A Generative Gaussian Graph to Learn the Topology of a Set of Points

Michaël Aupetit
CEA - DASE
BP 12 - 91680
Bruyères-Le-Châtel, France
aupetit@dase.bruyeres.cea.fr

Abstract - *We propose a generative model based on a Delaunay graph to learn the topology of a set of points. It uses the maximum likelihood principle to tune its parameters. This work is a first step towards a topological model of a set of points grounded on statistics.*

Key words - topology modelling, Delaunay graph, maximum likelihood, generative model, mixture model

1 Introduction

In many applications [10, 1, 3], it is intended to model the topology of a manifold[1] $\mathcal{M} \subset \Delta \subset \mathbb{R}^D$ only known through a finite set $\underline{v} \subset \mathcal{M}$ of M data points in a bounded domain Δ of a D-dimensional euclidean space. $\mathcal{M}$ is the support of the probability density function (pdf) p from which are drawn the data. In fact, this is not the topology of $\mathcal{M}$ which is of interest, but the topology of manifolds $\mathcal{M}^{prin}$ called "principal manifolds" of the distribution p (in reference to the definition of Tibshirani [12]) which can be viewed as the manifold $\mathcal{M}$ without the noise.

In this work, we assume that any sample of $\mathcal{M}$ is generated on $\mathcal{M}^{prin}$ and corrupted with additive spherical gaussian noise with mean 0 and variance σ^2_{noise}. So that p is the convolution over $\mathbb{R}^D$ of some unknown pdf p^{prin} with support $\mathcal{M}^{prin}$ and a Normal pdf $\mathcal{N}$ with isovariance σ^2_{noise}.

There are two families of approaches which deal with "topology" : on one hand, the "topology preserving" approaches based on nonlinear projection of the data in lower dimensional spaces with a constrained topology to allow visualization [5, 8]; on the other hand, the "topology modelling" approaches based on the construction of a structure whose topology is not constrained *a priori*, so it better accounts for that of the data [2, 10] at the expense of the visualisability. In this work, we expose a new approach part of the latter family.

A way to create an explicit model of the topology of $\mathcal{M}^{prin}$ is to define a graph $G(\underline{W}, \underline{E}(\underline{w}, \underline{v}))$, where the set of vertices $\underline{W}$ is in one to one correspondence with a set $\underline{w}$ of N_0 vector prototypes in $\mathbb{R}^D$ and $\underline{E}$ is a set of edges $\{i, j\}$ connecting W_i and W_j. If the prototypes are "well" located on the data distribution, and the graph is "well" chosen, then the topology of

[1] In this paper, $\mathcal{M}$ is called a manifold for short, it is in fact a set of manifolds which may have various intrinsic dimensions.

the graph accounts for the one of the manifold $\mathcal{M}^{prin}$. We assume such a "good" location may be obtained using a Vector Quantization (VQ) algorithm (*e.g.* the Neural-Gas [9]) leading the pdf of the prototypes to roughly approximate the pdf p^{prin} of the data distribution. The main problem we focus on, is the choice of a "good" graph.

Each prototype w_i is representative of the data for which it is the closest prototype among $\underline{w}$. All these data fall in a region $\mathcal{M}_i$ which is the intersection between the manifold $\mathcal{M}$ and the Voronoï cell $\mathcal{V}_i$ of w_i defined as [11]:

$$\mathcal{V}_i = \mathcal{V}_{(\Delta,\underline{w})}(w_i) = \{v \in \Delta \mid \forall w_j \in \underline{w}, \|v - wi\| \leq \|v - w_j\|\} \tag{1}$$

Let $DG = DG(\underline{W}, \underline{E}_{DG})$ be the Delaunay graph of $\underline{w}$, for which $\underline{E}_{DG}$ is defined as the set of edges which connect prototypes whose Voronoï cells share a common boundary:

$$\underline{E}_{DG} = \underline{E}_{DG}(\underline{w}) = \left\{\{i,j\} \in (1,\ldots,N_0)^2 \mid \mathcal{V}_i \cap \mathcal{V}_j \neq \emptyset\right\}. \tag{2}$$

A set of pieces of $\mathcal{M}$ which are connected to $\mathcal{M}_i$ is the set of manifolds $\mathcal{M}_j$ for which $\mathcal{V}_j$ shares a common boundary with $\mathcal{V}_i$ *i.e.* such that w_j is a Delaunay neighbor of w_i. The Delaunay graph is then an appealing candidate to model the topology of the data distribution. However, it remains to prune it in a relevant way.

Martinetz and Schulten [10] devised an algorithm called "Competitive Hebbian Learning" (CHL) which forms a Topology Representing Network (TRN). The CHL constructs a graph DG_{CHL} whose topology approximate that of $\mathcal{M}^{prin}$: it considers each datum of $\underline{v}$, looking for its closest and second closest prototypes in $\underline{w}$, and then creating an edge between both these prototypes:

$$\underline{E}_{CHL} = \underline{E}_{CHL}(\underline{w}, \underline{v}) = \left\{\{i,j\} \in (1,\ldots,N_0)^2 \mid \exists v \in \underline{v}, v \in \mathcal{V}_{\{i,j\}}\right\} \tag{3}$$

with $\mathcal{V}_{\{i,j\}} = \left\{v \in E \mid v \in \mathcal{V}_{(\Delta,\underline{w}\setminus w_i)}(w_j) \cap \mathcal{V}_{(\Delta,\underline{w}\setminus w_j)}(w_i)\right\}$.

The set $\mathcal{V}_{\{i,j\}}$ is called the 2^{nd}-order Voronoï cell associated to $\{w_i, w_j\}$. Such a cell exists for any couple of Delaunay neighbors and only for them (see Figure 2). Notice that DG_{CHL} is a graph, hence bears no information about the local dimension of the manifolds in the collection $\mathcal{M}$. However in most of the applications, it is sufficient to compute the topological information needed, *e.g.* pathwise connectedness is useful to compute shortest paths on manifolds [8, 15] or to explore the topology of high-dimensional labelled data [3].

Despite its simplicity and its low time complexity ($O(DMN_0)$) to model the topology, the CHL algorithm is prone to a series of limits (see Figure 2):

1. There is no energy function minimized during the construction of the graph. Consequently, there is neither way to check the quality of the graph obtained nor to compare which one of two graphs is the best. In general, visual inspection is necessary to insure the graph "looks like" the data distribution, which is not possible in dimension higher than 3.

2. If the data set is corrupted with noise, the connectedness of the support manifold including the noise $\mathcal{M}$ will be represented because one datum is sufficient to create an edge. It would be more interesting to filter out the noise in order to represent the connectedness of the principal manifolds $\mathcal{M}^{prin}$. Martinetz and Schulten proposed an aging process by associating an "age" to the edges and a kind of hit frequency high-pass

filtering to prune older edges. But setting the age threshold and the aging machinery to control the noise sensitivity, is neither trivial nor intuitive without any objective criterion to do so (visual inspection is still needed).

3. The 2^{nd}-order Voronoï regions $\mathcal{V}_{\{i,j\}}$, which are the regions of influences of the Delaunay edges, *i.e.* the regions which must contain at least one datum for the corresponding edge to be created, may have no intersection at all with the segment $[w_i w_j]$. In other words, data sampled from the segment $[w_i w_j]$ itself may not give rise to it.

4. The desired graph, as a subgraph of DG, may span a set of points and segments. However, the CHL without the aging machinery cannot keep isolated prototypes in order to represent isolated bumps in the distribution, because for $N_0 \geq 2$, there always exists a first and a second closest prototypes to a datum, which are then connected.

5. The area of the region of influence of an edge may be very tiny depending on the location of the prototypes, but this is not related trivially to the length of the corresponding segment. This may prevent an edge from being created in cases where it should obviously be.

Two other approaches are known:

- The Optimally Topology Preserving Map (OTPM) proposed by Bruske and Sommer [6] which uses just the same algorithm as TRN but with weaker hypotheses thanks to a definition of topology preservation proposed by Villmann *et al.* [14] which states that a graph optimally represents the topology of a set of points if it is identical to DG_{CHL}.

- The Robust TRN (RTRN) proposed by the author [2], which attempts to take into account the shape of the pdf of the data projected on the closest Delaunay segments to decide wether the corresponding edges have to be pruned or not.

The OTPM shares the same limits as the TRN, the definition of optimally topology preservation being self-referent, it cannot solve the item 1, and it ignores the aging process allowing to filter the noise in item 2. The RTRN has been proposed to deal with item 2 to 5, and defines another heuristic than the aging process to find the best model with noisy data. But none of these approaches answers the first crucial point: in order to deal with data in dimensions higher than 3 and to be confident with the result, we need an error function at least intuitively relevant, at best statistically relevant. RTRN opened up the way by considering the density of the data, and more intuitive regions of influence for the edges.
Here, we propose to go beyond all the previous limits by considering a statistical two-phase approach:

- First, the Delaunay graph (or an approximation of it) of the prototypes is constructed independently of the data set.

- Second, each edge and each vertex of the graph is the basis of a generative model so that the entire set of edges and vertices generates a mixture of gaussian density functions. The maximization of the likelihood of the data *wrt* the model, allows to tune the weights of this mixture and leads to the emergence of the expected graph through the edges with non-zero weights that remain at the end of the optimization process.

We first present the algorithm we use in section 2. Then we test it on toy problems in section 3 before the discussion and conclusion.

2 A Generative Gaussian Graph to learn topology

2.1 Computing the Delaunay graph

In that work, we consider a low dimensional space, so we can use available softwares to compute the Delaunay graph, such as Qhull [4].

2.2 Generative Gaussian Graph

We assume the data have been generated by some set of points and segments constituting the collection of manifolds $\mathcal{M}^{prin}$ which have been corrupted with additive spherical gaussian noise with mean 0 and unknown variance σ^2_{noise}. Then, we define a gaussian mixture model to account for the observed data, which is based on both gaussian kernels that we call "gaussian-points", and what we call "gaussian-segments", forming a "Generative Gaussian Graph" (GGG).

The value at point v of a normalized gaussian-point centered on a prototype $w_i \in \underline{w}$ with variance σ^2 is defined as: $g^0(v, w_i, \sigma) = (2\pi\sigma^2)^{-D/2}\exp(-\frac{(v-w_i)^2}{2\sigma^2})$

A normalized gaussian-segment is defined as the sum of an infinite number of gaussian-points evenly spread on a line segment. Thus, this is the integral of a gaussian-point along a line segment. The value at point v of the gaussian-segment $[w_a w_b]$ associated to the edge $\{a, b\}$ with variance σ^2 is:

$$\begin{aligned} g^1(v, \{w_a, w_b\}, \sigma) &= (2\pi\sigma^2)^{-\frac{D-1}{2}}\exp\left(-\frac{(v-v_p)^2}{2\sigma^2}\right) \cdot \frac{\int_{w_a}^{w_b}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{(v_p-w)^2}{2\sigma^2}\right)dw}{\|w_b - w_a\|} \\ &= (2\pi\sigma^2)^{-\frac{D-1}{2}}\exp\left(-\frac{(v-v_p)^2}{2\sigma^2}\right) \cdot \frac{\mathrm{erf}\left(\frac{n(v,w_a)}{\sigma\sqrt{2}}\right) - \mathrm{erf}\left(\frac{n(v,w_b)}{\sigma\sqrt{2}}\right)}{2\|w_b - w_a\|} \end{aligned} \tag{4}$$

where $\|.\|$ denotes the euclidean norm, v_p is the orthogonal projection of v on the straight line passing through w_a and w_b, and $\forall w_x \in \{w_a, w_b\}, n(v, w_x) = \frac{\langle v-w_x | w_b-w_a\rangle}{\|w_b-w_a\|}$. In the case where $w_a = w_b$, we set $g^1(v, \{w_a, w_b\}, \sigma) = g^0(v, w_a, \sigma)$.

The left part of the dot product accounts for the gaussian noise orthogonal to the line segment, and the right part for the gaussian noise integrated along the line segment.The functions g^0 and g^1 are positive and: $\int_{\mathbb{R}^D} g^0(v, w_i, \sigma)dv = \int_{\mathbb{R}^D} g^1(v, \{w_a, w_b\}, \sigma)dv = 1$, so they are both probability density functions[2]. A gaussian-point is associated to each prototype in $\underline{w}$ and a gaussian-segment to each segment $[w_a w_b]$ such that the edge $\{a, b\} \in \underline{E}_{DG}$.

The gaussian mixture is obtained by a weighting sum of the N_0 gaussian-points and N_1 gaussian-segments, such that the weights $\underline{\pi}$ sum to 1 and are non-negative:

[2]Proof: $g^1(v, \{w_a, w_b\}, \sigma) = \int_{[w_a w_b]} g^0(v, w, \sigma)dw/\|w_b - w_a\|$ and $\int_{\mathbb{R}^D} g^0(v, w, \sigma)dv = 1$ so $\int_{\mathbb{R}^D} g^1(v, \{w_a, w_b\}, \sigma)dv = \int_{[w_a w_b]}\int_{\mathbb{R}^D} g^0(v, w, \sigma)dvdw/\|w_b - w_a\| = \int_{[w_a w_b]} dw/\|w_b - w_a\| = 1$.

$$p(v|\underline{\pi}, \underline{w}, \sigma, DG) = \sum_{k=0}^{1} \sum_{i=1}^{N_k} \pi_i^k g^k(v, s_i^k, \sigma) \quad (5)$$

with $\sum_{k=0}^{1} \sum_{i=1}^{N_k} \pi_i^k = 1$ and $\forall i, k,\ \pi_i^k \geq 0$, where $s_i^0 = w_i$ and $s_i^1 = \{w_a, w_b\}$ such that $\{a, b\}$ is the i^{th} edge in $\underline{E}_{DG}$. The weights π_i^0 (resp.π_i^1) denote the probability that the datum v was drawn from the gaussian-point associated to w_i (resp.the gaussian-segment associated to the i^{th} edge of the DG).

2.3 Measure of quality

The function $p(v|\underline{\pi}, \underline{w}, \sigma, DG)$ is the probability density of v given the parameters of the model. We measure the likelihood of the data $\underline{v}$ *wrt* the parameters of the GGG model, *i.e.* the probability density of the data set $\underline{v}$ assuming the data are iid:

$$P = P(\underline{\pi}, \underline{w}, \sigma, DG) = \prod_{k=1}^{M} p(v_k|\underline{\pi}, \underline{w}, \sigma, DG) \quad (6)$$

2.4 Learning the topology by maximizing the likelihood

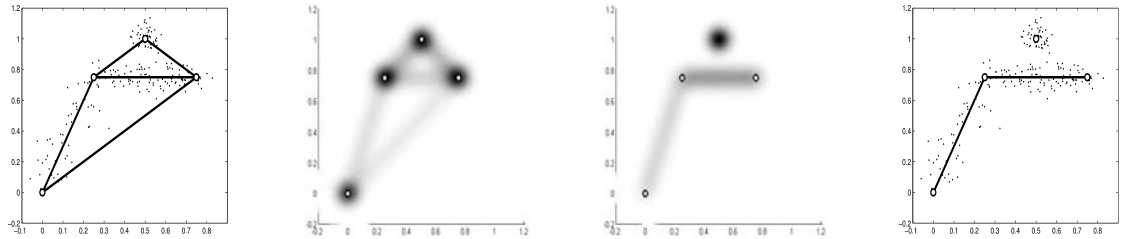
The core idea is to prune from the DG the edges for which there is no chance they generated the data. The algorithm is the following:

1. Initialize the location of the prototypes $\underline{w}$ using vector quantization [9].
2. Construct the Delaunay graph of the prototypes $DG(\underline{W}, \underline{E}_{DG}(\underline{w}))$.
3. Initialize the weights $\underline{\pi}$ to $1/(N_0 + N_1)$ to give equiprobability to each vertices and edges.
4. Given $\underline{w}$ and DG, find σ^* and $\underline{\pi}^*$ maximizing the likelihood P, or equivalently minimizing the log-likelihood $L = -\log(P)$, subject to $\sigma > 0$, $\pi \geq 0$ and $\sum_{\pi \in \underline{\pi}} \pi = 1$.
5. Prune the edges $e_i \in \underline{E}_{DG}$ associated to the gaussian segment with probability $\pi_i^1 = 0$ where $\pi_i^1 \in \underline{\pi}^*$.

The topology emerges from the edges with non-zero probabilities $\underline{\pi}$. This graph is the one which best models the topology of the data in the sense of the maximum likelihood *wrt* the parameters $\underline{\pi}$ and σ.

3 Experiments

In these experiments, we want to verify the relevance of the GGG to learn the topology in various noise conditions. The principle of the GGG is shown on the Figure 1. On the Figure 2, we show the comparison of the GGG to a CHL for which we filter out edges which have a number of hits lower than a threshold T. The data and prototypes are the same for both algorithms. We set T^* such that the graph obtained matches as close as possible the expected solution. We optimize separately σ and $\underline{\pi}$, by setting σ to different candidate values, and finding $\underline{\pi}$ using "fmincon" Matlab constraint optimization function. The conditions and conclusion of the experiments are given in the caption of the figures.



(a) Initial Delaunay graph (b) Corresponding initial GGG (c) Optimal GGG (d) Non-zero probability edges

Figure 1: **Principle of the Generative Gaussian Graph**: (a) Data drawn from an oblique segment, an horizontal one and an isolated point with respective density $\{0.25; 0.5; 0.25\}$. The prototypes are located at the extreme points of the segments, and at the isolated point. They are connected with edges from the Delaunay graph. (b) The corresponding initial Generative Gaussian Graph. (c) The optimal GGG obtained after optimization of the likelihood according to σ and $\underline{\pi}$. (d) The edges of the optimal GGG associated to non-zero probabilities model the topology of the data.

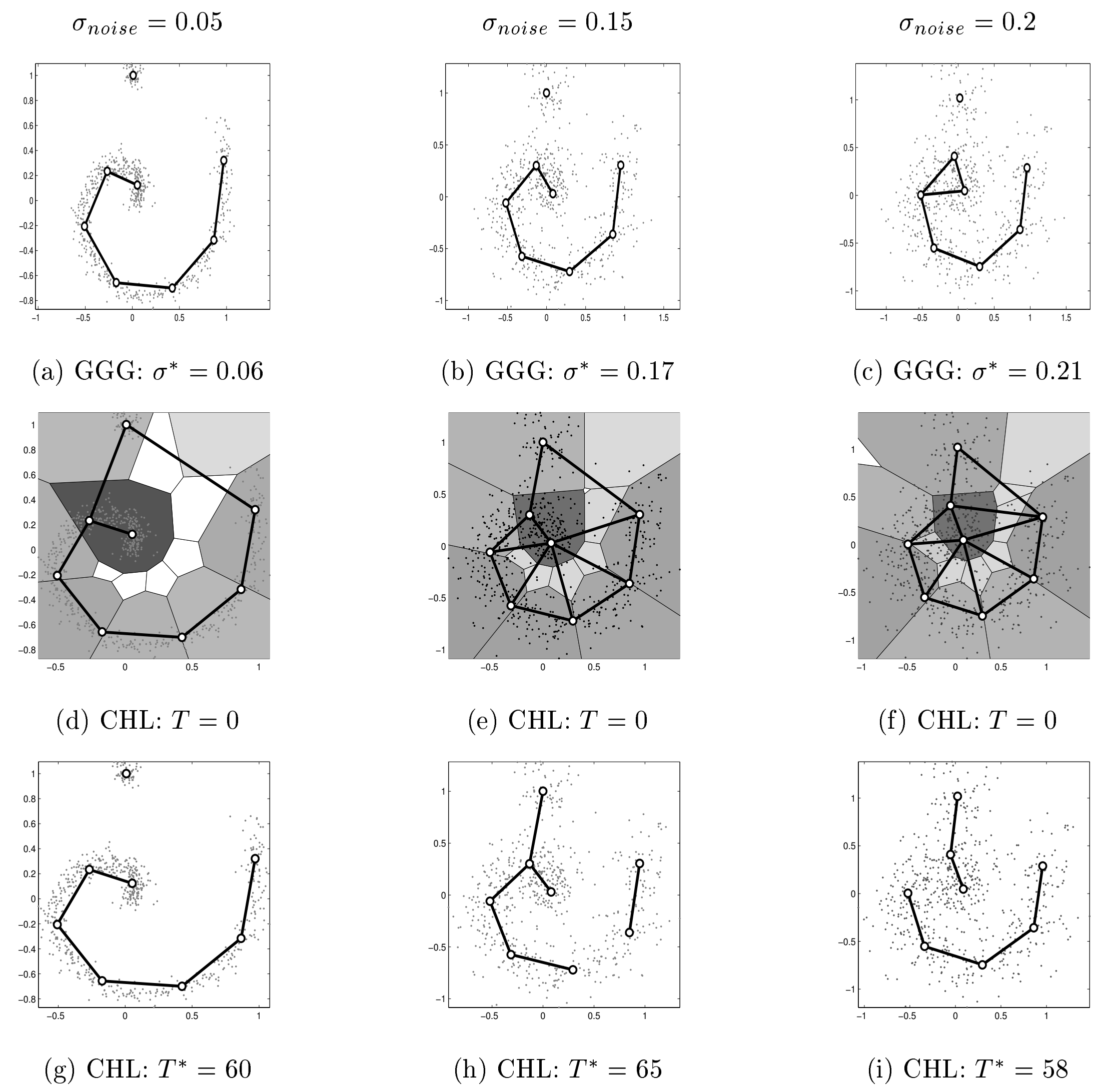
4 Discussion

We propose to model the topology of a data set, using a generative mixture model that we call Generative Gaussian Graph. Although its time complexity is higher than that of the Competitive Hebbian Learning, due to the learning phase based on maximization of the likelihood, GGG is an attempt to avoid the limits of the CHL for modelling topology:

1. The likelihood of a mixture model is maximized. The quality is assessed by the value of the likelihood, allowing comparisons. The higher is the likelihood of the model, the higher is the topological similarity between the generative manifold $\mathcal{M}^{prin}$ of the observed data and the generative manifold based on points and segments of the model.
2. No threshold is needed, and the noise is taken into account in the model so the topology of the principal manifold $\mathcal{M}^{prin}$ is modelled.
3. The region of influence of a segment surrounds it. The segments associated to edges of the graph are full part of the model.
4. The model can represent isolated bumps.
5. The area of the region of influence of a segment is proportional to its length.

Generative Gaussian Graph can be viewed as a generalization of gaussian mixtures to points and segments: a gaussian mixture is a GGG with no edge. GGG provides at the same time an estimation of the data distribution density more accurate than the gaussian mixture based on the same set of prototypes and the same noise isovariance hypothesis, and intrinsically an explicite model of the topology of the data set. Notice other generative models do not provide any insight about the topology of the data, except the Generative Topographic Map (GTM) [5], the revisited Principal Manifolds [12] or the mixture of Probabilistics Principal Component Analysers (PPCA) [13]. However, in the two former cases, the intrinsic dimension of the model is fixed *a priori* and not learned from the data, while in the latter the local intrinsic dimension is learned but the connectedness between the local models is not.
There are several ways to follow to carry on with this work:

- learning by maximum likelihood the location and the noise covariance of each prototype;



(a) GGG: $\sigma^* = 0.06$ (b) GGG: $\sigma^* = 0.17$ (c) GGG: $\sigma^* = 0.21$

(d) CHL: $T = 0$ (e) CHL: $T = 0$ (f) CHL: $T = 0$

(g) CHL: $T^* = 60$ (h) CHL: $T^* = 65$ (i) CHL: $T^* = 58$

Figure 2: **Learning the topology of a data set**: 600 data drawn from a spirale and an isolated point corrupted with additive gaussian noise with mean 0 and variance σ^2_{noise}. (a-c) The edges of the GGG with non-zero weights allow to recover the topology of the principal manifolds except for large noise variance (c) where a triangle was created at the center of the spirale. σ^* over-estimates σ_{noise} because the model is piecewise linear while the true manifolds are non-linear. (d-f) The CHL without threshold (T=0) is not able to recover the true topology of the data for even small σ_{noise}. In particular, the isolated bump cannot be recovered. The grey cells correspond to 2^{nd}-order Voronoï cells (darker cells contain more data). It shows these cells are not intuitively related to the edges they are associated to (*e.g.* they may have very tiny areas (e), and may partly (d) or never (f) contain the corresponding line segment). (g-h) The CHL with a threshold T allows to recover the topology of the data only for small noise variance (g) (Notice $T_1 < T_2 \Rightarrow DG_{CHL}(T_2) \subseteq DG_{CHL}(T_1)$). Moreover, setting T requires visual control and is not associated to the optimum of any energy function which prevents its use in higher dimensional space.

- extending the gaussian graph to gaussian simplicial complexes (surfaces, volumes...), in order to get the full topological information of the data set (intrinsic dimension > 1);
- extending the approach to non-linear segments, to better fit the shape of the principal manifolds $\mathcal{M}^{prin}$ and the density p^{prin} over them;
- considering the Expectation-Maximization algorithm [7] and the Bayesian framework;
- defining benchmarks and protocols for topology modelling in more than 3 dimensions, in order to compare the approaches and discuss their generalization capacities;
- exploring the conditions ensuring the existence of a maximum of the likelihood.

The Generative Gaussian Graph is an attempt to bridge the gap between Statistics and Computational Topology allowing the modelling of the topology of a data set to enter the framework of statistical learning theory.

References

[1] M. Aupetit. Induced voronoï kernels for principal manifolds approximation. *Advances in self-organizing maps, N. Allison, H. Yin, L. Allinson & J. Slack eds. Springer*, pages 54–60, 2001.

[2] M. Aupetit. Robust topology representing networks. *European Symp. on Artificial Neural Networks, Bruges (Belgium), d-side eds.*, pages 45–50, 2003.

[3] M. Aupetit and T. Catz. High-dimensional labeled data analysis with topology representing graphs. *Neurocomputing, Elsevier*, 63:139–169, 2005.

[4] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. Math. Software*, 22:469–483, 1996.

[5] C. M. Bishop, M. Svensén, and C. K. I. Williams. Gtm: the generative topographic mapping. *Neural Computation, MIT Press*, 10(1):215–234, 1998.

[6] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(5):572–575, 1998.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[8] J. A. Lee, A. Lendasse, and M. Verleysen. Curvilinear distance analysis versus isomap. *European Symp. on Artificial Neural Networks, Bruges (Belgium), d-side eds.*, pages 185–192, 2002.

[9] T. M. Martinetz, S. G. Berkovitch, and K. J. Schulten. "neural-gas" network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[10] T. M. Martinetz and K. J. Schulten. Topology representing networks. *Neural Networks, Elsevier London*, 7:507–522, 1994.

[11] A. Okabe, B. Boots, and K. Sugihara. *Spatial tessellations: concepts and applications of Voronoï diagrams.* John Wiley, Chichester, 1992.

[12] R. Tibshirani. Principal curves revisited. *Statistics and Computing*, (2):183–190, 1992.

[13] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[14] T. Villmann, R. Der, and T. Martinetz. A novel approach to measure the topology preservation of feature maps. *Int. Conf. on Artificial Neural Networks*, pages 289–301, 1994.

[15] M. Zeller, R. Sharma, and K. Schulten. Topology representing network for sensor-based robot motion planning. *World Congress on Neural Networks, INNS Press*, pages 100–103, 1996.