i - q_j^i)^2}{2\sigma^2})(I_1[(x_i - q_j^i)^2 + \sigma^2] + I_2)}{L_j \cdot g(x_i|(1,j);\sigma)} \Bigg] \\
\beta_{cj}^{d[\text{new}]} &= \frac{\sum_{i=1:c_i=c}^{M} p(d,j|x_i,c_i)}{\sum_{i=1}^{M} p(d,j|x_i,c_i)}
\end{aligned} \qquad (9)$$

$$I_1 = \sigma \sqrt{\frac{\pi}{2}} \left(\text{erf}\left(\frac{Q_j^i}{\sigma\sqrt{2}}\right) - \text{erf}\left(\frac{Q_j^i - L_j}{\sigma\sqrt{2}}\right) \right)$$

$$I_2 = \sigma^2 \left((Q_j^i - L_j) \exp\left(-\frac{(Q_j^i - L_j)^2}{2\sigma^2}\right) - Q_j^i \exp\left(-\frac{(Q_j^i)^2}{2\sigma^2}\right) \right)$$

EM updates

$$w_k^{[\text{new}]} = \frac{\sum_{i=1}^{M}\left[p(0,k|x_i)x_i + \sum_{j\in W_k} p(1,j|x_i)\frac{g^{O}(x_i|q_j^i;\sigma)}{L_j\cdot g_j^1(x_i;\sigma)}(-E_2 w_{b_j} + E_3 x_i)\right]}{\sum_{i=1}^{M}\left[p(0,k|x_i) + \sum_{j\in W_k} p(1,j|x_i)E_1\right]}$$

$$E_1 = \frac{\sigma^2}{L_j^2}[e^{\frac{-(Q_j)^2}{2\sigma^2}}(Q_j - 2L_j) - e^{\frac{-(Q_j-L_j)^2}{2\sigma^2}}(Q_j - L_j)] + \frac{1}{L_j^2}((L_j - Q_j)^2 + \sigma)I_1$$

$$E_2 = \frac{\sigma^2}{L_j^2}[e^{-\frac{(Q_j-L_j)^2}{2\sigma^2}}Q_j - e^{-\frac{(Q_j)^2}{2\sigma^2}}(Q_j - L_j)] - \frac{1}{L_j^2}(Q_j^2 - L_jQ_j + \sigma^2)I_1$$

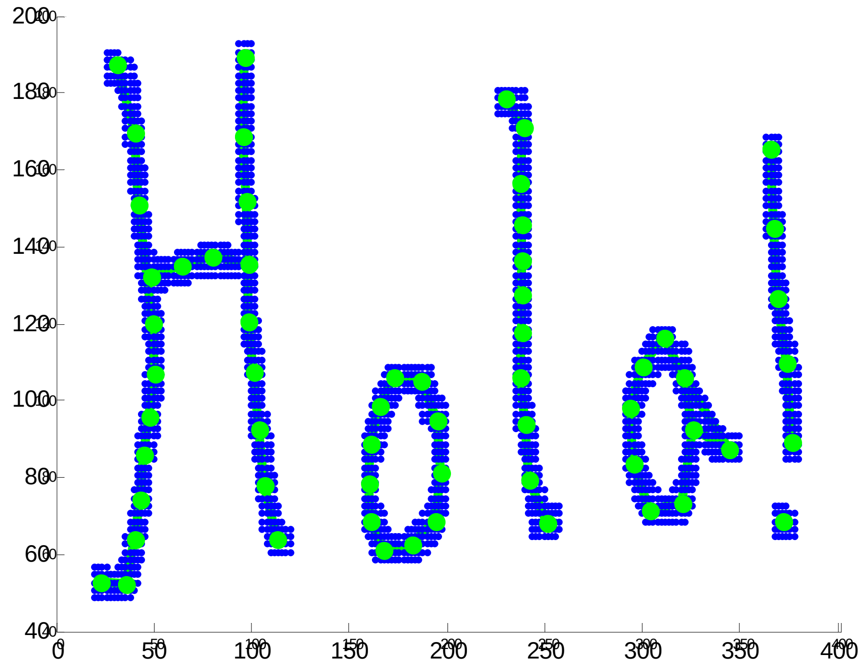
$$E_3 = \frac{1}{L_j}[e^{\frac{-(Q_j-L_j)^2}{\sigma^2}} - e^{\frac{-Q_j^2}{\sigma^2}} + (Q_j - L_j)I_1]$$

$$Q_j^i = \frac{\langle x_i - w_{a_j} | w_{b_j} - w_{a_j}\rangle}{L_j}$$

$$q_j^i = w_{a_j} + (w_{b_j} - w_{a_j})\frac{Q_j^i}{L_j}$$

Proposed approach : 3 steps

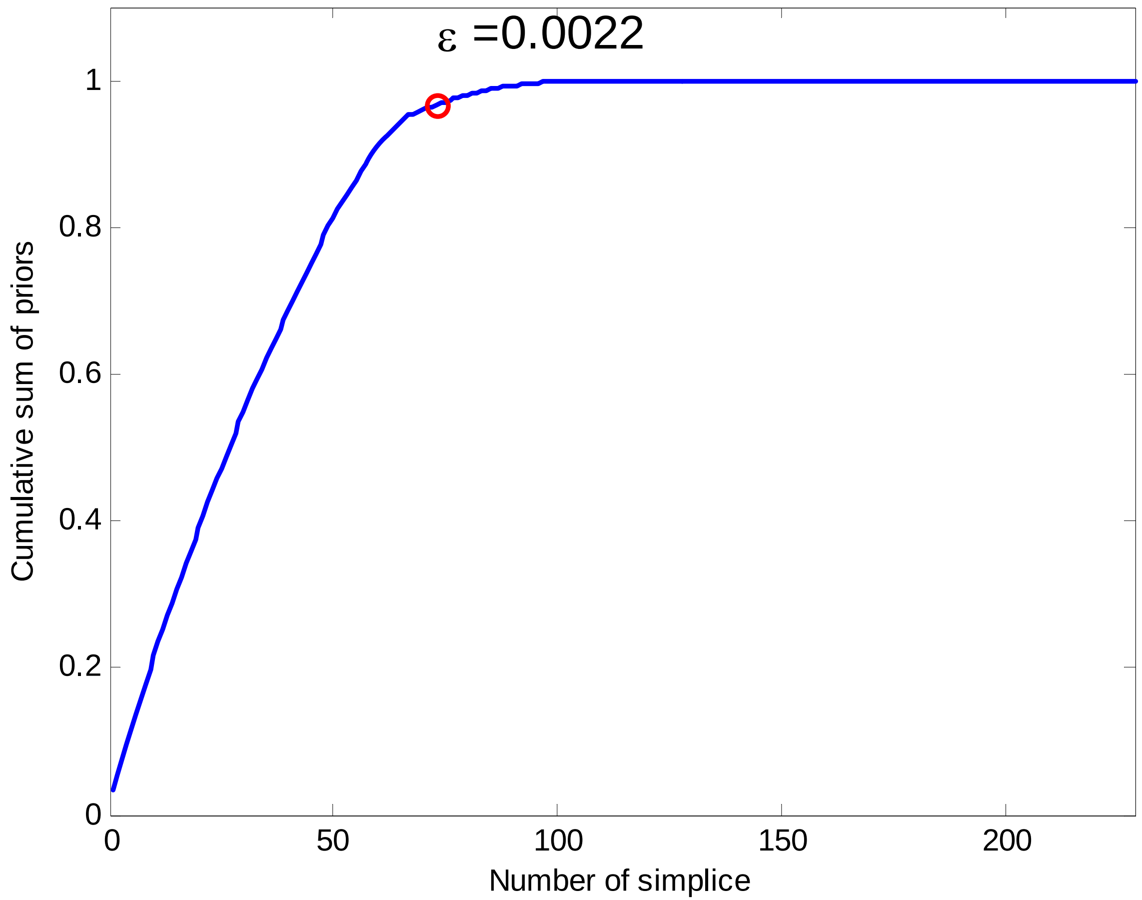
3. After the learning



Some components have a (quasi-) nul probability (weights):
They do not explain the data and can be prunned from the initial graph

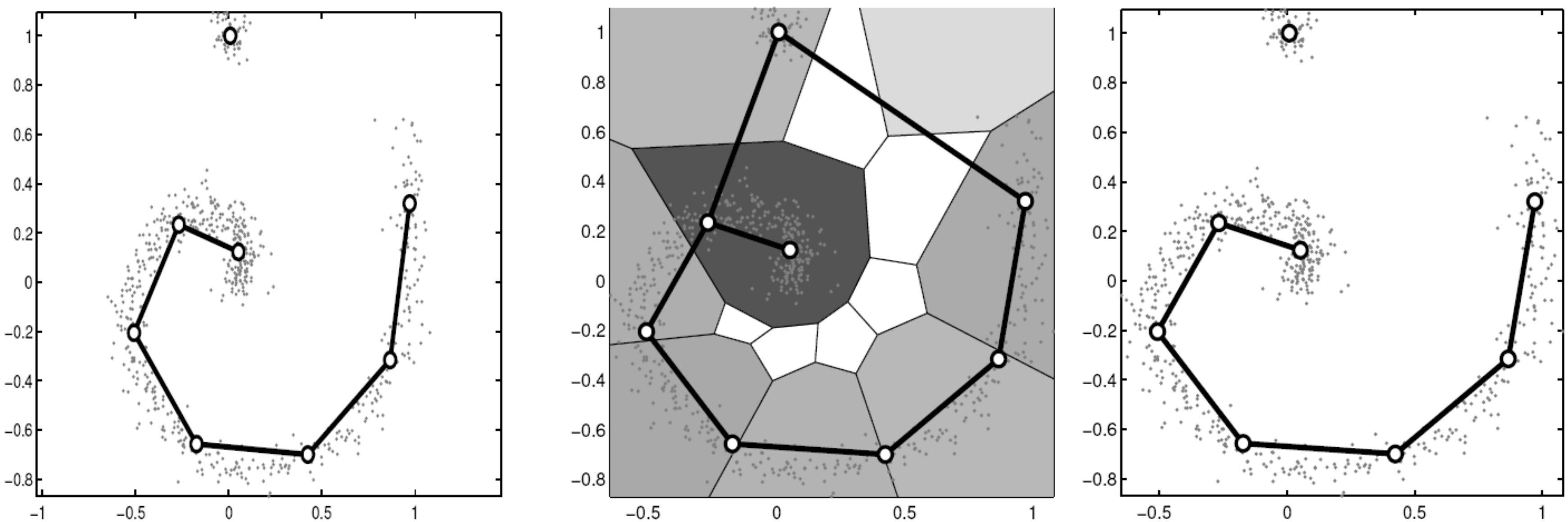
Threshold setting

~ « Cattell Scree Test »



Toy Experiment

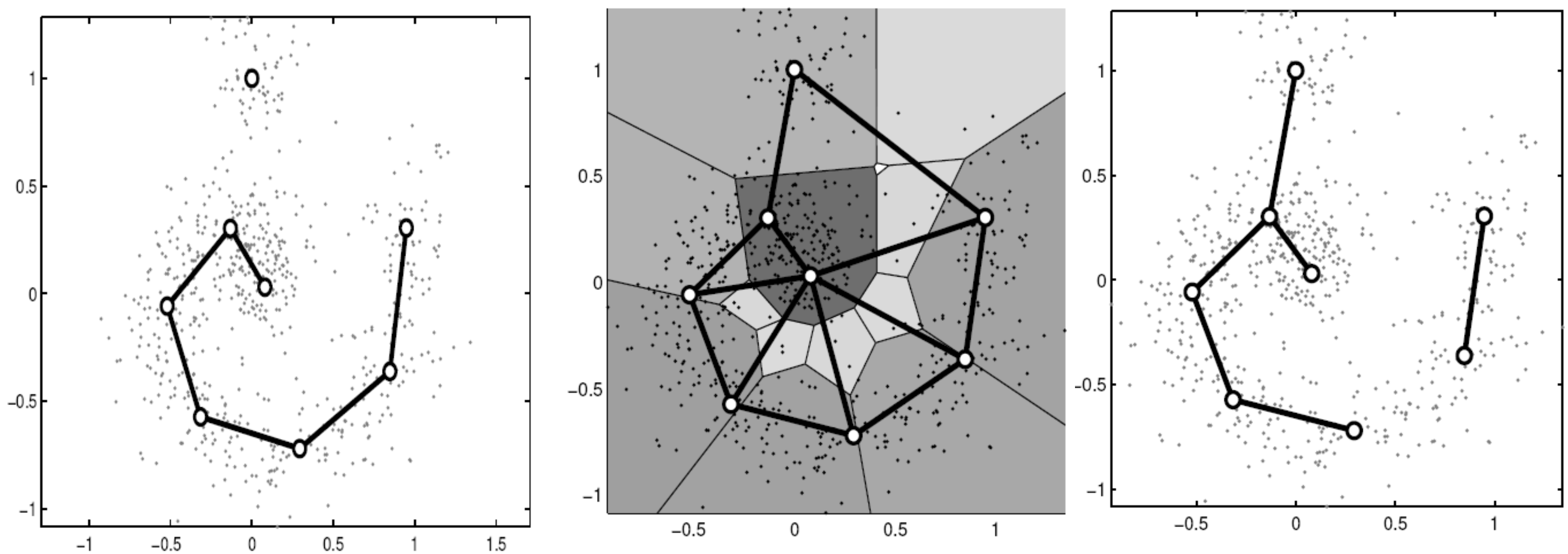
$$\sigma_{noise} = 0.05$$



(a) GGG: $\sigma^* = 0.06$ (d) CHL: $T = 0$ (g) CHL: $T^* = 60$

Seuillage sur le nombre de witness

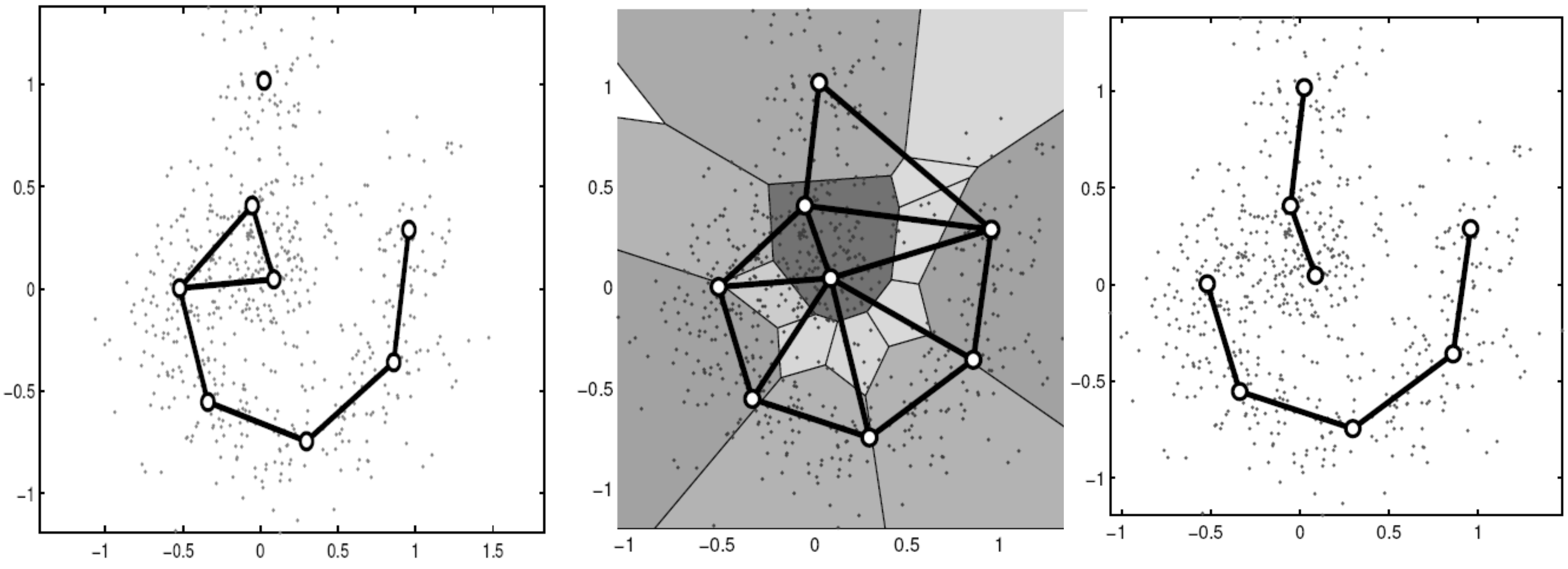
$$\sigma_{noise} = 0.15$$



(b) GGG: $\sigma^* = 0.17$ (e) CHL: $T = 0$ (h) CHL: $T^* = 65$

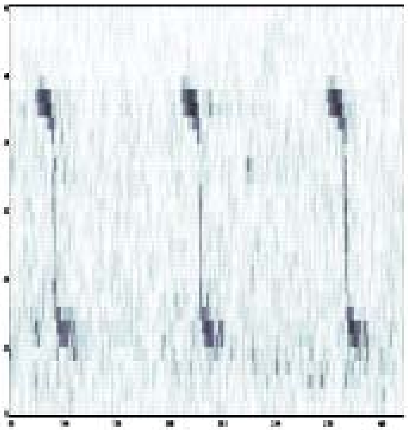
Toy Experiment

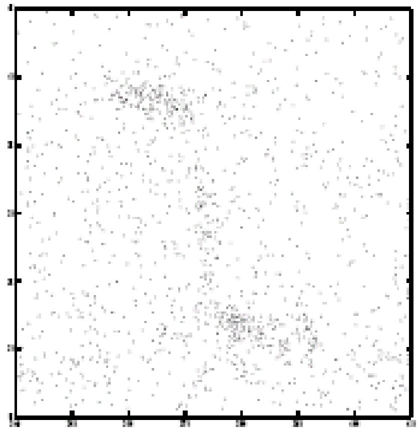
$$\sigma_{noise} = 0.2$$

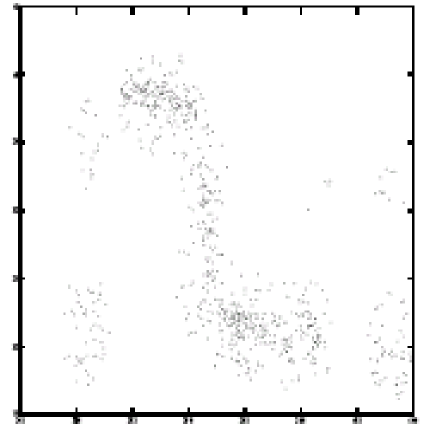


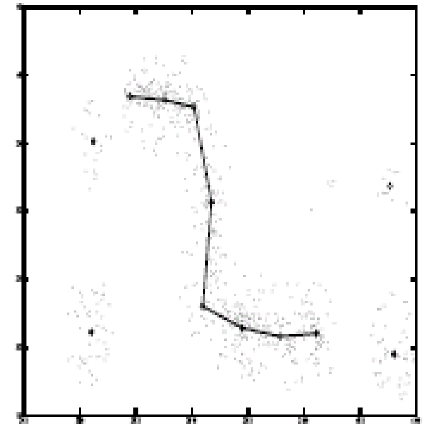
(c) GGG: $\sigma^* = 0.21$ (f) CHL: $T = 0$ (i) CHL: $T^* = 58$

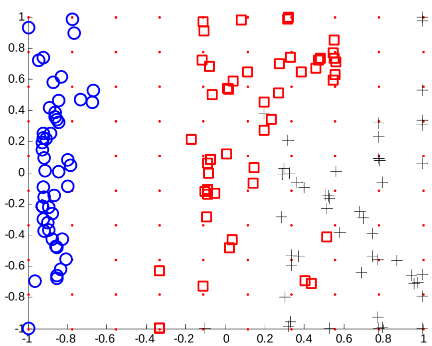
Other applications

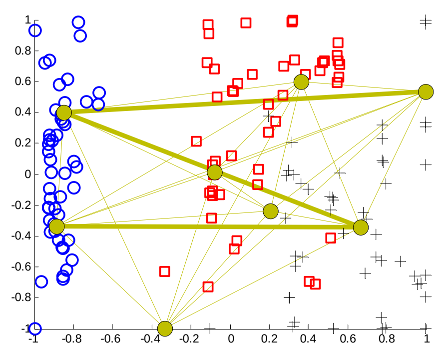


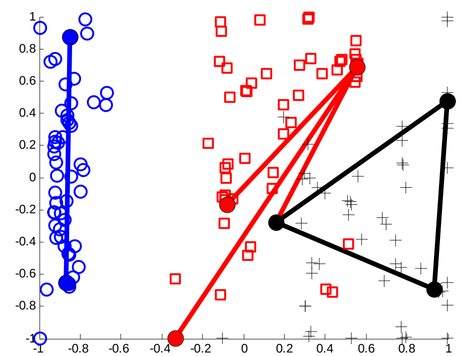


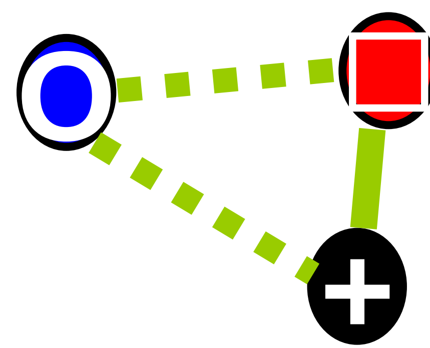












Comments

- **There is « no free lunch »**
  - **Time Complexity** $O(DN^3)$ (initial Delaunay graph)
  - **Slow convergence** (EM)
  - **Local optima**

Key Points

- **Statistical learning of the topology of a data set**
  - Assumption :
    - » Initial Delaunay graph is rich enough to contain a sub-graph having the same topology as the principal manifolds
  - Based on a statistical criterion (the likelihood) available in any dimension

- **« Generalized » Gaussian Mixture**
  - Can be seen as a generalization of the « Gaussian mixture » (no edges)
  - Can be seen as a finite mixture (number of « Gaussian-segment ») of an infinite mixture (Gaussian-segment)

- This preliminary work is an attempt to bridge the gap between Statistical Learning Theory and Computational Topology

Open questions

- **Validity of the assumption**

  « good » penalized-likelihood = « good » topology

- Theorem of «universal approximation» of manifold ?

Related works

- Publications NIPS 2005 (unsupervised) and ESANN 2007 (supervised: analysis of the iris and oil flow data sets)

- Workshop submission at NIPS on this topic

in collaboration with **F. Chazal** (INRIA Futurs) , **D. Cohen-Steiner** (INRIA Sophia), **S. Canu and G.Gasso** (INSA Rouen)

Thanks

Equations

$$p(x, c; \underline{\pi}, \underline{\beta}, \underline{w}, \sigma, DG) = \sum_{d=0}^{1} \sum_{j=1}^{N_d} \pi_j^d \beta_{cj}^d g(x|(d,j); \sigma)$$

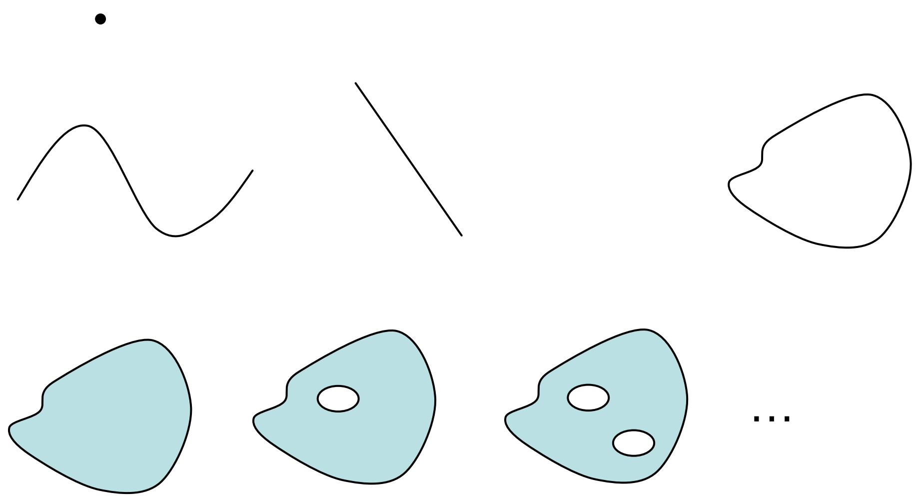
$$L(\underline{\pi}, \underline{\beta}, \underline{w}, \sigma, DG) = \prod_{i=1}^{M} p(x_i, c_i; \underline{\pi}, \underline{\beta}, \underline{w}, \sigma, DG)$$

$$\begin{aligned}
\pi_j^{d[\text{new}]} &= \frac{1}{M} \sum_{i=1}^{M} p((d,j)|x_i, c_i) \\
\sigma^{2[\text{new}]} &= \frac{1}{DM} \sum_{i=1}^{M} \Big[\sum_{j=1}^{N_0} p((0,j)|x_i, c_i)(x_i - w_j)^2 \\
&\quad + \sum_{j=1}^{N_1} p((1,j)|x_i, c_i) \frac{(2\pi\sigma^2)^{-D/2} \exp(-\frac{(x_i - q_j^i)^2}{2\sigma^2})(I_1[(x_i - q_j^i)^2 + \sigma^2] + I_2)}{L_j \cdot g(x_i|(1,j); \sigma)} \Big] \\
\beta_{cj}^{d[\text{new}]} &= \frac{\sum_{i=1:c_i=c}^{M} p(d,j|x_i, c_i)}{\sum_{i=1}^{M} p(d,j|x_i, c_i)}
\end{aligned} \qquad (9)$$

$$I_1 = \sigma\sqrt{\frac{\pi}{2}}\left(\text{erf}\left(\frac{Q_j^i}{\sigma\sqrt{2}}\right) - \text{erf}\left(\frac{Q_j^i - L_j}{\sigma\sqrt{2}}\right)\right)$$

$$I_2 = \sigma^2 \left((Q_j^i - L_j) \exp\left(-\frac{(Q_j^i - L_j)^2}{2\sigma^2}\right) - Q_j^i \exp\left(-\frac{(Q_j^i)^2}{2\sigma^2}\right) \right)$$

intrinsic dimension



$d_i=0$

$d_i=1$

$d_i=2$

Number of holes (Betti), connectedness…

…

Homeomorphism: topological equivalence