Robust shape inference from a sparse approximation of the Gaussian trimmed loglikelihood.*

Brécheteau, Claire
claire.brecheteau@ec-nantes.fr
Université de Nantes and École Centrale de Nantes, France

December 6, 2018

Abstract

Given a noisy sample of points lying around some shape $\mathcal{M}$, with possibly outliers or clutter noise, we focus on the question of recovering $\mathcal{M}$, or at least geometric and topological information about $\mathcal{M}$. Often, such inference is based on the sublevel sets of distance-like functions such as the function distance to $\mathcal{M}$, the distance-to-measure (DTM) or the k-witnessed distance. In this paper, we firstly widespread the concept of trimmed log-likelihood to probability distributions. This trimmed log-likelihood can be considered as a generalisation of the DTM.

A sparse approximation of the DTM, the m-power distance-to-measure (m-PDTM), has been introduced and studied by Brécheteau and Levrard in 2017. Its sublevel sets are unions of m balls, with m possibly much smaller than the sample size. By miming the construction of the m-PDTM from the DTM, we propose an approximation of the trimmed log-likelihood associated to the family of Gaussian distributions on $\mathbb{R}^d$. This approximation is sparse is the sense that its sublevel sets are unions of m ellipsoids.

We provide a Lloyd-type algorithm to compute the centers and covariance matrices associated to the ellipsoids. We improve our algorithm by allowing an additional noise parameter to wipe out some points, just as the trimmed m-means algorithm of Cuesta-Albertos et al. [9]. Our algorithm comes together with a heuristic to select this parameter. Some illustrations on different examples enhance that our algorithm is efficient in wiping out clutter noise, recovering the shape and recovering the homology of $\mathcal{M}$; this requiring a storage of only m points and covariance matrices.

1 Introduction

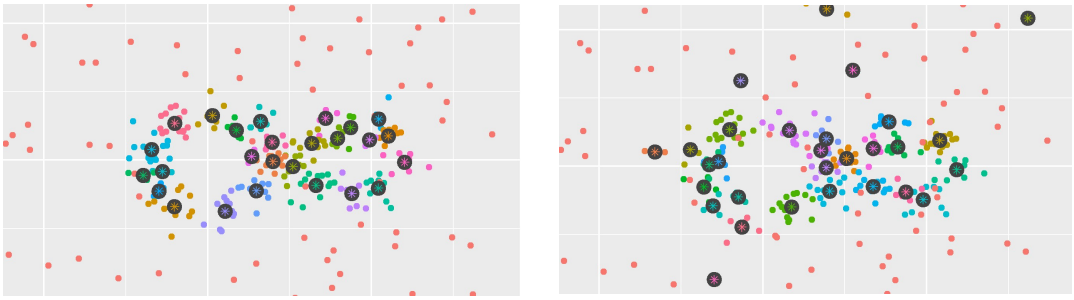
Let $\mathcal{M}$ be an unknown subset of $\mathbb{R}^d$, for instance a submanifold or a crossing manifold. A main objective in Topological Data Analysis (TDA) consists in infering geometric or topological information about $\mathcal{M}$ from a sample $\mathbb{X}$ of n points generated according to some distribution P, around $\mathcal{M}$. Originally, such information was recovered from the sub-level sets of $\mathrm{d}_{\mathbb{X}}$, the distance to the compact set $\mathbb{X}$, which are unions of n balls. In [10], Edelsbrunner et al. introduced a theory to investigate the persistence of the homology of unions of growing balls. Later on, other distance-like functions have been used to infer topological information about $\mathcal{M}$; see [14, 16, 6]. The distance-to-measure function (DTM) [14] have been used to face the presence of outliers in $\mathbb{X}$. Indeed, a

*This work was partially supported by the ANR project TopData and GUDHI and École Centrale Nantes

single outlier might cause the function $\mathrm{d}_{\mathbb{X}}$ to be drastically different from $\mathrm{d}_{\mathcal{M}}$, whereas the DTM remains close to $\mathrm{d}_{\mathcal{M}}$. The sublevel sets of the DTM are unions of $\binom{n}{k}$ balls for k, the number of nearest neighbours considered in the computation of the DTM. The sublevel sets of the k-witnessed distance, an approximation of the DTM introduced in [16], are unions of n balls. This number of balls was reduced to m in [6] with the m-PDTM function, which is more satisfactory for large datasets. Indeed taking m much smaller than n might reduce significantly the computational time of topological tools such as the persistent homology.

Bregman divergences [7] are alternatives to the Euclidean metric, more adapted in certain contexts, e.g. for data generated according to mixtures of Poisson, Gamma, binomial distributions, etc. Such situations are frequent in text mining or for daily water falls datasets, just to name a few. In this context, the upper-level sets of the density of P are unions of Bregman balls. Topological tools initially developped for the Euclidean metric have been adapted to Bregman divergences; from the Bregman-Voronoï tesselation in [4] to the tools for persistent homology in [11]. Anyway, such methods do not allow to take into account the variations of the data within different directions. For instance, approximating a polygonal line (made of m segments) with a union of ϵ-balls would require around $\frac{1}{\epsilon}$ balls. In this paper, we propose to face this issue by approximating $\mathcal{M}$ with a union of m ellipsoids, from noisy data generated around $\mathcal{M}$ and in presence of outliers.

The m-PDTM [6] is a function $x \mapsto \min_{i\in[\![1,m]\!]} \|x-\mu_i^*\|^2 + \omega_i^*$ that approximates the DTM from above. The optimal centers $(\mu_i^*)_{i\in[\![1,m]\!]}$ are solutions of a Bregman clustering algorithm, for some data-dependent Bregman divergence that is function of the DTM. In order to deal with outliers, the authors used the robust Bregman clustering method of [5] with the DTM-based Bregman divergence. In this paper, we introduce the function trimmed log-likelihood-to-measure (TLM), a generalisation of the DTM that coincides with the trimmed log-likelihood for discrete distributions P. We introduce the m-PLM, an approximation of the TLM, by miming the construction of the m-PDTM from the DTM. The m-PLM is a function of m centers $(\mu_i^*)_{i\in[\![1,m]\!]}$ and m covariance matrices $(\Sigma_i^*)_{i\in[\![1,m]\!]}$. Just like the m-PDTM, the optima (μ_i^*, Σ_i^*) can be approximated with an algorithm derived from the m-means algorithm [20] (also known as Lloyd's algorithm [19]). Note that a faster algorithm to approximate the optimal means for the m-means problem is available for large datasets [23]. It might be a challenge to adapt this type of algorithm to approximate the m-PLM. In order to make our method robust to outliers, we add a trimming step just as the trimmed m-means method of Cuesta-Albertos et al. in [9] and [5].

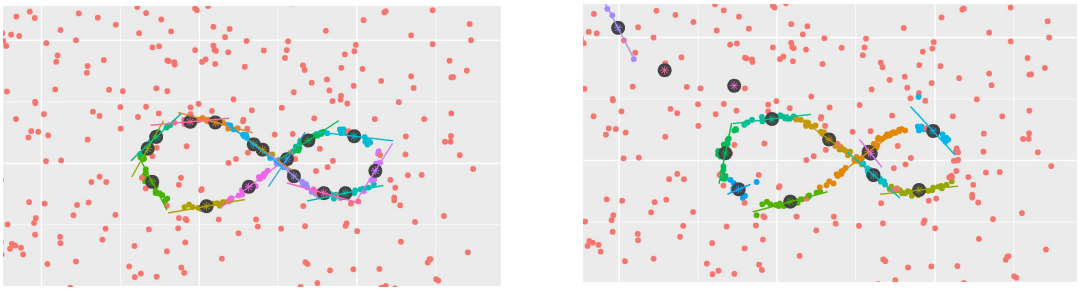


m-PDTM [6] Trimmed m-means [9]

Figure 1: Methods to approximate $\mathcal{M}$ with a set of m centers.

In order to approximate $\mathcal{M}$ with a set of m points, it is possible to use the method of

m-means [20]. Optimal centers for the m-means problem correspond to the support of the best discrete measure approximating the distribution P, in terms of the Wasserstein metric. In this case, the optimal m is of order $n^{\frac{d}{2(d+2)}}$ [17]. In presence of outliers in $\mathbb{X}$, the robust version of m-means for clustering tasks in the trimmed m-means algorithm [9]. Nonetheless, if this method is perfectly well-suited to make m clusters from data naturally divided in m clusters, it fails in approximating general shapes $\mathcal{M}$. The m-PDTM is efficient for this task, see Figure 1. In this context, the optimal parameter m holds for the m-PDTM is also of order $n^{\frac{d'}{2(d'+2)}}$, with d' the intrinsic dimension of $\mathcal{M}$. In this paper, we aim at recovering centers together with covariance matrices. The algorithm tclust [15] performs in clustering data from mixtures of Gaussian distributions, in presence of clutter noise. Nonetheless, just as trimmed m-means, tclust fails in approximating $\mathcal{M}$. The algorithm we propose in this paper is efficient for this task, see Figure 2. Our method can be considered as a robust version of the m-flats algorithm [17]. In TDA, it is primordial to wipe outliers out, see e.g. [8] and [1]. Our method performs well in a situation for which the method of [8] fails.



m-PLM t-clust algorithm [15]

Figure 2: Methods to approximate $\mathcal{M}$ with a set of m centers and covariance matrices.

In this paper, we introduce a continuous version of the trimmed likelihood as a generalization of the DTM, the trimmed likelihood to measure (TLM), in Section 2. In Section 3, we introduce an approximation of the TLM, the m-PLM, a function of type $L^m_{P,h} : x \mapsto \max_{i\in[\![1,m]\!]} \log(p_i(x)) + \omega_i$. The function p_i is the density of the Gaussian distribution $\mathcal{N}(m_h(\mu_i^*, \Sigma_i^*), \Sigma_i^*)$ and ω_i is a constant $\omega_h(\mu_i^*, \Sigma_i^*)$ that depend on some optimal parameters μ_i^* and Σ_i^*. Morally, $m_h(\mu_i^*, \Sigma_i^*)$ corresponds to the mean of the measure $\tilde{P}_{(\mu_i^*,\Sigma_i^*),h}$ defined as the restriction of P to the Σ_i^*-Mahalanobis ball centered at μ_i^* of P-mass h. The constant $2\omega_h(\mu_i^*, \Sigma_i^*)$ is a variance term, it corresponds to the expectation of the Σ_i^*-Mahalanobis distance to μ_i^* for $\tilde{P}_{(\mu_i^*,\Sigma_i^*),h}$. We determine the optimal parameters μ^* and Σ^* when P is Gaussian or uniform on a polygonal line. In Section 4, we develop a Lloyd-type algorithm to compute local optima (μ_i, Σ_i) for the TLM approximation problem. We adapt this algorithm to data corrupted by clutter noise by adding a trimming step, just as Cuesta-Albertos et al. in [9]. We propose a heuristic to select the optimal trimming parameter. Finally, in Section 5, we illustrate our method on two different examples that enhance that we perform in recovering $\mathcal{M}$ from the m-PLM and in getting rid of outliers.

2 The trimmed log-likelihood

In this section, we consider a family of distributions on $\mathbb{R}^d$, $\mathcal{F} = \{P_\theta \mid \theta \in \Theta\}$ for some parameter space Θ. In this paper, we mainly work with the family of Gaussian distributions $\mathcal{F}_\mathcal{N} = \{P_{(\mu,\Sigma)} = \mathcal{N}(\mu, \Sigma) \mid (\mu, \Sigma) \in \Theta = \mathbb{R}^d \times \Xi\}$. Here, Ξ denotes the set

of all covariance matrices on $\mathbb{R}^d$, that is the set of all symmetrical positive matrices on $\mathbb{R}^d$. We recall that the density of $\mathcal{N}(\mu, \Sigma)$ at any point $x \in \mathbb{R}^d$ is given by:

$$p_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}\|x-\mu\|_\Sigma\right).$$

Here, $\|.\|_\Sigma$ denotes the squared Mahalanobis norm, that is defined for every $y \in \mathbb{R}^d$ by $\|y\|_\Sigma = y^T\Sigma^{-1}y$, with y^T the transpose of the vector y and Σ^{-1} the inverse of Σ.

2.1 The discrete version

Let P be a probability distribution on $\mathbb{R}^d$, $\mathbb{X} = \{X_1, X_2, \ldots, X_n\}$ a n-sample from P and P_n the uniform distribution on $\mathbb{X}$. When $P \in \mathcal{F}$ (that is $P = P_{\theta^*}$ for some $\theta^* \in \Theta$), the notion of likelihood has been introduced by Fisher in 1922 [13, Section 6] to infer θ^* from $\mathbb{X}$. This function in defined on Θ by $l_n : \theta \mapsto \prod_{i=1}^n p_\theta(X_i)$, with p_θ the density of P_θ. An estimator of θ^* from the likelihood is given by the maximum likelihood estimator, $\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} l_n(\theta)$. This estimator is also the maximiser of the log-likelihood $\theta \mapsto \sum_{i=1}^n \log(p_\theta(X_i))$. We prefer to consider the normalised version L_n of the log-likelihood, defined for $\theta \in \Theta$ by $L_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log(p_\theta(X_i))$. The maximum likelihood estimator is not robust to noise and outliers. To face this problem, the trimmed log-likelihood has been introduced by Neykov and Neytchev 1990 [22]. It has been extensively investigated since then [18, 2, 21]. The trimmed log-likelihood depends on some parameter k and requires to store the elements of $\mathbb{X}$ as $(X^{(i)})_{i\in[\![1,n]\!]}$ so that: $\forall i \leq j$, $\log\left(p_\theta\left(X^{(i)}\right)\right) \geq \log\left(p_\theta\left(X^{(j)}\right)\right)$. It is defined by

$$L_{n,k} : \theta \mapsto \frac{1}{k}\sum_{i=1}^{k} \log\left(p_\theta\left(X^{(i)}\right)\right). \tag{1}$$

Trimming the log-likelihood makes the estimation of θ^* more robust when there are more than k signal points in $\mathbb{X}$. By signal points, we mean points generated according to P_{θ^*}. In this case, the (at most $n-k$) outliers are no longer involved in the computation of the trimmed log-likelihood, at least when θ is close to θ^*. Then, the minimizer of $L_{n,k}$ is more likely to be close to θ^* than $\hat{\theta}_{MLE}$.

In this paper, our target θ^* will not be approximated with the minimiser of the function $L_{n,k}$. Instead, we propose to approximate the function $L_{n,k}$ with a function L_θ. And our target will be estimated by the parameter θ for which L_θ is the best approximation of $L_{n,k}$ in some sense, cf. Section 3.

2.2 The continuous version and its relation with the distance-to-measure

For Q a probability distribution and f a function, we denote by $Qf(.)$, the expectation of f with respect to Q. With this notation, the log-likelihood rewrites as $L_n(\theta) = P_n \log(p_\theta(.))$. This *log-likelihood* naturally extends to any probability P by $L_P(\theta) = P\log(p_\theta(.))$.

As well, it is possible to extend the notion of trimmed log-likelihood, using the notion of sub-measure. A sub-measure Q of P is a positive measure such that $Q(A) \leq P(A)$ for every Borel set A. We use the notation $Q \preccurlyeq P$. The uniform distribution on $X^{(1)}, X^{(2)}, \ldots, X^{(k)}$, $\tilde{P}_{n,\theta,k}$, satisfies: $\frac{k}{n}\tilde{P}_{n,\theta,k} \preccurlyeq P_n$. Then, the trimmed likelihood defined in (1) rewrites as

$$L_{n,k}(\theta) = \tilde{P}_{n,\theta,k}\log(p_\theta(.)) = \sup\left\{\tilde{P}\log(p_\theta(.)) \mid \tilde{P} \text{ probability such that } \frac{k}{n}\tilde{P} \preccurlyeq P_n\right\}.$$

For h in $[0,1]$, we define the *trimmed log-likelihood to a measure* P as an extension of the trimmed likelihood, by

$$L_{P,h} : \theta \mapsto \sup \left\{ \tilde{P} \log(p_\theta(.)) \mid \tilde{P} \text{ probability such that } h\tilde{P} \preccurlyeq P \right\}. \tag{2}$$

A probability $\tilde{P}$ reaches the supremum if and only if $h\tilde{P} \preccurlyeq P$, P and $h\tilde{P}$ coincide on the upper-level set of P_θ of P-mass h, and $\tilde{P}$ is supported on the closure of this upper-level set. We denote by $\tilde{P}_{\theta,h}$ any such distribution.

When we consider the family of Gaussian distributions, the assertion $\tilde{P}_{\theta',h} \log(p_{\theta'}(.)) = \sup_{\theta\in\Theta} \tilde{P}_{\theta,h} \log(p_{\theta'}(.))$ extends in the following way.

Lemma 1.
If $\mathcal{F} = \mathcal{F}_{\mathcal{N}}$ is the family of Gaussian distributions, then for every $\Sigma \in \Xi$ and probability Q on $\mathbb{R}^d$ with expectation $Q.$, we have

$$\tilde{P}_{(Q.,\Sigma),h} \int \log(p_{(u,\Sigma)}(.)) \mathrm{d}Q(u) = \sup_{\theta\in\Theta} \tilde{P}_{\theta,h} \int \log(p_{(u,\Sigma)}(.)) \mathrm{d}Q(u).$$

Proof. We may write for $x \in \mathbb{R}^d$,

$$\int \log(p_{(u,\Sigma)}(x)) \mathrm{d}Q(u) = \left(-\frac{1}{2} \int \|Q. - u\|_\Sigma \mathrm{d}Q(u) \right) + \log(p_{(Q.,\Sigma)}(x)).$$

Moreover, since for every $\theta \in \Theta$, $\tilde{P}_{\theta,h}$ is the restriction of P to the upper-level set of the function $x \mapsto \log(p_\theta(x))$ with P-mass h, it comes that

$$\tilde{P}_{(Q.,\Sigma),h} \log(p_{(Q.,\Sigma)}(.)) = \sup_{\theta\in\Theta} \tilde{P}_{\theta,h} \log(p_{(Q.,\Sigma)}(.)).$$

□

We consider the family of Gaussian distributions with covariance matrix I_d the identity matrix on $\mathbb{R}^d$, $\mathcal{F} = \left\{ \mathcal{N}(\theta, I_d) \mid \theta \in \mathbb{R}^d \right\}$. In this context, the trimmed log-likelihood relates to the notion of distance-to-measure (DTM) [14]. The DTM with parameter h, $\mathrm{d}_{P,h}$ is a function defined on $\mathbb{R}^d$ by $\mathrm{d}^2_{P,h}(\theta) = \inf \left\{ \tilde{P}\|. - \theta\|^2 \mid \tilde{P} \text{ probability such that } h\tilde{P} \preccurlyeq P \right\}$. The upper-level sets of the density of $\mathcal{N}(\theta, I_d)$ are balls centered at θ. So, the distributions $\tilde{P}$ involved in the computation of the DTM coincide with the ones involved in the computation of the TLM. Moreover, for every $\theta \in \Theta$, $L_{P,h}(\theta) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\mathrm{d}^2_{P,h}(\theta)$.

When P is a distribution uniform on some manifold $\mathcal{M}$, Chazal et al. proved in [14] that the DTM was a good approximation of $\mathrm{d}_{\mathcal{M}}$, the distance to $\mathcal{M}$, even in presence of outliers. It follows on that the upper-level sets of $L_{P,h}$ approximate $\mathcal{M}$ well. We denote by $\mathrm{d}_{P,h,m}$, the m-PDTM [6]. Then, $L^m : \theta \mapsto \frac{d}{2}\log(2\pi) - \frac{1}{2}\mathrm{d}_{P,h,m}(\theta)$ is the best approximation of $L_{P,h}$ from below for the criterion $P|L_{P,h}(.) - f(.)|$ among all functions $f \in \mathcal{G}$. The family $\mathcal{G}$ contains all functions f not larger than $L_{P,h}$, of type $f : \theta \mapsto -\frac{1}{2}\min_{i\in[\![1,m]\!]} \left(\|\theta - c_i\|^2 + \omega_i \right)$ with $c_i \in \mathbb{R}^d$ and $\omega_i \in \mathbb{R}$. The upper-level sets of L^m are unions of m balls.

Inferring topological information from a set of $m << n$ balls requires very few memory and computational time. Nonetheless, unless $\mathcal{M}$ is a set of m points or balls, approximating $\mathcal{M}$ with a union of m balls may require m to be very large. For instance, an ϵ-Hausdorff approximation of the segment $\mathcal{M} = [0,1] \times \{0\}$ with a union of balls requires at least $m = \frac{1}{\epsilon}$ balls. In this case, approximating $\mathcal{M}$ with a single ellipse would be more satisfactory in terms of storage memory.

In the following, we develop a method to approximate $\mathcal{M}$ from $\mathbb{X}$ as a union of ellipsoids. To this aim, we enrich the model $\{\mathcal{N}(\theta, I_d) \mid \theta \in \mathbb{R}^d\}$ with covariance matrices and consider $L_{P,h}$ the trimmed log-likelihood associated to the family $\mathcal{F}_{\mathcal{N}} = \{P_{(\mu,\Sigma)} = \mathcal{N}(\mu,\Sigma) \mid (\mu,\Sigma) \in \Theta = \mathbb{R}^d \times \Xi\}$, the mass parameter h and the probability P.

3 Approximation of the Trimmed log-likelihood

In this section, we adapt the construction of the m-PDTM to the trimmed log-likelihood-to-measure $L_{P,h}$ associated to the family $\mathcal{F}_{\mathcal{N}}$. A first remark is that $L_{P,h}$ is no longer a function defined on $\mathbb{R}^d$. Indeed, it is defined on $\Theta = \mathbb{R}^d \times \Xi$. Thus, it makes no sense to approximate $L_{P,h}$ globally on Θ. Instead, we will find the best set of covariance matrices $(\Sigma_i^*)_{i\in[\![1,m]\!]}$ and the best decomposition of $\mathbb{R}^d$ as a union of non-intersecting cells $\mathbb{R}^d = \bigcup_{i\in[\![1,m]\!]} \mathcal{C}_i^*$ such that $L_{P,h}$ can be well approximated on $\bigcup_{i\in[\![1,m]\!]} \mathcal{C}_i^* \times \{\Sigma_i^*\}$.

3.1 The method

The trimmed log-likelihood-to-measure defined by Equation (2) expresses as

$$L_{P,h}(\theta) = \tilde{P}_{\theta,h} \log(p_\theta(.)) = \sup_{\theta' \in \Theta} \tilde{P}_{\theta',h} \log(p_\theta(.)).$$

We consider Θ^m, the set of all families of m pairs of centres and covariance matrices, or codebooks $\boldsymbol{\theta} = (\theta_i)_{i\in[\![1,m]\!]} = ((\mu_i, \Sigma_i))_{i\in[\![1,m]\!]}$. The method we develop and investigate in this paper consists in approximating the TLM, $L_{P,h}$ with a function $L_{P,h}^{\boldsymbol{\theta}} : \theta \mapsto \max_{i\in[\![1,m]\!]} \tilde{P}_{\theta_i,h} \log(p_\theta(.))$.

Such functions satisfy $L_{P,h}^{\boldsymbol{\theta}}(\theta) \leq L_{P,h}(\theta)$ for every $\theta \in \Theta$. Our approximation $L_{P,h}^{\boldsymbol{\theta}^*}$ of $L_{P,h}$ if defined as the function $L_{P,h}^{\boldsymbol{\theta}}$ whose expectation according to some well-chosen distribution is maximal. The optimal codebook $\boldsymbol{\theta}^*$ is defined as follows.

Definition 2.
The optimal codebook $\boldsymbol{\theta}^* = (\theta_i^*)_{i\in[\![1,m]\!]} = ((\mu_i^*, \Sigma_i^*))_{i\in[\![1,m]\!]}$ is defined as a maximizer of

$$\boldsymbol{\theta} \mapsto \int_{u\in\mathbb{R}^d} \left(\max_{i\in[\![1,m]\!]} \tilde{P}_{\theta_i,h} \log(p_{(u,\Sigma_i)}(.)) \right) \mathrm{d}P(u).$$

Equivalently, if $\Sigma(x)$ is defined for every $x \in \mathbb{R}^d$ as the matrix Σ_i for i such that $\tilde{P}_{\theta_i,h} \log(p_{(x,\Sigma_i)}(.))$ is maximal, then

$$\boldsymbol{\theta}^* \in \arg\max \left\{ \int_{u\in\mathbb{R}^d} L_{P,h}^{\boldsymbol{\theta}}((u, \Sigma(u)))\, \mathrm{d}P(u) \mid \boldsymbol{\theta} \in \Theta^m \right\}.$$

According to the following theorem, it is optimal to choose the same Σ_i in the measure $\tilde{P}_{\theta_i,h} = \tilde{P}_{(\mu_i,\Sigma_i),h}$ and in the likelihood $p_{(u,\Sigma_i)}$. Meaning that it is optimal to make $x \mapsto \Sigma(x)$ take values in $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_m\}$.

Theorem 3.
The codebook $\boldsymbol{\theta}^*$ is such that

$$\boldsymbol{\theta}^* \in \arg\max \left\{ \max_{(\Sigma_i')_{i\in[\![1,m]\!]} \in \Xi^m} \int_{u\in\mathbb{R}^d} \left(\max_{i\in[\![1,m]\!]} \tilde{P}_{\theta_i,h} \log(p_{(u,\Sigma_i')}(.)) \right) \mathrm{d}P(u) \mid \boldsymbol{\theta} \in \Theta^m \right\}.$$

Proof. For $\boldsymbol{\Sigma}' = (\Sigma_i')_{i\in[\![1,m]\!]} \in \Xi^m$, set $R_{\boldsymbol{\Sigma}'} : \boldsymbol{\theta} \mapsto \int_{u\in\mathbb{R}^d} \left(\max_{i\in[\![1,m]\!]} \tilde{P}_{\theta_i,h} \log(p_{(u,\Sigma_i')}(.)) \right) \mathrm{d}P(u)$.
We aim at proving that $\max_{\boldsymbol{\theta}\in\Theta^m} \max_{\boldsymbol{\Sigma}'\in\Xi^m} R_{\Sigma'}(\boldsymbol{\theta}) = \max_{\boldsymbol{\mu}\in(\mathbb{R}^d)^m} \max_{\boldsymbol{\Sigma}'\in\Xi^m} R_{\Sigma'}((\boldsymbol{\mu}, \boldsymbol{\Sigma}'))$.

We decompose $\mathbb{R}^d$ in cells $\mathcal{C}_i$ such that:

$$x \in \mathcal{C}_i \quad \Leftrightarrow \quad \forall j \in [\![1,m]\!],\ \tilde{P}_{(\mu_i,\Sigma_i),h} \log(p_{(u,\Sigma_i')}(x)) \geq \tilde{P}_{(\mu_j,\Sigma_j),h} \log(p_{(u,\Sigma_j')}(x)) \qquad (3)$$

Then, one may express P as $\sum_{i=1}^{m} P_i$ with P_i a positive measure supported on $\mathcal{C}_i$. According to Lemma 1, it comes that:

$$\begin{aligned} R_{\mathbf{\Sigma}'}\left(((\mu_i,\Sigma_i))_{i\in[\![1,m]\!]}\right) &= \sum_{i=1}^{m} P_i(\mathbb{R}^d)\int \tilde{P}_{(\mu_i,\Sigma_i),h}\log(p_{(u,\Sigma_i')}(.))\frac{1}{P_i(\mathbb{R}^d)}\mathrm{d}P_i(u) \\ &\leq \sum_{i=1}^{m} P_i(\mathbb{R}^d)\int \tilde{P}_{\left(\frac{P_i.}{P_i(\mathbb{R}^d)},{\Sigma'}_i\right),h}\log(p_{(u,\Sigma_i')}(.))\frac{1}{P_i(\mathbb{R}^d)}\mathrm{d}P_i(u) \\ &\leq R_{\mathbf{\Sigma}'}\left(\left(\left(\frac{P_i.}{P_i(\mathbb{R}^d)},{\Sigma'}_i\right)\right)_{i\in[\![1,m]\!]}\right). \end{aligned}$$

□

Our approximation of the trimmed log-likelihood is defined from θ^* as follows.

Definition 4.
The m power log-likelihood to the measure P (m-PLM), $L_{P,h}^m$ is defined by:

$$L_{P,h}^m : \theta\in\Theta \mapsto L_{P,h}^{\boldsymbol{\theta}^*}(\theta) = \max_{i\in[\![1,m]\!]} \tilde{P}_{\theta_i^*,h}\log(p_\theta(.)).$$

With a slight abuse of notation, we define its restriction to $\mathbb{R}^d$ by

$$L_{P,h}^m : x\in\mathbb{R}^d \mapsto \max_{i\in[\![1,m]\!]} \tilde{P}_{\theta_i^*,h}\log(p_{(x,\Sigma_i^*)}(.)).$$

3.2 Interpretation

The definition of the m-PLM suggests a decomposition of the space $\mathbb{R}^d$ into m cells $(\mathcal{C}_i^*)_{i\in[\![1,m]\!]}$. Each cell $\mathcal{C}_i^*$ is associated to a pair (μ_i^*,Σ_i^*) accordingly to (3). Morally, $x\in\mathcal{C}_i^*$ for a i such that the expectation of the logarithm of the density $\mathcal{N}(x,\Sigma_i^*)$ at $X\sim\tilde{P}_{\theta_i^*,h}$ is minimized. By symmetry, it is also the expectation (relatively to X) of the logarithm of the density $\mathcal{N}(X,\Sigma_i^*)$ at x. Recall that $\tilde{P}_{(\mu_i,\Sigma_i),h}$ is the restriction of P to the $\|\cdot\|_{\Sigma_i}$-ball centered at μ_i, with P-mass h.

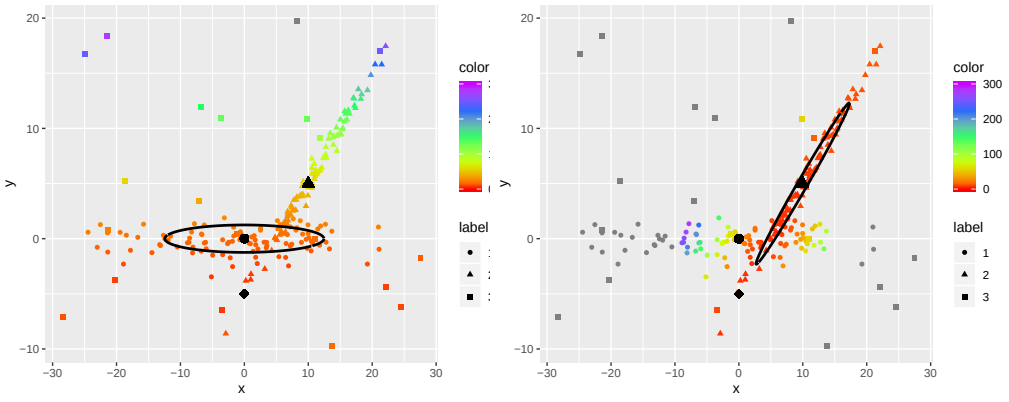


Figure 3: Illustration of the assignment phase

In Figure 3, we illustrate the procedure of decomposition of $\mathbb{R}^d$ into cells, when P is a uniform distribution on a set of $n=220$ points. We consider two couples (μ_1,Σ_1) and (μ_2,Σ_2). The center μ_1 is represented by a black disk and the center μ_2 by a black

triangle. We set the parameter $h = \frac{k}{n}$ to 0.25, so that $k = 55$. On the left Figure, we have represented $\mathcal{E}_1$, the (μ_1, Σ_1)-ellipse (sub-level set of $x \mapsto \|x - \mu_1\|_{\Sigma_1}$) that contains k points of the sample. The uniform distribution on this set of k points is $\tilde{P}_{n\,(\mu_1,\Sigma_1),h}$. On the left figure, we have represented $\mathcal{E}_2$, the (μ_2, Σ_2)-ellipse that contains k points.

Now, we consider some point x in $\mathbb{R}^2$, represented by a diamond. The point x is assigned to the cell $\mathcal{C}_i^*$ for which $\tilde{P}_{n\,\theta_i^*,h} \log(p_{(x,\Sigma_i^*)}(.))$ is maximal. On the left Figure, we have colored all points $y \in \mathbb{X}$ according to the value of the opposite of their log-likelihood to the distribution $\mathcal{N}(x, \Sigma_1)$, $-\log(p_{(x,\Sigma_1)}(y))$. We are interested in the mean of this value for the k points in the ellipse. The mean is 16.9. We do the same for the figure on the right, for the density $\mathcal{N}(x, \Sigma_2)$. The mean is 5.0. Since $5.0 < 16.9$, the point x is assigned to the cell $\mathcal{C}_2^*$. We define a cost at x by $C(x) = 5.0$. Almost all points are orange in $\mathcal{E}_1$ whereas they are almost all red in $\mathcal{E}_2$. Thus, it was clear that x was to be assigned to $\mathcal{C}_2^*$.

Once all of the points $x \in \mathbb{X}$ are assigned to a cluster, the total cost for $\boldsymbol{\theta} = (\mu_i, \Sigma_i)_{i\in[\![1,m]\!]}$ is given by $C_{\boldsymbol{\theta}} = \frac{1}{|\mathbb{X}|} \sum_{x\in X} C(x)$. The optimal codebook $\boldsymbol{\theta}^* = (\mu_i^*, \Sigma_i^*)_{i\in[\![1,m]\!]}$ is the minimizer of $\boldsymbol{\theta} \mapsto C_{\boldsymbol{\theta}}$.

An additional interpretation of the m-PLM follows on from the expression:

$$\tilde{P}_{\theta_i,h} \log(p_{(x,\Sigma_i)}(.)) = \log\left(\frac{1}{\sqrt{(2\pi)^d \det(\Sigma_i)}} \exp\left(-\frac{\|x - m(\tilde{P}_{\theta_i,h})\|_{\Sigma_i} + v(\tilde{P}_{\theta_i,h})}{2}\right)\right) \tag{4}$$

with $m(\tilde{P}_{\theta_i,h}) = \tilde{P}_{\theta_i,h}.$, the expectation of $\tilde{P}_{\theta_i,h}$ and $v(\tilde{P}_{\theta_i,h}) = \tilde{P}_{\theta_i,h}\|. - m(\tilde{P}_{\theta_i,h})\|_{\Sigma_i}$ its "variance" within the direction given by Σ_i.

According to Equation (4), $-C_i : x \mapsto P_{\theta_i,h}(\log(p_{(x,\Sigma_i)}(.)))$ coincides with the logarithm of the density of $\mathcal{N}(m(\tilde{P}_{\theta_i,h}), \Sigma_i)$ minus a term $v_i = \frac{1}{2} v(\tilde{P}_{\theta_i,h})$. The term v_i is large when the points in $\tilde{P}_{\theta_i,h}$ are far from its mean $m(\tilde{P}_{\theta_i,h})$, in terms of the Mahalanobis metric $\|\cdot\|_{\Sigma_i}$. As a consequence, any point $x \in \mathbb{R}^d$ will be assigned to a cluster i when the density of $\mathcal{N}(m(\tilde{P}_{\theta_i,h}), \Sigma_i)$ at x is large, and $\tilde{P}_{\theta_i,h}$ has a small $\|\cdot\|_{\Sigma_i}$-variance.

3.3 Some theoretical results – or performance of the method

In this section, we investigate the ability of the m-PLM to recover the sampling distribution P, on some examples. According to Lemma 1, the mean of P can be properly recovered from $L_{P,h}^m$. The following examples are evidences that the m-PLM allows to recover the right covariance matrice, up to a normalization factor. As a consequence, our method may be of interest to recover tangent spaces. We begin with the case of Gaussian distributions.

Theorem 5.
When $m = 1$, if $P = \mathcal{N}(\mu_0, \Sigma_0)$ with Σ_0 positive definite, the optimal parameters are given by $\mu^* = \mu_0$ and $\Sigma^* = \frac{d+\mathrm{d}^2_{\mathcal{N}(0,I_d),h}(0)}{d} \Sigma_0$.

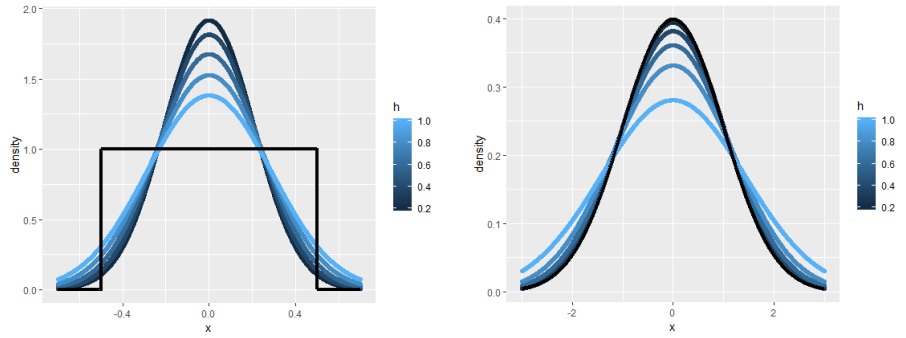
Proof. The proof of Theorem 5 is to be found in Section A. □

The following example of a uniform distribution on the segment $\left[-\frac{1}{2}, \frac{1}{2}\right] \times \{0\}$ of $\mathbb{R}^2$ provides a situation for which the optimal matrix Σ^* is degenerate (with determinant 0). The optimal matrix rewrites as $\Sigma^* = \begin{pmatrix} {\sigma^*}^2 & 0 \\ 0 & 0 \end{pmatrix}$, with σ^{*2} the optimal σ^2 for the m-PLM associated to the uniform distribution on $\left[-\frac{1}{2}, \frac{1}{2}\right]$, on $\mathbb{R}$ and some $\alpha \in \mathbb{R}$.

Example 6. *When $m = 1$, if P is uniform on the segment $\left[-\frac{1}{2}, \frac{1}{2}\right] \times \{0\}$, then the optimal parameter $\theta^* = (\mu^*, \Sigma^*)$ is given by $\mu^* = 0$ and $\Sigma^* = \begin{pmatrix} \frac{h^2+1}{12} & 0 \\ 0 & 0 \end{pmatrix}$.*

Moreover, the value $L^1_{P,h}((x, \Sigma^*)) = \tilde{P}_{\theta^*,h} \log(p_{(x,\Sigma^*)}(.))$ taken by the m-PLM at a point $x = (x_1, x_2)$ in $\mathbb{R}^2$ is $-\infty$ when $x_2 \neq 0$ and $+\infty$ when $x_2 = 0$. Consequently, all of the upper-level sets of the function $x \mapsto L^1_{P,h}((x, \Sigma^*))$ coincide with the line $\mathbb{R} \times \{0\}$.

Proof. The proof of Example 6 is to be found in Section B. □



Example 6 Example of Theorem 5, for $\mathcal{N}(0, 1)$

Figure 4: Optimal density $\mathcal{N}(\mu^*, \Sigma^*)$ with the m-PLM method.

As a consequence, our method allows to recover polygonal lines.

Theorem 7.
Let $\mathcal{M}$ be a union of c segments $(S_i)_{i \in [\![1,c]\!]}$. For every $i \in [\![1, c]\!]$, S_i is the segment centered at $\mu_i \in \mathbb{R}^d$ with length $L_i > 0$ and directed by a vector $v_i \in \mathbb{R}^d$ with $\|v_i\| = 1$. That is, $S_i = \left\{\mu_i + tL_i v_i \mid t \in \left[-\frac{1}{2}, \frac{1}{2}\right]\right\}$. Set $L = \sum_{i=1}^m L_i$. We assume that the vectors v_i are not collinear. If P is uniform on $\mathcal{M}$ and $h \leq \min_{i \in [\![1,m]\!]} L_i$, then the optimal parameters for the m-PLM are given by $\mu_i^* = \mu_i$ and $\Sigma_i^* = P_i \begin{pmatrix} L^2 \frac{h^2 + \left(\frac{L_i}{L}\right)^2}{12} & 0 \\ 0 & 0 \end{pmatrix} P_i^T$ where $P_i = \left[v_i, v_i^\perp\right]$ for $v_i^\perp$, a vector with norm 1 orthogonal to v_i.

The m-PLM $x \mapsto \max_{i \in [\![1,m]\!]} \tilde{P}_{\theta_i^*,h} \log(p_{(x,\Sigma_i^*)}(.))$ is equal to $+\infty$ when x is in the union of lines directed by the segments S_i and to $-\infty$ otherwise.

Proof. The proof of Theorem 7 is to be found in Section C. □

When dealing with a polygonal line $\mathcal{M}$, according to Theorem 7, the centers and directions are recovered by our method. Nonetheless, the upper-level sets of the m-PLM are unions of lines. Thus, approximating the set $\mathcal{M}$ with an upper-level set of the function m-PLM would be terrible in terms of the Hausdorff metric. Nonetheless, when $\mathcal{M}$ is a polygonal line, it is possible to recover $\mathcal{M}$ from the Σ_i^*s and the μ_i^*s. Indeed, if σ denotes the non-zero eigenvalue of Σ_i^*, the eigenvector associated to σ is v_i (or $-v_i$). Moreover, the P-mass of the cell $\mathcal{C}_i^*$ associated to (μ_i^*, Σ_i^*) is $\frac{L_i}{L}$, and L can be recovered from σ^* which is equal to $L^2 \frac{h^2 + \left(\frac{L_i}{L}\right)^2}{12}$. Thus, the segment S_i can be recovered. See Figure 5 for an illustration of this assertion on the example of Buchet et al. [8].

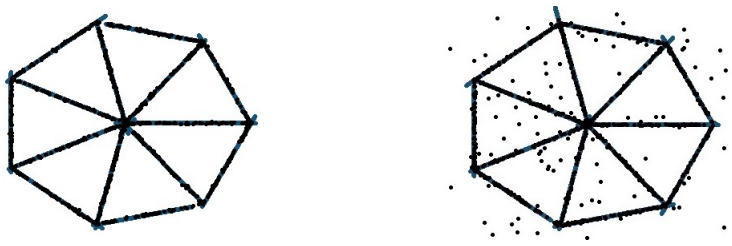


Figure 5: Illustration of the reconstruction method for a 1000-sample (left), with additionnal 100 points of clutter noise (right)

4 Algorithms

4.1 Algorithm to compute the m-PLM

The definition of the m-PLM is based on the computation of optimal centers $\boldsymbol{\mu}^* = (\mu_1^*, \mu_2^*, \ldots, \mu_m^*)$ in $\mathbb{R}^d$ (with a set of optimal covariance matrices $\boldsymbol{\Sigma}^*$). In this sense, we can relate it to the k-means problem [20], that consists in finding minimizers $(c_1^*, c_2^*, \ldots, c_m^*)$ of the k-means loss $(c_1, c_2, \ldots, c_m) \mapsto P \min_{i \in [\![1,m]\!]} \|. - c_i\|^2$. As for k-means, computing the optimal centers (i.e. computing the m-PLM) is not possible in practice, but it is possible to approximate it with a Lloyd-type algorithm. Such algorithms are made of mainly two steps. We consider m centers $c_1, c_2, \ldots, c_m$ in the space $\mathbb{R}^d$. The first step consists in splitting the space $\mathbb{R}^d$ into m cells $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m$ according to the m centers. A point is assigned to a cell $\mathcal{C}_i$ if it is closer to the center c_i that to any other center, for some metric or distortion. The second step consists in replacing, for every i, the center c_i by the barycentre of the points in the cell $\mathcal{C}_i$.

When considering the Euclidean metric, we recover the initial Lloyd's algorithm [19]. This algorithm was designed to compute a local minimum for the k-means loss. In our case, we consider a distortion of type $\|.\|_{\Sigma_i} + \omega_i$, where the matrix Σ_i for the Mahalanobis metric $\|.\|_{\Sigma_i}$ depends on the center c_i. The covariance matrix Σ_i is regularly updated, accordingly to Lemma 8. The measure Q and Q' in question correspond to the uniform measure on the ellipsoid centered at c_i in the direction Σ_i (i.e. the sublevel set of $x \mapsto \|x - c_i\|_{\Sigma_i}$) that contains $k = nh$ points of $\mathbb{X}$ and the uniform measure on the points in $\mathcal{C}_i \cap \mathbb{X}$. The new covariance matrix is a kind of mean of the covariance of the points in the ellipsoid and the covariance of the points in the cell. Moreover, the coefficient ω_i brings robustness. It plays a role of a variance term for the points in the ellipsoid, in the direction Σ_i. It will be large if the initial Σ_i has either a large determinant or is not well adapted to describe the points in the ellipsoid, see Section 3.2.

Lemma 8.
For every probability distributions Q and Q' on $\mathbb{R}^d$, a maximizer of

$$\psi : \Sigma \mapsto \int \int \log(p_{(u,\Sigma)}(v)) \mathrm{d}Q(u) \mathrm{d}Q'(v)$$

is given by:

$$\Sigma = \left[\int \int (v_i - u_i)(v_j - u_j) \mathrm{d}Q(u) \mathrm{d}Q'(v) \right]_{i,j \in [\![1,d]\!]},$$

where $u = (u_1, u_2, \ldots, u_d)$ and $v = (v_1, v_2, \ldots, v_d)$ are vectors in $\mathbb{R}^d$.

Proof. The proof of Lemma 8 is to be bound in Section D. □

The algorithm we propose to approximate $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ is the following.

Algorithm 1: Algorithm

```
Input : S = (X_1,X_2,... ,X_n) an n-sample from P; k and m;

# Initialization
h = k/n;
Sample mu_1, mu_2,... ,mu_m from S without replacement;
for i in 1..m:
  Sigma_i = I_d;
  theta_i = (mu_i,Sigma_i);
while the (mu_i,Sigma_i)s vary make the following two steps:
  # Decomposition in cells.
  for i in 1..m:
    C_i = [];
    c_i = P_{n,theta_i,h} . ;
    x_i = P_{n,theta_i,h} sqnorm(.-c_i,Sigma) ;
    d_i = log(det(Sigma_i));
    w_i = x_i + d_i
  for j in 1..n:
    Add X_j to the C_i (for i as small as possible) satisfying
for all l in 1..m
    sqnorm(X-c_i,Sigma_i) + w_i <= sqnorm(X-c_l,Sigma_l) + w_l;
  # Computation of the new centres and covariance matrices.
  for i in 1..m:
    mu_i = (1/|C_i|)sum(X , X in C_i);
    theta2_i = (mu_i,Sigma_i);
    for t,l in 1..d:
        for X=(X^1,X^2,... ,X^d) in C_i:
        cov(X,i,t,l) = P_{n,theta2_i,h} (X^t-.^t)(X^l-.^l);
        [Sigma_i]_{t,l} = 1/|C_i| sum(cov(X,i,t,l) , X in C_i);
    theta_i = (mu_i,Sigma_i);

Output : (mu_i)_{i in 1..m}, (Sigma_i)_{i in 1..m},
(C_i)_{i in 1..m}
```

In Algorithm 1, we use the notation and functions `I_d` for the identity matrix on $\mathbb{R}^d$, `sqnorm(x,Sigma)` the `Sigma`-Mahalanobis squared norm of `x` defined by `sqnorm(`x`,`Σ`)` $= \|x\|_\Sigma = x^T\Sigma^{-1}x$. Also, `P_{n,theta'_i,h}`$f$`(.)` corresponds to the expectation of f with respect to the distribution $\tilde{P}_{n\,\theta'_i,h}$. In particular, `P_{n,theta'_i,h} .` stands for the expectation of $\tilde{P}_{n\,\theta'_i,h}$. Moreover, `sum( f(X) , X in C_i)` represents the sum $\sum_{X\in C_i} f(X)$. And the `i`-th coordinate of any vector `X` in $\mathbb{R}^d$ is denoted by `X^i`.

Algorithm 1 allows to approximate the m-PLM, according to the following theorem.

Theorem 9.
Algorithm 1 converges to a local maximum of

$$(\boldsymbol{\mu},\boldsymbol{\Sigma}) \in (\mathbb{R}^d \times \Xi)^m \mapsto \frac{1}{n}\sum_{j=1}^{n} \max_{i\in[\![1,m]\!]} \tilde{P}_{n\,(\mu_i,\Sigma_i),h} \log(p_{(X_j,\Sigma_i)}(.)).$$

Proof. The proof of Theorem 9 is to be found in Section E. Its proof follows on from Lemma 8 and Lemma 1. □

4.2 Additional trimming step to wipe outliers out

We add some denoising parameter q, as an additional input for Algorithm 1, accordingly to [9, 5]. Just before the step `# Computation of the new centres and covariance matrices.`, we add the following step :

Algorithm 2: Trimming step

```
# Trimming step
for j in 1..n:
  l(X_j) = min(sqnorm(X_j - c_i , Sigma_i) + w_i | i in 1..m);
Sort [l(X_1),l(X_2),..., l(X_n)] in a non-increasing order;
The sorted vector is [l(X__1),l(X__2),... l(X__n)];
C_0 = [];
for i in 1..q:
  Set j the index such that X__i is in C_j;
  Remove X__i from C_j;
  Add X__i to C_0;
```

We aim at finding the best parameter q among a family of parameters `vect_q`. To this aim, we may apply Algorithm 1 together with the trimming step in Algorithm 2 many times for the different parameters q in `vect_q`. The term `Algorithm 2` stands for the Algorithm 1 with the additional trimming step. Moreover, the function `l` is defined in Algorithm 2.

Algorithm 3: Heuristic for the selection of q

```
Input : S = (X_1,X_2,...,X_n) ; k, m, vect_q, Ntimes

best_cost = []

for q in vect_q:
  cost = []
  for ntimes in 1..Ntimes:
    Apply Algorithm 2 to the input (S, k, m, q)
    cost[ntimes] = sum(l(X_j) | X_j not in C_0)
  best_cost[q] = min(cost)


Plot best_cost as a function of q

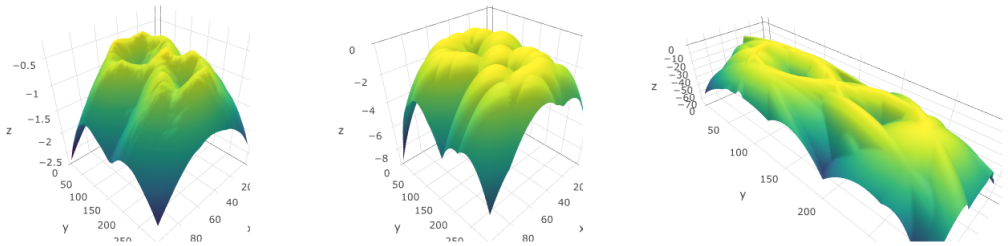
Output : parameter q at which there is a slope failure
 in the curve q -> best_cost(q).
```

5 Some illustrations of the method

5.1 Recovering the shape $\mathcal{M}$ and its homology.

We have sampled 200 points from the uniform measure on a sideways with radii $\sqrt{2}$ and $\sqrt{\frac{9}{8}}$, convolved with a Gaussian $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.15$. We have also added 100 points of clutter noise. We have launched our algorithm with the parameters $m = 30$ centers, $k = 20$ nearest neighbours (that is $h = 0.067$) and $q = 200$ points to be considered as signal. In Figure 6 we have plotted the sub-level sets of the opposite of the function DTM

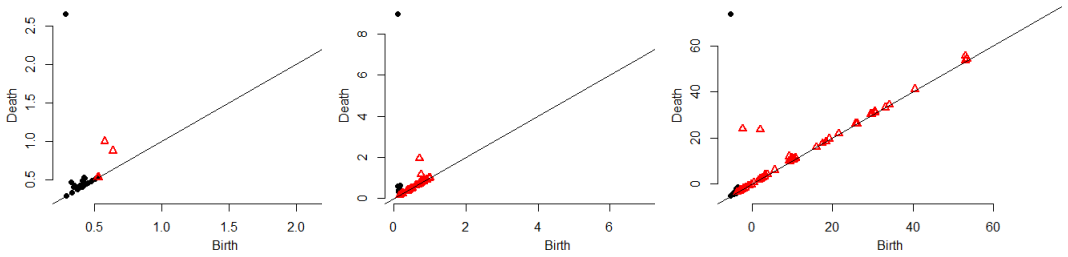
of [14] (with $h = 0.067$), the opposite of the m-PDTM of [6] (with $m = 30$, $h = 0.067$ and $q = 200$) and our m-PLM. Just like for the DTM, the two holes of the m-PLM are visible, and deep. An advantage of our method is that $\mathcal{M}$ can be recovered as an upper-level set of the m-PLM. Indeed, the top of the m-PLM is smooth, whereas the top of the DTM is more dented. Moreover, just as the m-PDTM, the m-PLM representation require the storage of only m centers (plus m covariance matrices), whereas the DTM requires to store around $\binom{n}{k}$ centers.



- DTM - m-PDTM m-PLM

Figure 6: Some distance functions for the sideways.

Persistent homology is a widespread tool in TDA to read the topological components of a shape, it was introduced by Edelsbrunner et al. in 2002, [10]. A persistence diagram associated to some function f stores the evolution of the homology (connected components, holes, voids...) of its sub-level sets. We used the R package TDA [12] to compute the persistent diagrams associated with the three distance functions in Figure 7: the DTM, the m-PDTM and minus the m-PLM. The connected components are represented by black points and the holes by red triangles. For each point, the absciss corresponds to the parameter b for which the component appears in $f^{-1}((-\infty, b])$ and its ordinate, the parameter d for which the component disappears in $f^{-1}((-\infty, d])$. The Figure 7 confirms that both methods perform in recovering the two holes of the sideways (the two red triangles further from the diagonal) and the fact that the sideways is connected (the single black point with ordinate $+\infty$). We may note that for the m-PLM, the red triangles are further from the diagonal than for other distance functions. Moreover, their birth time is smaller. It confirms that the true homology of $\mathcal{M}$ can be recover from an upper-level set of the m-PLM.

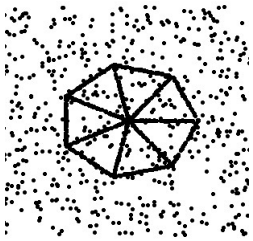


DTM m-PDTM m-PLM

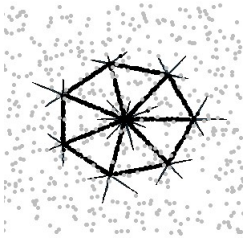
Figure 7: Persistent homology of three distance functions for the sideways.

5.2 The degenerated case

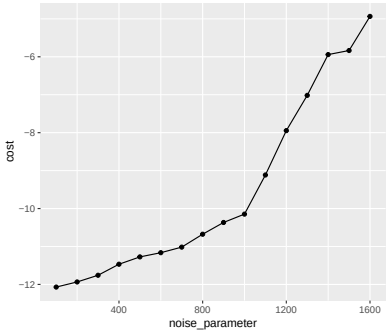
In this section, we consider the example of Buchet et al. in [8]. We sample 1000 points uniformly on a polygonal line $\mathcal{M}$ made of 14 segments disrupted with Gaussian noise $\mathcal{N}(0, \sigma^2)$, for $\sigma = 0.005$. We add 633 points of clutter noise, cf Figure 8. We launch our algorithm for the parameters $m = 14$, $k = 50$ and $q = 1000$. In Figure 8, we have colored the points of the sample that were considered as outliers by our algorithm in grey. We have also plotted some sublevel set of the m-PLM. Note that most of the outliers were indeed considered as outliers by the algorithm, except from some points that lie on the lines directed by the segments of $\mathcal{M}$. Thus our method allows to recover the true signal for this example for which the method of Buchet et al. was inefficient. Their method is free of parameter. Our is not, but it is possible to select the amount of points to keep q, according to the slope heuristic in Algorithm 3. According to this heuristic in Figure 8, the choice $q = 1000$ is the best one, which corresponds to the true number of signal points.



Sample of points



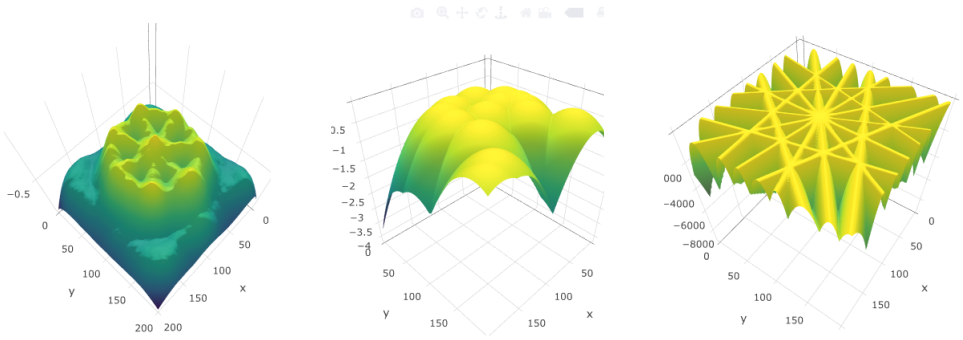
Upper-level sets of the m-PLM



Slope heuristic

Figure 8: Sample of 1000 points of signal with additional 633 points of cluter noise.

The fact that clutter noise lying on the lines directed by the segments of $\mathcal{M}$ are considered as signal points is a consequence of Theorem 7. Indeed, according to this theorem, the upper-level sets of the m-PLM are unions of lines. This property is illustrated by Figure 9, where we compared the m-PLM to the DTM and the m-PDTM.



- DTM - m-PDTM m-PLM

Figure 9: Sub-level sets for the example in [8]

6 Conclusion and perspectives

In this paper, we have developed tools to approximate structures $\mathcal{M}$ with a union of m ellipsoids, from a n-sample (n possibly much larger than m) generated around $\mathcal{M}$, in presence of clutter noise and outliers. It raises the question of the persistence methods for unions of ellipsoids; is it possible to adapt the TDA methods of Edelsbrunner et al. in [11] to this non-Bregman context? Moreover, we hope that this paper will motivate applications for (crossing)-manifold reconstruction tasks, extending for instance the work of Boissonnat et al. in [3]. Such reconstruction method would be adapted to noisy datasets, just like [1], but with the further advantage of small manifold storage cost. Moreover, we think that dealing with ellipsoids instead of balls would help dealing with intersection points for crossing manifolds.

On the other hand, we may note that for a uniform distribution on a polygonal line $\mathcal{M}$, our method turns out to be degenerated, accordingly to Theorem 7 and Figure 8. In this context, the upper-level sets of the m-PLM are unions of lines, which is a terrible approximation of $\mathcal{M}$ in terms of the Hausdorff metric. In future work, we propose to adapt this method to automatically select directions, so that this method perform well for large dimension d but low intrinsic dimension for the support of P. Such a method might face the remaining problem of the failure of our algorithm to automatically detect some outliers in Figure 8; the ones located on the line extending the segments of the original polygonal line.

Acknowledgements

I want to thank Clément Levrard, Bertrand Michel, Thomas Guyard and Marc Glisse for discussions on the material available in this paper or references. I am also extremely grateful to Luc Lehericy for the discussion about Lemma 10.

References

[1] Eddie Aamari and Clément Levrard. "Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction". In: *Discrete & Computational Geometry* 59.4 (2018), pp. 923–971.

[2] Tadeusz Bednarski and Brenton R. Clarke. "Trimmed likelihood estimation of location and scale of the normal distribution". In: *Australian Journal of Statistics* 35.2 (1993), pp. 141–153.

[3] Jean-Daniel Boissonnat and Arijit Ghosh. "Manifold Reconstruction Using Tangential Delaunay Complexes". In: *Discrete & Computational Geometry* 51.1 (2014), pp. 221–267.

[4] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. "Bregman Voronoi Diagrams". In: *Discrete & Computational Geometry* 44.2 (2010), pp. 281–307.

[5] Claire Brécheteau, Aurélie Fischer, and Clément Levrard. "Robust Bregman clustering". preprint. 2018.

[6] Claire Brécheteau and Clément Levrard. "The k-PDTM: A coreset for robust geometric inference". preprint. 2017.

[7] Lev M. Bregman. "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming". In: *USSR Computational Mathematics and Mathematical Physics* 7.3 (1967), pp. 200 –217. ISSN: 0041-5553.

[8] Mickael Buchet et al. “Declutter and Resample: Towards Parameter Free Denoising”. In: *33rd International Symposium on Computational Geometry (SoCG 2017)*. Ed. by Boris Aronov and Matthew J. Katz. Vol. 77. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, 23:1–23:16. ISBN: 978-3-95977-038-5.

[9] Juan. A. Cuesta-Albertos, Alfonso Gordaliza, and Carlos Matrán. “Trimmed k-means: an attempt to robustify quantizers”. In: *Ann. Statist.* 25.2 (Apr. 1997), pp. 553–576. DOI: 10.1214/aos/1031833664.

[10] H.E. Edelsbrunner, D.L. Letscher, and A. Zomorodian. “Topological persistence and simplification.” In: *Discrete Comput. Geom.* 28 (Nov. 2002), 511—533.

[11] Herbert Edelsbrunner and Hubert Wagner. “Topological Data Analysis with Bregman Divergences”. In: *Symposium on Computational Geometry*. Vol. 77. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017, 39:1–39:16.

[12] Brittany Fasy et al. “Introduction to the R package TDA”. In: (Nov. 2014).

[13] R.A. Fisher. “On the mathematical foundations of theoretical statistics”. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 222.594-604 (1922), pp. 309–368. ISSN: 0264-3952. DOI: 10.1098/rsta.1922.0009. eprint: http://rsta.royalsocietypublishing.org/content/222/594-604/309.full.pdf.

[14] David Cohen-Steiner Frédéric Chazal and Quentin Mérigot. “Geometric Inference for Probability Measures”. In: *Foundations of Computational Mathematics* 11.6 (2011), pp. 733–751.

[15] Luis A. García-Escudero et al. “A general trimming approach to robust cluster Analysis”. In: *Ann. Statist.* 36.3 (June 2008), pp. 1324–1345. DOI: 10.1214/07-AOS515.

[16] Leonidas J. Guibas, Quentin Mérigot, and Dmitriy Morozov. “Witnessed K-distance”. In: *Proceedings of the Twenty-seventh Annual Symposium on Computational Geometry*. SoCG ’11. Paris, France: ACM, 2011, pp. 57–64. ISBN: 978-1-4503-0682-9.

[17] Lorenzo Rosasco: Guillermo D. Cañas Tomaso A. Poggio. “Learning Manifolds with K-Means and K-Flats”. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems*. NIPS’12. 2012, pp. 2474–2482.

[18] Ali S. Hadi and Alberto Luceño. “Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms”. In: *Computational Statistics & Data Analysis* 25.3 (1997), pp. 251 –272. ISSN: 0167-9473.

[19] S. P. Lloyd. “Least squares quantization in PCM”. In: *IEEE Transactions on Information Theory* 28 (1982), pp. 129–137.

[20] J. MacQueen. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967, pp. 281–297.

[21] Ch.H. Müller. “Trimmed likelihood estimators in generalized linear models”. In: *Sixth International Conference on Computer Data Analysis and Modeling*. Vol. 2. eds. S. Aivazian, Y. Kharin, H. Rieder, Minsk. 2001, pp. 142–150.

[22] N. Neykov and Plamen Neytchev. “A robust alternative of the maximum likelihood estimators”. In: *COMPSTAT 1990 - Short Communications*. Jan. 1990, pp. 99–100.

[23] Michael Shindler, Alex Wong, and Adam Meyerson. "Fast and Accurate K-means for Large Datasets". In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS'11. Granada, Spain: Curran Associates Inc., 2011, pp. 2375–2383. ISBN: 978-1-61839-599-3.

A Proof of Theorem 5

The proof of Theorem 5 is based on the following Lemma.

Lemma 10.
For all $h \in [0,1]$, the distance to the measure $\mathcal{N}(0, I_d)$ at 0 for the metric $\|.\|_\Sigma$, $\mathrm{d}_{\mathcal{N}(0,1),h,\|.\|_\Sigma}$ is minimal for $\Sigma = I_d$ among all symmetric positive definite matrices Σ with determinant 1. The matrix I_d is the only minimizer when $h \in (0,1)$.

Proof. First recall that $\mathrm{d}^2_{\mathcal{N}(0,1),h,\|.\|_\Sigma}(0) = \frac{1}{h}\int_{l=0}^{h} \delta^2_{\mathcal{N}(0,1),l,\|.\|_\Sigma}(0)\mathrm{d}l$ with $\delta_{\mathcal{N}(0,1),l,\|.\|_\Sigma}(0)$ the radius of the $\|\cdot\|_\Sigma$-ball centered at 0 with mass l for $\mathcal{N}(0,1)$. In order to prove Lemma 10, it suffices to prove that for every $l > 0$, $\delta_{\mathcal{N}(0,1),l,\|.\|_\Sigma}(0) \geq \delta_{\mathcal{N}(0,1),l,\|.\|}(0)$ where $\|.\|$ denotes the Euclidean norm. Thus, it suffices to prove that:

$$\forall \delta > 0,\ \mathbb{P}\left(\sum_{i=1}^{n} \lambda_i X_i^2 < \delta\right) \leq \mathbb{P}\left(\sum_{i=1}^{n} X_i^2 < \delta\right) \tag{5}$$

when $\prod_{i=1}^n \lambda_i = 1$ and the λ_is are positive, where the probability $\mathbb{P}$ is computed according to X_1, X_2,..., X_n iid from $\mathcal{N}(0,1)$. Indeed, the matrix Σ^{-1} can be diagonalised in an orthogonal basis, and the λ_is correspond to its eigenvalues.

Equation (5) can be proved as follows. Set $\mathcal{B}_\lambda(\delta) = \left\{x_1, x_2, \ldots, x_n \in \mathbb{R} \mid \sum_{i=1}^n \lambda_i x_i^2 < \delta\right\}$, and $\mathcal{B}(\delta) = \left\{x_1, x_2, \ldots, x_n \in \mathbb{R} \mid \sum_{i=1}^n x_i^2 < \delta\right\}$. Since $\prod_{i=1}^n \lambda_i = 1$, after an integration by substitution, it follows that $\mathrm{Leb}(\mathcal{B}_\lambda(\delta)) = \mathrm{Leb}(\mathcal{B}(\delta))$, where Leb denotes the Lebesgue measure on $\mathbb{R}^d$. Then, for all $\delta > 0$, if $f(\delta)$ denotes the density of $P = \mathcal{N}(0, I_d)$ at any point $x = (x_1, x_2, \ldots, x_n)$ such that $\sum_{i=1}^n x_i^2 = \delta$, it holds:

$$\begin{aligned} P(\mathcal{B}_\lambda(\delta) \backslash \mathcal{B}(\delta)) &\geq f(\delta)\mathrm{Leb}(\mathcal{B}_\lambda(\delta) \backslash \mathcal{B}(\delta)) \\ &= f(\delta)\mathrm{Leb}(\mathcal{B}(\delta) \backslash \mathcal{B}_\lambda(\delta)) \\ &\geq P(\mathcal{B}(\delta) \backslash \mathcal{B}_\lambda(\delta)), \end{aligned}$$

which concludes.

□

The fact that $\mu^* = \mu_0$ is optimal is a direct consequence of Lemma 1.
For $\Sigma = AA^T \in \Xi$ positive definite, we may write

$$\begin{aligned} &\int \tilde{P}_{(\mu_0,\Sigma),h} \log(p_{(u,\Sigma)}(.))\mathrm{d}P(u) \\ &= -\int \tilde{P}_{(\mu_0,\Sigma),h} \frac{1}{2}\|u - .\|_\Sigma \mathrm{d}P(u) - \frac{1}{2}\log(\det(2\pi\Sigma)) \\ &= -\tilde{P}_{(\mu_0,\Sigma),h}\left(\int \frac{1}{2}\|\mu_0 - u\|_\Sigma \mathrm{d}P(u) - \frac{1}{2}\|\mu_0 - .\|_\Sigma\right) - \frac{1}{2}\log(\det(2\pi\Sigma)) \\ &= -\frac{1}{2}Tr(\Sigma^{-1}\Sigma_0) - \frac{1}{2}\tilde{P}_{(\mu_0,\Sigma),h}\|\mu_0 - .\|_\Sigma - \frac{1}{2}\log(\det(2\pi\Sigma)) \end{aligned}$$

where

$$\begin{aligned} \tilde{P}_{(\mu_0,\Sigma),h}\|\mu_0 - .\|_\Sigma &= \mathrm{d}^2_{P,h,\|.\|_\Sigma}(\mu_0) \\ &= \mathrm{d}^2_{\mathcal{N}(0,I_d),h,\|.\|_{((A^{-1}A_0)^T(A^{-1}A_0))^{-1}}}(0), \end{aligned}$$

since for every $\mu \in \mathbb{R}^d$,

$$\mathrm{d}_{\mathcal{N}(\mu_0,\Sigma_0),h,\|.\|_\Sigma}(\mu) = \mathrm{d}_{\mathcal{N}(0,I_d),h,\|.\|_{\left((A^{-1}A_0)^T(A^{-1}A_0)^{-1}\right)^{-1}}}(A_0^{-1}(\mu-\mu_0)). \tag{6}$$

Set $\tilde{\Sigma} = \left((A^{-1}A_0)^T(A^{-1}A_0)\right)^{-1}$, and $\Sigma' = \frac{\tilde{\Sigma}}{\det(\tilde{\Sigma})^{\frac{1}{d}}}$. Then it remains to minimise

$$\begin{aligned}
&\frac{Tr(\Sigma'^{-1})}{\det(\tilde{\Sigma})^{\frac{1}{d}}} + \mathrm{d}^2_{\mathcal{N}(0,I_d),h,\|.\|_{\tilde{\Sigma}}}(0) + \log(\det(\Sigma)) \\
&= \frac{Tr(\Sigma'^{-1})}{\det(\tilde{\Sigma})^{\frac{1}{d}}} + \frac{\mathrm{d}^2_{\mathcal{N}(0,I_d),h,\|.\|_{\Sigma'}}(0)}{\det(\tilde{\Sigma})^{\frac{1}{d}}} + d\log(\det(\tilde{\Sigma})^{\frac{1}{d}}) + \log(\det(\Sigma_0)).
\end{aligned}$$

Then the minimum is attained for $\det(\tilde{\Sigma})^{\frac{1}{d}} = \frac{Tr(\Sigma'^{-1})+\mathrm{d}^2_{\mathcal{N}(0,I_d),h,\|.\|_{\Sigma'}}(0)}{d}$ and according to Lemma 10, $\mathrm{d}^2_{\mathcal{N}(0,I_d),h,\|.\|_{\Sigma'}}(0)$ is minimal at $\Sigma' = I_d$. Moreover $Tr(\Sigma'^{-1})$ is also minimal for $\Sigma' = I_d$, according to the inequality of arithmetic and geometric means for the eigenvalues of Σ', which product equals to 1. Thus the minimum is unique and satisfies $\Sigma^* = \frac{d+\mathrm{d}^2_{\mathcal{N}(0,I_d),h,\|.\|}(0)}{d}\Sigma_0$.

B Proof of Example 6

Again, the fact that $\mu^* = 0$ is optimal is a direct consequence of Lemma 1. Now we aim at finding the matrix $\Sigma = \Sigma^*$ that minimizes

$$L(\Sigma) = \int \tilde{P}_{(0,\Sigma),h}\|u - .\|_\Sigma \mathrm{d}P(u) + \log(\det(2\pi\Sigma)).$$

Since 0 is the expectation of P, we may write

$$L(\Sigma) = \int \|u\|_\Sigma \mathrm{d}P(u) + \tilde{P}_{(0,\Sigma),h}\|.\|_\Sigma + \log(\det(2\pi\Sigma)).$$

We denote by σ the $(1,1)$-coefficient of the matrix Σ^{-1}. Then, $\int \|u\|_\Sigma \mathrm{d}P(u) = \frac{\sigma}{12}$ and

$$\begin{aligned}
\tilde{P}_{(0,\Sigma),h}\|.\|_\Sigma &= \frac{1}{h}\int_{u=-\frac{h}{2}}^{\frac{h}{2}} \sigma u^2 \mathrm{d}u \\
&= \sigma\frac{h^2}{12}.
\end{aligned}$$

It remains to minimize the function

$$\Sigma \mapsto \frac{\sigma}{12}\left(h^2+1\right) + \log(\det(\Sigma)).$$

Denote by λ_1 and λ_2 the eigenvalues of Σ. We can write $\Sigma = PDP^T$ for some orthogonal matrix P and D the diagonal matrix with coefficients λ_1 and λ_2. Then, σ is the top-left coefficient of the matrix $PD^{-1}P^T$. As a consequence, $\sigma = \frac{1}{\lambda_1}p_{1,1}^2 + \frac{1}{\lambda_2}p_{1,2}^2$. With $p_{1,1}^2 + p_{1,2}^2 = 1$ since P is orthogonal. Thus, the problem boils down to maximize the function $C\left(\frac{1}{\lambda_1}\alpha + \frac{1}{\lambda_2}(1-\alpha)\right) + \log(\lambda_1) + \log(\lambda_2)$ in $\alpha \in [0,1]$ and non negative λ_1 and λ_2, with $C = \frac{1}{12}\left(h^2+1\right)$.

When λ_1 and λ_2 are different, then optimal α is 0 if $\lambda_1 < \lambda_2$ and 1 otherwise.

The optimum is attained for $\alpha = 1$, $\lambda_2 \to 0$ and $\lambda_1 = C$, with a total cost going to $-\infty$. Then, $\Sigma = \begin{pmatrix} C & 0 \\ 0 & \lambda_2 \end{pmatrix}$, with λ_2 that goes to 0.

Now we can compute the value $L^1_{P,h}((x, \Sigma^*))$ taken by the m-PLM at a point $x = (x_1, x_2)$ in $\mathbb{R}^2$:

$$\begin{aligned} L^1_{P,h}((x,\Sigma^*)) &= \tilde{P}_{\theta^*,h} \log(p_{(x,\Sigma^*)}(.)) \\ &= -\frac{1}{2}\left(\|x\|_{\Sigma^*} + \tilde{P}_{\theta^*,h}\|.\|_{\Sigma^*} + \log(\det(2\pi\Sigma^*))\right) \\ &= \lim_{\lambda_2 \to 0} -\frac{1}{2}\left(\frac{12x_1^2}{1+h^2} + \frac{x_2^2}{\lambda_2} + \frac{h^2}{1+h^2} + 2\log(2\pi) + \log\left(\frac{1+h^2}{12}\right) + \log(\lambda_2)\right). \end{aligned}$$

As a consequence, $L^1_{P,h}((x, \Sigma^*)) = -\infty$ when $x_2 \neq 0$ and $L^1_{P,h}((x, \Sigma^*)) = +\infty$ when $x_2 = 0$. Thus, all of the upper-level sets of the function $x \mapsto L^1_{P,h}((x, \Sigma^*))$ coincide with the line $\mathbb{R} \times \{0\}$.

C Proof of Theorem 7

For simplicity, we assume that $L = \sum_{i=1}^k L_i = 1$.

We may write for $\theta_i = (\mu_i, \Sigma_i)$ with Σ_i any covariance matrix and μ_i the mean of the segment S_i,

$$\begin{aligned} &-2\int_{u\in\mathbb{R}^d} \sum_{i=1}^m \mathbb{1}_{S_i}(u)\tilde{P}_{\theta_i,h}\log(p_{(u,\Sigma_i)}(.))\mathrm{d}P(u) \\ &= \sum_{i=1}^m L_i \int_{t=-\frac{1}{2}}^{\frac{1}{2}} \log(\det(2\pi\Sigma_i)) + \tilde{P}_{\theta_i,h}\|\cdot - \mu_i - tL_iv_i\|_{\Sigma_i}\mathrm{d}t \\ &= \sum_{i=1}^m L_i \int_{t=-\frac{1}{2}}^{\frac{1}{2}} \left(\log(\det(2\pi\Sigma_i)) + \int_{u=-\frac{1}{2}}^{\frac{1}{2}} \|uhv_i - tL_iv_i\|_{\Sigma_i}\mathrm{d}u\right)\mathrm{d}t \\ &= \sum_{i=1}^m L_i \left(\log(\det(2\pi\Sigma_i)) + \frac{h^2\|v_i\|_{\Sigma_i}}{12} + \frac{L_i^2\|v_i\|_{\Sigma_i}}{12}\right) \end{aligned}$$

We optimize in Σ_i just as for the proof of Example 6, and get that $\Sigma_i^* = P_i \begin{pmatrix} \frac{h^2+L_i^2}{12} & 0 \\ 0 & 0 \end{pmatrix} P_i^T$ with P_i the matrix with first column v_i and second column, a vector $v_i^\perp$ with norm 1, orthogonal to v_i.

Just as for Example 6, we get that $\tilde{P}_{(\mu_i,\Sigma_i^*),h}\log(p_{(x,\Sigma_i^*)}(.))$ is $+\infty$ when x lies in the line directed by segment S_i, and $-\infty$ when it is not. Then, the partition of $\mathbb{R}^d$ made of the lines directed by the segments S_i is optimal. Indeed, if we consider another partition, then some cluster will be made of elements from two segments, S_1 and S_2 for instance. Then, the optimal Σ will have a finite determinant. Consequently, $\tilde{P}_{(\mu,\Sigma),h}\log(p_{(x,\Sigma)}(.))$ will be finite for every x in the cluster, thus less than $+\infty$, and not optimal.

When each cluster is a segment S_i, then, the optimal mean is μ_i, the mean of P restricted to S_i (accordingly to Lemma 1), and the optimal Σ_i is Σ_i^*. Moreover, we get that

$$\begin{aligned} &-2\int_{u\in\mathbb{R}^d} \max_{i\in[\![1,m]\!]} \tilde{P}_{\theta_i^*,h}\log(p_{(u,\Sigma_i^*)}(.))\mathrm{d}P(u) \\ &\qquad = -2\int_{u\in\mathbb{R}^d} \sum_{i=1}^m \mathbb{1}_{S_i}(u)\tilde{P}_{\theta_i^*,h}\log(p_{(u,\Sigma_i^*)}(.))\mathrm{d}P(u) = +\infty. \end{aligned}$$

D Proof of Lemma 8

We consider the function ψ defined by

$$\begin{aligned}\psi(\Sigma) &= \int\int \log(p_{(u,\Sigma)}(v))\mathrm{d}Q(u)\mathrm{d}Q'(v)\\ &= \int\int \frac{1}{2}\left(-\|v-u\|_\Sigma - \log(\det(2\pi\Sigma))\right)\mathrm{d}Q(u)\mathrm{d}Q'(v).\end{aligned}$$

The matrix Σ is symmetric with coefficients in $\mathbb{R}$, thus diagonalisable in an orthogonal basis. Thus, we can write $\Sigma = PDP^{-1} = PDP^T$ for some diagonal matrix D with coefficients λ_i on the diagonal. We use the notation $P = [p_{i,j}]_{i,j}$. Then,

$$\begin{aligned}-\psi(\Sigma) &= \int\int \frac{1}{2}\left((v-u)^T PD^{-1}P^T(v-u) + \log((2\pi)^d\det(PDP^{-1}))\right)\mathrm{d}Q(u)\mathrm{d}Q'(v)\\ &= \frac{1}{2}\int\int\left(\log(2\pi)^d + \sum_{i,j=1}^d (v_i-u_i)(v_j-u_j)\left(\sum_{k=1}^d p_{j,k}p_{i,k}\lambda_k^{-1}\right) + \sum_{k=1}^d\log(\lambda_k)\right)\mathrm{d}Q(u)\mathrm{d}Q'(v)\\ &= \frac{1}{2}\left(\log(2\pi)^d + \sum_{k=1}^d\lambda_k^{-1}\left[\int\int\sum_{i,j=1}^d (v_i-u_i)(v_j-u_j)p_{j,k}p_{i,k}\mathrm{d}Q(u)\mathrm{d}Q'(v)\right] + \log(\lambda_k)\right)\end{aligned}$$

Thus, $\psi(\Sigma)$ is maximal when $\lambda_k = \lambda_k^*$, where:

$$\lambda_k^* = \int\int\sum_{i,j=1}^d (v_i-u_i)(v_j-u_j)p_{j,k}p_{i,k}\mathrm{d}Q(u)\mathrm{d}Q'(v) = [P^TAP]_{k,k},$$

where A is defined by:

$$A = \left[\int\int (v_i-u_i)(v_j-u_j)\mathrm{d}Q(u)\mathrm{d}Q'(v)\right]_{i,j}.$$

Thus, it remains to minimise the function

$$\tilde{\psi} : P \mapsto Tr(\log(P^TAP)) = \sum_{k=1}^d \log\left(\sum_{i,j=1}^d\left[\int\int (v_i-u_i)(v_j-u_j)\mathrm{d}Q(u)\mathrm{d}Q'(v)\right]p_{j,k}p_{i,k}\right).$$

Again, we can diagonalise A in an orthonormal basis: $A = P_0D_0P_0^T = P_0D_0P_0^{-1}$. By setting $\tilde{P} = P^TP_0$, we get

$$\begin{aligned}\tilde{\psi}(P_0\tilde{P}^{-1}) &= Tr\log(\tilde{P}D_0\tilde{P}^T)\\ &= \sum_{l=1}^d\log(\sum_{k=1}^d\tilde{p}_{l,k}^2\lambda_k^0)\\ &\geq \sum_{l=1}^d\sum_{k=1}^d\tilde{p}_{l,k}^2\log(\lambda_k^0)\\ &= \sum_{k=1}^d\log(\lambda_k^0)\sum_{l=1}^d\tilde{p}_{l,k}^2\\ &= Tr\log[I_dD_0I_d^T].\end{aligned}$$

We used concavity of the function log, and the fact that $\tilde{P}$ is orthogonal, thus, $\sum_{l=1}^d\tilde{p}_{l,k}^2 = \sum_{k=1}^d\tilde{p}_{l,k}^2 = 1$. Thus, we can choose $\tilde{P} = I_d$, that is, $P = P_0$. Moreover, $\lambda_k^* = [P_0^TAP_0]_{k,k} = \lambda_k^0$. Thus, $\Sigma = P_0D_0P_0^T = A$ is a minimizer of ψ.

E Proof of Theorem 9

For every $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots \mu_k) \in \left(\mathbb{R}^d\right)^m$ and $\boldsymbol{\Sigma} = (\Sigma_1, \Sigma_2, \dots \Sigma_m) \in \Xi^m$, we note $\theta_i = (\mu_i, \Sigma_i)$, $\mathcal{C}_i = \mathcal{C}(\theta_i)$ the set of elements $X \in \mathbb{X}$ such that for every j, $\tilde{P}_{n\,\theta_i,h} \log(p_{(X,\Sigma_i)}(.)) \geq \tilde{P}_{n\,\theta_j,h} \log(p_{(X,\Sigma_j)}(.))$. Moreover, we set $\theta'_i = \left(\frac{1}{|\mathcal{C}_i|} \sum_{X \in \mathcal{C}_i} X, \Sigma_i\right)$,

$$\Sigma'_i \in \arg\max \left\{ \frac{1}{|\mathcal{C}_i|} \sum_{X \in \mathcal{C}_i} \tilde{P}_{n\,\theta'_i,h} \log(p_{(X,\Sigma)}(.)) \mid \Sigma \in \mathcal{X} \right\},$$

an expression for Σ'_i is given by Lemma 8 and $\theta''_i = \left(\frac{1}{|\mathcal{C}_i|} \sum_{X \in \mathcal{C}_i} X, \Sigma'_i\right)$.Then,

$$\begin{aligned}
&\frac{1}{n} \sum_{X \in \mathbb{X}} \max_{i \in [\![1,m]\!]} \tilde{P}_{n\,\theta_i,h} \log(p_{(X,\Sigma_i)}(.)) \\
&\leq \frac{1}{n} \sum_{i=1}^{m} \sum_{X \in \mathcal{C}_i} \tilde{P}_{n\,\theta'_i,h} \log(p_{(X,\Sigma_i)}(.)) \\
&\leq \frac{1}{n} \sum_{i=1}^{m} \sum_{X \in \mathcal{C}_i} \tilde{P}_{n\,\theta'_i,h} \log(p_{(X,\Sigma'_i)}(.)) \\
&\leq \frac{1}{n} \sum_{i=1}^{m} \sum_{X \in \mathcal{C}_i} \tilde{P}_{n\,\theta''_i,h} \log(p_{(X,\Sigma'_i)}(.)) \\
&\leq \frac{1}{n} \sum_{X \in \mathbb{X}} \max_{i \in [\![1,m]\!]} \tilde{P}_{n\,\theta''_i,h} \log(p_{(X,\Sigma'_i)}(.)).
\end{aligned}$$

We used Lemma 1.