Robust anisotropic power-functions-based filtrations for clustering

Claire Brécheteau

▶ **To cite this version:**

Claire Brécheteau. Robust anisotropic power-functions-based filtrations for clustering. 2019. hal-02397100v1

HAL Id: hal-02397100

https://hal.science/hal-02397100v1

Preprint submitted on 6 Dec 2019 (v1), last revised 25 Mar 2020 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Robust anisotropic power-functions-based filtrations for clustering.

Claire Brécheteau
Laboratoire de Mathématiques Jean Leray, École Centrale Nantes, France
https://www.math.sciences.univ-nantes.fr/~brecheteau/

Abstract

We consider robust power-distance functions that approximate the distance function to a compact set, from a noisy sample. We pay particular interest to robust power-distance functions that are anisotropic, in the sense that their sublevel sets are unions of ellipsoids, and not necessarily unions of balls. Using persistence homology on such power-distance functions provides robust clustering schemes. We investigate such clustering schemes and compare the different procedures on synthetic and real datasets. In particular, we enhance the good performance of the anisotropic method for some cases for which classical methods fail.

2012 ACM Subject Classification Theory of computation → Unsupervised learning and clustering

Keywords and phrases Power functions, Filtrations, Hierarchical Clustering, Ellipsoids

Funding *Claire Brécheteau*: [funding]

Acknowledgements I am extremely grateful to Samuel Tapie, for his suggestion to use tangency of ellipsoids at their intersection point, to derive the expression of their intersection radius.

Lines 500

1 Introduction

Often data can be represented as a point cloud $\mathbb{X}$ in a Euclidean space $\mathbb{R}^d$. Grouping data into clusters as homogeneous and well-separated as possible is the purpose of clustering. When no label is know in advance, we talk about unsupervised clustering. Topological data analysis (TDA) tools are designed to understand the shape of the data. Thereby, such tools may help to understand the shape of clusters in which to group the data. In this paper, we develop and study a TDA-based unsupervised clustering scheme. In addition, our method detects and removes points that do not really belong to any cluster; the outliers.

Clustering datasets is of extreme importance in multiple domains including medicine and social networks among others. The classical k-means method clusters data into isotropic clusters. In particular, the trimmed version of k-means of [14] that removes outliers, supplies balls-shaped clusters. These two algorithms have been extended by [2, 5] for Bregman-balls-shaped clusters, see also tclust [17] for ellipsoidal clusters. Such methods are well-suited for data generated according to mixtures of distributions which sublevel-set are Bregman balls themselves. For more general datasets, for instance, a sample of point from a disconnected manifold, these methods are no longer appropriate. Spectral clustering methods [26] perform such tasks, but are not robust to outliers. DBSCAN [19] is an algorithm based on a fixed upper-level set of an approximation of the density, and consequently, does not provide a multiscale information. Via a dendrogram, the classical single-linkage hierarchical clustering algorithm provides such a multiscale information. The dendrogram encodes information about the connectivity of unions of balls centered at points in $\mathbb{X}$, or equivalently, of the sublevel sets of the distance function to $\mathbb{X}$. For a fixed radius r, the Čech complex is a simplicial complex defined as the collection of simplices (vertex, edge, triangle, tetrahedron) for which the r-balls centered at the vertices have a non-empty common intersection. We


© C. Brécheteau;
licensed under Creative Commons License CC-BY


call 1-skeleton its subcomplex (a graph) that contains only vertices and edges. The non-decreasing family of such graphs indexed by $r \in \mathbb{R}$ is called a filtration. Single-linkage is a persistence-based method since is based on the persistence, prominence or equivalently lifetime of the connected components into this graph filtration, however, it is not robust to outliers. The algorithm ToMATo in [12] is robust and persistence-based. Indeed, it is based on a graph filtration built from a neighborhood graph and a (robust) distance-like function whose values guide the appearance of vertices and edges in the graph filtration. An example of robust distance function that Chazal et al. consider in [12] is given by the distance-to-measure (DTM) [11]. Note that the graph is a priori not intrinsic to the distance function, which may cause bad clustering. For instance, an edge that links two vertices with small distance-function value but intersects an area with large distance function value, may link two clusters that should not be. This problem was the cause of failure of the single-linkage method for data corrupted by outliers. Alternative filtrations that do not suffer from this problem are the DTM-filtration [1], or the power filtrations [7], based on the 1-skeleton of the Čech filtration associated to the sublevel sets of a power distance function: a function of type $x \mapsto \min_{i\in I} \|x - m_i\|^2 + \omega_i$ for some $(m_i)_{i\in I}$ in $\mathbb{R}^d$ and $(\omega_i)_{i\in I}$ in $\mathbb{R}$. Some approximations of the DTM that are power functions have been introduced and studied in the literature: the k-witnessed distance [18], the c-PDTM [6] whose sublevel sets are unions of c balls, and the c-PLM [4] whose sublevel sets are unions of c ellipsoids, with c possibly much smaller than the sample size. The last two functions are robust to outliers since their construction is based on the principle of trimmed least squares [25].

Contributions

By replacing balls with ellipsoids, we enlarge the notion of weighted Čech filtration into the anisotropic weighted Čech filtration. We derive an expression for the radius of intersection of two ellipsoids. We introduce a clustering algorithm based on persistence. Such a clustering algorithm can be run from any graph filtration, in particular, from the 1-skeleton of the anisotropic weighted Čech filtration, which corresponds to the filtration of sublevel sets of an anisotropic power function. We experiment this algorithm on the filtration of the c-PLM [4].

Practical interests

A clustering algorithm based on the persistence filtration of the sublevel sets of a power function is pertinent since unlike ToMATo, the graph is intrinsic to the distance function. So, no additional parameters are required for the algorithm. The main advantage of using an anisotropic power function is that its sublevel sets are ellipsoids. Much less ellipsoids are required than balls to Hausdorff-approximate a compact manifold with intrinsic dimension smaller than the ambient dimension. The clustering scheme can also be applied to decompose a set of points generated on a polygonal lines into segments. Once the ellipsoids computed, the persistence algorithm runs very fast, in c^2, with c, the number of ellipsoids. Most importantly, the robustness of the persistence algorithm relies on the robustness of the distance function. The c-PLM [4] is robust to outliers. The guaranty for the clustering method follows from the $\|\cdot\|_\infty$-distance closeness between the power distance function and the distance function to the underlying manifold $\mathcal{X}$, relatively to the minimal distance between the connected components of $\mathcal{X}$. Note that such a proximity condition is sufficient but not necessary, as illustrated by the different numerical examples, with the c-PLM.

Organisation of the paper

In Section 2, we recall the notions of power function and weighted Čech filtration, the filtration of the nerves of its sublevel sets, that we extend to anisotropic power functions. We prove some stability and approximation properties for such filtrations. Examples of robust power filtrations are also displayed. The main clustering algorithm, Algorithm 1 is given in Section 3. This algorithm applies to any filtration of graphs, including the graph filtrations obtained as the 1-skeleton of a weighted Čech filtration. We enumerate other types of filtrations that fit into this framework. Finally, we implement Algorithm 1 with the robust anisotropic aforementioned power function in Section 4. We compare this method to other clustering methods on synthetic and real datasets.

2 Power-functions-based filtrations for robust clustering

In the sequel, we will recall the notion of filtration for subsets of $\mathbb{R}^d$ and for simplicial complexes. We will consider a class of functions for which filtrations associated to sublevel sets are easily represented by filtrations of simplicial complexes, making the evolution of their connected components tractable: the power functions. In addition, we will give an example of robust power-functions [6] that can be built from a probability distribution or a pointset $\mathbb{X}$. Their sublevel sets are unions of c balls, with c possibly much smaller than the size of $\mathbb{X}$. Most importantly, we will also give an example of a robust anisotropic power-function, whose sublevel sets are unions of c ellipsoids [4]. Both of these power functions will be considered in the next sections for clustering purposes.

2.1 Generalities on filtrations

Consider T, a subset of $\mathbb{R}$ that represents time. We equip $\mathbb{R}^d$ with the Euclidean norm $\|\cdot\|$. A filtration indexed by T is a family $(V^t)_{t\in T}$ of subsets of $\mathbb{R}^d$, that is non-decreasing for the inclusion relation, in the sense that $\forall t \leq t', V^t \subset V^{t'}$. A typical example of filtration is given by the filtration of the sub-level sets of a function $f : \mathbb{R}^d \mapsto \mathbb{R}$: $\left(f^{-1}((-\infty, t])\right)_{t\in T}$. For any simplex S with vertex set given by $\mathbb{X}$, a finite set of points, the notion of filtration extends as follows. A filtration of simplicial complexes of S is a non-decreasing family $(S^t)_{t\in T}$ of subcomplexes of S, meaning that for every $t \leq t'$, any simplex of S^t is also a simplex of $S^{t'}$.

It is possible to compare two filtrations with the notion of interleaving. The interleaving pseudo-distance between two filtrations $(V^t)_{t\in T}$ and $(W^t)_{t\in T}$ is defined as the smallest $\epsilon > 0$ such that $(V^t)_{t\in T}$ and $(W^t)_{t\in T}$ are ϵ-interleaved, meaning that: $\forall t \in T$, $V^t \subset W^{t+\epsilon}$ and $W^t \subset V^{t+\epsilon}$. The same definition extends to simplicial complexes. It should be noted that for two functions f and g such that $\|f - g\|_\infty \leq \epsilon$, their sub-level-sets filtrations are ϵ-interleaved.

We will see in Section 3 that the notion of interleaving is primordial, since it measures the difference of topology between two filtrations. In particular, for our purpose of clustering based on sub-level sets of two functions, clustering stability will be guarantied from the closeness of the functions.

2.2 Power-functions-based filtrations

In this paper, we consider classes of functions whose sub-level sets filtration has a sparse representation, the power functions. The sublevel sets of these functions can be represented by simplicial complexes in so-called weighted Čech filtrations. We will consider two types of power functions, the isotropic and the anisotropic ones.

2.2.1 The isotropic case

An isotropic power function is a function $f_{\mathbf{m},\boldsymbol{\omega}} : \mathbb{R}^d \to \mathbb{R}$ defined from an index set $I = [\![1, c]\!]$, a family of centers $\mathbf{m} = (m_i)_{i\in I}$ in $\mathbb{R}^d$ and a family of weights $\boldsymbol{\omega} = (\omega_i)_{i\in I}$ in $\mathbb{R}$ by $f_{\mathbf{m},\boldsymbol{\omega}} : x \mapsto \min_{i\in I} \|x - m_i\|^2 + \omega_i$. A simple example of power function is the squared Euclidean distance function to a set of points $\mathbb{X}$, $\mathrm{d}^2_{\mathbb{X}} : x \in \mathbb{R}^d \mapsto \min_{m\in\mathbb{X}} \|x - m\|^2$. The sublevel sets of $f_{\mathbf{m},\boldsymbol{\omega}}$, $V^t_{\mathbf{m},\boldsymbol{\omega}} = f^{-1}_{\mathbf{m},\boldsymbol{\omega}}((-\infty, t])$, are unions of at most c balls $\mathcal{B}^t_i = \overline{\mathrm{B}}(m_i, \sqrt{t - \omega_i})$ with $\overline{\mathrm{B}}(m, r) = \{x \in \mathbb{R}^d \mid \|x - m\| \leq r\}$. Note that $\mathcal{B}^t_i$ is empty for $t < \omega_i$. A simple computation enhances that two balls $\mathcal{B}^t_i$ and $\mathcal{B}^t_j$ intersect if and only if $t \geq t_{i,j}$ with $t_{i,j} = \frac{(\omega_j-\omega_i)^2+2(\omega_j+\omega_i)\|m_j-m_i\|^2+\|m_j-m_i\|^4}{4\|m_j-m_i\|^2}$. The information of connectivity of $V^t_{\mathbf{m},\boldsymbol{\omega}}$ can be encoded in a graph $\mathcal{G}^t_{\mathbf{m},\boldsymbol{\omega}}$, composed of vertices indexed by I. The graph $\mathcal{G}^t_{\mathbf{m},\boldsymbol{\omega}}$ contains vertices $i \in I$ that satisfy $\omega_i \leq t$ and for every $i, j \in I$, the edge between the vertices i and j is in $\mathcal{G}^t_{\mathbf{m},\boldsymbol{\omega}}$ when $t_{i,j} \leq t$. Indeed, $\mathcal{G}^t_{\mathbf{m},\boldsymbol{\omega}}$ and $V^t_{\mathbf{m},\boldsymbol{\omega}}$ have the same number of connected components, and m_i and m_j are in the same connected component in $V^t_{\mathbf{m},\boldsymbol{\omega}}$ if and only if i and j are in the same connected component in $\mathcal{G}^t_{\mathbf{m},\boldsymbol{\omega}}$.

More generally, the topological information of $V^t_{\mathbf{m},\boldsymbol{\omega}}$ (number of connected components, loops, voids etc.) can be encoded in the weighted Čech complex $\mathrm{Cech}_{\mathbf{m},\boldsymbol{\omega}}(t)$, defined as the nerve of the union of balls $(\mathcal{B}^t_i)_{i\in I}$: $\mathrm{Cech}_{\mathbf{m},\boldsymbol{\omega}}(t) = \{\sigma \subset I \mid \bigcap_{i\in\sigma} \mathcal{B}^t_i \neq \emptyset\}$, [1, 7, 3]. According to the Nerve Lemma [20, Corollary 4G.3], any sublevel set $V^t_{\mathbf{m},\boldsymbol{\omega}}$ is homotopic to $\mathrm{Cech}_{\mathbf{m},\boldsymbol{\omega}}(t)$ and thus contains the same topological information. For computational reasons, the weighted Vietoris-Rips filtration is frequently considered as a provably good surrogate for the weighted Čech filtration $(\mathrm{Cech}_{\mathbf{m},\boldsymbol{\omega}}(t))_{t\in T}$. The weighted Vietoris-Rips complex $\mathrm{VR}_{\mathbf{m},\boldsymbol{\omega}}(t)$ is the flag complex of $\mathcal{G}^t_{\mathbf{m},\boldsymbol{\omega}}$ ($\mathcal{G}^t_{\mathbf{m},\boldsymbol{\omega}}$ is the 1-skeleton of the weighted Čech complex): $\mathrm{VR}_{\mathbf{m},\boldsymbol{\omega}}(t) = \{\sigma \subset I \mid \forall i, j \in \sigma, \mathcal{B}^t_i \cap \mathcal{B}^t_j \neq \emptyset\}$. Indeed, as a direct consequence of [3, Theorem 3.2] which is a generalization of the non-weighted case in [15, Theorem 2.5.], if the weights in $\boldsymbol{\omega}$ are non-negative, then these two filtrations are interleaved:

$$\forall 0 < t' \leq \frac{d+1}{2d} t, \mathrm{VR}_{\mathbf{m},\boldsymbol{\omega}}(t') \subset \mathrm{Cech}_{\mathbf{m},\boldsymbol{\omega}}(t) \subset \mathrm{VR}_{\mathbf{m},\boldsymbol{\omega}}(t). \tag{1}$$

We will see that these notions can all be extended to anisotropic power functions.

2.2.2 The anisotropic case

Consider an index set $I = [\![1, c]\!]$, families of centers $\mathbf{m} = (m_i)_{i\in I}$ in $\mathbb{R}^d$, of weights $\boldsymbol{\omega} = (\omega_i)_{i\in I}$ in $\mathbb{R}$ and of matrices $\boldsymbol{\Sigma} = (\Sigma_i)_{i\in I}$ in $\mathcal{M}_d$, the set of definite positive symmetric matrices. An anisotropic power function is a function $f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}} : \mathbb{R}^d \to \mathbb{R}$ defined from I, $\mathbf{m}$, $\boldsymbol{\omega}$ and $\boldsymbol{\Sigma}$ by $f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}} : x \mapsto \min_{i\in I} \|x - m_i\|^2_{\Sigma_i^{-1}} + \omega_i$. For any matrix $\Sigma \in \mathcal{M}_d$, $\|\cdot\|_{\Sigma^{-1}}$ denotes the Σ-Mahalanobis distance, defined for every $x \in \mathbb{R}^d$ by $\|x\|^2_{\Sigma^{-1}} = x^T\Sigma^{-1}x$. The sublevel sets of $f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$, $V^t_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}} = f^{-1}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}((-\infty, t])$, are unions of at most c ellipsoids $\mathcal{E}^t_i = \overline{\mathrm{B}}_{\Sigma_i}(m_i, \sqrt{t - \omega_i})$ with $\overline{\mathrm{B}}_\Sigma(m, r) = \{x \in \mathbb{R}^d \mid \|x - m\|_{\Sigma^{-1}} \leq r\}$. Again, $\mathcal{E}^t_i$ is empty for $t < \omega_i$. The value of $t_{i,j}$ above which the two ellipsoids $\mathcal{E}^t_i$ and $\mathcal{E}^t_j$ intersect is given below.

▶ **Proposition 1.** *Consider two ellipsoids $\mathcal{E}^t_i = \overline{\mathrm{B}}_{\Sigma_i}(m_i, \sqrt{t - \omega_i})$ and $\mathcal{E}^t_j = \overline{\mathrm{B}}_{\Sigma_j}(m_j, \sqrt{t - \omega_j})$ with $\omega_i \leq \omega_j$ in $\mathbb{R}$, m_i and m_j in $\mathbb{R}^d$, $\Sigma_i = P_iD_iP_i^T$ and $\Sigma_j = P_jD_jP_j^T$ in $\mathcal{M}_d$, with two positive diagonal matrices D_i and D_j, and two orthogonal matrices Σ_i and Σ_j accordingly to the spectral theorem. Set $\tilde{\Sigma} = \sqrt{D_i}P_i^T\Sigma_j^{-1}P_i\sqrt{D_i}$, the orthogonal and diagonal matrices $\tilde{P}$ and $\tilde{D} = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$ such that $\tilde{\Sigma} = \tilde{P}\tilde{D}\tilde{P}^T$ and $\tilde{m} = \tilde{P}^T\sqrt{D_i^{-1}}P_i^T(m_j - m_i)$. Ellipsoids $\mathcal{E}^t_i$ and $\mathcal{E}^t_j$ intersect if and only if $t \geq t_{i,j}$, with $t_{i,j}$ defined as follows. If $\|\tilde{m}\| \leq \sqrt{\omega_j - \omega_i}$,*

then $t_{i,j} = \omega_j$. *If* $\|\tilde{m}\| > \sqrt{\omega_j - \omega_i}$, *then* $t_{i,j} = \omega_j + \sum_{k=1}^{d} \left(\frac{\lambda \tilde{m}_k}{\lambda + \lambda_k}\right)^2 \lambda_k$, *where* λ *is the unique solution of the following equation:*

$$\sum_{k=1}^{d} \frac{\lambda_k - \lambda^2}{(\lambda + \lambda_k)^2} \lambda_k \tilde{m}_k^2 = \omega_j - \omega_i. \tag{2}$$

The proof of Proposition 1 is to be found in Section A. It is based on the fact that the ellipsoids $\mathcal{E}_i^t$ and $\mathcal{E}_j^t$ are tangent at their intersection point, and the corresponding gradients are orthogonal. In the context of isotropy (i.e. for $\Sigma_i = \Sigma_j = I_d$, the identity matrix of $\mathbb{R}^d$), $\tilde{m} = m_j - m_i$ and when $\|m_j - m_i\| > \sqrt{\omega_j - \omega_i}$, Equation (2) has a unique positive solution given by $\lambda = \frac{\omega_i - \omega_j + \|m_j - m_i\|^2}{\omega_j - \omega_i + \|m_j - m_i\|^2}$. The merging time is then $t_{i,j} = \frac{(\omega_j - \omega_i)^2 + 2(\omega_j + \omega_i)\|m_j - m_i\|^2 + \|m_j - m_i\|^4}{4\|m_j - m_i\|^2}$. This is consistent with the previous section. Furthermore, in the context of anisotropy, it is also possible to define $\mathcal{G}^t_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$, $\mathrm{Cech}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}(t)$ and $\mathrm{VR}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}(t)$, the anisotropic counterparts of $\mathcal{G}^t_{\mathbf{m},\boldsymbol{\omega}}$, $\mathrm{Cech}_{\mathbf{m},\boldsymbol{\omega}}(t)$ and $\mathrm{VR}_{\mathbf{m},\boldsymbol{\omega}}(t)$. Note that the nerve lemma still applies, since unions of ellipsoids are contractible. Although this paper is mostly based on the study of connected components for the purpose of clustering, anisotropic weighted Čech and Vietoris-Rips filtrations are primordial to have a tractable good estimation of the topology of compact sets from good approximations of the compact sets with finite unions of ellipsoids. In fact, just as their isotropic counterparts (1), these filtrations are interleaved, provided that the eigenvalues of the matrices in $\boldsymbol{\Sigma}$ belong to some interval $[\lambda_{\min}, \lambda_{\max}]$ with $\lambda_{\min}$ positive.

▶ **Proposition 2.** *If* $\boldsymbol{\omega}$ *is a set on non-negative weights in* $\mathbb{R}$. *Then, for every* $t > 0$ *and* $0 < t' \leq \frac{\lambda_{\min}}{\lambda_{\max}} \frac{d+1}{2d} t$, *we get that*

$$\mathrm{VR}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}(t') \subset \mathrm{Cech}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}(t) \subset \mathrm{VR}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}(t). \tag{3}$$

The condition of non-negative weights is not too restrictive since for general weights, the proposition can be applied with $\boldsymbol{\omega} - \min_{i \in I} \omega_i$, $t - \min_{i \in I} \omega_i$ and $t' - \min_{i \in I} \omega_i$ instead of $\boldsymbol{\omega}$, t and t'. So, for general weights, the condition is actually: $\min_{i \in I} \omega_i < t' \leq \frac{\lambda_{\min}}{\lambda_{\max}} \frac{d+1}{2d} t + \left(1 - \frac{\lambda_{\min}}{\lambda_{\max}} \frac{d+1}{2d}\right) \min_{i \in I} \omega_i$. As noted in [15], when $\lambda_{\min} = \lambda_{\max}$ and the weights in $\boldsymbol{\omega}$ are null, the term $\frac{\lambda_{\min}}{\lambda_{\max}} \frac{d+1}{2d}$ is optimal. Equality for the left inclusion is obtained for $\mathbf{m}$, the vertices of a regular d-simplex. The proof of Proposition 2 is available in Section B.

Often, less ellipsoids than balls are required to describe a compact set, for a fixed level of precision (e.g. for the Hausdorff distance). For instance, a segment in $\mathbb{R}^2$, and more generally, any d'-dimensional submanifold in $\mathbb{R}^d$, with $d' < d$. For this reason, isotropic Čech and Vietoris-Rips filtrations are pertinent tools to compute and store the topological information about the compact set $\mathcal{X}$ efficiently. The requisite condition is that we dispose of an anisotropic power function that is a good approximation of the squared Euclidean distance to the compact set $\mathcal{X}$, $\mathrm{d}^2_{\mathcal{X}}$. Such examples of functions are discussed in the sequel.

2.3 Examples of filtrations based on robust power functions

2.3.1 Isotropic robust power functions

Set $\mathbb{X}$, a set of n points generated on the neighborhood of a compact subset $\mathcal{X}$ of $\mathbb{R}^d$. In order to face the non robustness of the distance function to $\mathbb{X}$, Chazal et al. have introduced the notion of distance-to-measure (DTM), in [11]. The distance-to-measure function is a counterpart of the distance to the compact set $\mathbb{X}$ that is robust to noise and in particular to outliers. Its robustness depends on some parameter $k \in [\![1, n]\!]$, the number of nearest-neighbors $X^1, X^2, \ldots, X^k$ of x in $\mathbb{X}$ that are used to estimate the distance function of

x to $\mathbb{X}$. The distance-to-measure function $\mathrm{d}_{\mathbb{X},k}$ with parameter $k \in [\![1,n]\!]$ is defined by $\mathrm{d}^2_{\mathbb{X},k} : x \mapsto \frac{1}{k}\sum_{i=1}^k \|x - X^i\|^2 = \|x - m_{x,k}\|^2 + v_{x,k}$ with $m_{x,k} = \sum_{i=1}^k X^i$, the mean of the k nearest neighbours of x in $\mathbb{X}$ and $v_{x,k} = \frac{1}{k}\sum_{i=1}^k \|X^i - m_{x,k}\|^2$ their variance. Note that $\mathrm{d}_{\mathbb{X},1}$ coincides with $\mathrm{d}_{\mathbb{X}}$ and suffers from non robustness, whereas $\mathrm{d}_{\mathbb{X},n}(x)$ is the distance of x to the barycenter of the point cloud $\mathbb{X}$, up to some factor, which is very robust, but very poor in terms of topological information. Guibas et al. noted in [18] that the distance-to-measure is actually a weighted power function, according to its alternative expression:

$$\mathrm{d}^2_{\mathbb{X},k}(x) = \inf_{y\in\mathbb{R}^d} \|x - m_{y,k}\|^2 + v_{y,k}. \tag{4}$$

Indeed, the mean distance between x and its k nearest neighbors is necessarily smaller than the mean distance between x and the k neighbors of any other point $y \in \mathbb{R}^d$. This infimum is actually a minimum over a set of c points $\mathbf{y} = (y_i)_{i\in[\![1,c]\!]}$ in $\mathbb{R}^d$, with c of order $\binom{n}{k}$. In [18], Guibas et al. introduced and studied the k-witnessed distance, a power approximation of the DTM defined by replacing $\mathbb{R}^d$ by the sample $\mathbb{X}$ in (4). The sublevel sets of the k-witnessed distance are unions of n balls. An approximation of the DTM with c (possibly much smaller than n) balls has been introduced in [6]. This approximation, the c-PDTM, consists in replacing $\mathbb{R}^d$ by a set $\mathbf{y}_{c,k}$ of c points in $\mathbb{R}^d$. This set $\mathbf{y}_{c,k}$ is obtained as a minimum for a "k-means"-type criterion [24] : $\mathbf{y}_{c,k} \in \arg\min_{\mathbf{y}||\mathbf{y}|=c} \sum_{i=1}^n \min_{y\in\mathbf{y}} \|X_i - m_{y,k}\|^2 + v_{y,k}$, with $|\mathbf{y}|$, the cardinal of $\mathbf{y}$. Morally, the optimal set $\mathbf{y}_{c,k}$ is chosen such that on average, on $\mathbb{X}$, the power function $x \mapsto \min_{y\in\mathbf{y}} \|x - m_{y,k}\|^2 + v_{y,k}$ takes small values. Note that the graph of such a function is necessarily above the graph of the DTM. According to [6], for a sample on a regular d'-manifold, c can be chosen of order $n^{\frac{d'}{d'+4}}$, which is much smaller than n. For such a c, the m-PDTM is a good approximation of the distance to $\mathcal{X}$, in spite of noise.

2.3.2 An anisotropic robust power function

An anisotropic version of the k-PDTM has been introduced in [4], the k-power likelihood to measure (k-PLM). It consists in replacing Euclidean norms with Mahalanobis norms. For every $x \in \mathbb{R}^d$ and $\Sigma \in \mathcal{M}_d$, set X^1, X^2,... X^k the k-nearest neighbors of x in $\mathbb{X}$, for the Σ^{-1}-Mahalanobis norm: $\|X^i - x\|_{\Sigma^{-1}} \leq \|X^j - x\|_{\Sigma^{-1}}$ for every $i \leq j$. Denote by $m_{x,\Sigma,k}$ their mean, and by $v_{x,\Sigma,k} = \frac{1}{k}\sum_{i=1}^k \|X^i - m_{x,\Sigma,k}\|^2_{\Sigma^{-1}}$ their variance, relative to the Σ-Mahalanobis norm. Set $\boldsymbol{\theta}_{c,k}$, a family of c pairs $(y,\Sigma) \in \mathbb{R}^d \times \mathcal{M}_d$ that minimizes (or which criterion is as close as possible to the optimal criterion, in case of non existence of a minimum) the following "k-means"-type criterion $R_{c,k}$ among all $\boldsymbol{\theta}$s of cardinal c: $R_{c,k}(\boldsymbol{\theta}) = \sum_{i=1}^n \min_{(y,\Sigma)\in\boldsymbol{\theta}} \|X_i - m_{y,\Sigma,k}\|^2_{\Sigma^{-1}} + v_{y,\Sigma,k} + \log(\det(\Sigma))$. The term $\log(\det(\Sigma))$ prevents optimal covariances matrices to be degenerated, with Σ^{-1} going to 0. In some sense, minimizing such a criterion boils down to fit Gaussian distributions to the data set $\mathbb{X}$, at best. The c-PLM is the power function defined from $\boldsymbol{\theta}_{c,k}$ by: $x \mapsto \min_{(y,\Sigma)\in\boldsymbol{\theta}_{c,k}} \|x - m_{y,\Sigma,k}\|^2_{\Sigma^{-1}} + v_{y,\Sigma,k}$. A modification of the criterion $R_{c,k}$ has been introduced in [4], to remove some datapoints ($|\mathbb{X}| - sig$ for some parameter sig), when $\mathbb{X}$ is corrupted with outliers. The criterion is given by $R_{c,k,sig}(\boldsymbol{\theta}) = \min_{(i_1,i_2,\ldots,i_{sig})\in[\![1,c]\!]} \sum_{j=1}^{sig} \min_{(y,\Sigma)\in\boldsymbol{\theta}} \|X_{i_j} - m_{y,\Sigma,k}\|^2_{\Sigma^{-1}} + v_{y,\Sigma,k} + \log(\det(\Sigma))$.

In [4], it has been proved that Lloyd-type iterative algorithms [22] provide a local minimum $\tilde{\boldsymbol{\theta}}_{c,k}$ for the criterion $R_{c,k}$, but also a local minimum $\tilde{\boldsymbol{\theta}}_{c,k,sig}$ for the criterion $R_{c,k,sig}$. These algorithms run in $n\log(n)c \times it$, with it the number of iterations of the algorithm. Roughly, the algorithms consists, given $\boldsymbol{\theta} = (\mathbf{y}, \boldsymbol{\Sigma})$, to split the space $\mathbb{R}^d$ into weighted $\boldsymbol{\Sigma}$-curved Voronoi cells, to replace centers $\mathbf{y}$ by the barycenters of the cells, and to update the set of matrices $\boldsymbol{\Sigma}$ by some close formula that depends on the points on the cell and the points in

the ellipsoids. To compute $\tilde{\boldsymbol{\theta}}_{c,k,sig}$, a trimming step is added at each iteration. Actually, for the purpose of clustering, disposing of a local minimum is enough, as enhanced in the numerical illustration section. Indeed, the parameter $Threshold$ of our clustering algorithm, Algorithm 1, will make possible the removing of bad centers in $\tilde{\boldsymbol{\theta}}_{c,k}$ or in $\tilde{\boldsymbol{\theta}}_{c,k,sig}$.

3 Persistence-based clustering from power-functions-based filtrations

3.1 Persistence for power-functions-based filtrations

Set $f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}} : x \in \mathbb{R}^d \mapsto \min_{i\in I} \|x - m_i\|^2_{\Sigma_i^{-1}} + \omega_i$, an anisotropic power-function indexed by a finite set $I = [\![1, c]\!]$ and such that the ω_is are sorted in non-decreasing order. As above-mentioned, the sublevel sets $V^t = f^{-1}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}((-\infty, t])$ are unions of at most c ellipsoids $\mathcal{E}_i^t = B_{\Sigma_i}(m_i, \sqrt{t - \omega_i})$, non empty as soon as $t \geq \omega_i$. In particular, each sublevel set of $f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$ contains at most c connected components. Each connected component of V^t, V_i^t can be indexed by some element $i \in I$: the smallest i such that m_i belongs to the component. With a language abuse, we call connected component V_i, the family of connected components $(V_i^t)_{t\in T}$ that gets born at time $t = b_i = \omega_i$ and dies at a time $t = d_i$ when V_i^t merges with another connected component V_j^t for some $j \leq i$. Note that $d_1 = \infty$. The lifetime of the component V_i^t, $d_i - b_i$, is called persistence or prominence of the component i. We can represent this merging information with a barcode or a dendrogram. In these two representations, each line is associated to a component V_i, with length $d_i - b_i$, and begins at the height given by b_i. The dendrogram is obtained from the barecode by linking the bars associated to merging components, at a height given by the merging time.

When $\mathbf{m}$ is a point set $\mathbb{X}$, $\Sigma_i = I_d$ and $\omega_i = 0$ for every i, clustering points accordingly to the connected components of V^t boils down to the classical single-linkage clustering procedure, with $t > 0$, calibrated in accordance with the dendrogram. Such a clustering procedure suffers from non-robustness to outliers. In this paper, we consider an adjacent procedure, based on the idea of ToMATo algorithm [12], that consists in taking into account the prominence of components. To be precise, in the clustering scheme, we decide that a component V_i cannot merge with another component V_j at a time t larger than $\omega_i + Stop$, for some fixed threshold $Stop$. In other words, components with large prominence will never die in this clustering procedure. This is the purpose of Algorithm 1 in the next section.

In order to better visualize the prominence of the components, we represent their lifetimes in a persistence diagram. A persistence diagram is a multiset of points $(b_i, d_i) \in \mathbb{R}^2$ that lie above the diagonal $b = d$. Each point (b_i, d_i) is associated to a connected component V_i. The notion of persistence diagram was actually introduced by Edelsbrunner et al. in [16], in the broader framework of homology, and allows to compute lifetimes of additionnal features such as loops, voids etc. It is defined for filtrations that are regular enough, on triangulable spaces such as $\mathbb{R}^d$. The proper notion of regularity is the notion of q-tameness [9]. In [7, Proposition 3.5], Buchet et al. proved that the distance-to-measure function is q-tame. The proof of [7] can be straightforwardly adjusted for the distance function to a compact set and most importantly, for anisotropic power functions, provided that the eigenvalues of the matrices Σ_i are all positive. A sketch of proof is given in the Appendix, in Section C.

Since distance to compact sets, distance-to-measure and anisotropic power functions are q-tame, the persistence diagrams associated to their filtrations are well defined. They can be compared with the bottleneck distance, a distance between two diagrams D and D' defined by the minimal value of $\max_{x\in D, y\in D'} |y - \phi(x)|_\infty$ among functions ϕ that pair points in D with points in D', with some points potentially paired to diagonal points. Diagrams associated to interleaved filtrations are close, according to the following theorem.

▶ **Theorem 3** (Stability of persistence diagrams [9, 10, 13]). *If two filtrations V and W are q-tame and ϵ-interleaved, then the persistence diagrams of these filtrations are ϵ-close in bottleneck distance.*

According to Proposition 3, the persistence diagram of any anisotropic power function $f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$ that is $\epsilon - \|\cdot\|_\infty$ close to $\mathrm{d}_{\mathcal{X}}$, will be ϵ-bottleneck close to the persistence diagram of the sublevel sets of $\mathrm{d}_{\mathcal{X}}$. Consequently, prominence of the connected components of $\mathcal{X}$ can be deduced from the diagram associated to $f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$, for ϵ small enough. This bottleneck closeness occurs with large probability for a regular manifold $\mathcal{X}$ for the k-PDTM built from a noisy sample from $\mathcal{X}$, according to [6]. No such result has been proved yet for the k-PLM function. Anyway, intuitively, its sublevel sets are good approximations of the manifold $\mathcal{X}$, with the advantage that they are made of less ellipsoids, and that these ellipsoids are oriented accordingly to the manifold, i.e. with large eigenvalues on the tangent space and small eigenvalues on its orthogonal. This will be confirmed in the numerical illustrations section.

By construction, the persistence diagram (for connected components) associated to the filtration of the sublevel sets of $f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$ coincides with the persistence diagram associated to the anisotropic weighted Čech complex $\mathrm{Cech}(f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}})$. Consequently, we can forget about the ellipsoids and focus on the simplicial complex filtration, which can be computed and stored efficiently, in a $c \times c$ matrix $\mathrm{Mat} = (t_{i,j})_{i,j\in I}$. Such a matrix would contain the times of appearance of vertices and of merging of connected components in $\mathrm{Cech}(f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}})$. The clustering scheme of this paper exposed just below is based on such a merging matrix Mat.

3.2 An algorithm for persistence-based clustering

Consider $(\mathcal{G}^t)_{t\in\mathbb{R}}$ a filtration of sub-graphs of $\mathcal{G}$, a graph with c nodes. Based on this filtration, we define an algorithm, strongly inspired from the ToMATo algorithm [12]. The clustering scheme is guided by the persistence of the connected components in $(\mathcal{G}^t)_{t\in\mathbb{R}}$, and preserves components with large prominence. We assume that the nodes of $\mathcal{G}$ are labeled such that the node labeled i gets born before the node labeled j, when $i \leq j$. The procedure is as follows. A connected component gets born when an edge gets born, with the same label. A component will change of label at each time t for which it merges with a component with smaller label in $\mathcal{G}^t$, unless its prominence is larger than some parameter *Stop*. The prominence of an edge or a component is defined as the lifetime of the component in the filtration. In other words, the prominence of an edge is the elapsed time between the birth of the edge and the time t such that an edge with smaller index is present in its connected component in $\mathcal{G}^t$. The resulting clustering is given by the label of the edges at time $t = +\infty$. It will contain exactly labels of edges with a prominence larger than *Stop*. In this clustering scheme, it is also possible to decide that edges born after some time parameter *Threshold* are not relevant, and decide to remove them from the clustering procedure.

This procedure is implemented in Algorithm 1. This algorithm requires a merging matrix $\mathrm{Mat} = (t_{i,j})_{i,j\in I}$, with $I = [\![1, c]\!]$. We define its coefficients by $t_{i,i}$, the birth time of the node i in the filtration $(\mathcal{G}^t)_{t\in T}$; for $i > j$, $t_{i,j}$ the birth time of the edge $[i, j]$ and for $i < j$, $t_{i,j} = \infty$. The vector *Color* contains the resulting clustering, the vector *Birth*, the birth

time of the components and $Death$ their death time. Note that $Death[1]$ will always be $+\infty$.

Algorithm 1 Persistence-based Clustering Algorithm

Data: Mat, Threshold, Stop
Result: Color, Birth, Death
`Initialization` ;
c ← max{i | Mat[i,i] ≤ Threshold} ;
Mat ← Mat[1:c,1:c] ;
Birth ← [Mat[i,i] for i in 1:c] ;
Death ← [∞ for i in 1:c] ;
indice ← 1 ;
I ← 1 ;
time ← Mat[I,I] ;
Color ← [] ;
while *time* < ∞ **do**
  if *time = Mat[I,I]* **then**
    `Component I appears` ;
    indice ← indice + 1 ;
    Mat[I,I] ← ∞ ;
    Color[I] ← I;
  else
    (col_max, col_min) ← (max(Color[I],Color[J]) , min(Color[I],Color[J]));
    if *time - Birth[col_max]* ≤ *Stop* **then**
      `Components col_max and col_min merge` ;
      Replace all entries col_max by col_min in Color ;
      Death[col_max] ← time ;
    else
      `Component col_max will never die` ;
    end
    Mat[i,j] ← ∞ for every i, j ≤ min(indice,c) such that
    (Color[i],Color[j]) ∈ {(col_min, col_max), (col_max, col_min)};
  end
  I,J ← arg $\min_{i,j\leq \min(indice,c)}$ Mat[i,j] ;
  time ← Mat[I,J]
end

In practice, just as Chazal et al. [12], we recommend to run Algorithm 1 twice. A first time to measure the prominence of the components, and a second time with the parameters $Stop$ and $Threshold$ chosen in accordance with the prominence values. The selection of the parameters $Stop$ and $Threshold$ for the second passage can be done from the persistence diagram of the connected components given by the set of points $(Birth[i], Death[i])$, with $Birth$ and $Death$ given by the first passage. If γ points are well-separated from the other points in the diagram, then we will select $Stop$ such that there are γ clusters after the second passage. By well-separated, we mean that for some $b_{\min} < b_{\max}$, these points are above the line $d = b + b_{\max}$ whereas the other points are below the line $d = b + b_{\min}$, with $b_{\max} - b_{\min}$ large enough. In this context, we set $Stop = \frac{b_{\min}+b_{\max}}{2}$. In addition, if there is a group of points in the diagram, that are below the line $d = b + b_{\min}$, on the very right of the diagram, then these points correspond to components that can be considered as noise and can be removed. In these points are separated to the others with a line $b = b_s$, then we can chose $Threshold = d_s$.

When the graph filtration $(\mathcal{G}_t)_{t\in T}$ corresponds to the filtration of the sublevel sets of some power function $f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$, the matrix Mat has coefficients given by $t_{i,i} = \omega_i$ and for $i > j$, $t_{i,j}$ given by Proposition 1: the intersecting time of the ellipsoids $\mathcal{E}_i^t$ and $\mathcal{E}_j^t$. Recall that when the matrices are all equal to the identity, then $t_{i,j} = \frac{(\omega_j-\omega_i)^2+2(\omega_j+\omega_i)\|m_j-m_i\|^2+\|m_j-m_i\|^4}{4\|m_j-m_i\|^2}$.

Giving a sense to a minimal prominence $Stop^*$ for a good clustering is possible for distance functions. For instance, for the sublevel-sets filtration of $\mathrm{d}_{\mathcal{X}}$, $Stop^*$ is given by half of the minimal distance between two disjointed connected components of $\mathcal{X}$. Consequently, for any $\epsilon - \|\cdot\|_\infty$-close approximation of $\mathrm{d}_{\mathcal{X}}$, taking $Stop = Stop^* - \epsilon$ leads to a perfect clustering, provided that $2\epsilon < Stop^*$. A power function is not homogeneous to a distance function, but to the square of a distance function. Therefore, for positive weights $\boldsymbol{\omega}$, it could be more appropriate to consider the filtration of sublevel sets of $\sqrt{f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}}$ instead of $f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$.

The power function used to approximate $\mathrm{d}^2_{\mathcal{X}}$ from a set of points $\mathbb{X}$ is based on a family $\mathbf{m}$ of c centers, possibly different from $\mathbb{X}$. Algorithm 1 assigns labels to the centers in $\mathbf{m}$. Clustering points in $\mathbb{X}$ is made accordingly to these labels and to the Voronoi decomposition of $\mathbb{R}^d$, based on $\mathbf{m}$, $\boldsymbol{\omega}$ and $\boldsymbol{\Sigma}$: $x \in \mathbb{X}$ has the same label as m_i if $\|x - m_i\|^2_{\Sigma_i^{-1}} + \omega_i \leq \|x - m_j\|^2_{\Sigma_j^{-1}} + \omega_j$ for every j such that m_j and m_i are not born after parameter $Threshold$. This parameter removes centers in $\mathbf{m}$. To deal with outliers, it is also possible to remove points (i.e. assign a label 0) in the sample $\mathbb{X}$. For this, we remove points x for which $f_{\mathbf{m}^*,\boldsymbol{\omega}^*,\boldsymbol{\Sigma}^*}(x)$ is the largest, with $\mathbf{m}^*, \boldsymbol{\omega}^*, \boldsymbol{\Sigma}^*$ the parameters that were not removed by the algorithm. Indeed, $f_{\mathbf{m}^*,\boldsymbol{\omega}^*,\boldsymbol{\Sigma}^*}$ is supposed to approximate $\mathrm{d}^2_{\mathcal{X}}$. For the k-PLM function, the parameter $Threshold$ is primordial. Indeed, algorithm that computes $\boldsymbol{\theta}_{c,k}$ actually converges to a local minimum of the criterion. Consequently, it happens that ellipsoids $\mathcal{E}_i$ are actually far from the support. For such ellipoids, the weight ω_i will necessarily be large, from its definition. Since other ellipsoids should have small weight, such bad ellipsoids will be removed for a suitable choice of parameter $Threshold$ that should be calibrated from the persistence diagram.

3.3 Connection to other persistence-based clustering methods

In the sequel, we display different graph filtrations, to be used for persistence-based clustering, with Algorithm 1. For each of these filtrations, we give a summarize of the corresponding matrices Mat, in Table 1, with the convention that $t_{i,i} \leq t_{j,j}$ when $i \leq j$.

ToMATo Algorithm [12] is based on a graph filtration that is constructed from a graph $\mathcal{G}$ and some function f defined on the nodes of the graph $\mathcal{G}$. Morally, $\mathcal{G}^t$ is the sub-graph of $\mathcal{G}$ that contains the node i is and only if $f(i) \leq t$, and the edge $[i, j]$ if and only if i and j are in $\mathcal{G}^t$. Given a set of points $\mathbb{X}$ in $\mathbb{R}^d$, Chazal et al. have mostly studied this method for $\mathcal{G}$, a Rips graph of $\mathbb{X}$, and for $f(i)$, the DTM to $\mathbb{X}$ at X_i.

The DTM-filtration [1] corresponds to the 1-skeleton of the nerve of the union of balls $\left(\bigcup_{x\in\mathbb{X}} \overline{\mathrm{B}}(x, r_t(x))\right)_{t>0}$ with $r_t(x) = -\infty$ for $t < \mathrm{d}_{\mathbb{X},k}(x)$ and $r_t(x) = (t^p - \mathrm{d}^p_{\mathbb{X},k}(x))^{\frac{1}{p}}$ for $t \geq \mathrm{d}_{\mathbb{X},k}(x)$, for some $p \geq 1$ and with the convention that $\overline{\mathrm{B}}(x, -\infty)$ is empty. In Table 1, we give the coefficients for $p = 1$. The DTM-filtration with $p = 2$ was actually introduced in [7], leading to what we call Power filtration, which is actually the sublevel-sets filtration of the square of a power distance. We also consider additional power-functions-based filtrations, from the k-witnessed distance [18], the c-PDTM [6] and the c-PLM.

Table 1 Coefficients of Mat for the different methods, for $f = \mathrm{d}_{\mathbb{X},k}$, the DTM to $\mathbb{X}$ with number of nearest neighbors parameter k.

Method	$t_{i,i}$	$t_{i,j}$ for $i < j$
ToMATo	$f(i)$	$\max(f(i), f(j))(\mathbb{1}_{[i,j]\in\mathcal{G}})^{-1}$
DTM-filtration	$f(i)$	$\left(\frac{\|X_i - X_j\| + f(i) + f(j)}{2}\right)\mathbb{1}_{\|X_i - X_j\| > \lvert f(i) - f(j)\rvert} + f(i)\mathbb{1}_{f(i)-f(j)\geq\|X_i - X_j\|}$
$f_{\mathbf{m},\boldsymbol{\omega}}$	ω_i	$\frac{(\omega_j-\omega_i)^2+2(\omega_j+\omega_i)\|m_j-m_i\|^2+\|m_j-m_i\|^4}{4\|m_j-m_i\|^2}$
$\sqrt{f_{\mathbf{m},\boldsymbol{\omega}}}$	$\sqrt{\omega_i}$	$\sqrt{\frac{(\omega_j-\omega_i)^2+2(\omega_j+\omega_i)\|m_j-m_i\|^2+\|m_j-m_i\|^4}{4\|m_j-m_i\|^2}}$
$f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$	ω_i	Given by Proposition 1

Power filtration	$\sqrt{f_{\mathbf{m},\boldsymbol{\omega}}}$ with $\mathbf{m} = \mathbb{X}$ and $\boldsymbol{\omega} = (f^2(x))_{x\in\mathbb{X}}$
Witnessed	$\sqrt{f_{\mathbf{m},\boldsymbol{\omega}}}$ with $(\mathbf{m},\boldsymbol{\omega}) = (m_{x,k}, v_{x,k})_{x\in\mathbb{X}}$
c-PDTM	$f_{\mathbf{m},\boldsymbol{\omega}}$ with $(\mathbf{m},\boldsymbol{\omega}) = (m_{y,k}, v_{y,k})_{y\in\mathbf{y}_{c,k}}$
c-PLM	$f_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$ with $(\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}) = (m_{y,\Sigma,k}, v_{y,\Sigma,k} + \log(\det(\Sigma)), \Sigma)_{(y,\Sigma)\in\theta_{c,k}}$

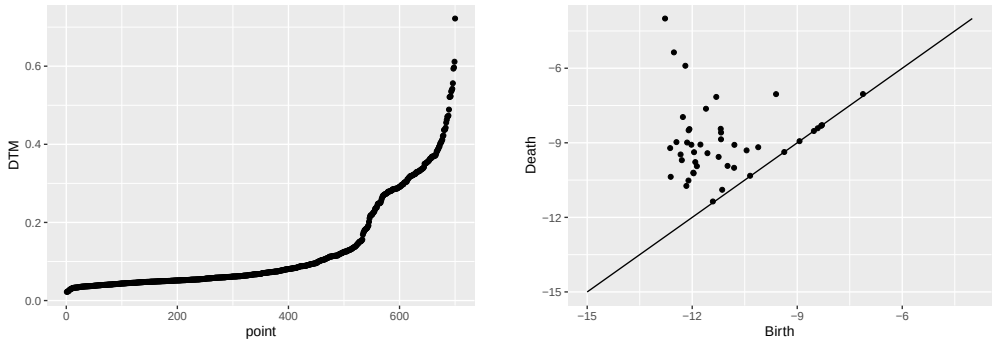
4 Numerical Illustrations

4.1 A complete illustration of the method

We consider a set of $N = 500$ points $\mathbb{Y} = (Y_i)_{i\in[\![1,N]\!]}$ uniformly sampled on a set $\mathcal{X}$. We corrupt this points with Gaussian noise : $\mathbb{Z} = (Z_i)_{i\in[\![1,N]\!]}$, N variables of distribution $\mathcal{N}(0,\sigma^2)$, with $\sigma = 0.02$. We observe a sample of N points $(X_i = Y_i + Z_i)_{i\in[\![1,N]\!]}$, corrupted by $Nnoise = 200$ outliers $(X_i)_{i\in[\![N+1,N+Nnoise]\!]}$ uniformly sampled on the square $[-1.5, 2.5]\times[-1.5, 2.5]$. This results in an $(N + Nnoise)$-sample $\mathbb{X}$. In the sequel, we expose the clustering method based on Algorithm 1 with the sublevel sets of an approximation of the c-PLM. The parameters are set to $c = 50$ centers, $k = 10$ nearest neighbors, $sig = 520$ points to consider as signal, and $it = 100$ iterations of the Lloyd-type algorithm to compute a local optimum of the criterion $R_{c,k,sig}$, $\tilde{\boldsymbol{\theta}}_{c,k,sig}$. We ran this Lloyd-type algorithm $ntimes = 10$ times and chose $\tilde{\boldsymbol{\theta}}_{c,k,sig}$ that minimizes $R_{c,k,sig}$ among the 10 ones.

The criterion $sig = 520$ was chosen according to the following heuristic. Outliers have a large DTM-value, in comparison to signal points. Therefore, we compute the vector of DTM values, $[\mathrm{d}_{\mathbb{X},k}(X_i), i \in [\![1, N + Nnoise]\!]]$ and sort it by non-decreasing order. We represent these sorted values and select sig as the point for which the curve's slope changes. The sorted values are represented in Figure 1. Note that this step can also be performed by plotting the values of $f_{\tilde{\boldsymbol{\theta}}_{c,k}}$ for some $\tilde{\boldsymbol{\theta}}_{c,k}$ obtained as a first approximation of the PLM, without trimming.

We ran Algorithm 1 a first time with the parameters $Threshold = \infty$ and $Stop = \infty$. We obtained two vectors $Birth$ and $Death$ and represented the points $(Birth[i], Death[i])_{i\in[\![1,c]\!]}$ in the persistence diagram, in Figure 1. From this diagram, it is possible to recover the true number of clusters, 3. Indeed, 3 points are well-separated from the other ones with a large band, parallel to the diagonal. Then, we selected $Stop = 5$ as the vertical distance between a line parallel to the diagonal that separates the 3 points from the others, and the diagonal. We ran Algorithm 1 a second time with the parameters $Threshold = \infty$ and $Stop = 5$. The resulting clustering is represented in Figure 2. A sublevel set of the function $f_{\tilde{\boldsymbol{\theta}}_{c,k}}$ is represented by the union of ellipses in this figure. Note that some ellipses are badly placed. Therefore, we use the parameter $Threshold$ to remove them. In the Figure 1, 7 points are on the right side, separated from the other points. We select $Threshold$ as the abscissa of a vertical line separating these 7 points from the others: $Threshold = -9.49$. We ran Algorithm 1 with the parameters $Threshold = -9.49$ and $Stop = 5$. The resulting clustering is represented in Figure 2. The bad ellipses have been removed.

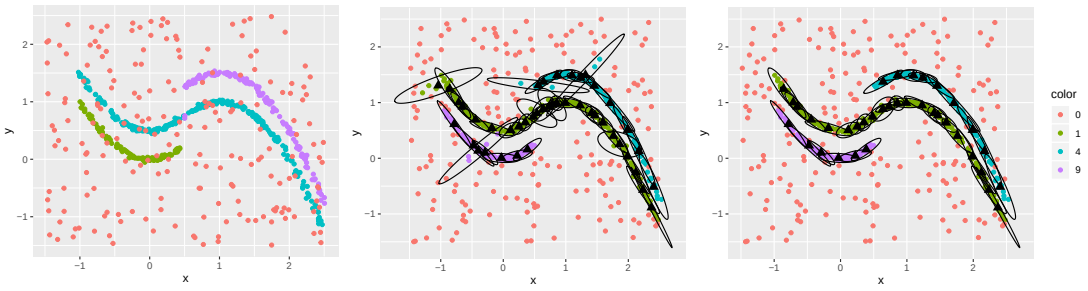


Sorted vector of DTM on points in $\mathbb{X}$, to approximate the number of outliers

Persistence diagram, to select *Stop* and *Threshold*

Figure 1 Parameters selection heuristics

The color of a point x in Figure 2 is given by the label in *Color* (returned by the algorithm) of its assigned center y_i in $\tilde{\theta}_{c,k}$ (the center y_i such that $f_{\tilde{\boldsymbol{\theta}}_{c,k}}(x) = \|x - m_{y_i,\Sigma_i,k}\|^2_{\Sigma^{-1}} + v_{y_i,\Sigma_i,k} + \log(\det(\Sigma_i))$). Set $\tilde{\boldsymbol{\theta}}'_{c,k,sig}$, the family $\tilde{\boldsymbol{\theta}}_{c,k,sig}$ without the centered removed because of *Threshold* in Algorithm 1. In the final clustering, we can assign the label 0 to the sig' points with smaller value for $f_{\tilde{\boldsymbol{\theta}}'_{c,k,sig}}$, with sig' calibrated with the same heuristic as in Figure 1, with $f_{\tilde{\boldsymbol{\theta}}'_{c,k,sig}}$ in spite of the DTM. Note that for very large datasets, since computing $\tilde{\theta}_{c,k}$ may take some time, it is possible to compute a $\tilde{\theta}_{c,k}$ from a random sub-sample of $\mathbb{X}$, to make the clustering procedure on the centers of $\tilde{\theta}_{c,k}$, and then to go back to $\mathbb{X}$ by clustering points with the label of the center of its curved weighted Voronoi cell.



Original labels

Threshold $= \infty$

Threshold = -9.49

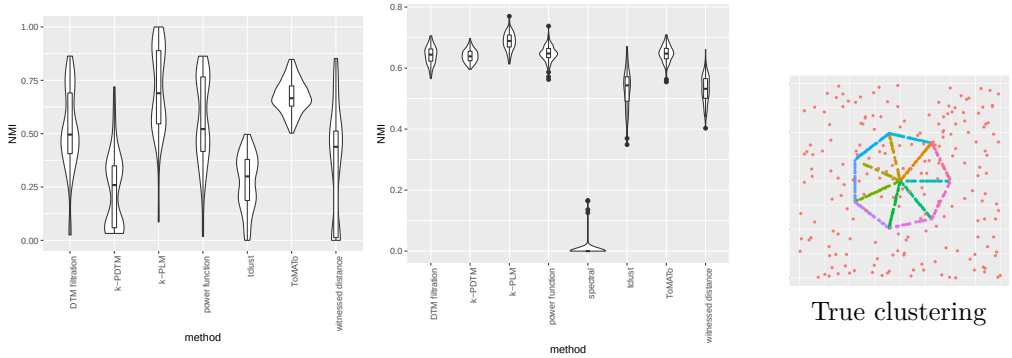
Figure 2 Two resulting clusterings, with ellipsoids

The false discovery rate FDR is defined as the proportion of signal points that are considered as outliers by the algorithm. In this example, for the two values of *Threshold*, we got exactly the same false discovery rate: FDR = 0.022. Nonetheless, the proportion of outliers that are considered as signal points by the algorithm equals 0.062 for $Threshold = \infty$ and 0.03 for $Threshold = -9.49$. It means that the improved method that removes bad means is better to detect outliers. For both methods, we computed the normalized mutual information NMI, a classical tool that measures the clustering performance (NMI = 1 for perfect clustering, NMI = 0 for a terrible clustering). The NMI between the true labels (0,1,2 or 3 ; 0 for outliers) and the label returned by the algorithm is 0.8052 when $Threshold = \infty$ and 0.8587 when $Threshold = -9.49$. We also computed the NMI between the true labels

(1,2,3) and the labels returned by the algorithm for non-outliers points, considered as signal. For both value of $Threshold$, $NMI = 0.9766712$, this is an almost perfect clustering.

4.2 Comparison of the different methods on synthetic datasets

We compared different clustering methods on two generated datasets : the previous dataset with 3 curves, and datapoints from a polygonal curve of 14 segments, as in [8]. We set parameters to $N = 500$, $Nnoise = 200$, $\sigma = 0.02$ is the standard deviation for the corrupting Gaussian distribution, $c = 50$, $k = 10$, $it = 100$ and at each step, $Threshold$ is chosen such that 10 means are removed from centers $\tilde{\boldsymbol{\theta}}_{c,k,sig}$ for the c-PLM. For the Tomato algorithm, the parameter $r = 0.12$ for the Rips graph is such that the graph is not connected (for the first example), thus there are more than 3 clusters. We used the function dbscan from the R packages dbscan [19], with parameters $eps = 0.15$ and $minPts = 10$; tclust and specc from the tclust [17] and kernlab [21] R packages. The shape of the violin for the c-PLM in the first example suggests that more that half of the clustering, among the 100, were almost perfect.



Example of the 3 curves

Example of the 14 segments

True clustering

Figure 3 Violin plots representing the NMI computed on signal points, detected as signal points.

4.3 Applications to real datasets

4.3.1 Recovering fleas species, based on 6 measurements

We picked the dataset flea from the R-package tourr [27], initially from [23]. This dataset contains records of 6 measurements for 74 males insects from the Palaeartic, from three different species : Heptapotamica, Concinna, Heikertingeri. The variables correspond to measurements on the tarsus, the aedeagus and the head. We normalized data so that the mean and variance of each of the 6 variables are respectively 0 and 1. In Table 2, we computed the NMI between the true species and the clustering returned by different methods. We ran each algorithm 10 times with at most 100 iterations. For every k-nearest-neighbours-based algorithm, we set $k = 10$. For ToMATo, we set $r = 1.9$ so that the graph is connected ; for the c-PLM and the c-DTM, $c = 50$ and for dbscan, $eps = 1.5$ and $minPts = 10$. The 3-PDTM and 3-PLM methods consists in clustering data according to the weighted Voronoi cells given by the optimal centers and covariance matrices.

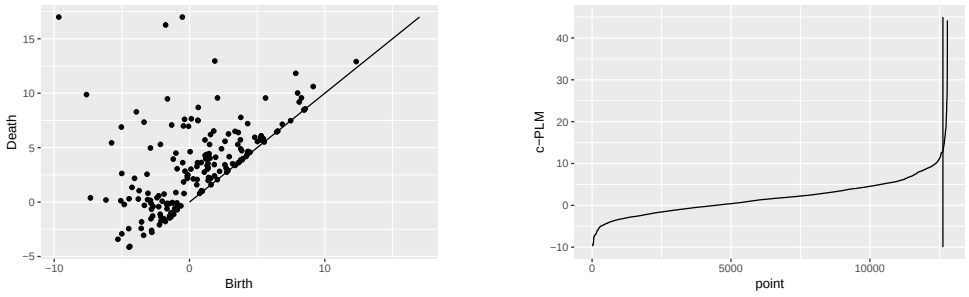
Table 2 NMI between clustering of fleas and their true specie

Without Algorithm 1	k-means 0.825	tclust 0.769	DBSCAN 0.647	Spectral clustering 1	3-PLM 1	3-PDTM 1
With Algorithm 1	ToMATo 0.628	Witnessed 0.906	power 1	DTM-filtration 1	c-PLM hier. 1	c-PDTM hier. 1

4.3.2 Clustering a earthquake dataset

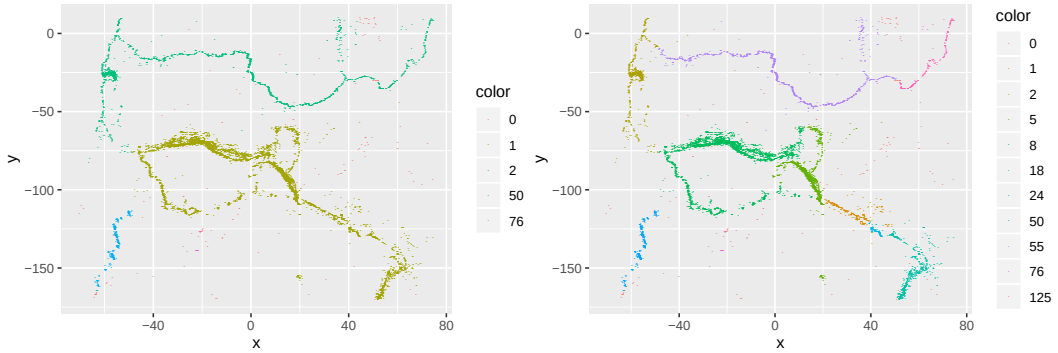
We consider a set of 12790 points representing the longitude and latitude of measured earthquakes of magnitude non smaller than 5.0, between the 01/01/1970 and the 01/01/2010. This dataset was picked from the website http://earthquake.usgs.gov/earthquakes/eqarchives/epic/.

We used Algorithm 1 with an approximation of the c-PLM based on a sub-sample of 2000 points from the dataset, with parameters $c = 200$, $k = 10$ and for $it = 50$ iterations. We restricted matrices Σ to have eigenvalues smaller than 50 by thresholding them. The persistence diagram in Figure 4 suggests that the dataset has 4 or 10 clusters. Moreover, the curve of the sorted values of the c-PLM approximation on the pointset in Figure 4 suggests to keep $sig = 12250$ points as signal points. See Figure 5 for the corresponding clustering.



Persistence diagram | Number of signal points selection

Figure 4 Parameters selection heuristics.



Clustering, 4 groups | Clustering, 10 groups

Figure 5 Earthquake clustering with Algorithm 1, for the c-PLM function.

References

1. Hirokazu Anai, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hiroya Inakoshi, Raphaël Tinarrage, and Yuhei Umeda. DTM-Based Filtrations. In Gill Barequet and Yusu Wang, editors, *35th International Symposium on Computational Geometry (SoCG 2019)*, volume 129 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 58:1–58:15, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. URL: http://drops.dagstuhl.de/opus/volltexte/2019/10462, doi:10.4230/LIPIcs.SoCG.2019.58.
2. Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, December 2005. URL: http://dl.acm.org/citation.cfm?id=1046920.1194902.
3. Gregory Bell, Austin Lawson, Joshua Martin, James Rudzinski, and Clifford Smyth. Weighted persistent homology. *Involve*, 12(5):823–837, 2019. URL: https://doi.org/10.2140/involve.2019.12.823, doi:10.2140/involve.2019.12.823.
4. Claire Brécheteau. Robust shape inference from a sparse approximation ofthe gaussian trimmed loglikelihood. unpublished, 2018.
5. Claire Brécheteau, Aurélie Fischer, and Clément Levrard. Robust bregman clustering. submitted, 2018.
6. Claire Brécheteau and Clément Levrard. A k-points-based distance for robust geometric inference. submitted, 2017.
7. Mickaël Buchet, Frédéric Chazal, Steve Oudot, and Donald R Sheehy. Efficient and robust persistent homology for measures. *Computational Geometry*, 58:70–96, 2016.
8. Mickael Buchet, Tamal K. Dey, Jiayuan Wang, and Yusu Wang. Declutter and Resample: Towards Parameter Free Denoising. In Boris Aronov and Matthew J. Katz, editors, *33rd International Symposium on Computational Geometry (SoCG 2017)*, volume 77 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 23:1–23:16, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
9. F. Chazal, V. de Silvia, M. Glisse, and S. Oudot. *The structure and stability of persistence modules*. Springer International Publishing, 2016.
10. Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry*, SCG '09, pages 237–246, New York, NY, USA, 2009. ACM. URL: http://doi.acm.org/10.1145/1542362.1542407, doi:10.1145/1542362.1542407.
11. Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric Inference for Measures based on Distance Functions. *Foundations of Computational Mathematics*, 11(6):733–751, 2011. URL: https://hal.inria.fr/inria-00383685, doi:10.1007/s10208-011-9098-0.
12. Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *J. ACM*, 60(6):41:1–41:38, November 2013. URL: http://doi.acm.org/10.1145/2535927, doi:10.1145/2535927.
13. David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, Jan 2007. URL: https://doi.org/10.1007/s00454-006-1276-5, doi:10.1007/s00454-006-1276-5.
14. Juan. A. Cuesta-Albertos, Alfonso Gordaliza, and Carlos Matrán. Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.*, 25(2):553–576, 04 1997. URL: https://doi.org/10.1214/aos/1031833664, doi:10.1214/aos/1031833664.
15. Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebr. Geom. Topol.*, 7(1):339–358, 2007. URL: https://doi.org/10.2140/agt.2007.7.339, doi:10.2140/agt.2007.7.339.
16. H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

17 Heinrich Fritz, Luis A. Garcia-Escudero, and Agustin Mayo-Iscar. tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47(12):1–26, 2012. URL: http://www.jstatsoft.org/v47/i12/.

18 Leonidas J. Guibas, Quentin Mérigot, and Dmitriy Morozov. Witnessed k-distance. In *Proceedings of the Twenty-seventh Annual Symposium on Computational Geometry*, SoCG '11, pages 57–64, New York, NY, USA, 2011. ACM. URL: http://doi.acm.org/10.1145/1998196.1998205, doi:10.1145/1998196.1998205.

19 Michael Hahsler, Matthew Piekenbrock, and Derek Doran. dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, 91(1):1–30, 2019. doi:10.18637/jss.v091.i01.

20 Allen Hatcher. *Algebraic topology*. Cambridge Univ. Press, Cambridge, 2000. URL: https://cds.cern.ch/record/478079.

21 Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. URL: http://www.jstatsoft.org/v11/i09/.

22 S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

23 Alexander A. Lubischew. On the use of discriminant functions in taxonomy. pages 455–477, 1962.

24 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*, pages 281–297, Berkeley, Calif., 1967. University of California Press. URL: https://projecteuclid.org/euclid.bsmsp/1200512992.

25 P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. 1987.

26 Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec 2007.

27 Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja. tourr: An R package for exploring multivariate data with projections. *Journal of Statistical Software*, 40(2):1–18, 2011. URL: http://www.jstatsoft.org/v40/i02/.

A Proof of Theorem 1

For ease of exposition, we set $i = 1$ and $j = 2$, with the assumption that $\omega_1 \leq \omega_2$. In order to make the problem simpler, we first transform ellipsoid $\mathcal{E}_1^t$ into a sphere by noting that

$$\|x-m_1\|^2_{\Sigma_1^{-1}} = (x-m_1)^T P_1 D_1^{-1} P_1^T (x-m_1) = \left(\sqrt{D_1^{-1}} P_1^T (x-m_1)\right)^T \left(\sqrt{D_1^{-1}} P_1^T (x-m_1)\right).$$

Therefore, we set $m_1' = \sqrt{D_1^{-1}} P_1^T m_1$ and $y = \sqrt{D_1^{-1}} P_1^T x - m_1'$. With this notation, it comes that $y^T y \leq t - \omega_1$ is the new equation of $\mathcal{E}_1^t$, and $(y - m_2')^T \tilde{\Sigma} (y - m_2') \leq t - \omega_2$ the new equation of $\mathcal{E}_2^t$, with $\tilde{\Sigma} = \sqrt{D_1} P_1^T \Sigma_2^{-1} P_1 \sqrt{D_1}$ and $m_2' = \sqrt{D_1^{-1}} P_1^T m_2 - m_1'$. Set $\tilde{m} = \tilde{P}^T m_2'$ and $z = \tilde{P}^T y$, for $\tilde{P}$ orthogonal and $\tilde{D}$ diagonal matrices such that $\tilde{\Sigma} = \tilde{P} \tilde{D} \tilde{P}^T$. With these notations, $\mathcal{E}_1^t$ has equation $z^T z \leq t - \omega_1$ and $\mathcal{E}_2^t$ has equation $(z - \tilde{m})^T \tilde{D} (z - \tilde{m}) \leq t - \omega_2$ in some coordinate system.

The ellipsoid $\mathcal{E}_1^t$ appears at time $t = \omega_1$, this is before $\mathcal{E}_2^t$ that appears at time $t = \omega_2$. Consequently, when $\|\tilde{m}\| \leq \sqrt{\omega_2 - \omega_1}$, the first time t for which $\mathcal{E}_1^t$ and $\mathcal{E}_2^t$ intersect is given by $t_{1,2} = \omega_2$.

From now on, we may assume that $\|\tilde{m}\| > \sqrt{\omega_2 - \omega_1}$. Ellipsoids $\mathcal{E}_1^t$ and $\mathcal{E}_2^t$ merge at a time $t_{1,2} > \omega_2$, and their intersection is given by some point $z \in \mathbb{R}^d$ that satisfies:

- $\sum_i (z_i - \tilde{m}_i)^2 \lambda_i = t_{1,2} - \omega_2$

- $\sum_i z_i^2 = t_{1,2} - \omega_1$
- $\exists \lambda > 0$, $\lambda z = -\tilde{D}(z - \tilde{m})$.

The last assumption comes from the fact that $\mathcal{E}_1^{t_{1,2}}$ and $\mathcal{E}_2^{t_{1,2}}$ are tangent at their intersection point z. Moreover, the tangent space of $\mathcal{E}_1^{t_{1,2}}$ at z is orthogonal to $2z$, the gradient of $\tilde{z} \mapsto \tilde{z}^T\tilde{z}$ at z ; and the tangent space of $\mathcal{E}_2^{t_{1,2}}$ at z is orthogonal to $2\tilde{D}(z - \tilde{m})$. Then, necessarily, λ satisfies the following equation (c.f. Equation (2)):

$$\sum_{i=1}^{d} \frac{\lambda_i - \lambda^2}{(\lambda + \lambda_i)^2} \lambda_i \tilde{m}_i^2 = \omega_2 - \omega_1.$$

The eigenvalues of $\tilde{\Sigma}$, $(\lambda_i)_{i\in[\![1,d]\!]}$ are positive. For every $a > 0$, $f_a : \lambda \mapsto \frac{a-\lambda^2}{(\lambda+a)^2} a$ is decreasing, with $\sum_{i=1}^d f_{\lambda_i}(0)\tilde{m}_i^2 = \|\tilde{m}\|^2 > \omega_2 - \omega_1$ and $\lim_{\lambda\to+\infty} \sum_{i=1}^d f_{\lambda_i}(\lambda)\tilde{m}_i^2 = -\sum_{i=1}^d \lambda_i \tilde{m}_i^2 < 0$. Consequently, Equation (2) has a unique solution λ. It comes that the ellipsoids $\mathcal{E}_1^t$ and $\mathcal{E}_2^t$ intersect at time $t_{1,2} = \omega_2 + \sum_{i=1}^d \left(\frac{\lambda\tilde{m}_i}{\lambda+\lambda_i}\right)^2 \lambda_i$ for this unique λ.

B Proof of Theorem 2

Since the anisotropic weighted Vietoris-Rips complex $\mathrm{VR}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}(t)$ is the flag complex of the 1-skeleton of the weighted anisotropic Čech complex $\mathrm{Cech}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}(t)$, $\mathcal{G}^t_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}$, the second inclusion of (3) is trivially satisfied. We now focus on the first inclusion of (3). For every $i \in I$ and $t > 0$, we have that

$$\overline{\mathrm{B}}\left(m_i, \sqrt{\lambda_{\min}}\sqrt{t-\omega_i}\right) \subset \overline{\mathrm{B}}_{\Sigma_i}\left(m_i, \sqrt{t-\omega_i}\right) \subset \overline{\mathrm{B}}\left(m_i, \sqrt{\lambda_{\max}}\sqrt{t-\omega_i}\right). \tag{5}$$

Set $t > 0$ and $0 < t' \leq \frac{\lambda_{\min}}{\lambda_{\max}} \frac{d+1}{2d} t$. If $\sigma \in \mathrm{VR}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}(t')$, then, for every $i, j \in \sigma$, the intersection $\mathrm{B}_{\Sigma_i}(m_i, \sqrt{t'-\omega_i}) \cap \mathrm{B}_{\Sigma_j}(m_j, \sqrt{t'-\omega_j})$ is nonempty and according to (5), $\mathrm{B}(m_i, \sqrt{\lambda_{\max}}\sqrt{t'-\omega_i}) \cap \mathrm{B}(m_j, \sqrt{\lambda_{\max}}\sqrt{t'-\omega_j})$ is nonempty. According to the Vietoris-Rips theorem of [3, Theorem 3.2], $\bigcap_{i\in\sigma} \mathrm{B}(m_i, \sqrt{\lambda_{\max}}\sqrt{\frac{2d}{d+1}}\sqrt{t'-\omega_i})$ is nonempty and since the weights are non-negative and $\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}\frac{2d}{d+1}} \geq 1$, it comes that $\bigcap_{i\in\sigma} \mathrm{B}(m_i, \sqrt{\lambda_{\min}}\sqrt{t-\omega_i})$ is nonempty, and finally, according to (5), $\sigma \in \mathrm{Cech}_{\mathbf{m},\boldsymbol{\omega},\boldsymbol{\Sigma}}(t)$.

C Sketch of proof for the q-tameness of power functions

Any function f of this type is both continuous and proper (the pre-image of any compact set is included in a union of ellipsoids and thus compact).

Since $\mathbb{R}^d$ is triangulable, there is some homeomorphism h between $\mathbb{R}^d$ and some locally finite simplicial complex C, then for any fixed t, we can define a complex C_t that contains $f^{-1}((-\infty, t])$, and $f \circ h_{|C_t}$ is continuous. Just as [7, Proposition 3.5], the conclusion follows from [9, Theorem 2.22] that states that the filtration of the sublevel sets of a continuous function defined on a finite polyhedron is q-tame.