

Regularization of Mixture Models for Robust Principal Graph Learning

Tony Bonnaire, Aurélien Decelle and Nabila Aghanim

Abstract—A regularized version of Mixture Models is proposed to learn a principal graph from a distribution of D -dimensional data points. In the particular case of manifold learning for ridge detection, we assume that the underlying manifold can be modeled as a graph structure acting like a topological prior for the Gaussian clusters turning the problem into a maximum *a posteriori* estimation. Parameters of the model are iteratively estimated through an Expectation-Maximization procedure making the learning of the structure computationally efficient with guaranteed convergence for any graph prior in a polynomial time. We also embed in the formalism a natural way to make the algorithm robust to outliers of the pattern and heteroscedasticity of the manifold sampling coherently with the graph structure. The method uses a graph prior given by the minimum spanning tree that we extend using random sub-samplings of the dataset to take into account cycles that can be observed in the spatial distribution.

Index Terms—Gaussian Mixture Models, Expectation-Maximization, Graph regularization, Principal Graph, Manifold Learning.

1 INTRODUCTION AND CONTRIBUTIONS

DATA often come as a spatially organized set of discrete D -dimensional data points $\mathbf{X} = \{\mathbf{x}_i\}_{i=0,\dots,N}$ with $\mathbf{x}_i \in \mathcal{D} \subset \mathbb{R}^D$ sampled from an unknown probability distribution p . These datapoints are usually not spreading uniformly over the entire \mathbb{R}^D space but often result from the sampling of a lower dimensional manifold whose topology and characteristics are linked with the process that generated the data. Capturing this information, may it be for visual, geometrical or topological analyses of the dataset requires the application of non-linear methods that are parts of the rich field of manifold learning [1, 2]. Revealing and extracting patterns in such datasets with the least prior knowledge is the core of unsupervised machine learning and is usually essential to understand the underlying hidden structure, build models and make predictions. In many applications, data even appear as standing on a one-dimensional structure as, for instance, GPS measurements collected by vehicles standing on the road network [3], large-scale galaxy distribution describing the filamentary structure of the cosmic web [4] or vessel networks transporting blood through the human body [5].

The problem of estimating a one-dimensional manifold approximating the underlying distribution of \mathbf{X} is a particular case of dimensionality reduction [6] also known as *ridge detection* or *principal curve extraction*. The seminal work of [7] provides an intuitive definition of a principal curve as the smooth self-consistent and non-intersecting curve with

finite length passing in “the middle” of the point cloud distribution. As such, principal curves are a generalization of principal components and relax the straight line condition to describe non-linear relations between variables of multidimensional datasets. Since then, several works followed to obtain an uniform convergence bound [8], give more suitable definitions of these lines as ridges of a probability density [e.g. 9, 10] and extend them to higher dimensions (principal surfaces or volumes).

To allow the description of more flexible structures than those imposed by curves and to bypass their inability to represent self-intersecting or cycling topologies, a formulation relying on graph theory to model the one-dimensional structure was introduced in [11] and extended in [12]. This latter is based on predefined rules for growing the graph and includes regularization terms to limit its complexity. However, this model come with a large number of parameters and with no guaranteed convergence to which the double optimization scheme of [13] offers an alternative for the learning of a tree structure. In this landscape of methods, only a few address the problem of estimating a principal graph with a proper handling of outliers, which considerably complicates the learning of the graph structure. A built-in robustness to outliers is proposed in [14] and [15] by discarding from the update of a node position all datapoints beyond a robustness radius R_0 . However, the choice of R_0 is not trivial, scale-dependent and require a careful tuning to deal with uniform background noise [15].

In this work, we use a mixture model to approximate the data distribution p and regularize it over a graph structure to constrain Gaussian centroids to pave this approximation of the underlying manifold. The proposed method extends the original presentation of [9], that makes use of smooth differentiable curves, to a more general representation of the dataset given by a graph structure. Examples of ridge extractors based on the Subspace Constrained Mean-Shift algorithm [SCMS, 10, 16] offer a mathematically well-defined

• T. Bonnaire and N. Aghanim are with Université Paris-Saclay, CNRS, Institut d’astrophysique spatiale, 91405, Orsay, France.

• T. Bonnaire and A. Decelle are with Université Paris-Saclay, CNRS, Laboratoire de recherche en informatique, 91190, Gif-sur-Yvette, France.

• A. Decelle is with Departamento de Física Teórica I, Universidad Complutense, 28040 Madrid, Spain

framework to extract the set of modes and ridges from a kernel density estimate of p using a modified version of the mean-shift algorithm [17, 18]. Yet, after convergence of the projected points, it is not trivial to link them to obtain a continuous structure. In the proposed model, this difficulty is overcome by the embedded prior graph structure.

The mixture model formulation offers a natural handling of heteroscedastic patterns through the learning of the local variances of graph nodes, coherently with the local spatial extent of the data. This provides a way to take into account the multi-scale features appearing in many datasets. In contrast, the fixed-scale optimizations proposed by the SCMS or the SimplePPT [13] algorithms do not permit the description of the local width of the extracted features, even though several scales can be combined. The second advantage of the proposed probabilistic framework is an explicit distribution for outliers, freeing the algorithm from pre-processing steps as required by previous implementations when applied to real-world data [19, 20]. We propose here to bypass this difficulty by directly handling outliers in the mixture model through an added uniform background component, making the extraction of the pattern robust to undesirable points with no prior threshold nor processing. It is also more flexible than the fixed trimming radius proposed in [14, 15].

The overall optimization is based on the Expectation-Maximization [EM, 21] procedure, allowing us to propose a computationally efficient algorithm for the learning of a principal graph with guaranteed convergence for any prior graph construction. The presented formulation goes beyond the principal graph construction of [13, 22]. In particular, the same update equation for node positions can be derived by our formalism in the case of a fixed-variance, the absence of background noise and with a minimum spanning tree graph prior.

Finally, we introduce a graph construction able to take into account cycling underlying topologies. By exploiting the instability of the minimum spanning tree (MST) to random sub-samplings, we build an average graph consisting in the union of edges from the MST and of high probability edges. We show in several applications that this graph embeds cycles representing one-dimensional holes in the dataset independently of their scales or sizes and without giving them a formal definition. This construction is compared to what is proposed by the framework of persistent homology [23]. Even though mathematically well-defined, this framework does not allow the easy extraction of relevant edges independently of their size and create spurious edges irrelevant for the underlying pattern that are not captured in the proposed empirical construction.

The paper is organized such that Section 2 sets notations and provides technical information about mixture models, Expectation-Maximization algorithm and graph theory. Section 3 presents the regularized mixture model embedding the multiple prior distributions, gives explicit equations for parameter estimation, details the algorithm and reviews some of its technical aspects. It also describes the graph construction based on the average of minimum spanning trees. Finally, Section 4 provides and discuss some direct applications of the algorithm on real-life problematics in which the proposed features are relevant.

2 TECHNICAL BACKGROUND

2.1 Mixture models and Expectation-Maximization algorithm

Observed data are usually not directly drawn from known probability distributions but come with multiple modes and complex shapes. Mixture models tackle this complexity through a linear combination of K known laws with unknown parameters by writing the probability of a datapoint \mathbf{x} as

$$p(\mathbf{x} | \Theta) = \sum_{k=1}^K \pi_k f(\mathbf{x}, \theta_k), \quad (1)$$

with $\Theta = \{\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K\}$ the set of model parameters, $\pi_k \geq 0$ the mixing coefficient of component k such that $\sum_{k=1}^K \pi_k = 1$ and $f(\mathbf{x}, \theta_k)$ the probability distribution describing the k^{th} component. The particular Gaussian case of mixture models can be obtained by considering $f(\mathbf{x}, \theta_k) = \mathcal{N}(\mathbf{x}, \theta_k)$, where $\mathcal{N}(\mathbf{x}, \theta_k)$ is the Gaussian distribution centered on μ_k and with variance Σ_k . Thanks to their ability to easily model complex data and their tractability, Gaussian mixture models (GMMs) are the most used mixture distributions in the machine learning literature, mainly as a clustering algorithm to separate a dataset into K components.

What makes mixture models so used in practice is the EM algorithm [21], allowing the maximization of the log-likelihood in a polynomial time. Assuming an independent and identically distributed set of data points $\mathbf{X} \in \mathbb{R}^{N \times D}$, the log-likelihood of the model defined by Eq. (1) can be written

$$\mathcal{L}(\Theta; \mathbf{X}) = \sum_{i=1}^N \log \left(\sum_{k=1}^K \pi_k f(\mathbf{x}_i, \theta_k) \right), \quad (2)$$

which cannot be analytically optimized due to the summation inside the logarithm function. To counter that difficulty, the idea behind EM is to optimize iteratively a simpler function acting like a lower-bound on this log-likelihood. Introducing a set of latent variables $\mathbf{Z} = \{z_i\}_{i=1}^N$ encoding the partition of the dataset such that $z_i \in \{1, \dots, K\}$ denotes which of the K Gaussian components x_i belongs to, the log-likelihood conditioned on the value of \mathbf{Z} can hence be expressed as

$$\mathcal{L}(\Theta; \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^N \log (\pi_{z_i} f(\mathbf{x}_i, \theta_{z_i})). \quad (3)$$

EM states that (3) can be maximized through an iterative procedure involving two alternating steps. Assuming we stand at the iteration t , the E-step uses the current values of the parameters $\Theta^{(t)}$ to estimate the posterior distribution of the latent variables $p(\mathbf{Z} | \mathbf{X}, \Theta^{(t)})$ that will allow the computation of the expectation over the completed log-likelihood

$$Q(\Theta, \Theta^{(t)}) = \mathbb{E}_{\mathbf{Z} | \mathbf{X}, \Theta^{(t)}} \{ \mathcal{L}(\Theta; \mathbf{X}, \mathbf{Z}) \} \quad (4)$$

The M-step then updates parameters by maximizing (4) as

$$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}} Q(\Theta, \Theta^{(t)}). \quad (5)$$

By alternating both Expectation and Maximization steps, EM guarantees the convergence toward a local maximum of the log-likelihood (2).

2.2 Graph theory

Let us consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ as an ensemble of vertices \mathcal{V} , edges $\mathcal{E} = \{(i, j) | i \sim j\}$ where $i \sim j$ denotes that nodes indices i and j are linked together and edge weights $\mathcal{W} = \{w_{ij}\}_{(i,j) \in \mathcal{V}^2}$ such that $\forall (i, j) \in \mathcal{V}^2, w_{ij} \geq 0$ being non-zero when i and j are linked. In this work, all considered graphs are undirected meaning that \mathcal{E} is an unordered set. To further use graphs in linear algebra contexts, we rely on matrix representations given by the adjacency matrix \mathbf{A} , a symmetric $|\mathcal{V}| \times |\mathcal{V}|$ matrix encoding whether two vertices are linked or not, i.e. takes value 1 at position (i, j) if $i \sim j$ and 0 otherwise.

In this work, we assume that a spatial graph \mathcal{G} embedded in \mathbb{R}^D with K vertices at positions $\{\boldsymbol{\mu}_k\}_{k=1}^K$ can be computed to geometrically approximate the data space. In particular, we use $\|\boldsymbol{\mu}\|_{\mathcal{G}}^2$ with $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)^T$ as a measure of the smoothness of the graph structure [24] to constrain the mixture model. Mathematically, this quantity is similar to the usually used L_2 -norm to achieve that purpose, but computed over the non-Euclidean structure of the graph,

$$\begin{aligned} \|\boldsymbol{\mu}\|_{\mathcal{G}}^2 &= \sum_{i=1}^K \sum_{j=1}^K a_{ij} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 \\ &= 2 \text{Tr}\{\boldsymbol{\mu}^T \mathbf{L} \boldsymbol{\mu}\}, \end{aligned} \quad (6)$$

where $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is the Laplacian matrix, symmetric in the undirected case, with \mathbf{D} a diagonal $K \times K$ matrix such that $d_{kk} = \sum_{j=1}^K a_{kj}$ is the degree of vertex k .

Usually, for a given dataset, the underlying graph modeling of the manifold is not known and not obvious to build. There are many different kind of graphs that can be used to approximate the discrete data structure based on topological priors one has at hand. In this work, we particularly focus on two kinds of structures with different topologies to model the data: the MST and an extension of it allowing the representation of cycles in the structure that we present in Sect. 3.3.

The MST of a collection of K vertices corresponds to the set of $K - 1$ edges with a tree topology (assumes no self loop and the existence of a unique path between any pair of nodes) minimizing the total sum of edge weights to reach all vertices. We can formulate it as an integer programming problem

$$\mathbf{A}_{\text{MST}} = \underset{\mathbf{A} \in \mathcal{A}}{\operatorname{argmin}} \sum_{i=1}^K \sum_{j=1}^K a_{ij} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 \quad (7)$$

where \mathcal{A} is the set of matrices such that $\mathbf{A} = \mathbf{A}^T$, $a_{ij} \in \{0, 1\}$, $\sum_{i=1}^K \sum_{j=1}^K a_{ij} = 2(K - 1)$, $\forall k \in \{1, \dots, K\}$, $a_{kk} = 0$ and $\forall S \subseteq \mathcal{V}, \sum_{i \in S} \sum_{j \in S} a_{ij} \leq |S| - 1$. The first two constraints encode the undirected graph definition, the third one denotes a graph with $K - 1$ edges, the fourth the absence of self loops in the structure and the last one requires that any subset of $|S|$ vertices has at most $|S| - 1$ edges. By relaxing the integer constraint to $a_{ij} > 0$, the problem can be solved in quasi-linear time with the number of edges using Kruskal's algorithm [25].

3 METHOD

3.1 Regularized mixture model

In what follows, we focus on the particular case of GMMs hence assuming that the sampling noise around the ridge is Gaussian. From the principal curve point of view, these Gaussian clusters are the projected points standing in the middle of the data distribution. One of the drawbacks of usual mean square projection definitions of principal curves is their sensitivity to outliers, tending to bias the curve because all datapoints have the same weights. Real-world datasets, however, often come with outliers or noisy measurements that are not part of the underlying pattern that generated the data we aim to extract. In the mixture model context, similarly to what is done in [19] for clustering purposes, we tackle this problem by assuming an explicit form for the distribution of outliers. To make the representation robust to these noisy measurements, an additional uniform background distribution is incorporated leading the mixture model to

$$p(\mathbf{x} | \Theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}, \boldsymbol{\theta}_k) + \alpha \rho(\mathbf{x}), \quad (8)$$

where $\sum_{k=1}^K \pi_k + \alpha = 1$ and $\rho(\mathbf{x}) = [\int_{\mathbb{R}^D} 1_{\mathcal{D}}(\mathbf{x}) d\mathbf{x}]^{-1}$ the uniform background distribution with \mathcal{D} the support of the dataset. α being a parameter of the model estimated during the learning, the sole prior information required by the framework about outliers is $\rho(\mathbf{x})$, a constant estimating the inverse of the data support volume. In the rest of the paper, we rely on the estimate of the volume of the convex hull of the N data points. The uniform background distribution for outliers is not an unique choice and could be replaced by any other fitting a prior knowledge. In the absence of external information about a positional dependency for outliers, the uniform distribution is however a natural choice in the sense it maximizes the entropy over the closed domain \mathcal{D} .

Because D -dimensional datasets often lie on a lower dimensional support $d < D$, we can include this knowledge into the model as a prior constraining the parameter space Θ . In the context of ridge extraction, $d = 1$ and, keeping the underlying topology as general as possible, it is assumed that a graph \mathcal{G} can approximate this low-dimensional manifold from which \mathbf{X} is drawn. The topological prior \mathcal{G} can be embedded as a penalty term on the log-likelihood adding a smoothness condition on the set of centers $\boldsymbol{\mu}$ through $\|\boldsymbol{\mu}\|_{\mathcal{G}}^2$ defined in Eq. (6). Formally, we write

$$\log p(\boldsymbol{\mu}) = -\frac{\lambda_\mu}{2} \|\boldsymbol{\mu}\|_{\mathcal{G}}^2, \quad (9)$$

where $\lambda_\mu \geq 0$ is a parameter controlling the force of the penalty that we later refer to as "regularization parameter". This constraint is very similar to the one introduced in [9] where the curvature measure of the differentiable curve is used instead to regularize the log-likelihood. Here, the represented structure and its topology are made more general by using a graph \mathcal{G} . The same type of constraints as in [26, 9] can be obtained by restricting the graph to a chain topology with elements of the Laplacian matrix $l_{ij} = 2\delta_{i,i} - \delta_{i,j+1} - \delta_{i,j-1}$, where $\delta_{i,j}$ denotes the Kronecker delta function. The non-Euclidean nature of the approximated manifold is encoded in the Laplacian matrix \mathbf{L} and

$\|\mu\|_{\mathcal{G}}^2$ hence plays the role of $\int_{\mathcal{G}} \|\nabla \mu_k\|^2$, a measure of the manifold smoothness [27] when approximated by \mathcal{G} . From a pure Bayesian view, Eq. (9) also corresponds to a Gaussian prior on the edge weights distribution with a precision λ_μ .

Natural manifestations of heteroscedastic patterns, such as trees or blood vessels, exhibit a linear evolution of the sampling size without sudden variations. For instance, the width of a tree is continuously expanding from small branches to the trunk and the size of the vessel network is smoothly enlarging from capillaries to wide arteries. For these physically motivated reasons, the local size of the underlying continuous structure as measured by variances of the model can be considered to evolve smoothly along the graph. To incorporate this idea in the formalism, we use again the non-Euclidean proximity measure on the graph structure for variances update. To this end, an additional prior distribution on variances is used, based on the local neighborhood of a node. In what follows, we pave the point-cloud distribution with spherical Gaussian clusters such that $\forall k \in \{1, \dots, K\}, \Sigma_k = \sigma_k^2 \mathbf{I}_D$. To obtain a closed-form expression of the new update equation, we use the conjugate prior for variances of a Gaussian likelihood. Formally, we rely on the inverse-Gamma distribution distribution with shape parameter $1 + \lambda_\sigma$ and scale parameter $\lambda_\sigma \sigma_{\mathcal{N}_k}^2$ defined such that the mode of the distribution is located at the mean variance of the neighboring nodes, namely $\sigma_{\mathcal{N}_k}^2 = 1/|\mathcal{N}_k| \sum_{i \in \mathcal{N}_k} \sigma_i^2$ with $\mathcal{N}_k = \{i \mid a_{ik} = 1\}$ and $|\mathcal{N}_k| = d_{kk}$ the degree of node k . Mathematically, we write

$$\log p(\sigma_k^2) = -\lambda_\sigma [\log \sigma_k^2 + \sigma_{\mathcal{N}_k}^2 / \sigma_k^2] + \text{cte}, \quad (10)$$

the prior distribution for the σ_k^2 parameter and use the same $\lambda_\sigma \geq 0$ to constrain all components.

Finally, a prior distribution is added for mixing coefficients to avoid singular solutions to happen in situations in which nodes of the graph are paving almost empty regions leading to $\pi_k \rightarrow 0$. This can be achieved by assuming a Gaussian prior centered on uniform coefficients on the set of datapoints not represented by the background noise, namely $(1 - \alpha)/K$. Hence, we have

$$\log p(\pi_k) = -\frac{\lambda_\pi}{2} \left[\frac{1 - \alpha}{K} - \pi_k \right]^2 + \text{cte}, \quad (11)$$

where $\lambda_\pi \geq 0$ controls the force of this prior.

The full prior distribution on the parameter set Θ can hence be written as the summation of all individual terms,

$$\log p(\Theta) = \log p(\mu) + \sum_{k=1}^K \log p(\sigma_k^2) + \sum_{k=1}^K \log p(\pi_k). \quad (12)$$

Optimal values of the parameters are now obtained by maximizing the posterior distribution defined as $\log p(\Theta | \mathbf{x}) \propto \log p(\mathbf{x} | \Theta) + \log p(\Theta)$ turning the problem into a maximum *a posteriori* estimation of the parameters for which the EM algorithm can still be used.

As the additional term only depends on Θ , the E-step remains unchanged while the M-step now seeks to solve

$$\Theta^{(t+1)} = \operatorname{argmax}_{\Theta} Q(\Theta, \Theta^{(t)}) + \log p(\Theta). \quad (13)$$

For the log-likelihood as defined by the particular mixture model (8) and the prior defined by Eq. (12), responsibilities of Gaussian and uniform background components,

respectively noted p_{ik} and p_i^{bkg} , can be computed during the E-step as $p(\mathbf{Z} | \mathbf{X}, \Theta^{(t)})$ using Bayes' theorem and the current parameter values $\Theta^{(t)}$

$$\begin{cases} p_i^{\text{bkg}} = \frac{\alpha \rho(\mathbf{x}_i)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_i | \boldsymbol{\theta}_j) + \alpha \rho(\mathbf{x}_i)}, \\ p_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\theta}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_i | \boldsymbol{\theta}_j) + \alpha \rho(\mathbf{x}_i)}. \end{cases} \quad (14)$$

Update equations for each parameter are then derived during the M-step of Eq. (13) as

$$\begin{cases} \alpha^{(t+1)} = \frac{1}{N} \sum_{i=1}^N p_i^{\text{bkg}}, \\ \pi_k^{(t+1)} = \frac{1/N \sum_{i=1}^N p_{ik} + \lambda_\pi (1 - \alpha^{(t+1)}) / K}{1 + \lambda_\pi}, \\ \boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^N \mathbf{x}_i p_{ik} / \sigma_k^2 + 2\lambda_\mu \sum_{j=1}^K a_{kj} \boldsymbol{\mu}_j^{(t+1)}}{\sum_{i=1}^N p_{ik} / \sigma_k^2 + 2\lambda_\mu \sum_{j=1}^K a_{kj}}, \\ \sigma_k^{(t+1)} = \left[\frac{\sum_{i=1}^N p_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2 + 4\lambda_\sigma \sigma_{\mathcal{N}_k}^2}{D \sum_{i=1}^N p_{ik} + 4\lambda_\sigma} \right]^{1/2}. \end{cases} \quad (15)$$

When not specified, parameters not indexed by time correspond to time t . This is especially important in the E-step and in the computation of $\boldsymbol{\mu}_k^{(t+1)}$. It is indeed interesting to note the contribution of all $\boldsymbol{\mu}_j^{(t+1)}$ in the update of $\boldsymbol{\mu}_k^{(t+1)}$ in Eq. (15). It is hence more convenient to write this update equation matricially for $\boldsymbol{\mu} \in \mathbb{R}^{K \times D}$.

$$\boldsymbol{\mu}^{(t+1)} = [\boldsymbol{\Gamma} \mathbf{S}^{-1} + 2\lambda_\mu \mathbf{L}]^{-1} \mathbf{S}^{-1} \mathbf{R}^T \mathbf{X}, \quad (16)$$

where \mathbf{S} is a diagonal $K \times K$ matrix with $s_{kk} = \sigma_k^2$, $\boldsymbol{\Gamma}$ a diagonal $K \times K$ matrix of average data points explained by the k^{th} component, i.e. $\gamma_{kk} = \sum_{i=1}^N p_{ik}$ and $\mathbf{R} \in \mathbb{R}^{N \times K}$ the responsibility matrix for Gaussian components such that $r_{ik} = p_{ik}$.

3.2 Algorithm and illustration

Algorithm 1 Graph regularized mixture model

Input: Data: $\mathbf{X} \in \mathbb{R}^{N \times D}$, hyper-parameters: $\mathcal{Y} = \{\lambda_\mu, \lambda_\sigma, \lambda_\pi\}$, initialization: $\Theta^{(0)}, \mathcal{G}^{(0)}$.
Output: $\Theta^{(t)}$ and $\mathcal{G}^{(t)}$.
while $\mathcal{R}_t \geq \epsilon$ **do**
 Compute the adjacency matrix \mathbf{A} of $\mathcal{G}^{(t)}$
 E-step: Compute responsibilities p_{ik} and p_i^{bkg} using equations (14)
 M-step: Compute new parameters Θ_t based on responsibilities using equations (15) and (16)
 Compute the increment in log posterior $\mathcal{R}_t = \log p(\Theta^{(t)} | \mathbf{X}) - \log p(\Theta^{(t-1)} | \mathbf{X})$
 Optional: Update graph topology by recomputing $\mathcal{G}^{(t)}$ on the new set $\{\boldsymbol{\mu}_k\}_{k=1}^K$
end while

Algorithm (1) sums up all the steps for the learning of the proposed graph regularized mixture model (GRMM) from a given set of measurements \mathbf{X} and a graph prior \mathcal{G} . It extends the GMM by embedding a topological prior through \mathcal{G} constraining the position of centers, propose an handling of outliers and the learning of the local widths. In our model, the topology of the graph can be either fixed or updated during the optimization. This is particularly important since the prior approximation on the noisy pattern with outliers may not be suitable to describe the final smooth underlying structure. For instance, a bifurcation in the prior graph may not be relevant in the final one. Consequently, it may be useful to update the topological prior based on the current estimate of $\{\mu\}_{k=1}^K$. When updating the topology at each iteration, \mathbf{A} cannot be understood as a prior anymore but has to be included as a parameter in the model making the chosen graph construction itself acting like a prior. In the case of fitting an MST topology, we can embed the integer programming optimization problem definition of the MST from Eq. (7) in the regularized log-likelihood hence leading to estimate both Θ and \mathbf{A} in the M-step which simply involves computing the new graph \mathcal{G}_t at iteration t given a graph construction process and the positions $\{\mu\}_{k=1}^K$.

The top right panel of Fig. 1 shows the graph learned by the GRMM Algorithm (1) for a dataset made of three branches with linearly evolving standard deviations $\sigma \in [0.015, 0.15]$ and 25% background noise uniformly added in the square. The resulting graph visually shows a smoothly evolving adaptive scale along the structure. Even though the graph is initialized with only $K = 100$ nodes taken randomly over \mathbf{X} , there are no final nodes standing outside of the pattern and the resulting estimate of the background level is 25%. The bottom left panel of Fig. 1 illustrates the principal graph learned by the SimplePPT algorithm on the same dataset and using the same parameters and initialization as our construction. Many branches are falling in the background noise and, because of its fixed-variance scheme, it fails at capturing the ridges of both the large and small variance branches. Choosing a large value for σ_0^2 biases the graph in the small branch, while a low one irremediably creates a lot of spurious branches in the large-variance part of the pattern, even in the absence of outliers. The sensitivity of the graph to outliers is handled by the ElPiGraph framework [15] through a trimming radius R_0 discarding distant datapoints from the update of a node position. In the bottom right panel of Fig. 1 can be found three realizations of this algorithm with different values of R_0 , showing the robustness of ElPiGraph to uniform background noise with a fine-tuned value. The choice of R_0 however remains scale-dependent and can alter the recovered structures in case of heteroscedastic patterns. It is also not obvious to tune in real applications and comes in addition to other hyper-parameters already complex to choose. Moreover, the resulting graph do not embed a description of the local width as opposed to the GRMM which fits all parts of the pattern in its center. Note nonetheless that the ElPiGraph algorithm handles better the dense area at the center of the dataset when $R_0 = 0.3$ with a single bifurcation thanks to an added regularization term to the cost function minimizing the overall graph complexity (length as Eq. (9) but also bending).

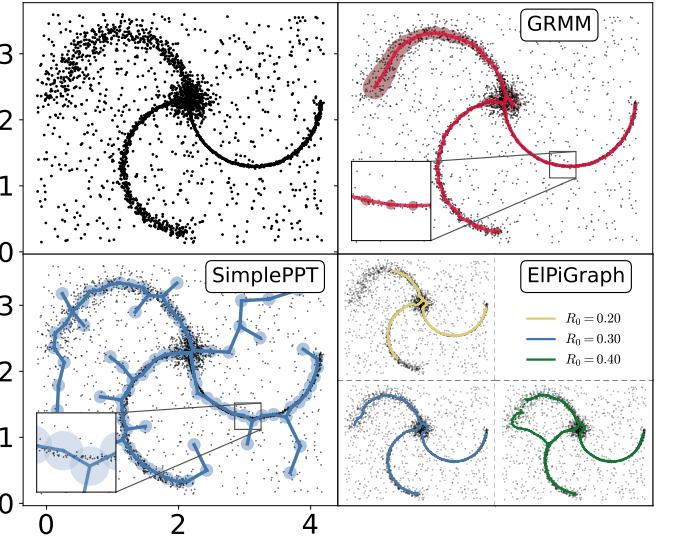


Fig. 1. Illustrative comparison of three procedures to learn a principal graph on an artificial dataset made of $N = 2400$ datapoints with three branches with linearly evolving standard deviations converging to a spherical Gaussian cluster shown in the top left panel. Top right is the principal graph learned by the proposed GRMM algorithm (1) with $K = 100$ initialized randomly over \mathbf{X} , $\mathcal{Y} = \{5/\sigma_0^2, 10, 1\}$ and $\sigma_0^2 = 0.01$. Bottom left is the one from the SimplePPT algorithm [13, 22] with $\sigma = \sigma_0$ and identical initialization nodes. In both cases, the shaded areas correspond to the $1-\sigma_k$ circles centered on μ_k . The bottom right panel is the result from the ElPiGraph procedure [15] with 100 nodes initialized with 70 taken randomly over \mathbf{X} and for different values of the trimming radius R_0 with the same elasticity parameters $\lambda = 0.01$, the length constraint and $\mu = 0.01$ the bending constraint.

3.3 Average graph prior

Although the MST exhibits some nice features, especially the scale invariance and the absence of free parameter, it has a limited representative power for general datasets due to its tree topology that cannot represent cycles. By using a non-cycling topology, the regularized graph becomes even more inaccurate if the data embeds holes because the optimization shortens graph extremities hence emphasizing the absence of loops. Several graph constructions allow such a feature, as for instance the k -nearest neighbors graph. This solution would however create a lot of redundancy in the structure and long-range edges for isolated nodes standing in outliers.

Because the MST results from a global minimization of the total Euclidean distance as seen in Eq. (7), the obtained preferential path is very sensitive to random removals of data points. We exploit this particular aspect by merging the idea of a graph with the minimum total length and the handling of one-dimensional holes in the dataset to propose an empirical construction based on the computation of MSTs obtained from a set of B realizations of random sub-samplings of the set of nodes. Formally, the average adjacency matrix is $\bar{\mathbf{A}} = \sum_{b=1}^B \mathbf{A}_b / B$, where \mathbf{A}_b is the adjacency matrix of the MST computed the b^{th} random subsampling of μ using a fraction K_b/K of points. This approach, proposed in the context of pattern extraction to provide an uncertainty measurement [20, 28, 15], is exploited in the present case to combine all the realizations to obtain a single graph construction. This operation is also done *a priori* on non-regularized MSTs hence reducing the computational

cost by avoiding the need of running Algorithm (1) multiple times.

The averaged adjacency matrix \bar{A} represents the frequency of appearance of an edge during B realizations of MST and thus the probability, for each pair of centers, to be linked. To illustrate the method, we build a discrete Voronoi dataset consisting of intersecting straight lines surrounding low-density areas visible on the top panel of Fig. 2. The bottom panel shows the distribution of edge probabilities for several ratios K_b/K . For a large range of values of this ratio, we clearly distinguish two populations of edges: those with high probabilities and those with much lower ones. In the high probability mode, we retrieve, in the three cases, all $K - 1$ edges of the original MST computed over μ plus 27 new ones which corresponds to the exact number of closed cycles in the artificial dataset. These additional edges are shown in bold dark blue in the top panel of Fig. 2. When the ratio K_b/K becomes smaller, the high probability mode tends to be centered at lower and lower probability until the pattern is so much altered by the sub-sampling that it is not retrieved. The high probability mode then becomes centered at lower and lower values. The proposed method hence allows to take advantage of the inherent instability of the MST to recover additional edges inducing cycles independently of their scale. It solely and indirectly depends on the edge length required to close the cycle without imposing neither a formal definition for the cycle nor a hard threshold in edge length.

Even though observed in the example of Fig. 2, the retrieval of the $K - 1$ edges of the MST is not theoretically guaranteed. To ensure it, the resulting set of edges is obtained by the union of all MST edges and those in the high probability mode leading to the non-singular symmetric adjacency matrix

$$(A)_{ij} = \max \left((A_{\text{MST}})_{ij}, (\bar{A}_{>m})_{ij} \right), \quad (17)$$

with $\bar{A}_{>m}$ the thresholded average adjacency matrix \bar{A} at level m such that it isolates the high probability mode only. By doing so, the proposed graph construction is an extension of the MST containing all its edges with additional ones sharing similar probabilities of appearance during all the random computations.

For the sake of illustrative comparison, we add, as the turquoise blue line in the top panel of Fig. 2, the result of the cycling topology as defined by the framework of persistent homology [29]. In particular, we use the 1D-homologically persistent skeleton of [1D HoPeS, 23] built as the MST completed with edges that are creating persistent 1D homologies in the dataset. After the extraction of those critical edges, it is not trivial to obtain a persistence threshold that captures the desired cycles. Here, we rely on the bootstrap procedure that, together with the stability of the persistence diagram allows a statistically well-defined computation of a threshold [30]. The high sensitivity of this construction to outliers creates some high-scale loops leading to spurious branches corrupting the representation of the underlying pattern. Even though these undesired cycles could be handled using the sub-level sets of the distance-to-measure function of [31, 30], this does not allow an easy extraction of a graph on the inputted datapoints as

required by our application. A pre-processing of the data by removing points in low-density areas or a careful fine-tuning of the persistence threshold are also possible but this is at odds with the proposed formalism meant to handle such datapoints. One hint that these spurious branches come from the noisy measurements is that, when varying the randomness of the sampling, a different pattern is drawn (plotted as the orange line on the top panel of Fig. 2) while the one traced by the proposed average graph remains the same. It is also worth emphasizing that, in both realizations of the persistent skeleton, the two small-scale cycles at the center of the dataset are not captured while they are observed in our average graph prior. Since the persistence is linked with the cycle size, the incapacity in representing small cycles is caused by the too high persistence threshold imposing a lower-bound on the scales of the detected cycles.

3.4 Convergence and time complexity

Convergence toward a local maximum and monotonic increase of the log posterior through iterations are guaranteed by the EM procedure when fixing $\forall t, \mathcal{G}^{(t)} = \mathcal{G}^{(0)}$ [see 32, for a detailed analysis of EM convergence properties]. When updating the topology at each iteration and using a prior based on the MST, convergence as well as monotonic increase of the regularized likelihood are still guaranteed [22], but it does not remain true for any general graph construction if not based on optimization procedures such as defined by Eq. (7).

The E-step equation (14) requires the computation of the responsibilities, taking $O(NDK)$ operations to complete, while the M-step equations (15) respectively require $O(NK)$, $O(NK)$, $O(K^3)$ and $O(NKD)$ operations. Considering T iterations for the algorithm to converge and a complexity of C_g for the graph structure computation, algorithm (1) needs $O(T[K^3 + NKD + C_g])$ operations. C_g acts on K and can take very different forms depending on the used topological approximation. For the MST, it takes $O(K \log K)$ operations to build the Delaunay Tessellation and $O(K^{1+D/2} \log K)$ to find the MST with Kruskal's algorithm. Naturally, the proposed average graph being obtained from B random sub-samplings of the MST, it requires $O(B C_{\text{MST}}(K_b) + C_{\text{MST}}(K))$ operations, with $C_{\text{MST}}(K)$ the complexity of getting an MST over K vertices. To give a broad idea of the computational time of the algorithm, the principal graph from Fig. 2 is obtained in less than three minutes on a modern laptop with our single-core Python implementation.

3.5 Hyper-parameters

Hyper-parameters of the full model are K and $\mathcal{Y} = (\lambda_\mu, \lambda_\sigma, \lambda_\pi)$. K denotes the number of Gaussian components used in the mixture model while $\lambda_\mu, \lambda_\sigma, \lambda_\pi$ are all related with shapes of prior distributions on the parameter indicated as subscripts.

λ_μ controls the force of smoothness constraint and corresponds to the precision of the prior Gaussian distribution put on edge weights. From Eq. (9), and emphasized by the upper left quadrant of Fig. 3 showing regularized graphs obtained with increasing values of λ_μ from 0 to $200/\sigma_0^2$, we see that the higher λ_μ , the more $\{\mu_k\}_{k=1}^K$ are pulling

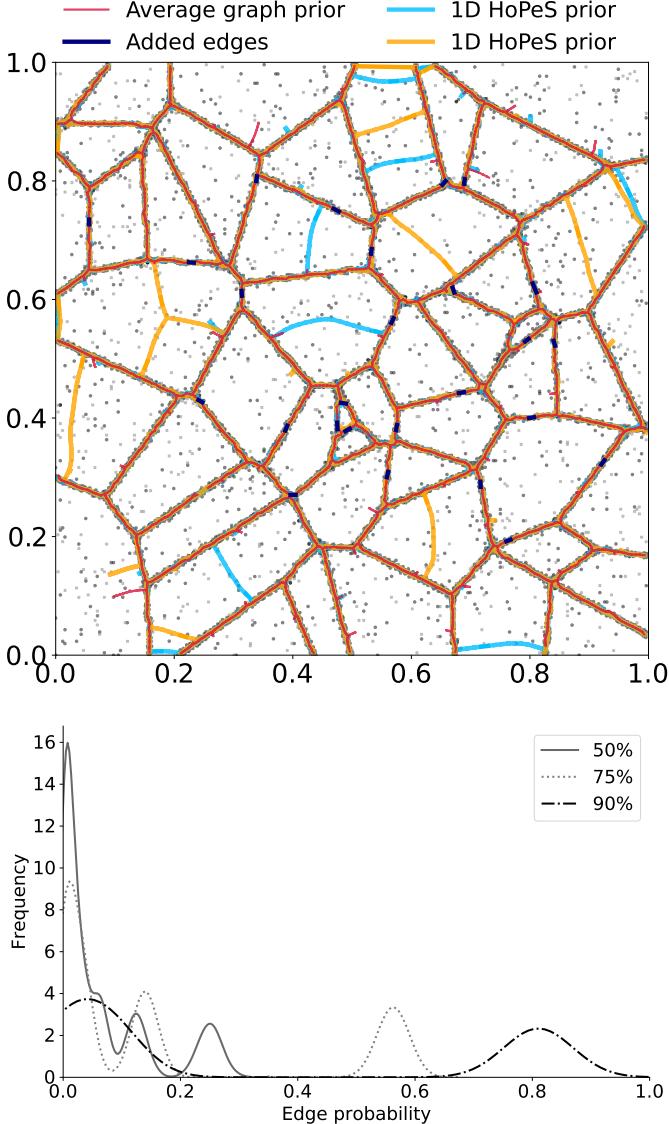


Fig. 2. (top) Black points and gray crosses are two sampling realizations of artificial 2D datasets obtained from a Voronoi pattern with $N = 9249$ points. Red edges are those from the MST over the final regularized structure with $K = 2800$ nodes and bold dark blue ones are those added by the high probability mode with $B = 500$, $K_b/K = 0.75$, and thresholded at $m = 0.35$. The turquoise blue and orange lines are the set of edges from the regularized graph respectively computed from the two noisy realizations with the 1D HoPeS prior [23]. (bottom) Probability distributions of edge probabilities $(\bar{A})_{ij}$ for several ratios K_b/K .

themselves along the graph structure leading to constrain the full graph length $\sum_{ij} w_{ij}$ by shortening its extremities. The bottom left panel of Fig. 3 shows the final distribution of edge weights, in other words the distances between linked nodes in the graph, w_{ij} . We clearly observe the mode of the distribution being translated to lower values when λ_μ increases, indicating lower and lower distances between connected nodes. At a fixed value of λ_μ , a large increase of the number of nodes K can induce over-fitting. Sticking with a similar extracted pattern would hence require the increase of λ_μ as well. This is shown for instance when comparing the top right panel of Fig. 1 and the top left panel of Fig. 3. One way to reduce the interlink between these two

parameters and the dependency of the pattern on K is the growing grammar proposed in [12, 15] in which the graph is initialized with a small value of K to then add some nodes iteratively using predefined rules. It also has the advantage to reduce the computational cost at early iterations.

λ_σ acts on the shape of the inverse-Gamma prior distribution of Gaussian variances σ_k^2 . Equation (15) teaches us that, as for other hyper-parameters, when $\lambda_\sigma \rightarrow 0$, the prior is canceled and variances are updated through the usual EM equation. When λ_σ increases, the shape of the inverse-Gamma law is more and more constrained leading σ_k^2 to be updated mainly through $\sigma_{N_k}^2$ and, $\forall k$, $\sigma_k^2 \simeq \sigma_0^2$ with σ_0^2 the value used to initialize the algorithm. Eventually, a high enough value for this parameter thus leads to a fixed scale version of the algorithm, similar to those presented in [22, 28], as shown in the middle bottom panel of Fig. 3 where the distribution of $\log_{10} \sigma_k^2$ is more and more centered around $\log_{10} \sigma_0^2$. As illustrated in the upper middle quadrant of Fig. 3, when λ_σ increases, the structure shows a smoothly evolving variance along the graph structure with multiple scales described at the same time. This is also highlighted by the lower panel where the variance distributions obtained from regularized graphs with small values of λ_σ are spreading over several order of magnitudes and show three distinct modes corresponding to the characteristic scales of the three branches in the dataset. These estimates tend to be more and more biased toward σ_0 when λ_σ increases.

Finally, λ_π acts on the amplitude π_k , corresponding to the proportion of points being represented by the component k and controls the force of the prior of uniformly distributed Gaussian mixtures. This parameter has a very low impact on the overall result but can avoid singular solutions to happen in very low-density regions paved by graph nodes. The upper right panel of Fig. 3 shows this small impact despite a large range of tested values, from 0 to 10. Although the obtained distributions of mixing weights π_k are clearly different, as seen in the bottom right panel tending to be centered at $(1 - \alpha)/K$ when λ_π increases, the obtained graph structure remains unchanged, and so is the local extent measured by variances.

Even though the impact of these hyper-parameters is interpretable and that there is a wide range of values providing similar results, their quality still depend on the settings. To the best of our knowledge, there is no well-defined method to choose regularization parameters independently from user tests or external information. In that regard, the adaption of the modified objective function from [33] in the context of principal graphs learning could be of interest to propose regularization-free models.

3.6 Initialization

There are four sets of parameters to initialize, namely $\{\mu_k\}_{k=1}^K$, $\{\sigma_k\}_{k=1}^K$, $\{\pi_k\}_{k=1}^K$ and α . A simple and direct strategy to initialize positions of Gaussian components is to choose $\mu^{(0)} = \mathbf{X}$. By doing so, we ensure that the observed point cloud distribution is well paved by centers. Note however that, from Sect. 3.4, when $K \simeq N$, the complexity scales with N^3 . For large datasets, it may be interesting to first reduce the complexity by initializing the

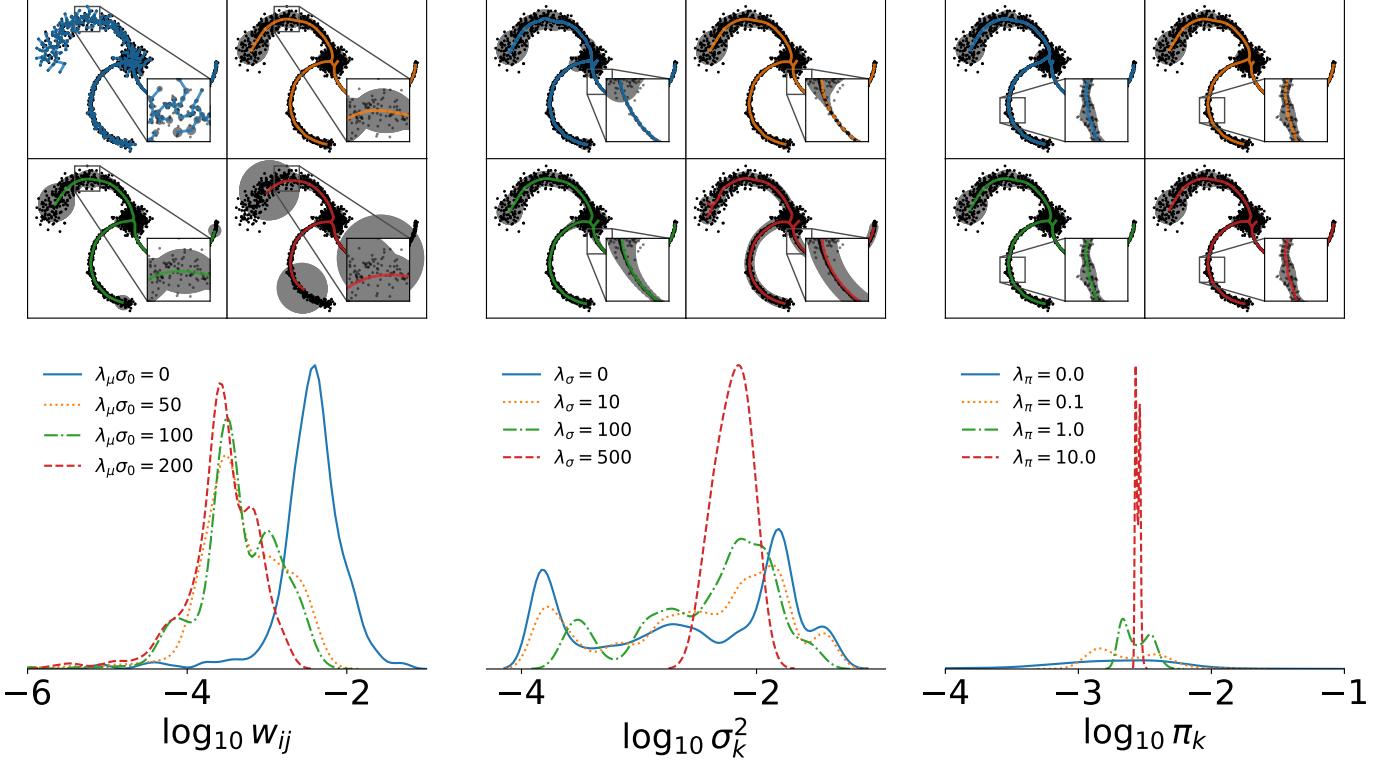


Fig. 3. Illustration of the impact of hyper-parameters $\mathcal{Y} = \{\lambda_\mu, \lambda_\sigma, \lambda_\pi\}$ in Algorithm (1) for an artificial dataset made of three branches with linearly evolving standard deviations converging to a spherical Gaussian cluster. In all cases, $K = 350$ nodes are used with the same random initialization. Black dots are data points, colored lines are the set of edges learned by Algorithm (1) and gray shaded areas correspond to the $1-\sigma_k$ circles centered on μ_k . (top row) From left to right, quadrants corresponds to several values of λ_μ , λ_σ and λ_π . (bottom row) Probability distributions of the impacted parameters. From left to right: edge weights $w_{ij} = \|\mu_i - \mu_j\|_2$, variances of components σ_k^2 , and mixing coefficients π_k .

model with $K \ll N$ using for instance sub-samplings, noise reduction techniques or simple clustering methods like the K-Means or fiducial GMM algorithms. When no prior knowledge on the local size of structures, variances can be initialized as $\forall k \in \{1, \dots, K\}, \sigma_k^{(0)} = \sigma_0^{(0)}$. In this context, $\sigma_0^{(0)}$ can be chosen as a prior guess on the average size of structures or through rules borrowed from density estimation methods [34]. The α parameter is the evaluation of the outlier level in the dataset that should not be paved by the Gaussian components. Its value depends on the application and data at hand. In our experiments, we fix $\alpha^{(0)} = 0.10$ and then let it adjust itself during the learning. As an example, results from Fig. 3 were obtained by starting with this guess and then quickly converges toward a value numerically indistinguishable from 0. Finally, mixing coefficients are initially assumed to be uniformly distributed and $\forall k \in \{1, \dots, K\}, \pi_k^{(0)} = (1 - \alpha^{(0)}) / K$.

One of the known major drawback of the EM algorithm is its dependency on parameter initialization due to its tendency to be easily trapped in local maxima of complexly shaped posterior distributions. To counter that problem, solutions based on simulated annealing (SA) were introduced [35]. [36] shows how SA can be used jointly with regularization of GMMs to learn a heteroscedastic graph representation independently of the initialization. However, this solution comes with a sizable increase of the computational cost and, in all our experiments, we did not find it necessary to obtain satisfactory results.

4 APPLICATIONS

In what follows, we focus on two applications where the newly introduced features, namely the handling of background noise, the adaptive scale and the cycling topology, are of interest. Figures 1 to 3 already introduced example of applications for general manifold denoising purposes in case of heteroscedastic sampling. These approaches can generally be used as pre-processing for clustering or to reduce the dimensionality of the data. To illustrate the proposed method, we tackle two problems in very different fields, namely the automatic detection of filaments in the galaxy distribution and the identification of road networks from noisy GPS measurements.

4.1 Filamentary structure detection in galaxy distribution

In observational cosmology, a recurrent problem is the identification and extraction of cosmic structures from the spatial distribution of galaxies in the universe. This distribution, far from being uniform shows multiple modes and vast empty areas. Among these structures is the striking filamentary pattern linking together regions of high density. One key issue in this problem is to efficiently detect these filaments from a point cloud distribution of observed galaxies to further assess astrophysical hypotheses and improve our understanding of phenomena driving the structure formation of the universe. Here, we rely on a large-scale simulation,

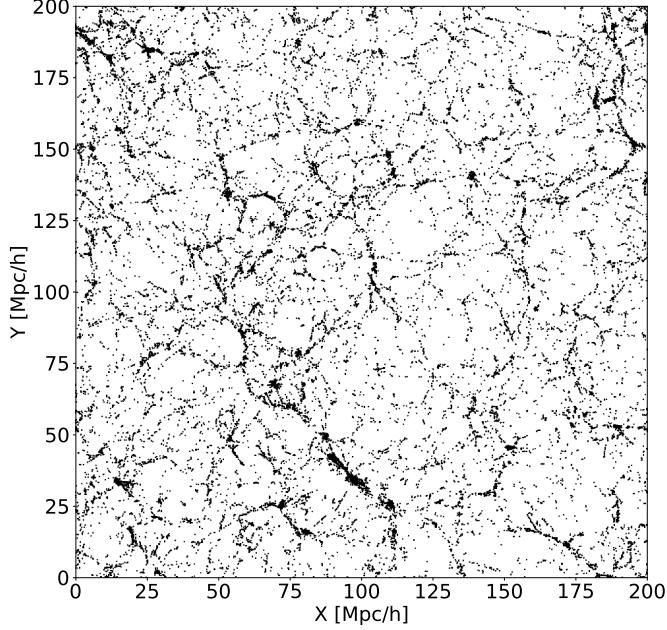


Fig. 4. A 2D distribution of $N = 31205$ galaxies obtained from the IllustrisTNG¹ simulation [37], where massive particles are evolved forward in time from the primordial universe to present time through gravity equations giving birth to a structured pattern visible on the thin slice of 0.5 Mpc/h width of Fig. 4. Such data are known to require particularly elaborated method to be processed with pattern exhibiting structures of various densities and sizes, from large arteries to tenuous tubular filaments, embedded in a noisy background of galaxies standing in quasi-empty areas. In this cosmological context, the MST definition has been applied in the past [38] and provides a parameter-free network linking all galaxies together. As we will see, the proposed extensions brings a smooth behavior for branches of the graph (filaments), the description of their local extent, and the representation of cycles in the network.

Figure 5 shows the result of the GRMM Algorithm (1) applied on the 2D slice of Fig. 4, together with the probability distribution of edge frequencies to build the average graph presented in Sect. 3.3. Although the dataset presents complex patterns and long-range features, we still observe two populations and, among edges with high probabilities, we add to the MST 132 edges that are, by construction, creating cycles in the datasets. The visual inspection of the obtained cycling graph on the zoomed top-right panel of Fig. 5 shows that the approach is closing the graph in relevant places, resulting in a much richer structure than if we had used the MST only. The gray shaded regions illustrate the local size of the filamentary pattern learned by the algorithm. Variances $\{\sigma_k^2\}$ are following the radial extent of datapoints but also show a coherent smooth evolution.

The obtained principal modeling the cosmological filamentary pattern leads to a smooth estimate of the ridge passing through galaxies which is not always the case in other extractors [4]. Additionally, our method provides

the connection between the Gaussian components through the graph structure allowing to define individual filaments which is not naturally allowed, for instance, for projected points of SCMS-based methods [20]. To a further extent, the resulting graph also allows the extraction of geometrical information on the structure, such as the local width or length of the pattern and an estimate of the local density, that are of high interest for astrophysical studies.

Finally, the probabilistic setup of the mixture model can provide a denoised view of the pattern. By keeping only the set of datapoints such as $\{\mathbf{x}_i \mid \sum_{k=1}^K p_{ik} > p_i^{\text{bkg}}\}$, i.e. those with a probability of being generated by the Gaussian components paving the pattern higher than being generated by the background noise, we remove datapoints in which there is no identified spatial structure. Such datapoints are shown in gray on Fig. 5 leading to a cleaned version where remains only galaxies inside the filamentary network plotted in black on the figure.

4.2 Road network extraction from noisy GPS measurements

Producing street maps from sampled positions of traveling vehicles is one of the objectives of map reconstruction. In this case, the set of measurements is obtained from sampled GPS positions of taxis in Berlin², a classical benchmark dataset of road network reconstruction [3]. These data are especially challenging due to their high level of noise and outliers but also because of the uneven sampling of several trajectories. Main roads are indeed traced by hundreds of trajectories, while some low-level ones are traced only by a few.

When applying the GRMM Algorithm (1) with the average graph prior on the Berlin dataset, we obtain the graph from the top panel of Fig. 6. Note that no pre-processing of the dataset nor post-processing of the resulting graph structure have been carried out. We see that the principal graph could use some topological refinements to make it look smoother, removing waving edges and closing some of the remaining dangling branches. For more details about particular processing of inputs and outputs in the context of graph structure learning for road networks, see [39]. The obtained graph passes through the entire dataset, correctly paving regions with both high and low densities. We can also observe that the proposed algorithm recovers a large variety of cycle sizes, from very small ones, as for instance in the red shaded region with intersections or roundabouts, but also larger ones, as in the top left corner. When comparing it with the ground truth map obtained from OpenStreetMap³ in the middle panel, we clearly distinguish that some roads are not well-traced, with several outliers due to the poor quality of sampled data, as it is the case in the top right corner. The GRMM however succeeds in proposing a robust version that does not take into account all such datapoints.

The proposed algorithm offers the possibility to estimate the local variance around the inferred principal graph. In the present dataset for instance, it is possible to investigate further small portions of roads where the estimated σ_k is

1. <https://www.tng-project.org/>

2. <http://mapconstruction.org/>
3. <https://www.openstreetmap.org/>

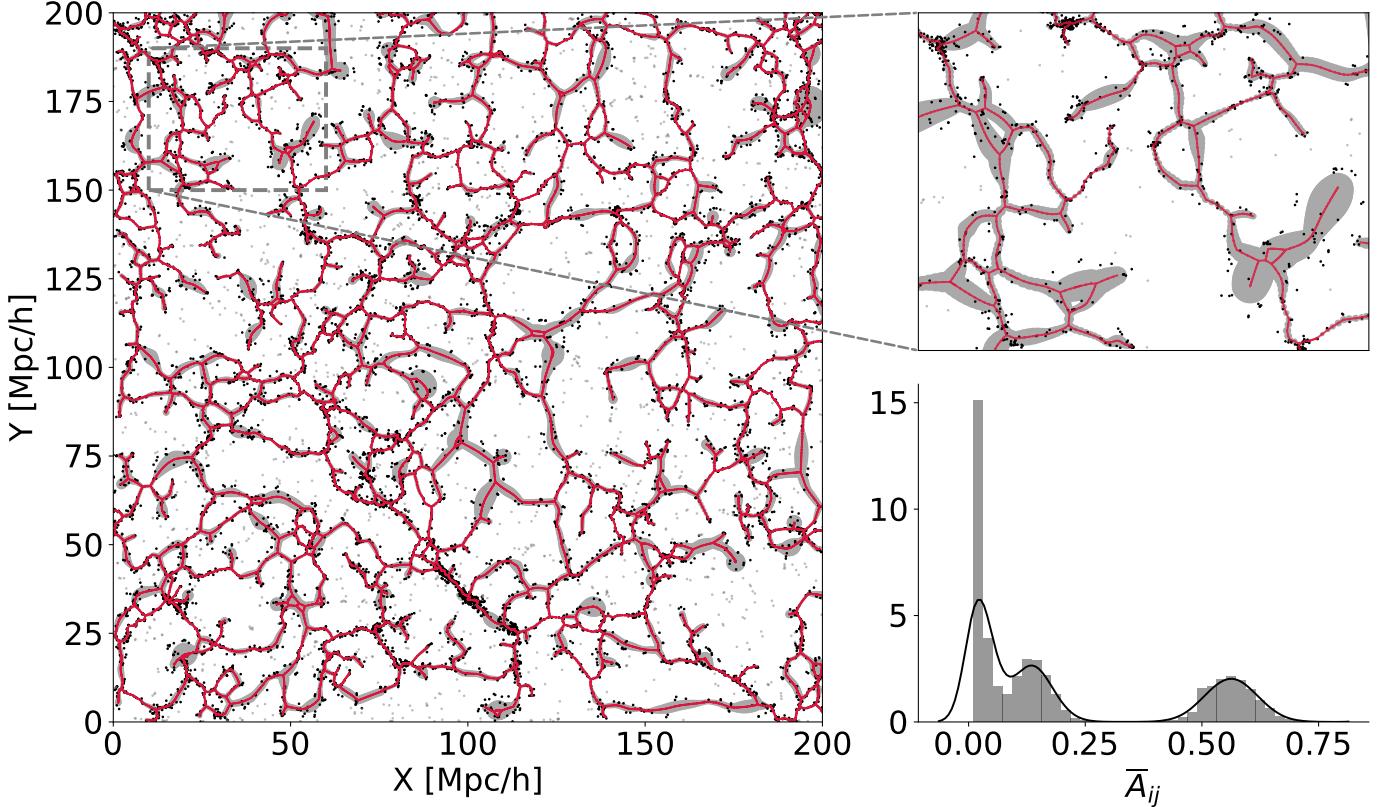


Fig. 5. (left) Black (resp. gray) points are those with $\sum_{k=1}^K p_{ik} > p_i^{\text{bkg}}$ (resp. $<$). The red skeleton corresponds to the regularized graph obtained from Algorithm (1) with the average graph prior and settings $K = 14000$, $\sigma_0^2 = 1$, $\mathcal{Y} = \{10, 5, 1\}$. Gray shaded areas are showing the learned $1-\sigma_k$ circles around graph nodes. (top right) Zoom over a 60 Mpc/h region. Color code is the same as for the left panel. (bottom right) Probability distribution function of $(\bar{A})_{ij}$ for minimum spanning trees obtained from $B = 500$ random sub-samplings of Gaussian centers with $K_b/K = 0.75$.

high or has a broad distribution to spot roads that are either wider or noisier than others. A quantitative comparison to other existing methods is beyond the scope of the present paper as this application only serves an illustrative purpose. The bottom panel of Fig. 6 reports the distribution of variances for nodes standing in rectangular regions of the top panel. When using variances of Gaussian clusters as a proxy of the road width, we conclude that main roads with multiple lanes like those in the green or purple rectangles have larger sampling standard deviation, by up to a factor of two, than the low-level road in the blue rectangle. These estimates of road widths however, depend on the quality of the sampling and some roads can appear wider than they actually are because of spurious points artificially increasing the estimate locally or at the extremities of some branches of the graph. This can be seen when inspecting the variance of the σ_k distributions where the purple one is much larger than the others because of the local noise in the sampling. Finally, the proposed cycling graph captures correctly some topological features of the road network, such as the most prominent roundabouts or road intersections highlighted in shaded reddish areas on the two top panels of Fig. 6. The representation of such features would not be permitted by tree-based topologies.

5 CONCLUSIONS

We present an approach for the learning of principal graphs from point cloud datasets based on mixture models and graph regularization. The proposed formalism provides a robustness to outliers of the pattern and can represent structures of various sizes that are estimated iteratively during the learning. We showed that the added features of heteroscedastic learning and outliers embedding come with only little modifications and additional time complexity of already existing procedures and compared it to state-of-the-art methods. Beyond that, one of our contributions is the introduction of a stochastic procedure to infer the principal graph, where the construction of many MSTs on sub-sampled datasets generate a graph with cycles of several scales, features exhibited in many real-world datasets. Not only the overall procedure is simple to implement but it also relies on fast and well-established algorithms, namely EM for the mixture model and Kruskal for the MSTs, making it particularly appealing from a computational point of view. The whole optimization scheme is meant to converge toward a local maximum of the log-posterior for any graph prior. The method also shows a stable behavior for a large range of model's hyper-parameters and initialization, giving freedom to the user to retrieve the underlying patterns without requiring a cautious fine-tuning. Finally, we illustrated the importance of these newly introduced characteristics in two applications using point-cloud distributions as input,

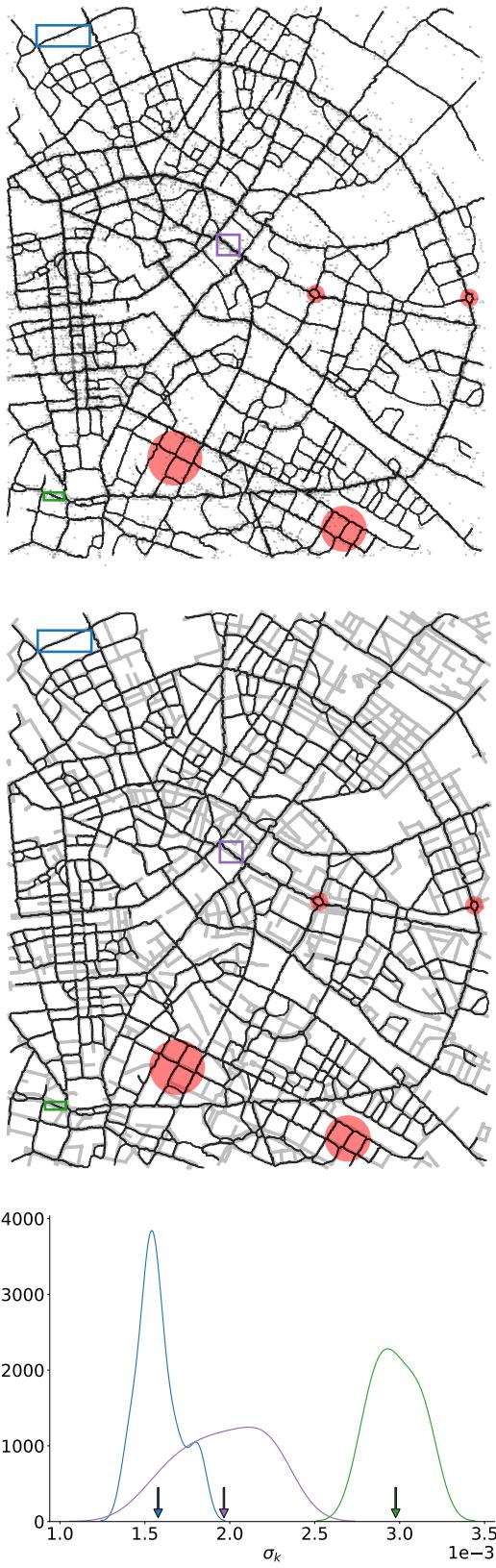


Fig. 6. (top) Black lines are edges of the regularized graph of Algorithm (1) overplotted on the raw datapoints. The graph was obtained using the average graph topology with $K = 7300$ and $\sigma_0 = 0.003$, $\mathcal{Y} = (10/\sigma_0^2, 5, 1)$. Red circles highlight features that could not be caught by a non-cycling topology. (bottom) Probability distributions of variances for nodes in the corresponding circled colored regions of the top panel.

namely the reconstruction of road networks from vehicles position and the detection of the filamentary pattern of the cosmic web. In both applications, even though the data show a high level of complexity with different local sampling densities, loops of various sizes, noises, features of different scales and that no pre-processing were employed to remove spurious measurements, the obtained cycling networks visually trace the expected underlying patterns. We leave for further investigations the adaption of the learning of locally higher dimensional embeddings that could be done using different graph priors as exploited in other manifold learning frameworks [2, 11].

ACKNOWLEDGMENTS

T.B. thanks Dr. Victor Bonjean for fruitful discussions. This research has been supported by the funding for the ByoPiC project from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program grant number ERC-2015-AdG 695561. A.D. was supported by the Comunidad de Madrid and the Complutense University of Madrid (Spain) through the Atracción de Talento program (Ref. 2019-T1/TIC-13298).

REFERENCES

- [1] S. T. Roweis and L. K. Saul, “Nonlinear Dimensionality Reduction by Locally Linear Embedding,” *Science* (80-), vol. 290, no. December, pp. 2323–2326, 2000.
- [2] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.
- [3] M. Ahmed, S. Karagiorgou, D. Pfoser *et al.*, “A comparison and evaluation of map construction algorithms using vehicle tracking data,” *Geoinformatica*, vol. 19, pp. 601–632, 2015.
- [4] N. I. Libeskind, R. van de Weygaert, M. Cautun *et al.*, “Tracing the cosmic web,” *Mon. Not. R. Astron. Soc.*, vol. 473, no. 1, pp. 1195–1217, 2017.
- [5] S. Moccia, E. D. Momi, S. E. Hadji *et al.*, “Blood vessel segmentation algorithms – Review of methods, datasets and evaluation metrics,” *Comput. Methods Programs Biomed.*, vol. 158, no. February, pp. 71–91, 2018.
- [6] L. Van Der Maaten, E. Postma, and J. Van den Herik, “Dimensionality reduction: a comparative review,” *J. Mach. Learn. Res.*, vol. 10, pp. 66–71, 2009.
- [7] T. Hastie and W. Stuetzle, “Principal curves,” *J. Am. Stat. Assoc.*, vol. 84, pp. 502–516, 1989.
- [8] B. Kegl, A. Krzyzak, T. Linder *et al.*, “Learning and design of principal curves,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 281–297, 2000.
- [9] R. Tibshirani, “Principal curves revisited,” *Stat. Comput.*, vol. 2, no. 4, pp. 183–190, 1992.
- [10] U. Ozertem and D. Erdogmus, “Locally Defined Principal Curves and Surfaces,” *J. Mach. Learn. Res.*, vol. 12, pp. 1249–1286, 2011.
- [11] A. Gorban and A. Zinovyev, “Elastic Principal Graphs and Manifolds and their Practical Applications,” *Computing*, vol. 75, pp. 359–379, 2005.

- [12] A. N. Gorban and A. Y. Zinovyev, "Principal graphs and manifolds," in *Handb. Res. Mach. Learn. Appl. trends algorithms, methods, Tech.*, 2009, pp. 28–59.
- [13] Q. Mao, L. Yang, L. Wang *et al.*, "SimplePPT: A Simple Principal Tree Algorithm," *Proc. 2015 SIAM Int. Conf. Data Min.*, pp. 792–800, 2015.
- [14] A. N. Gorban, E. M. Mirkes, and A. Zinovyev, "Robust principal graphs for data approximation," *arXiv e-prints*, vol. 2, no. 1, pp. 1–16, 2016.
- [15] L. Albergante, E. M. Mirkes, H. Chen *et al.*, "Robust and scalable learning of complex dataset topologies via elpigraph," *Entropy*, vol. 22, 2020.
- [16] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli *et al.*, "Nonparametric ridge estimation," *Ann. Stat.*, vol. 42, no. 4, pp. 1511–1545, 2014.
- [17] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inf. Theory*, vol. 21, no. 4, pp. 32–40, 1975.
- [18] Y. Cheng, "Mean Shift , Mode Seeking , and Clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–799, 1995.
- [19] D. C. Stanford and A. E. Raftery, "Finding Curvilinear Features in Spatial Point Patterns : Principal Curve Clustering with Noise," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 6, pp. 601–609, 2000.
- [20] Y.-c. Chen, C. R. Genovese, and L. Wasserman, "Generalized Mode and Ridge Estimation," *arXiv e-prints*, pp. 1–12, 2014.
- [21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. R. Stat. Soc.*, vol. 39, pp. 1–38, 1977.
- [22] Q. Mao, W. Li, T. Ivor W. *et al.*, "Principal Graph and Structure Learning Based on Reversed Graph Embedding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2227–2241, 2017.
- [23] V. Kurlin, "A One-Dimensional Homologically Persistent Skeleton of an Unstructured Point Cloud in Any Metric Space," in *Eurographics Assoc.*, vol. 34, no. 5. Graz, Austria: Eurographics Association, 2015, pp. 253–262.
- [24] A. J. Smola and R. Kondor, *Kernels and Regularization on Graphs*. Springer Berlin Heidelberg, 2003.
- [25] J. B. Kruskal, "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem," *Proc. Am. Math. Soc.*, vol. 7, pp. 48–50, 1956.
- [26] A. L. Yuille, "Generalized Deformable Models, Statistical Physics, and Matching Problems," *Neural Comput.*, vol. 2, pp. 1–24, 1990.
- [27] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," in *Adv. Neural Inf. Process. Syst. 14*. Press, MIT, 2001, pp. 585—591.
- [28] T. Bonnaire, N. Aghanim, A. Decelle *et al.*, "T-ReX : a graph-based filament detection method," *Astron. Astrophys.*, vol. 637, p. A18, 2020.
- [29] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological Persistence and Simplification," *Discret. Comput. Geom.*, vol. 28, p. 511, 2002.
- [30] F. Chazal, B. Fasy, B. Lecci *et al.*, "Robust Topological Inference: Distance To a Measure and Kernel Distance," *J. Mach. Learn. Res.*, vol. 18, pp. 1–40, 2018.
- [31] F. Chazal, D. Cohen-Steiner, and Q. Mérigot, "Geometric Inference for Probability Measures," *Found. Comput. Math.*, vol. 11, no. 6, pp. 733–751, 2011.
- [32] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. Wiley1997, 1997.
- [33] S. Gerber and R. Whitaker, "Regularization-Free Principal Curve Estimation," *J. Mach. Learn. Res.*, vol. 14, no. 3, pp. 1285–1302, 2013.
- [34] N.-B. Heidenreich, A. Schindler, and S. Sperlich, "Bandwidth selection for kernel density estimation: a review of fully automatic selectors," *Adv. Stat. Anal.*, vol. 97, no. 4, pp. 403–433, 2013.
- [35] N. Ueda and R. Nakano, "Deterministic Annealing EM Algorithm," *Neural Networks*, vol. 11, pp. 271–282, 1998.
- [36] T. Bonnaire, A. Decelle, and N. Aghanim, "Cascade of Phase Transitions for Multi-Scale Clustering," *Phys. Rev. E*, vol. 103, no. 1, p. 012105, 2021.
- [37] D. Nelson, V. Springel, A. Pillepich *et al.*, "The IllustrisTNG Simulations : Public Data Release," *Comput. Astrophys. Cosmol.*, vol. 6, 2019.
- [38] J. D. Barrow, S. P. Bhavsar, and D. H. Sonoda, "Minimal spanning trees, filaments and galaxy clustering," *Mon. Not. R. Astron. Soc.*, vol. 216, pp. 17–35, 1985.
- [39] J. Huang, M. I. N. Deng, J. Tang *et al.*, "Automatic Generation of Road Maps from Low Quality GPS Trajectory Data via Structure Learning," *IEEE Access*, vol. 6, pp. 71 965–71 975, 2018.