An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists

Frédéric Chazal and Bertrand Michel

October 10, 2017

Abstract

Topological Data Analysis (TDA) is a recent and fast growing field providing a set of new topological and geometric tools to infer relevant features for possibly complex data. This paper is a brief introduction, through a few selected topics, to basic fundamental and practical aspects of TDA for non experts.

1 Introduction and motivation

Topological Data Analysis (TDA) is a recent field that emerged from various works in applied (algebraic) topology and computational geometry during the first decade of the century. Although one can trace back geometric approaches for data analysis quite far in the past, TDA really started as a field with the pioneering works of Edelsbrunner et al. (2002) and Zomorodian and Carlsson (2005) in persistent homology and was popularized in a landmark paper in 2009 Carlsson (2009). TDA is mainly motivated by the idea that topology and geometry provide a powerful approach to infer robust qualitative, and sometimes quantitative, information about the structure of data - see, e.g. Chazal (2017).

TDA aims at providing well-founded mathematical, statistical and algorithmic methods to infer, analyze and exploit the complex topological and geometric structures underlying data that are often represented as point clouds in Euclidean or more general metric spaces. During the last few years, a considerable effort has been made to provide robust and efficient data structures and algorithms for TDA that are now implemented and available and easy to use through standard libraries such as the Gudhi library (C++ and Python) Maria et al. (2014) and its R software interface Fasy et al. (2014a). Although it is still rapidly evolving, TDA now provides a set of mature and efficient tools that can be used in combination or complementary to other data sciences tools.

The TDApipeline. TDA has recently known developments in various directions and application fields. There now exist a large variety of methods inspired by topological and geometric approaches. Providing a complete overview of all these existing approaches is beyond the scope of this introductory survey. However, most of them rely on the following basic and standard pipeline that will serve as the backbone of this paper:

1. The input is assumed to be a finite set of points coming with a notion of distance - or similarity - between them. This distance can be induced by the metric in the ambient space (e.g. the Euclidean metric when the data are embedded in $\mathbb{R}^d$) or come as an intrinsic metric defined by a pairwise distance matrix. The definition of the metric on the data is usually given as an input or guided by the application. It is however important to notice that the choice of the metric may be critical to reveal interesting topological and geometric features of the data.

2. A "continuous" shape is built on top of the data in order to highlight the underlying topology or geometry. This is often a simplicial complex or a nested family of simplicial complexes, called a filtration, that reflects the structure of the data at different scales. Simplicial complexes can be seen as higher dimensional generalizations of neighboring graphs that are classically built on top of data in many standard data analysis or learning algorithms. The challenge here is to define such structures that are proven to reflect relevant information about the structure of data and that can be effectively constructed and manipulated in practice.

3. Topological or geometric information is extracted from the structures built on top of the data. This may either results in a full reconstruction, typically a triangulation, of the shape underlying the data from which topological/geometric features can be easily extracted or, in crude summaries or approximations from which the extraction of relevant information requires specific methods, such as e.g. persistent homology. Beyond the identification of interesting topological/geometric information and its visualization and interpretation, the challenge at this step is to show its relevance, in particular its stability with respect to perturbations or presence of noise in the input data. For that purpose, understanding the statistical behavior of the inferred features is also an important question.

4. The extracted topological and geometric information provides new families of features and descriptors of the data. They can be used to better understand the data - in particular through visualization- or they can be combined with other kinds of features for further analysis and machine learning tasks. Showing the added-value and the complementarity (with respect to other features) of the information provided by TDA tools is an important question at this step.

TDA **and statistics.** Until very recently, the theoretical aspects of TDA and topological inference mostly relied on deterministic approaches. These deterministic approaches do not take into account the random nature of data and the intrinsic variability of the topological quantity they infer. Consequently, most of the corresponding methods remain exploratory, without being able to efficiently distinguish between information and what is sometimes called the "topological noise".

A statistical approach to TDA means that we consider data as generated from an unknown distribution, but also that the inferred topological features by TDA methods are seen as estimators of topological quantities describing an underlying object. Under this approach, the unknown object usually corresponds to the support of the data distribution (or at least is close to this support). However, this support does not always have a physical existence; for instance, galaxies in the universe are organized along filaments but these filaments do not physically exist.

The main goals of a statistical approach to topological data analysis can be summarized as the following list of problems:

Topic 1: proving consistency and studying the convergence rates of TDA methods.

Topic 2: providing confidence regions for topological features and discussing the significance of the estimated topological quantities.

Topic 3: selecting relevant scales at which the topological phenomenon should be considered, as a function of observed data.

Topic 4: dealing with outliers and providing robust methods for TDA.

TDA **in data science.** On the application side, many recent promising and successful results have demonstrated the interest of topological and geometric approaches in an increasing number

of fields such has, e.g., material science Kramar et al. (2013); Nakamura et al. (2015) 3D shape analysis Skraba et al. (2010); Turner et al. (2014b), multivariate time series analysis Seversky et al. (2016), biology Yao et al. (2009) , chemistry Lee et al. (2017) or sensor networks De Silva and Ghrist (2007) to name a few. It is beyond the scope to give an exhaustive list of applications of TDA. On another hand, most of the successes of TDA result from its combination with other analysis or learning techniques - see Section 5.9 for a discussion and references. So, clarifying the position and complementarity of TDA with respect to other approaches and tools in data science is also an important question and an active research domain.

The overall objective of this survey paper is two-fold. First, it intends to provide data scientists with a brief and comprehensive introduction to the mathematical and statistical foundations of TDA. For that purpose, the focus is put on a few selected, but fundamental, tools and topics: simplicial complexes (Section 2) and their use for exploratory topological data analysis (Section 3), geometric inference Section 4) and persistent homology theory (Section 5) that play a central role in TDA. Second, this paper also aims at providing a short practical introduction to the Gudhi library, in particular its Python version that allows to easily implement and use the TDA tools presented in this paper (Section 6). Our goal is to quickly provide the data scientist with a few basic keys - and relevant references - to get a clear understanding of the basics of TDA to be able to start to use TDA methods and software for his own problems and data.

2 Metric spaces, covers and simplicial complexes

As topological and geometric features are usually associated to continuous spaces, data represented as finite sets of observations, do not directly reveal any topological information per se. A natural way to highlight some topological structure out of data is to "connect" data points that are close to each other in order to exhibit a global continuous shape underlying the data. Quantifying the notion of closeness between data points is usually done using a distance (or a dissimilarity measure), and it often turns out to be convenient in TDA to consider data sets as discrete metric spaces or as samples of metric spaces.

Metric spaces. Recall that a metric space (M, ρ) is a set M with a function $\rho : M \times M \to \mathbb{R}_+$, called a distance, such that for any $x, y, z \in M$:
i) $\rho(x, y) \geq 0$ and $\rho(x, y) = 0$ if and only if $x = y$,
ii) $\rho(x, y) = \rho(y, x)$ and,
iii) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.
Given a a metric space (M, ρ), the set $\mathcal{K}(M)$ of its compact subsets can be endowed with the so-called *Hausdorff distance*: given two compact subsets $A, B \subseteq M$ the Hausdorff distance $d_H(A, B)$ between A and B is defined as the smallest non negative number δ such that for any $a \in A$ there exists $b \in B$ such that $\rho(a, b) \leq \delta$ and for any $b \in B$, there exists $a \in A$ such that $\rho(a, b) \leq \delta$ - see Figure 1. In other words, if for any compact subset $C \subseteq M$, we denote by $d(., C) : M \to \mathbb{R}_+$ the distance function to C defined by $d(x, C) := \inf_{c \in C} \rho(x, c)$ for any $x \in M$, then one can prove that the Hausdorff distance between A and B is defined by any of the two following equalities:

$$\begin{aligned} d_H(A, B) &= \max\{\sup_{b \in B} d(b, A), \sup_{a \in A} d(a, B)\} \\ &= \sup_{x \in M} |d(x, A) - d(x, B)| = \|d(., A) - d(., B)\|_\infty \end{aligned}$$

It is a basic and classical result that the Hausdorff distance is indeed a distance on the set of compact subsets of a metric space. From a TDA perspective it provides a convenient way to quantify the proximity between different data sets issued from the same ambient metric space. However, it sometimes occurs in that one has to compare data set that are not sampled from the same ambient space. Fortunately, the notion of Hausdorff distance can be generalized to the

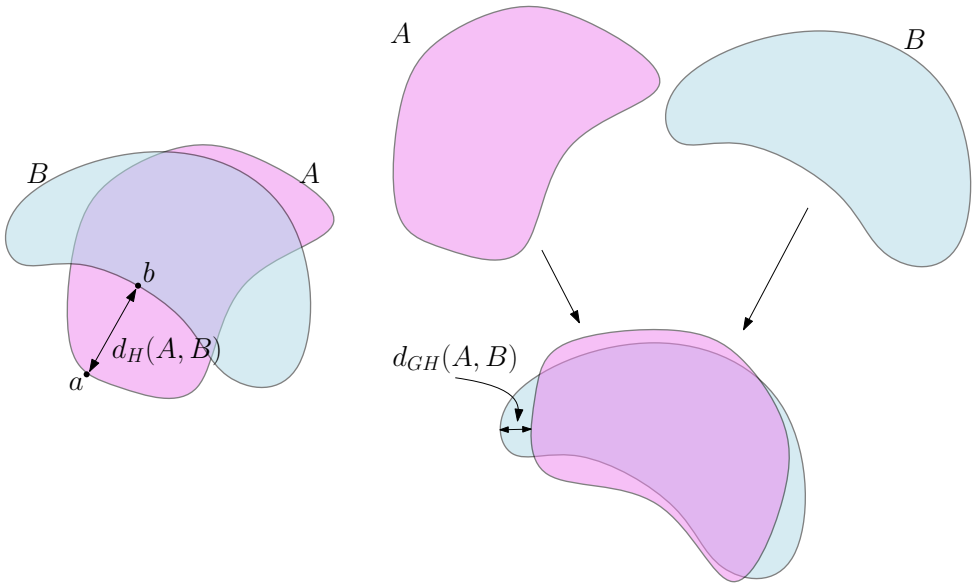


Figure 1: Right: the Hausdorff distance between two subsets A and B of the plane. In this example, $d_H(A,B)$ is the distance between the point a in A which is the farthest from B and its nearest neighbor b on B. Left: The Gromov-Hausdorff distance between A and B. A can been rotated - this is an isometric embedding of A in the plane - to reduce its Hausdorff distance to B. As a consequence, $d_{GH}(A,B) \leq d_H(A,B)$.

comparison of any pair of compact metric spaces, giving rise to the notion of *Gromov-Hausdorff distance.*

Two compact metric spaces (M_1, ρ_1) and (M_2, ρ_2) are *isometric* if there exists a bijection $\phi : M_1 \to M_2$ that preserves distances, i.e. $\rho_2(\phi(x), \phi(y)) = \rho_1(x, y)$ for any $x, y \in M_1$. The Gromov-Hausdorff distance measures how far two metric space are from being isometric.

Definition 1. *The Gromov-Haudorff distance $d_{GH}(M_1, M_2)$ between two compact metric spaces is the infimum of the real numbers $r \geq 0$ such that there exists a metric space (M, ρ) and two compact subspaces $C_1, C_2 \subset M$ that are isometric to M_1 and M_2 and such that $d_H(C_1, C_2) \leq r$.*

The Gromov-Hausdorff distance will be used later, in Section 5, for the study of stability properties persistence diagrams.

Connecting pairs of nearby data points by edges leads to the standard notion of neighboring graph from which the connectivity of the data can be analyzed, e.g. using some clustering algorithms. To go beyond connectivity, a central idea in TDA is to build higher dimensional equivalent of neighboring graphs by not only connecting pairs but also $(k+1)$-uple of nearby data points. The resulting objects, called simplicial complexes, allow to identify new topological features such as cycles, voids and their higher dimensional counterpart.

Geometric and abstract simplicial complexes. Simplicial complexes can be seen as higher dimensional generalization of graphs. They are mathematical objects that are both topological and combinatorial, a property making them particularly useful for TDA.

Given a set $\mathbb{X} = \{x_0, \cdots, x_k\} \subset \mathbb{R}^d$ of $k+1$ affinely independent points, the *k-dimensional simplex* $\sigma = [x_0, \cdots x_k]$ spanned by $\mathbb{X}$ is the convex hull of $\mathbb{X}$. The points of $\mathbb{X}$ are called the *vertices* of σ and the simplices spanned by the subsets of $\mathbb{X}$ are called the *faces* of σ. A *geometric simplicial complex* K in $\mathbb{R}^d$ is a collection of simplices such that:
i) any face of a simplex of K is a simplex of K,
ii) the intersection of any two simplices of K is either empty or a common face of both.

The union of the simplices of K is a subset of $\mathbb{R}^d$ called the underlying space of K that inherits from the topology of $\mathbb{R}^d$. So, K can also be seen as a topological space through its underlying space. Notice that once its vertices are known, K is fully characterized by the combinatorial description of a collection of simplices satisfying some incidence rules.

Given a set V, an *abstract simplicial complex* with vertex set V is a set $\tilde{K}$ of finite subsets of V such that the elements of V belongs to $\tilde{K}$ and for any $\sigma \in \tilde{K}$ any subset of σ belongs to $\tilde{K}$. The elements of $\tilde{K}$ are called the faces or the simplices of $\tilde{K}$. The dimension of an abstract simplex is just its cardinality minus 1 and the dimension of $\tilde{K}$ is the largest dimension of its simplices. Notice that simplicial complexes of dimension 1 are graphs.

The combinatorial description of any geometric simplicial K obviously gives rise to an abstract simplicial complex $\tilde{K}$. The converse is also true: one can always associate to an abstract simplicial complex $\tilde{K}$, a topological space $|\tilde{K}|$ such that if K is a geometric complex whose combinatorial description is the same as $\tilde{K}$, then the underlying space of K is homeomorphic to $|\tilde{K}|$. Such a K is called a *geometric realization* of $\tilde{K}$. As a consequence, abstract simplicial complexes can be seen as topological spaces and geometric complexes can be seen as geometric realizations of their underlying combinatorial structure. So, one can consider simplicial complexes at the same time as combinatorial objects that are well-suited for effective computations and as topological spaces from which topological properties can be inferred.

Building simplicial complexes from data. Given a data set, or more generally a topological or metric space, there exist many ways to build simplicial complexes. We present here a few classical examples that are widely used in practice.

A first example, is an immediate extension of the notion of α-neighboring graph. Assume that we are given a set of points $\mathbb{X}$ in a metric space (M, ρ) and a real number $\alpha \geq 0$. The *Vietoris-Rips complex* $\mathrm{Rips}_\alpha(\mathbb{X})$ is the set of simplices $[x_0, \ldots, x_k]$ such that $d_{\mathbb{X}}(x_i, x_j) \leq \alpha$ for all (i, j). It follows immediately from the definition that this is an abstract simplicial complex. However, in general, even when $\mathbb{X}$ is a finite subset of $\mathbb{R}^d$, $\mathrm{Rips}_\alpha(\mathbb{X})$ does not admit a geometric realization in $\mathbb{R}^d$; in particular, it can be of dimension higher than d.

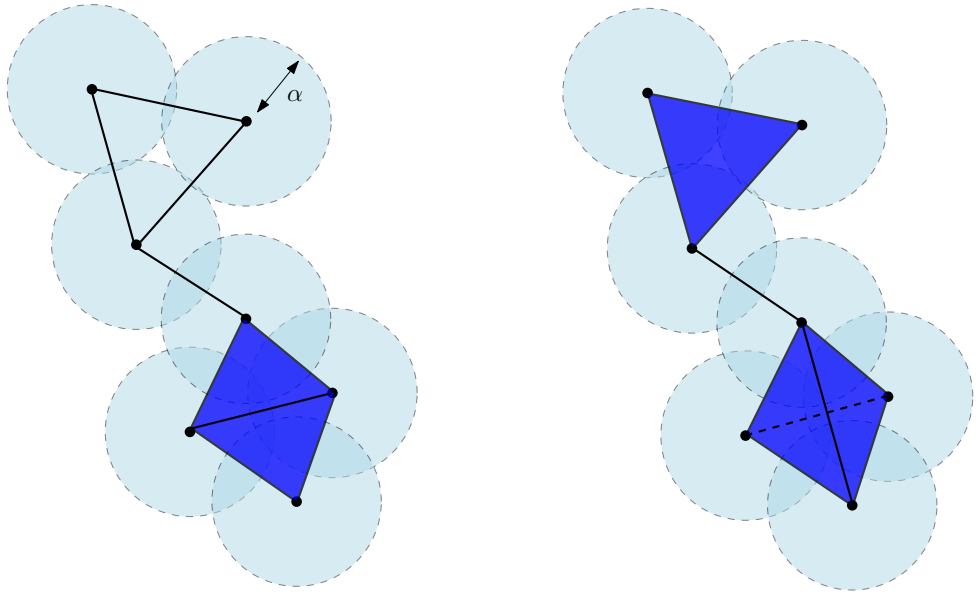


Figure 2: The Čech complex $\mathrm{Cech}_\alpha(\mathbb{X})$ (left) and the and Vietoris-Rips $\mathrm{Rips}_{2\alpha}(\mathbb{X})$ (right) of a finite point cloud in the plane $\mathbb{R}^2$. The bottom part of $\mathrm{Cech}_\alpha(\mathbb{X})$ is the union of two adjacent triangles, while the bottom part of $\mathrm{Rips}_{2\alpha}(\mathbb{X})$ is the tetrahedron spanned by the four vertices and all its faces. The dimension of the Čech complex is 2. The dimension of the Vietoris-Rips complex is 3. Notice that this later is thus not embedded in $\mathbb{R}^2$.

Closely related to the Vietoris-Rips complex is the *Čech complex* $\mathrm{Cech}_\alpha(\mathbb{X})$ that is defined as the set of simplices $[x_0, \ldots, x_k]$ such that the $k+1$ closed balls $B(x_i, \alpha)$ have a non-empty intersection. Notice that these two complexes are related by

$$\mathrm{Rips}_\alpha(\mathbb{X}) \subseteq \mathrm{Cech}_\alpha(\mathbb{X}) \subseteq \mathrm{Rips}_{2\alpha}(\mathbb{X})$$

and that, if $\mathbb{X} \subset \mathbb{R}^d$ then $\mathrm{Cech}_\alpha(\mathbb{X})$ and $\mathrm{Rips}_{2\alpha}(\mathbb{X})$ have the same 1-dimensional skeleton, i.e. the same set of vertices and edges.

The nerve theorem. The Čech complex is a particular case of a family of complexes associated to covers. Given a *cover* $\mathcal{U} = (U_i)_{i \in I}$ of $\mathbb{M}$, i.e. a family of sets U_i such that $\mathbb{M} = \cup_{i \in I} U_i$, the *nerve of* $\mathcal{U}$ is the abstract simplicial complex $C(\mathcal{U})$ whose vertices are the U_i's and such that

$$\sigma = [U_{i_0}, \cdots, U_{i_k}] \in C(\mathcal{U}) \text{ if and only if } \bigcap_{j=0}^{k} U_{i_j} \neq \emptyset.$$

Given a cover of a data set, where each set of the cover can be, for example, a local cluster or a grouping of data points sharing some common properties, its nerve provides a compact and global combinatorial description of the relationship between these sets through their intersection patterns - see Figure 3.

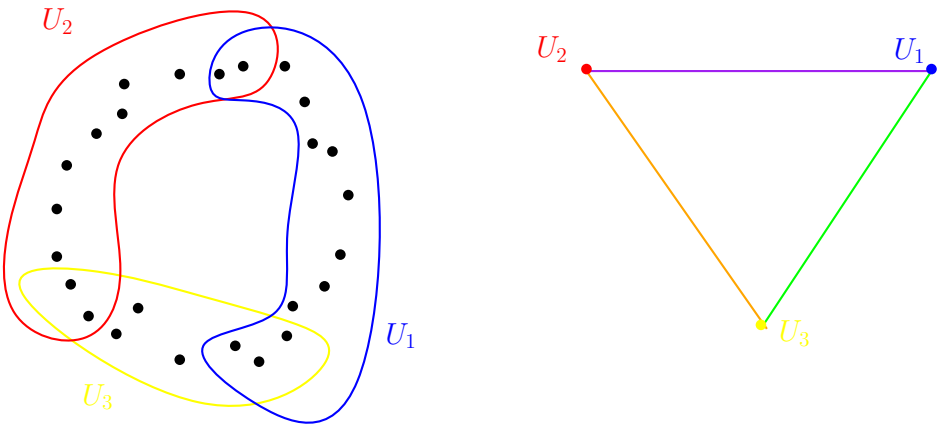


Figure 3: The nerve of a cover of a set of sampled points in the plane.

A fundamental theorem in algebraic topology, relates, under some assumptions, the topology of the nerve of a cover to the topology of the union of the sets of the cover. To be formally stated, this result, known as the Nerve Theorem, requires to introduce a few notions.

Two topological spaces X and Y are usually considered as being the same from a topological point of view if they are *homeomorphic*, i.e. if there exist two continuous bijective maps $f : X \to Y$ and $g : Y \to X$ such that $f \circ g$ and $g \circ f$ are the identity map of Y and X respectively. In many cases, asking X and Y to be homeomorphic turns out to be a too strong requirement to ensure that X and Y share the same topological features of interest for TDA. Two continuous maps $f_0, f_1 : X \to Y$ are said to be *homotopic* is there exists a continuous map $H : X \times [0,1] \to Y$ such that for any $x \in X$, $H(x,0) = f_0(x)$ and $H(x,1) = g(x)$. The spaces X and Y are then said to be *homotopy equivalent* if there exist two maps $f : X \to Y$ and $g : Y \to X$ such that $f \circ g$ and $g \circ f$ are homotopic to the identity map of Y and X respectively. The maps f and g are then called *homotopy equivalent*. The notion of homotopy equivalence is weaker than the notion of homeomorphism: if X and Y are homeomorphic then they are obviously homotopy equivalent, the converse being not true. However, spaces that are homotopy equivalent still share many topological invariant, in particular they have the same homology - see Section 4.
A space is said to be *contractible* if it is homotopy equivalent to a point. Basic examples of

contractible spaces are the balls and, more generally, the convex sets in $\mathbb{R}^d$. Open covers whose all elements and their intersections are contractible have the remarkable following property.

Theorem 1 (Nerve theorem)**.** *Let $\mathcal{U} = (U_i)_{i\in I}$ be a cover of a topological space X by open sets such that the intersection of any subcollection of the U_i's is either empty or contractible. Then, X and the nerve $C(\mathcal{U})$ are homotopy equivalent.*

It is easy to verify that convex subsets of Euclidean spaces are contractible. As a consequence, if $\mathcal{U} = (U_i)_{i\in I}$ is a collection of convex subsets of $\mathbb{R}^d$ then $C(\mathcal{U})$ and $\cup_{i\in I}U_i$ are homotopy equivalent. In particular, if $\mathbb{X}$ is a set of points in $\mathbb{R}^d$, then the Čech complex $\text{Cech}_\alpha(\mathbb{X})$ is homotopy equivalent to the union of balls $\cup_{x\in\mathbb{X}}B(x,\alpha)$.

The Nerve Theorem plays a fundamental role in TDA: it provide a way to encode the topology of continuous spaces into abstract combinatorial structures that are well-suited for the design of effective data structures and algorithms.

3 Using covers and nerves for exploratory data analysis and visualization: the Mapper algorithm

Using the nerve of covers as a way to summarize, visualize and explore data is a natural idea that was first proposed for TDA in Singh et al. (2007), giving rise to the so-called Mapper algorithm.

Definition 2. *Let $f : X \to \mathbb{R}^d$, $d \geq 1$, be a continuous real valued function and let $\mathcal{U} = (U_i)_{i\in I}$ be a cover of $\mathbb{R}^d$. The pull back cover of X induced by $(f,\mathcal{U})$ is the collection of open sets $(f^{-1}(U_i))_{i\in I}$. The refined pull back is the collection of connected components of the open sets $f^{-1}(U_i)$, $i \in I$.*

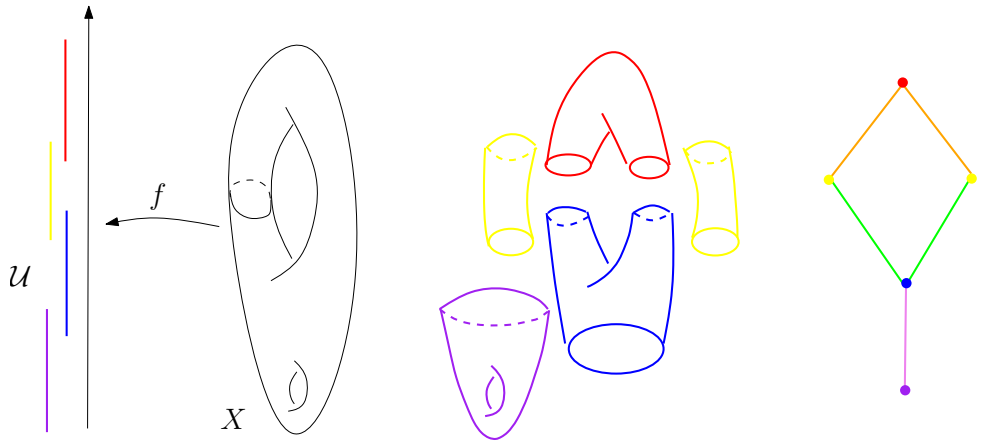


Figure 4: The refined pull back cover of the height function on a surface in $\mathbb{R}^3$ and its nerve

The idea of the Mapper algorithm is, given a data set $\mathbb{X}$ and well-chosen real valued function $f : \mathbb{X} \to \mathbb{R}^d$, to summarize $\mathbb{X}$ through the nerve of the refined pull back of a cover $\mathcal{U}$ of $f(\mathbb{X})$. For well-chosen covers $\mathcal{U}$ (see below), this nerve is a graph providing an easy and convenient way to visualize the summary of the data. It is described in Algorithm 1 and illustrated on a simple example in Figure 5.

The Mapper algorithm is very simple but it raises several questions about the various choices that are left to the user and that we briefly discuss in the following.

The choice of f. The choice of the function f, sometimes called the filter or lens function, strongly depends on the features of the data that one expect to highlight. The following ones are among the ones more or less classically encountered in the literature:

Algorithm 1 The Mapper algorithm

Input: A data set $\mathbb{X}$ with a metric or a dissimilarity measure between data points, a function $f : \mathbb{X} \to \mathbb{R}$ (or $\mathbb{R}^d$), and a cover $\mathcal{U}$ of $f(\mathbb{X})$.
for each $U \in \mathcal{U}$, decompose $f^{-1}(U)$ into clusters $C_{U,1}, \cdots, C_{U,k_U}$.
Compute the nerve of the cover of X defined by the $C_{U,1}, \cdots, C_{U,k_U}$, $U \in \mathcal{U}$
Output: a simplicial complex, the nerve (often a graph for well-chosen covers $\rightarrow$ easy to visualize):
- a vertex $v_{U,i}$ for each cluster $C_{U,i}$,
- an edge between $v_{U,i}$ and $v_{U',j}$ iff $C_{U,i} \cap C_{U',j} \neq \emptyset$

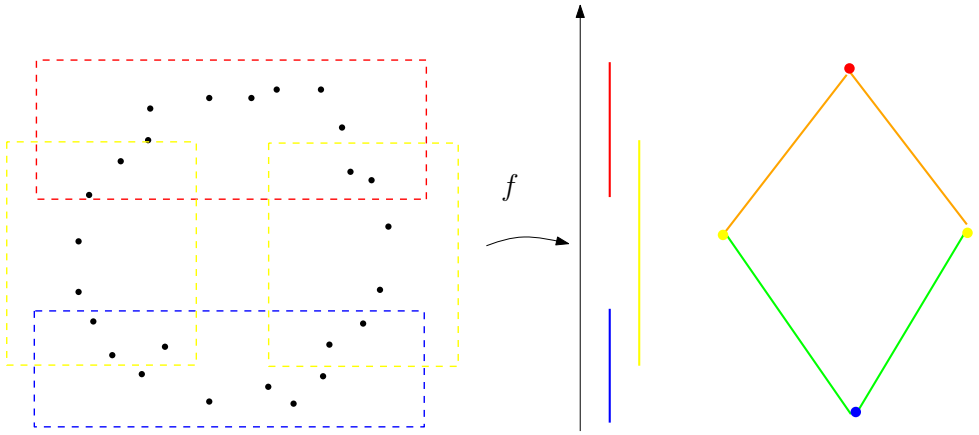


Figure 5: The mapper algorithm on a point cloud sampled around a circle.

- density estimates: the mapper complex may help to understand the structure and connectivity of high density areas (clusters).

- PCA coordinates or coordinates functions obtained from a non linear dimensionality reduction (NLDR) technique, eigenfunctions of graph laplacians,... may help to reveal and understand some ambiguity in the use of non linear dimensionality reductions.

- The centrality function $f(x) = \sum_{y \in \mathbb{X}} d(x, y)$ and the eccentricity function $f(x) = \max_{y \in \mathbb{X}} d(x, y)$, appears sometimes to be good choices that do not require any specific knowledge about the data.

- For data that are sampled around 1-dimensional filamentary structures, the distance function to a given point allows to recover the underlying topology of the filamentary structures Chazal et al. (2015c).

The choice of the cover $\mathcal{U}$. When f is a real valued function, a standard choice is to take $\mathcal{U}$ to be a set of regularly spaced intervals of equal length $r > 0$ covering the set $f(\mathbb{X})$. The real r is sometimes called the *resolution* of the cover and the percentage g of overlap between two consecutive intervals is called the the *gain* of the cover - see Figure 6. Note that if the gain g is chosen below 50%, then every point of the real line is covered by at most 2 open sets of $\mathcal{U}$ and the output nerve is a graph. It is important to notice that the output of the Mapper is very sensitive to the choice of $\mathcal{U}$ and small changes in the resolution and gain parameters may results in very large changes in the output, making the method very instable. A classical strategy consists in exploring some range of parameters and select the ones that turn out to provide the most informative output from the user perspective.

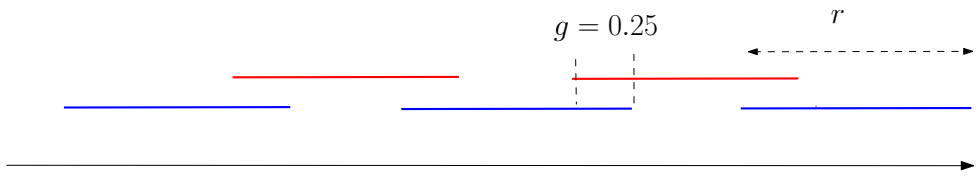


Figure 6: An example of a cover of the real line with resolution r and the gain $g = 25\%$.

The choice of the clusters. The Mapper algorithm requires to cluster the preimage of the open sets $U \in \mathcal{U}$. There are two strategies to compute the clusters. A first strategy consists in applying, for each $U \in \mathcal{U}$, a cluster algorithm, chosen by the user, to the premimage $f^{-1}(U)$. A second, more global, strategy consists in building a neighboring graph on top of the data set $\mathbb{X}$, e.g. k-NN graph or ε-graph, and, for each $U \in \mathcal{U}$, taking the connected components of the subgraph with vertex set $f^{-1}(U)$.

Theoretical and statistical aspects of Mapper. Based on the results on stability and the structure of Mapper proposed in Carrière and Oudot (2015), advances towards a statistically well-founded version of Mapper have been obtained recently in Carrière et al. (2017). Unsurprisingly, the convergence of Mapper depends on both the sampling of the data and the regularity of the filter function. Moreover, subsampling strategies can be proposed to select a complex in a Rips filtration at a convenient scale, as well as the resolution and the gain for defining the Mapper graph. Other approaches have been proposed to study and deal with the instabilities of the Mapper algorithm in Dey et al. (2016, 2017).

Data Analysis with Mapper. As an exploratory data analysis tool, Mapper has been successfully used for clustering and feature selection. The idea is to identify specific structures in the Mapper graph (or complex), in particular loops and flares. These structures are then used to identify interesting clusters or to select features or variable that best discriminate the data in these structures. Applications on real data, illustrating these techniques, may be found, for example, in Lum et al. (2013); Yao et al. (2009).

4 Geometric reconstruction and homology inference

Another way to build covers and use their nerves to exhibit the topological structure of data is to consider union of balls centered on the data points. In this section, we assume that $\mathbb{X}_n = \{x_0, \cdots, x_n\}$ is a subset of $\mathbb{R}^d$ sampled i.i.d. according to a probability measure μ with compact support $M \subset \mathbb{R}^d$. The general strategy to infer topological information about M from μ proceeds in two steps that are discussed in the following of this section:

1. $\mathbb{X}_n$ is covered by a union of balls of fixed radius centered on the x_i's. Under some regularity assumptions on M, one can relate the topology of this union of balls to the one of M;

2. From a practical and algorithmic perspective, topological features of M are inferred from the nerve of the union of balls, using the Nerve Theorem.

In this framework, it is indeed possible to compare spaces through *isotopy equivalence*, a stronger notion than homeomorphism: $X \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}^d$ are said to be *(ambient) isotopic* if there exists a continuous family of homeomorphisms $H : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, H continuous, such that for any $t \in [0,1]$, $H_t = H(t,.) : \mathbb{R}^d \to \mathbb{R}^d$ is an homeomorphism, H_0 is the identity map in $\mathbb{R}^d$ and $H_1(X) = Y$. Obviously, if X and Y are isotopic, then they are homeomorphic. The converse is not true: a knotted and an unknotted circles in $\mathbb{R}^3$ are not homeomorphic (notice that although this claim seems rather intuitive, its formal proof requires the use of some non obvious algebraic topology tools).

Union of balls and distance functions. Given a compact subset K of $\mathbb{R}^d$, and a non negative real number r, the union of balls of radius r centered on K, $K^r = \cup_{x\in K} B(x,r)$, called the r-offset of K, is the r-sublevel set of the distance function $d_K : \mathbb{R}^d \to \mathbb{R}$ defined by $d_K(x) = \inf_{y\in K} \|x-y\|$; in other words, $K^r = d_k^{-1}([0,r])$. This remark allows to use differential properties of distance functions and to compare the topology of the offsets of compact sets that are close to each other with respect to the Hausdorff distance.

Definition 3 (Hausdorff distance)**.** *The Hausdorff distance between two compact subsets K, K' of $\mathbb{R}^d$ is defined by*

$$d_H(K,K') = \|d_K - d_{K'}\|_\infty = \sup_{x\in\mathbb{R}^d} |d_K(x) - d_{K'}(x)|.$$

In our setting, the considered compact sets are the data set $\mathbb{X}_n$ and of the support M of the measure μ. When M is a smooth compact submanifold, under mild conditions on $d_H(\mathbb{X}_n, M)$, for some well-chosen r, the offsets of $\mathbb{X}_n$ are homotopy equivalent to M, Chazal and Lieutier (2008a); Niyogi et al. (2008) - see Figure 7 for an illustration. These results extend to larger classes of compact sets and leads to stronger results on the inference of the isotopy type of the offsets of M, Chazal et al. (2009c,d). They also lead to results on the estimation of other geometric and differential quantities such as normals Chazal et al. (2009c), curvatures Chazal et al. (2008) or boundary measures Chazal et al. (2010) under assumptions on the Haussdorff distance between the underlying shape and the data sample.

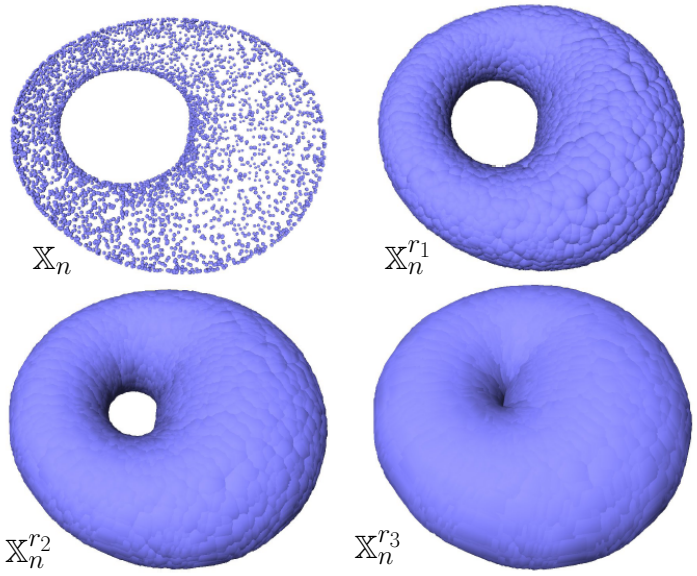


Figure 7: The example of a point cloud $\mathbb{X}_n$ sampled on the surface of a torus in $\mathbb{R}^3$ (top left) and its offsets for different values of radii $r_1 < r_2 < r_3$. For well chosen values of the radius (e.g. r_1 and r_2), the offsets are clearly homotopy equivalent to a torus.

These results rely on the 1-semiconcavity of the squared distance function d_K^2, i.e. the convexity of the function $x \to \|x\|^2 - d_K^2(x)$, and can be naturally stated in the following general framework.

Definition 4. *A function $\phi : \mathbb{R}^d \to \mathbb{R}_+$ is* distance-like *if it is proper (the pre-image of any compact set in $\mathbb{R}$ is a compact set in $\mathbb{R}^d$) and $x \to \|x\|^2 - \phi^2(x)$ is convex.*

Thanks to its semiconcavity, a distance-like function ϕ have a well-defined, but not continuous, gradient $\nabla\phi : \mathbb{R}^d \to \mathbb{R}^d$ that can be integrated into a continuous flow Petrunin (2007) that allows to track the evolution of the topology of its sublevel sets and to compare it to the one of the sublevel sets of close distance-like functions.

Definition 5. *Let ϕ be a distance-like function and let $\phi^r = \phi^{-1}([0,r])$ be the r-sublevel set of ϕ.*

- *A point $x \in \mathbb{R}^d$ is called* α-critical *if $\|\nabla_x \phi\| \leq \alpha$. The corresponding value $r = \phi(x)$ is also said to be* α-critical.
- *The* weak feature size *of ϕ at r is the minimum $r' > 0$ such that ϕ does not have any critical value between r and $r + r'$. We denote it by $\mathrm{wfs}_\phi(r)$. For any $0 < \alpha < 1$, the* α-reach *of ϕ is the maximum r such that $\phi^{-1}((0,r])$ does not contain any α-critical point.*

An important property of a distance-like function ϕ is the topology of their sublevel sets ϕ^r can only change when r crosses a 0-critical value.

Lemma 1 (Isotopy Lemma Grove (1993))**.** *Let ϕ be a distance-like function and $r_1 < r_2$ be two positive numbers such that ϕ has no 0-critical point, i.e. points x such that $\nabla\phi(x) = 0$, in the subset $\phi^{-1}([r_1, r_2])$. Then all the sublevel sets $\phi^{-1}([0,r])$ are isotopic for $r \in [r_1, r_2]$.*

As an immediate consequence of the Isotopy Lemma, all the sublevel sets of ϕ between r and $r + \mathrm{wfs}_\phi(r)$ have the same topology. Now the following reconstruction theorem from Chazal et al. (2011b) provides a connection between the topology of the sublevel sets of close distance-like functions.

Theorem 2 (Reconstruction Theorem)**.** *Let ϕ, ψ be two distance-like functions such that $\|\phi - \psi\|_\infty < \varepsilon$, with $\mathrm{reach}_\alpha(\phi) \geq R$ for some positive ε and α. Then, for every $r \in [4\varepsilon/\alpha^2, R - 3\varepsilon]$ and every $\eta \in (0, R)$, the sublevel sets ψ^r and ϕ^η are homotopy equivalent when*

$$\varepsilon \leq \frac{R}{5 + 4/\alpha^2}.$$

Under similar but slightly more technical conditions the Reconstruction Theorem can be extended to prove that the sublevel sets are indeed homeomorphic and even isotopic Chazal et al. (2009c, 2008).

Coming back to our setting, and taking for $\phi = d_M$ and $\psi = d_{\mathbb{X}_n}$ the distance functions to the support M of the measure μ and to the data set $\mathbb{X}_n$, the condition $\mathrm{reach}_\alpha(d_M) \geq R$ can be interpreted as regularity condition on M[1]. The Reconstruction Theorem combined with the Nerve Theorem tell that, for well-chosen values of r, η, the η-offsets of M are homotopy equivalent to the nerve of the union of balls of radius r centered on $\mathbb{X}_n$, i.e the Cech complex $\mathrm{Cech}_r(\mathbb{X}_n)$.

From a statistical perspective, the main advantage of these results involving Hausdorff distance is that the estimation of the considered topological quantities boil down to support estimation questions that have been widely studied - see Section 4.1.

The above results provide a mathematically well-founded framework to infer the topology of shapes from a simplicial complex built on top of an approximating finite sample. However, from a more practical perspective it raises raise two issues. First, the Reconstruction Theorem requires a regularity assumption through the α-reach condition that may not always be satisfied and, the choice of a radius r for the ball used to build the Čech complex $\mathrm{Cech}_r(\mathbb{X}_n)$. Second, $\mathrm{Cech}_r(\mathbb{X}_n)$ provides a topologically faithfull summary of the data, through a simplicial complex that is usually not well-suited for further data processing. One often needs easier to handle topological descriptors, in particular numerical ones, that can be easily computed from the complex. This second issue is addressed by considering the homology of the considered simplicial complexes in the next paragraph, while the first issue will be addressed in the next section with the introduction of persistent homology.

[1] As an example, if M is a smooth compact submanifold then $\mathrm{reach}_0(\phi)$ is always positive and known as the reach of M Federer (1959).

Homology in a nutshell. Homology is a classical concept in algebraic topology providing a powerful tool to formalize and handle the notion of topological features of a topological space or of a simplicial complex in an algebraic way. For any dimension k, the k-dimensional "holes" are represented by a vector space H_k whose dimension is intuitively the number of such independent features. For example the 0-dimensional homology group H_0 represents the connected components of the complex, the 1-dimensional homology group H_1 represents the 1-dimensional loops, the 2-dimensional homology group H_2 represents the 2-dimensional cavities,...

To avoid technical subtleties and difficulties, we restrict the introduction of homology to the minimum that is necessary to understand its usage in the following of the paper. In particular we restrict to homology with coefficients in $\mathbb{Z}_2$, i.e. the field with two elements 0 and 1 such that $1 + 1 = 0$, that turns out to be geometrically a little bit more intuitive. However, all the notions and results presented in the sequel naturally extend to homology with coefficient in any field. We refer the reader to Hatcher (2001) for a complete and comprehensible introduction to homology and to Ghrist (2017) for a recent concise and very good introduction to applied algebraic topology and its connections to data analysis.

Let K be a (finite) simplicial complex and let k be a non negative integer. The *space of k-chains* on K, $C_k(K)$ is the set whose elements are the formal (finite) sums of k-simplices of K. More precisely, if $\{\sigma_1, \cdots \sigma_p\}$ is the set of k-simplices of K, then any k-chain can be written as

$$c = \sum_{i=1}^{p} \varepsilon_i \sigma_i \quad \text{with} \quad \varepsilon_i \in \mathbb{Z}_2.$$

If $c' = \sum_{i=1}^{p} \varepsilon'_i \sigma_i$ is another k-chain and $\lambda \in \mathbb{Z}_2$, the sum $c + c'$ is defined as $c + c' = \sum_{i=1}^{p} (\varepsilon_i + \varepsilon'_i)\sigma_i$ and the product $\lambda.c$ is defined as $\lambda.c = \sum_{i=1}^{p} (\lambda.\varepsilon_i)\sigma_i$, making $C_k(K)$ a vector space with coefficients in $\mathbb{Z}_2$. Since we are considering coefficient in $\mathbb{Z}_2$, geometrically a k-chain can be seen as a finite collection of k-simplices and the sum of two k-chains as the symmetric difference of the two corresponding collections[2].

The *boundary* of a k-simplex $\sigma = [v_0, \cdots, v_k]$ is the $(k-1)$-chain

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \cdots, \hat{v}_i, \cdots, v_k]$$

where $[v_0, \cdots, \hat{v}_i, \cdots, v_k]$ is the $(k-1)$-simplex spanned by all the vertices except v_i [3]. As the k-simplices form a basis of $C_k(K)$, ∂_k extends as a linear map from $C_k(K)$ to $C_{k-1}(K)$ called the *boundary operator*. The kernel $Z_k(K) = \{c \in C_k(K) : \partial_k = 0\}$ of ∂_k is called the *space of k-cycles of K* and the image $B_k(K) = \{c \in C_k(K) : \exists c' \in C_{k+1}(K), \partial_{k+1}(c') = c\}$ of ∂_{k+1} is called the *space of k-boundaries* of K. The boundary operators satisfy the fundamental following property:

$$\partial_{k-1} \circ \partial_k \equiv 0 \quad \text{for any} \quad k \geq 1.$$

In other words, any k-boundary is a k-cycle, i.e. $B_k(K) \subseteq Z_k(K) \subseteq C_k(K)$. These notions are illustrated on Figure 8.

Definition 6 (Simplicial homology group and Betti numbers)**.** *The k^{th} (simplicial) homology group of K is the quotient vector space*

$$H_k(K) = Z_k(K)/B_k(K).$$

The k^{th} Betti number of K is the dimension $\beta_k(K) = \dim H_k(K)$ of the vector space $H_k(K)$.

[2]Recall that the symmetric difference of two sets A and B is the set $A \Delta B = (A \setminus B) \cup (B \setminus A)$.

[3]Notice that as we are considering coefficients in $\mathbb{Z}_2$, here $-1 = 1$ and thus $(-1)^i = 1$ for any i.

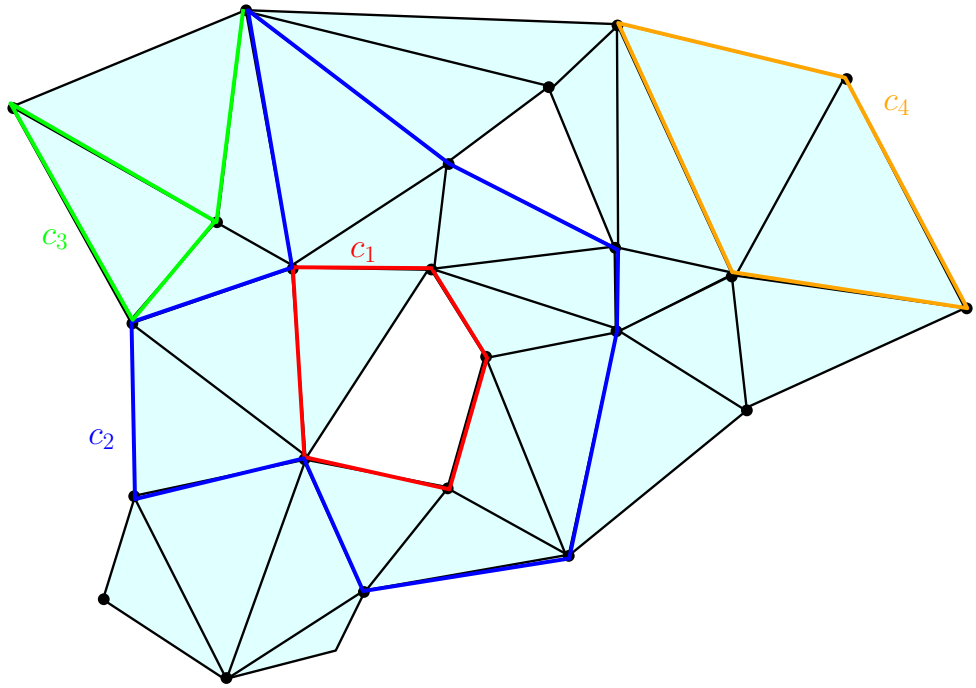


Figure 8: Some examples of chains, cycles and boundaries on a 2-dimensional complex K: c_1, c_2 and c_4 are 1-cycles; c_3 si a 1-chain but not a 1-cycle; c_4 is the 1-boundary, namely the boundary of the 2-chain obtained as the sum of the two triangles surrounded by c_4; The cycles c_1 and c_2 span the same element in $H_1(K)$ as their difference is the 2-chain represented by the union of the triangles surrounded by the union of c_1 and c_2.

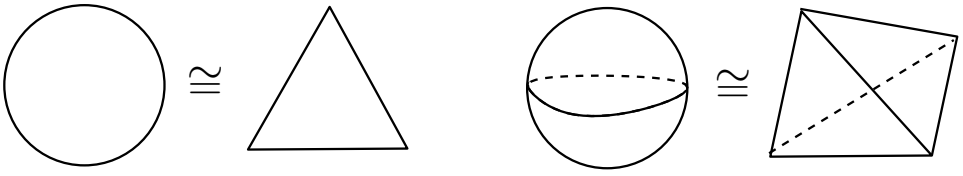
Two cycles $c, c' \in Z_k(K)$ are said to be *homologous* if they differ by a boundary, i.e. is there exists a $(k+1)$-chain d such that $c' = c + \partial_{k+1}(d)$. Two such cycles give rise to the same element of H_k. In other words, the elements of $H_k(K)$ are the equivalence classes of homologous cycles.

Simplicial homology groups and Betti numbers are topological invariants: if K, K' are two simplicial complexes whose geometric realizations are homotopy equivalent, then their homology groups are isomorphic and their Betti numbers are the same.

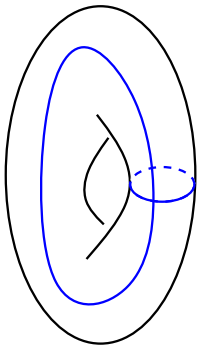
Singular homology is another notion of homology that allows to consider larger classes of topological spaces. It is defined for any topological space X similarly to simplicial homology except that the notion of simplex is replaced by the notion of *singular simplex* which is just any continuous map $\sigma : \Delta_k \to X$ where Δ_k is the standard k-dimensional simplex. The space of k-chains is the vector space spanned by the k-dimensional singular simplices and the boundary of a simplex σ is defined as the (alternated) sum of the restriction of σ to the $(k-1)$-dimensional faces of Δ_k. A remarkable fact about singular homology it that it coincides with simplicial homology whenever X is homeomorphic to the geometric realization of a simplicial complex. This allows us, in the sequel of this paper, to indifferently talk about simplicial or singular homology for topological spaces and simplicial complexes.

Observing, that if $f : X \to Y$ is a continuous map, then for any singular simplex $\sigma : \Delta_k \to X$ in X, $f \circ \sigma : \Delta_k \to Y$ is a singular simplex in Y, one easily deduces that continuous maps between topological spaces canonically induce homomorphisms between their homology groups. In particular, if f is an homeomorphism or an homotopy equivalence, then it induces an isomorphism between $H_k(X)$ and $H_k(Y)$ for any non negative integer k. As an example, it follows from the Nerve Theorem that for any set of points $X \subset \mathbb{R}^d$ and any $r > 0$ the r-offset X^r and the Čech complex $\mathrm{Cech}_r(X)$ have isomorphic homology groups and the same Betti numbers.

As a consequence, the Reconstruction Theorem 2 leads to the following result on the estimation of Betti numbers.



$\beta_0 = 1,\ \beta_1 = 1,\ \beta_2 = 0$ $\beta_0 = 1,\ \beta_1 = 0,\ \beta_2 = 1,\ \beta_3 = 0$



Two dimensional torus: $\beta_0 = 1$, $\beta_1 = 2,\ \beta_2 = 1,\ \beta_3 = 0$

Figure 9: The Betti numbers of the circle (top left), the 2-dimensional sphere (top right) and the 2-dimensional torus (bottom). The blue curves on the torus represent two independent cycles whose homology class is a basis of its 1-dimensional homology group.

Theorem 3. *Let $M \subset \mathbb{R}^d$ be a compact set such that $\mathrm{reach}_\alpha(d_M) \geq R > 0$ for some $\alpha \in (0,1)$ and let $\mathbb{X}$ be a finite set of point such that $d_H(M,\mathbb{X}) = \varepsilon < \frac{R}{5+4/\alpha^2}$. Then, for every $r \in [4\varepsilon/\alpha^2, R-3\varepsilon]$ and every $\eta \in (0,R)$, the Betti numbers of $\mathrm{Cech}_r(\mathbb{X})$ are the same as the ones of M^η.*
In particular, if M is a smooth m-dimensional submanifold of $\mathbb{R}^d$, then $\beta_k(\mathrm{Cech}_r(\mathbb{X})) = \beta_k(M)$ for any $k = 0, \cdots, m$.

From a practical perspective, this result raises three difficulties: first, the regularity assumption involving the α-reach of M may be too restrictive; second, the computation of the nerve of an union of balls requires they use of a tricky predicate testing the emptiness of a finite union of balls; third the estimation of the Betti numbers relies on the scale parameter r whose choice may be a problem.

To overcome these issues, Chazal and Oudot (2008) establishes the following result that offers a solution to the two first problems.

Theorem 4. *Let $\mathbb{M} \subset \mathbb{R}^d$ be a compact set such that $\mathrm{wfs}(M) = \mathrm{wfs}_{d_M}(0) \geq R > 0$ and let $\mathbb{X}$ be a finite set of point such that $d_H(M,\mathbb{X}) = \varepsilon < \frac{1}{9}\mathrm{wfs}(M)$. Then for any $r \in [2\varepsilon, \frac{1}{4}(\mathrm{wfs}(M) - \varepsilon)]$ and any $\eta \in (0,R)$,*

$$\beta_k(X^\eta) = \mathrm{rk}\left(H_k(\mathrm{Rips}_r(\mathbb{X})) \to H_k(\mathrm{Rips}_{4r}(\mathbb{X}))\right)$$

where $rk(H_k(\mathrm{Rips}_r(\mathbb{X})) \to H_k(\mathrm{Rips}_{4r}(\mathbb{X})))$ denotes the rank of the homomorphism induced by the (continuous) canonical inclusion $\mathrm{Rips}_r(\mathbb{X}) \hookrightarrow \mathrm{Rips}_{4r}(\mathbb{X})$.

Although this result leaves the question of the choice of the scale parameter r open, it is proven in Chazal and Oudot (2008) that a multiscale strategy whose description is beyond the scope of this paper provides some help to identify the relevant scales at which Theorem 4 can be applied.

4.1 Statistical aspects of Homology inference

According to the stability results presented in the previous section, a statistical approach to topological inference is strongly related to the problem of *distribution support estimation* and

level sets estimation under the Hausdorff metric. A large number of methods and results are available for estimating the support of a distribution in statistics. For instance, the Devroye and Wise estimator (Devroye and Wise, 1980) defined on a sample $\mathbb{X}_n$ is also a particular offset of $\mathbb{X}_n$. The convergence rates of both $\mathbb{X}_n$ and the Devroye and Wise estimator to the support of the distribution for the Hausdorff distance is studied in Cuevas and Rodríguez-Casal (2004) in $\mathbb{R}^d$. More recently, the minimax rates of convergence of manifold estimation for the Hausdorff metric, which is particularly relevant for topological inference, has been studied in Genovese et al. (2012). There is also a large literature about level sets estimation in various metrics (see for instance Cadre, 2006; Polonik, 1995; Tsybakov et al., 1997) and more particularly for the Hausdorff metric in Chen et al. (2015). All these works about support and level sets estimation shine light on the statistical analysis of topological inference procedures.

In the paper Niyogi et al. (2008), it is shown that the homotopy type of Riemannian manifolds with reach larger than a given constant can be recovered with high probability from offsets of a sample on (or close to) the manifold. This paper was probably the first attempt to consider the topological inference problem in terms of probability. The result of Niyogi et al. (2008) is derived from a retract contraction argument and on tight bounds over the packing number of the manifold in order to control the Hausdorff distance between the manifold and the observed point cloud. The homology inference in the noisy case, in the sense the distribution of the observation is concentrated around the manifold, is also studied in Niyogi et al. (2008, 2011). The assumption that the geometric object is a smooth Riemannian manifold is only used in the paper to control in probability the Hausdorff distance between the sample and the manifold, and is not actually necessary for the "topological part" of the result. Regarding the topological results, these are similar to those of Chazal et al. (2009d); Chazal and Lieutier (2008b) in the particular framework of Riemannian manifolds. Starting from the result of Niyogi et al. (2008), the minimax rates of convergence of the homology type have been studied by Balakrishna et al. (2012) under various models, for Riemannian manifolds with reach larger than a constant. In contrast, a statistical version of Chazal et al. (2009d) has not yet been proposed.

More recently, following the ideas of Niyogi et al. (2008), Bobrowski et al. (2014) have proposed a robust homology estimator for the level sets of both density and regression functions, by considering the inclusion map between nested pairs of estimated level sets (in the spirit of Theorem 4 above) obtained with a plug-in approach from a kernel estimators.

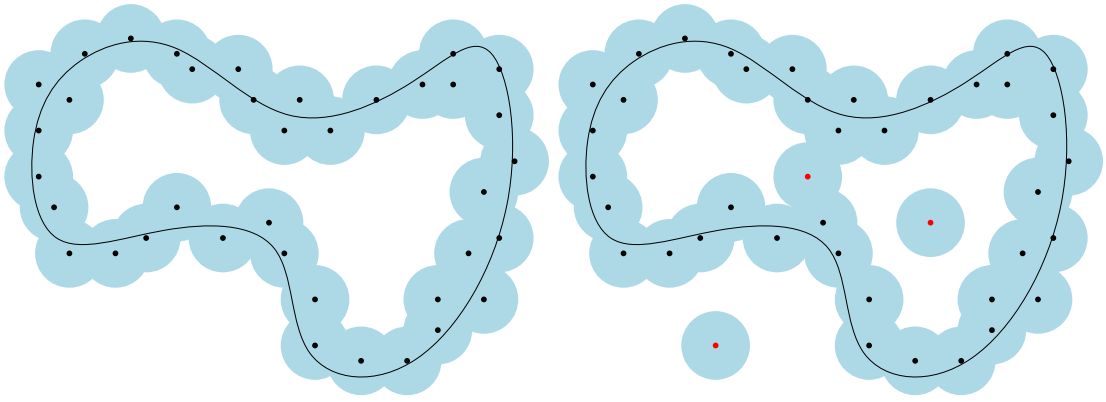


Figure 10: The effect of outliers on the sublevel sets of distance functions. Adding just a few outliers to a point cloud may dramatically change its distance function and the topology of its offsets.

4.2 Going beyond Hausdorff distance : distance to measure

It is well known that distance-based methods in TDA may fail completely in the presence of outliers. Indeed, adding even a single outlier to the point cloud can change the distance function

dramatically, see Figure 10 for an illustration. To answer this drawback, Chazal et al. (2011b) have introduced an alternative distance function which is robust to noise, the *distance-to-a-measure*.

Given a probability distribution P in $\mathbb{R}^d$ and a real parameter $0 \leq u \leq 1$, the notion of distance to the support of P may be generalized as the function

$$\delta_{P,u} : x \in \mathbb{R}^d \mapsto \inf\{t > 0\,;\, P(B(x,t)) \geq u\},$$

where $B(x,t)$ is the closed Euclidean ball of center x and radius t. To avoid issues due to discontinuities of the map $P \to \delta_{P,u}$, the distance-to-measure function (DTM) with parameter $m \in [0,1]$ and power $r \geq 1$ is defined by

$$d_{P,m,r}(x) : x \in \mathbb{R}^d \mapsto \left(\frac{1}{m}\int_0^m \delta_{P,u}^r(x)du\right)^{1/r}. \tag{1}$$

A nice property of the DTM proved in Chazal et al. (2011b) is its stability with respect to perturbations of P in the Wasserstein metric. More precisely, the map $P \to d_{P,m,r}$ is $m^{-\frac{1}{r}}$-Lipschitz, i.e. if P and $\tilde{P}$ are two probability distributions on $\mathbb{R}^d$, then

$$\|d_{P,m,r} - d_{\tilde{P},m,r}\|_\infty \leq m^{-\frac{1}{r}} W_r(P,\tilde{P}) \tag{2}$$

where W_r is the Wasserstein distance for the Euclidean metric on $\mathbb{R}^d$, with exponent r[4]. This property implies that the DTM associated to close distributions in the Wasserstein metric have close sublevel sets. Moreover, when $r = 2$, the function $d^2_{P,m,2}$ is semiconcave ensuring strong regularity properties on the geometry of its sublevel sets. Using these properties, Chazal et al. (2011b) show that, under general assumptions, if $\tilde{P}$ is a probability distribution approximating P, then the sublevel sets of $d_{\tilde{P},m,2}$ provide a topologically correct approximation of the support of P.

In practice, the measure P is usually only known through a finite set of observations $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ sampled from P, raising the question of the approximation of the DTM. A natural idea to estimate the DTM from $\mathbb{X}_n$ is to plug the empirical measure P_n instead of P in the definition of the DTM. This "plug-in strategy" corresponds to computing the distance to the empirical measure (DTEM). For $m = \frac{k}{n}$, the DTEM satisfies

$$d^r_{P_n,k/n,r}(x) := \frac{1}{k}\sum_{j=1}^{k} \|x - \mathbb{X}_n\|^r_{(j)}\,,$$

where $\|x - \mathbb{X}_n\|_{(j)}$ denotes the distance between x and its j-th neighbor in $\{X_1, \ldots, X_n\}$. This quantity can be easily computed in practice since it only requires the distances between x and the sample points. The convergence of the DTEM to the DTM has been studied in Chazal et al. (2014a) and Chazal et al. (2016b).

The introduction of DTM has motivated further works and applications in various directions such as topological data analysis (Buchet et al., 2015a), GPS traces analysis (Chazal et al., 2011a), density estimation (Biau et al., 2011), hypothesis testing Brécheteau (2017), clustering (Chazal et al., 2013) just to name a few. Approximations, generalizations and variants of the DTM have also been considered in (Buchet et al., 2015b; Guibas et al., 2013; Phillips et al., 2014).

5 Persistent homology

Persistent homology is a powerful tool to compute, study and encode efficiently multiscale topological features of nested families of simplicial complexes and topological spaces. It does not

[4]See Villani (2003) for a definition of the Wasserstein distance

only provide efficient algorithms to compute the Betti numbers of each complex in the considered families, as required for homology inference in the previous section, but also encodes the evolution of the homology groups of the nested complexes across the scales.

5.1 Filtrations

A *filtration of a simplicial complex* K is a nested family of subcomplexes $(K_r)_{r\in T}$, where $T \subseteq \mathbb{R}$, such that for any $r, r' \in T$, if $r \leq r'$ then $K_r \subseteq K_{r'}$, and $K = \cup_{r\in T} K_r$. The subset T may be either finite or infinite. More generally, a *filtration* of a topological space $\mathbb{M}$ is a nested family of subspaces $(M_r)_{r\in T}$, where $T \subseteq \mathbb{R}$, such that for any $r, r' \in T$, if $r \leq r'$ then $M_r \subseteq M_{r'}$ and, $M = \cup_{r\in T} M_r$. For example, if $f : \mathbb{M} \to \mathbb{R}$ is a function, then the family $M_r = f^{-1}((-\infty, r])$, $r \in \mathbb{R}$ defines a filtration called the sublevel set filtration of f.

In practical situations, the parameter $r \in T$ can often be interpreted as a scale parameter and filtrations classically used in TDA often belong to one of the two following families.

Filtrations built on top of data. Given a subset $\mathbb{X}$ of a compact metric space (M, ρ), the families of Rips-Vietoris complexes $(\mathrm{Rips}_r(\mathbb{X}))_{r\in\mathbb{R}}$ and and Čech complexes $(\mathrm{Cech}_r(\mathbb{X}))_{r\in\mathbb{R}}$ are filtrations [5]. Here, the parameter r can be interpreted as a resolution at which one considers the data set $\mathbb{X}$. For example, if $\mathbb{X}$ is a point cloud in $\mathbb{R}^d$, thanks to the Nerve theorem, the filtration $(\mathrm{Cech}_r(\mathbb{X}))_{r\in\mathbb{R}}$ encodes the topology of the whole family of unions of balls $\mathbb{X}^r = \cup_{x\in\mathbb{X}} B(x, r)$, as r goes from 0 to $+\infty$. As the notion of filtration is quite flexible, many other filtrations have been considered in the literature and can be constructed on top of data, such as e.g. the so called witness complex popularized in TDA by De Silva and Carlsson (2004).

Sublevel sets filtrations. Functions defined on the vertices of a simplicial complex give rise to another important example of filtration: let K be a simplicial complex with vertex set V and $f : V \to \mathbb{R}$. Then f can be extended to all simplices of K by $f([v_0, \cdots, v_k]) = \max\{f(v_i) : i = 1, \cdots, k\}$ for any simplex $\sigma = [v_0, \cdots, v_k] \in K$ and the family of subcomplexes $K_r = \{\sigma \in K : f(\sigma) \leq r\}$ defines a filtration call the sublevel set filtration of f. Similarly, one can define the upperlevel set filtration of f.

In practice, even if the index set is infinite, all the considered filtrations are built on finite sets and are indeed finite. For example, when $\mathbb{X}$ is finite, the Vietoris-Rips complex $\mathrm{Rips}_r(\mathbb{X})$ changes only at a finite number of indices r. This allows to easily handle them from an algorithmic perspective.

5.2 Starting with a few examples

Given a filtration $\mathrm{Filt} = (F_r)_{r\in T}$ of a simplicial complex or a topological space, the homology of F_r changes as r increases: new connected components can appear, existing component can merge, loops and cavities can appear or be filled, etc... Persistent homology tracks these changes, identifies the appearing features and associates a life time to them. The resulting information is encoded as a set of intervals called a *barcode* or, equivalently, as a multiset of points in $\mathbb{R}^2$ where the coordinate of each point is the starting and end point of the corresponding interval.

Before giving formal definitions, we introduce and illustrate persistent homology on a few simple examples.

Example 1. Let $f : [0, 1] \to \mathbb{R}$ be the function of Figure 11 and let $(F_r = f^{-1}((-\infty, r]))_{r\in\mathbb{R}}$ be the sublevel set filtration of f. All the sublevel sets of f are either empty or a union of interval, so the only non trivial topological information they carry is their 0-dimensional homology, i.e. their number of connected components. For $r < a_1$, F_r is empty, but at $r = a_1$ a first connected

[5] we take here the convention that for $r < 0$, $\mathrm{Rips}_r(\mathbb{X}) = \mathrm{Cech}_r(\mathbb{X}) = \emptyset$

component appears in F_{a_1}. Persistent homology thus registers a_1 as the birth time of a connected component and start to keep track of it by creating an interval starting at a_1. Then, F_r remains connected until r reaches the value a_2 where a second connected component appears. Persistent homology starts to keep track of this new connected component by creating a second interval starting at a_2. Similarly, when r reaches a_3, a new connected component appears and persistent homology creates a new interval starting at a_3. When r reaches a_4, the two connected components created at a_1 and a_3 merges together to give a single larger component. At this step, persistent homology follows the rule that this is the most recently appeared component in the filtration that dies: the interval started at a_3 is thus ended at a_4 and a first persistence interval encoding the lifespan of the component born at a_3 is created. When r reaches a_5, as in the previous case, the component born at a_2 dies and the persistent interval (a_2, a_5) is created. The interval created at a_1 remains until the end of the filtration giving rise to the persistent interval (a_1, a_6) if the filtration is stopped at a_6, or $(a_1, +\infty)$ if r goes to $+\infty$ (notice that in this later case, the filtration remains constant for $r > a_6$). The obtained set of intervals encoding the span life of the different homological features encountered along the filtration is called the *persistence barcode* of f. Each interval (a, a') can be represented by the point of coordinates (a, a') in $\mathbb{R}^2$ plane. The resulting set of points is called the *persistence diagram* of f. Notice that a function may have several copies of the same interval in its persistence barcode. As a consequence, the persistence diagram of f is indeed a multi-set where each point has an integer valued multiplicity. Last, for technical reasons that will become clear in the next section, one adds to the persistence all the points of the diagonal $\Delta = \{(b, d) : b = d\}$ with an infinite multiplicity.

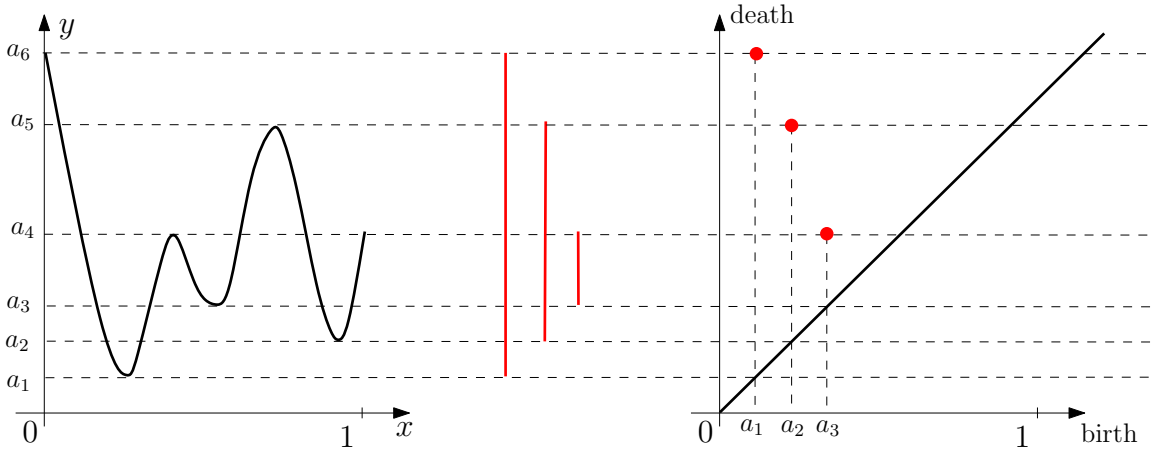


Figure 11: The persistence barcode and the persistence diagram of a function $f : [0, 1] \to \mathbb{R}$.

Example 2. Let now $f : M \to \mathbb{R}$ be the function of Figure 12 where M is a 2-dimensional surface homeomorphic to a torus, and let $(F_r = f^{-1}((-\infty, r]))_{r\in\mathbb{R}}$ be the sublevel set filtration of f. The 0-dimensional persistent homology is computed as in the previous example, giving rise to the red bars in the barcode. Now, the sublevel sets also carry 1-dimensional homological features. When r goes through the height a_1, the sublevel sets F_r that were homeomorphic to two discs become homeomorphic to the disjoint union of a disc and an annulus, creating a first cycle homologous to σ_1 on Figure 12. A interval (in blue) representing the birth of this new 1-cycle is thus started at a_1. Similarly, when r goes through the height a_2 a second cycle, homologous to σ_2 is created, giving rise to the start of a new persistent interval. These two created cycles are never filled (indeed they span $H_1(M)$) and the corresponding intervals remains until the end of the filtration. When r reaches a_3, a new cycle is created that is filled and thus dies at a_4, giving rise to the persistence interval (a_3, a_4). So, now, the sublevel set filtration of f gives rise to two barcodes, one for 0-dimensional homology (in red) and one for 1-dimensional homology (in blue). As previously, these two barcodes can equivalently be represented as diagrams in the

plane.

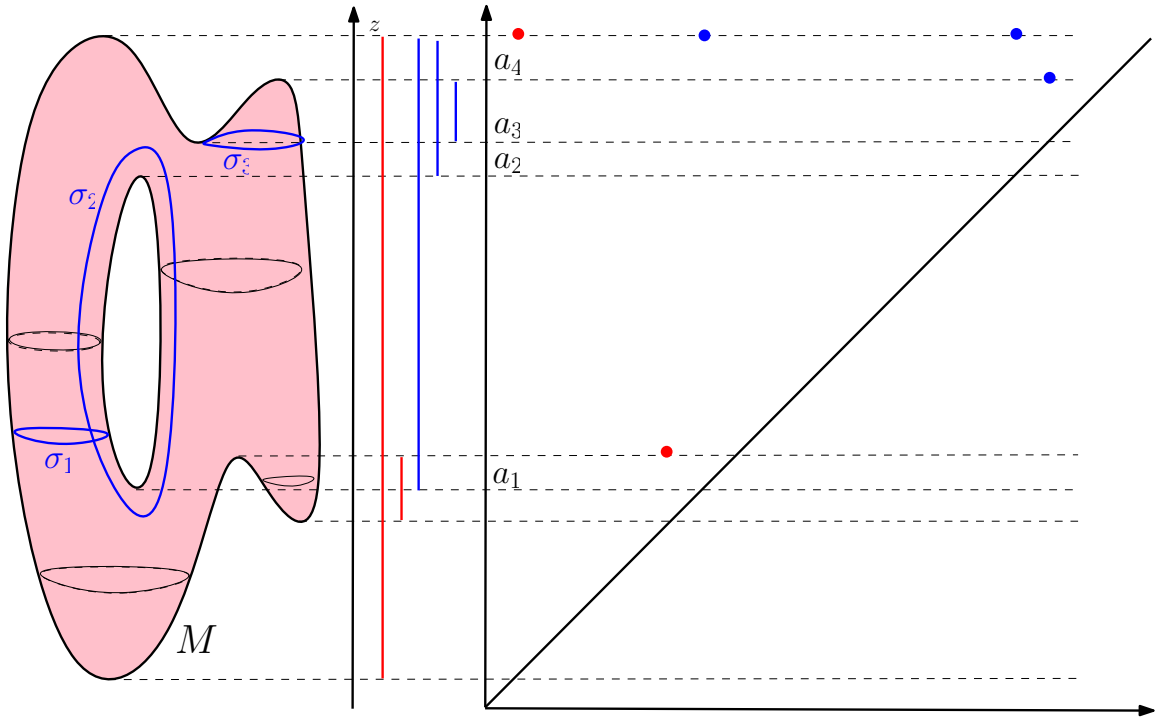


Figure 12: The persistence barcode and the persistence diagram of the height function (projection on the z-axis) defined on a surface in $\mathbb{R}^3$.

Example 3. In this last example we consider the filtration given by a union of growing balls centered on the finite set of points C in Figure 13. Notice that this is the sublevel set filtration of the distance function to C, and thanks to the Nerve Theorem, this filtration is homotopy equivalent to the Čech filtration built on top of C. Figure 13 shows several level sets of the filtration:
a) For the radius $r = 0$, the union of balls is reduced to the initial finite set of point, each of them corresponding to a 0-dimensional feature, i.e. a connected component; an interval is created for the *birth* for each of these features at $r = 0$.
b) Some of the balls started to overlap resulting in the *death* of some connected components that get merged together; the persistence diagram keeps track of these deaths, putting an end point to the corresponding intervals as they disappear.
c) New components have merged giving rise to a single connected component and, so, all the intervals associated to a 0-dimensional feature have been ended, except the one corresponding to the remaining components; two new 1-dimensional features, have appeared resulting in two new intervals (in blue) starting at their birth scale.
d) One of the two 1-dimensional cycles has been filled, resulting in its death in the filtration and the end of the corresponding blue interval.
e) all the 1-dimensional features have died, it only remains the long (and never dying) red interval. As in the previous examples, the final barcode can also be equivalently represented as a persistence diagram where every interval (a, b) is represented by the the point of coordinate (a, b) in $\mathbb{R}^2$. Intuitively the longer is an interval in the barcode or, equivalently the farther from the diagonal is the corresponding point in the diagram, the more persistent, and thus relevant, is the corresponding homological feature across the filtration. Notice also that for a given radius r, the k-th Betti number of the corresponding union of balls is equal of the number of persistence intervals corresponding to k-dimensional homological features and containing r. So,

the persistence diagram can be seen as a multiscale topological signature encoding the homology of the union of balls for all radii as well as its evolution across the values of r.

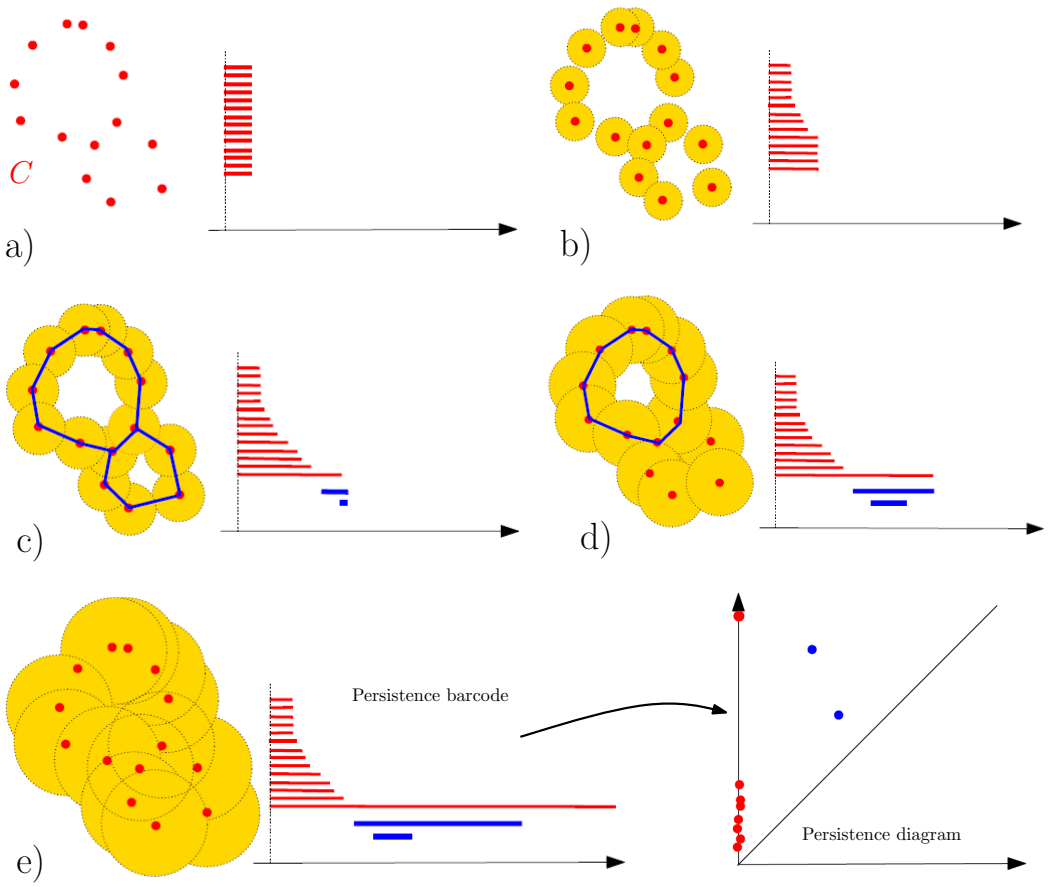


Figure 13: The sublevel set filtration of the distance function to a point cloud and the "construction" of its persistence barcode as the radius of balls increases.

5.3 Persistent modules and persistence diagrams

Persistent diagrams can be formally and rigorously defined in a purely algebraic way. This requires some care and we only give here the basic necessary notions, leaving aside technical subtelties and difficulties. We refer the readers interested in a detailed exposition to Chazal et al. (2016a).

Let $\text{Filt} = (F_r)_{r\in T}$ be a filtration of a simplicial complex or a topological space. Given a non negative integer k and considering the homology groups $H_k(F_r)$ we obtain a sequence of vector spaces where the inclusions $F_r \subset F_{r'}$, $r \leq r'$ induce linear maps between $H_k(F_r)$ and $H_k(F_{r'})$. Such a sequence of vector spaces together with the linear maps connecting them is called a *persistence module*.

Definition 7. *A persistence module $\mathbb{V}$ over a subset T of the real numbers $\mathbb{R}$ is an indexed family of vector spaces $(V_r \mid r \in T)$ and a doubly-indexed family of linear maps $(v_s^r : V_r \to V_s \mid r \leq s)$ which satisfy the composition law $v_t^s \circ v_s^r = v_t^r$ whenever $r \leq s \leq t$, and where v_r^r is the identity map on V_r.*

In many cases, a persistence module can be decomposed into a direct sum of *intervals* modules $\mathbb{I}_{(b,d)}$ of the form

$$\cdots \to 0 \to \cdots \to 0 \to \mathbb{Z}_2 \to \cdots \to \mathbb{Z}_2 \to 0 \to \cdots$$

where the maps $\mathbb{Z}_2 \to \mathbb{Z}_2$ are identity maps while all the other maps are 0. Denoting b (resp. d) the infimum (resp. supremum) of the interval of indices corresponding to non zero vector spaces, such a module can be interpreted as a feature that appears in the filtration at index b and disappear at index d. When a persistence module $\mathbb{V}$ can be decomposed as a direct sum of interval modules, one can show that this decomposition is unique up to reordering the intervals (see (Chazal et al., 2016a, Theorem 2.7)). As a consequence, the set of resulting intervals is independent of the decomposition of $\mathbb{V}$ and is called the *persistence barcode* of $\mathbb{V}$. As in the examples of the previous section, each interval (b, d) in the barcode can be represented as the point of coordinates (b, d) in the plane $\mathbb{R}^2$. The disjoint union of these points, together with the diagonale $\Delta = \{x = y\}$ is a multi-set called the the *persistence diagram* of $\mathbb{V}$.

The following result, from (Chazal et al., 2016a, Theorem 2.8), give some necessary conditions for a persistence module to be decomposable as a direct sum of interval modules.

Theorem 5. *Let $\mathbb{V}$ be a persistence module indexed by $T \subset \mathbb{R}$. If T is a finite set or if all the vector spaces V_r are finite-dimensional, then $\mathbb{V}$ is decomposable as a direct sum of interval modules.*

As both conditions above are satisfied for the persistent homology of filtrations of finite simplicial complexes, an immediate consequence of this result is that the persistence diagrams of such filtrations are always well-defined.

Indeed, it is possible to show that persistence diagrams can be defined as soon as the following simple condition is satisfied.

Definition 8. *A persistence module $\mathbb{V}$ indexed by $T \subset \mathbb{R}$ is* q-tame *if for any $r < s$ in T, the rank of the linear map $v_s^r : V_r \to V_s$ is finite.*

Theorem 6 (Chazal et al. (2009a, 2016a))**.** *If $\mathbb{V}$ is a q-tame persistence module, then it has a well-defined persistence diagram. Such a persistence diagram* $\mathrm{dgm}(\mathbb{V})$ *is the union of the points of the diagonal Δ of $\mathbb{R}^2$, counted with infinite multiplicity, and a multi-set above the diagonal in $\mathbb{R}^2$ that is locally finite. Here, by locally finite we mean that for any rectangle R with sides parallel to the coordinate axes that does not intersect Δ, the number of points of* $\mathrm{dgm}(\mathbb{V})$*, counted with multiplicity, contained in R is finite.*

The construction of persistence diagrams of q-tame modules is beyond the scope of this paper but it gives rise to the same notion as in the case of decomposable modules. It can be done either by following the algebraic approach based upon the decomposability properties of modules, or by adopting a measure theoretic approach that allows to define diagrams as integer valued measures on a space of rectangles in the plane. We refer the reader to Chazal et al. (2016a) for more informations. Although persistence modules encountered in practice are decomposable, the general framework of q-tame persistence module plays a fundamental role in the mathematical and statistical analysis of persistent homology. In particular, it is needed to ensure the existence of limit diagrams when convergence properties are studied - see Section 5.7.

A filtration $\mathrm{Filt} = (F_r)_{r \in T}$ of a simplicial complex or of a topological space is said to be tame if for any integer k, the persistence module $(H_k(F_r) \mid r \in T)$ is q-tame. Notice that the filtrations of finite simplicial complexes are always tame. As a consequence, for any integer k a persistence diagram denoted $\mathrm{dgm}_k(\mathrm{Filt})$ is associated to the filtration Filt. When k is not explicitly specified and when there is no ambiguity, it is usual to drop the index k in the notation and to talk about "the" persistence diagram dgm(Filt) of the filtration Filt. This notation has to be understood as "$\mathrm{dgm}_k(\mathrm{Filt})$ for some k".

5.4 Persistence landscapes

The persistence landscape has been introduced in Bubenik (2015) as an alternative representation of persistence diagrams. This approach aims at representing the topological information encoded

in persistence diagrams as elements of an Hilbert space, for which statistical learning methods can be directly applied.

The persistence landscape is a collection of continuous, piecewise linear functions $\lambda\colon \mathbb{N}\times\mathbb{R}\to\mathbb{R}$ that summarizes a persistence diagram dgm - see Figure 14. The landscape is defined by considering the set of functions created by tenting each point $p=(x,y)=\left(\frac{\alpha_{\text{birth}}+\alpha_{\text{death}}}{2},\frac{\alpha_{\text{death}}-\alpha_{\text{birth}}}{2}\right)$ representing a birth-death pair $(\alpha_{\text{birth}},\alpha_{\text{death}})\in$ dgm as follows:

$$\Lambda_p(t)=\begin{cases} t-x+y & t\in[x-y,x] \\ x+y-t & t\in(x,x+y] \\ 0 & \text{otherwise}\end{cases}=\begin{cases} t-\alpha_{\text{birth}} & t\in[\alpha_{\text{birth}},\frac{\alpha_{\text{birth}}+\alpha_{\text{death}}}{2}] \\ \alpha_{\text{death}}-t & t\in(\frac{\alpha_{\text{birth}}+\alpha_{\text{death}}}{2},\alpha_{\text{death}}] \\ 0 & \text{otherwise.}\end{cases}$$

The persistence landscape of dgm is a summary of the arrangement of piecewise linear curves obtained by overlaying the graphs of the functions $\{\Lambda_p\}_p$. Formally, the persistence landscape of dgm is the collection of functions

$$\lambda_{\text{dgm}}(k,t)=\underset{p}{\text{kmax}}\ \Lambda_p(t),\quad t\in[0,T],k\in\mathbb{N}, \tag{3}$$

where kmax is the kth largest value in the set; in particular, 1max is the usual maximum function. Given $k\in\mathbb{N}$, the function $\lambda_{\text{dgm}}(k,.):\mathbb{R}\to\mathbb{R}$ is called the *k-th landscape* of dgm. It is not difficult to see that the map that associate to each persistence diagram its corresponding landscape is injective. In other words, formally no information is lost when a persistence diagram is represented through its persistence landscape.

The advantage of the persistence landscape representation is two-fold. First, persistence diagrams are mapped as elements of a functional space, opening the door to the use of a broad variety of statistical and data analysis tools for further processing of topological features - see, e.g. Bubenik (2015); Chazal et al. (2015b) and Section 5.8. Second, and fundamental from a theoretical perspective, the persistence landscapes share the same stability properties as persistence diagrams - see Section 5.6. Following the same ideas, other alternatives to persistence diagrams have been proposed, such as, for instance, the persistence images Adams et al. (2017) - see Section 5.9.

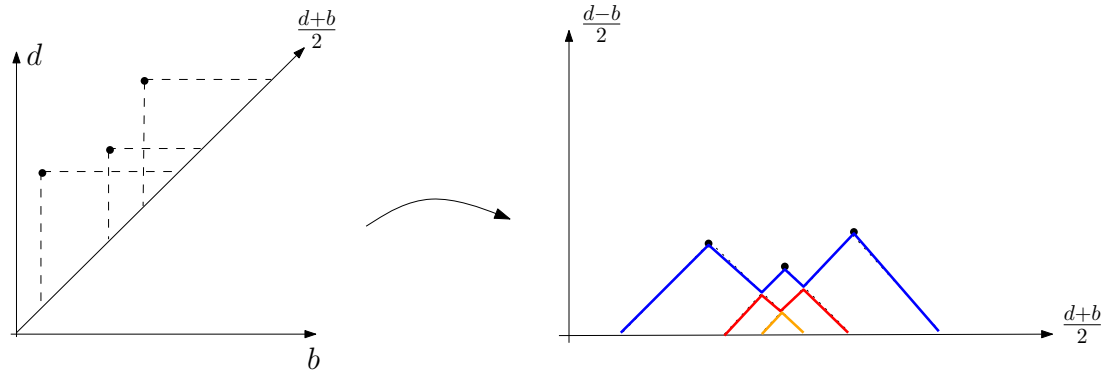


Figure 14: An example of persistence landscape (right) associated to a persistence diagram (left). The first landscape is in blue, the second one in red and the last one in orange. All the other landscapes are zero.

5.5 Metrics on the space of persistence diagrams

To exploit the topological information and topological features inferred from persistent homology, one needs to be able to compare persistence diagrams, i.e. to endow the space of persistence diagrams with a metric structure. Although several metrics can be considered, the most fundamental one is known as the *bottleneck distance*.

Recall that a persistence diagram is the union of a discrete multi-set in the half-plane above the diagonal Δ and, for technical reasons that will become clear below, of Δ where the point of Δ are counted with infinite multiplicity. A *matching* - see Figure 15 - between two diagrams dgm_1 and dgm_2 is a subset $m \subseteq \mathrm{dgm}_1 \times \mathrm{dgm}_2$ such that every points in $\mathrm{dgm}_1 \setminus \Delta$ and $\mathrm{dgm}_2 \setminus \Delta$ appears exactly once in m. In other words, for any $p \in \mathrm{dgm}_1 \setminus \Delta$, and for any $q \in \mathrm{dgm}_2 \setminus \Delta$, $(\{p\} \times \mathrm{dgm}_2) \cap m$ and $(\mathrm{dgm}_1 \times \{q\}) \cap m$ each contains a single pair. The *Bottleneck distance* between dgm_1 and dgm_2 is then defined by

$$\mathrm{d_b}(\mathrm{dgm}_1, \mathrm{dgm}_2) = \inf_{\text{matching } m} \max_{(p,q)\in m} \|p - q\|_\infty.$$

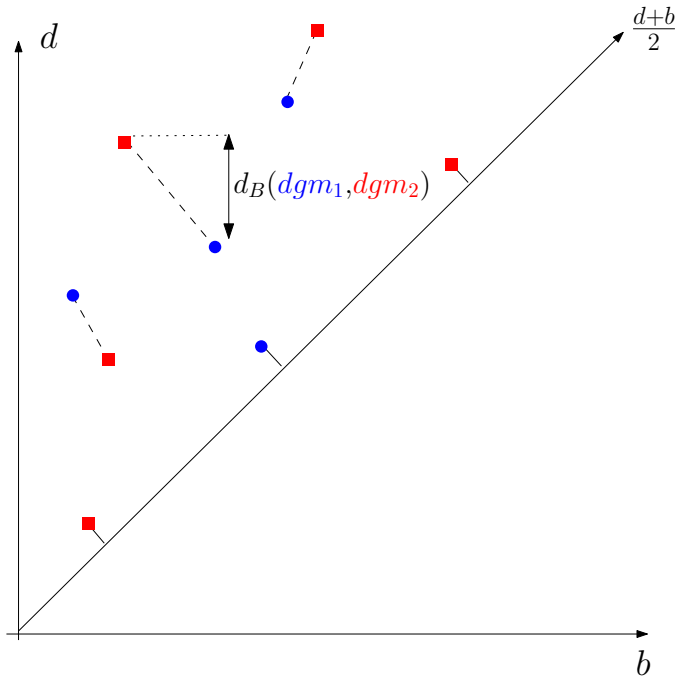


Figure 15: A perfect matching and the Bottleneck distance between a blue and a red diagram. Notice that some points of both diagrams are matched to points of the diagonal.

The practical computation of the bottleneck distance boils down to the computation of a perfect matching in a bipartite graph for which classical algorithms can be used.

The bottleneck metric is a L_∞-like metric. It turns out to be the natural one to express stability properties of persistence diagrams presented in Section 5.6, but it suffers from the same drawbacks as the usual L_∞ norms, i.e. it is completely determined by the largest distance among the pairs and do not take into account the closeness of the remaining pairs of points. A variant, to overcome this issue, the so-called Wasserstein distance between diagrams is sometimes considered. Given $p \geq 1$, it is defined by

$$W_p(\mathrm{dgm}_1, \mathrm{dgm}_2)^p = \inf_{\text{matching } m} \sum_{(p,q)\in m} \|p - q\|_\infty^p.$$

Useful stability results for persistence in the W_p metric exist among the literature, in particular Cohen-Steiner et al. (2010), but they rely on assumptions that make them consequences of the stability results in the bottleneck metric.

5.6 Stability properties of persistence diagrams

A fundamental property of persistence homology is that persistence diagrams of filtrations built on top of data sets turn out to be very stable with respect to some perturbations of the data. To

formalize and quantify such stability properties, we first need to precise the notion of perturbation that are allowed.

Rather than working directly with filtrations built on top of data sets, it turns out to be more convenient to define a notion of proximity between persistence module from which we will derive a general stability result for persistent homology. Then, most of the stability results for specific filtrations will appear as a consequence of this general theorem. To avoid technical discussions, from now on we assume, without loss of generality, that the considered persistence modules are indexed by $\mathbb{R}$.

Definition 9. *Let $\mathbb{V}, \mathbb{W}$ be two persistence modules indexed by $\mathbb{R}$. Given $\delta \in \mathbb{R}$, a homomorphism of degree δ between $\mathbb{V}$ and $\mathbb{W}$ is a collection Φ of linear maps $\phi_r : V_r \to W_{r+\delta}$, for all $r \in \mathbb{R}$ such that for any $r \leq s$, $\phi_s \circ v_s^r = w_{s+\delta}^{r+\delta} \circ \phi_r$.*

An important example of homomorphism of degree δ is the *shift endomorphism* $1_{\mathbb{V}}^{\delta}$ which consists of the families of linear maps $(v_{r+\delta}^r$. Notice also that homomorphisms of modules can naturally be composed: the composition of a homomorphism Ψ of degree δ between $\mathbb{U}$ and $\mathbb{V}$ and a homomorphism Φ of degree δ' between $\mathbb{V}$ and $\mathbb{W}$ naturally gives rise to a homomorphism $\Phi\Psi$ of degree $\delta + \delta'$ between $\mathbb{U}$ and $\mathbb{W}$.

Definition 10. *Let $\delta \geq 0$. Two persistence modules $\mathbb{V}, \mathbb{W}$ are δ-interleaved if there exists two homomorphism of degree δ, Φ, from $\mathbb{V}$ to $\mathbb{W}$ and Ψ, from $\mathbb{W}$ to $\mathbb{V}$ such that $\Psi\Phi = 1_{\mathbb{V}}^{2\delta}$ and $\Phi\Psi = 1_{\mathbb{W}}^{2\delta}$.*

Although it does not define a metric on the space of persistence modules, the notion of closeness between two persistence module may be defined as the smallest non negative δ such that they are δ-interleaved. Moreover, it allows to formalize the following fundamental theorem Chazal et al. (2009a, 2016a).

Theorem 7 (Stability of persistence)**.** *Let $\mathbb{V}$ and $\mathbb{W}$ be two q-tame persistence modules. If $\mathbb{V}$ and $\mathbb{W}$ are δ-interleaved for some $\delta \geq 0$, then*

$$\mathrm{d_b}(\mathrm{dgm}(\mathbb{V}), \mathrm{dgm}(\mathbb{W})) \leq \delta.$$

Although purely algebraic and rather abstract, this result is a efficient tool to easily establish concrete stability results in TDA. For example we can easily recover the first persistence stability result that appeared in the literature (Cohen-Steiner et al., 2005).

Theorem 8. *Let $f, g : M \to \mathbb{R}$ be two real-valued functions defined on a topological space M that are q-tame, i.e. such that the sublevel sets filtrations of f and g induce q-tame modules at the homology level. Then for any integer k,*

$$\mathrm{d_b}(\mathrm{dgm}_k(f), \mathrm{dgm}_k(g)) \leq \|f - g\|_\infty = \sup_{x \in M} |f(x) - g(x)|$$

where $\mathrm{dgm}_k(f)$ (resp. $\mathrm{dgm}_k(g)$) is the persistence diagram of the persistence module $(H_k(f^{-1}(-\infty, r]))|r \in \mathbb{R})$ (resp. $(H_k(g^{-1}(-\infty, r]))|r \in \mathbb{R})$ where the linear maps are the one induced by the canonical inclusion maps between sublevel sets.

Proof. Denoting $\delta = \|f - g\|_\infty$ we have that for any $r \in \mathbb{R}$, $f^{-1}(-\infty, r]) \subseteq g^{-1}(-\infty, r + \delta])$ and $g^{-1}(-\infty, r]) \subseteq f^{-1}(-\infty, r + \delta])$. This interleaving between the sublevel sets of f induces a δ-interleaving between the persistence modules at the homology level and the result follows from the direct application of Theorem 7. □

Theorem 7 also implies a stability result for the persistence diagrams of filtrations built on top of data.

Theorem 9. *Let* $\mathbb{X}$ *and* $\mathbb{Y}$ *be two compact metric spaces and let* Filt($\mathbb{X}$) *and* Filt($\mathbb{Y}$) *be the Vietoris-Rips of Čech filtrations built on top* $\mathbb{X}$ *and* $\mathbb{Y}$*. Then*

$$d_b\left(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X})), \mathrm{dgm}(\mathrm{Filt}(\mathbb{Y}))\right) \leq 2d_{GH}(\mathbb{X}, \mathbb{Y})$$

where dgm(Filt($\mathbb{X}$)) *and* dgm(Filt($\mathbb{Y}$)) *denote the persistence diagram of the filtrations* Filt($\mathbb{X}$) *and* Filt($\mathbb{X}$)*.*

As we already noticed in the Example 3 of Section 5.2, the persistence diagrams can be interpreted as multiscale topological features of $\mathbb{X}$ and $\mathbb{Y}$. In addition, Theorem 9 tells us that these features are robust with respect to perturbations of the data in the Gromov-Hausdorff metric. They can be used as discriminative features for classification or other tasks - see, for example, Chazal et al. (2009b) for an application to non rigid 3D shapes classification.

From the definition of persistence landscape, we immediately observe that $\lambda(k, \cdot)$ is one-Lipschitz and thus similar stability properties are satisfied for the landscapes as for persistence diagrams.

Proposition 1. [Bubenik 2015] *Let* $\mathbb{X}$ *and* $\tilde{\mathbb{X}}$ *be two compact sets. For any* $t \in \mathbb{R}$ *and any* $k \in \mathbb{N}$*, we have:*
(i) $\lambda_{\mathbb{X}}(k,t) \geq \lambda_{\mathbb{X}}(k+1,t) \geq 0$.
(ii) $|\lambda_{\mathbb{X}}(k,t) - \lambda_{\tilde{\mathbb{X}}}(k,t)| \leq \mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X})), \mathrm{dgm}(\mathrm{Filt}(\tilde{\mathbb{X}})))$.

5.7 Statistical aspects of persistent homology

Persistence homolog by itself does not take into account the random nature of data and the intrinsic variability of the topological quantity they infer. We now present a statistical approach to persistent homology, which means that we consider data as generated from an unknown distribution. We start with several consistency results on persistent homology inference.

5.7.1 Estimation of the persistent homology of a metric space

Assume that we observe n points $(X_1, \ldots, X_n)$ in a metric space (M, ρ) drawn i.i.d. from an unknown probability measure μ whose support is a compact set denoted $\mathbb{X}_\mu$. The Gromov-Hausdorff distance allows us to compare $\mathbb{X}_\mu$ with compact metric spaces not necessarily embedded in M. In the following, an *estimator* $\widehat{\mathbb{X}}$ of $\mathbb{X}_\mu$ is a function of $X_1 \ldots, X_n$ that takes values in the set of compact metric spaces and which is measurable for the Borel algebra induced by $\mathrm{d_{GH}}$.

Let Filt($\mathbb{X}_\mu$) and Filt($\widehat{\mathbb{X}}$) be two filtrations defined on $\mathbb{X}_\mu$ and $\widehat{\mathbb{X}}$. Starting from Theorem 9; an natural strategy for estimating the persistent homology of Filt($\mathbb{X}_\mu$) consists in estimating the support $\mathbb{X}_\mu$. Note that in some cases the space M can be unknown and the observations $X_1 \ldots, X_n$ are then only known through their pairwise distances $\rho(X_i, X_j)$, $i, j = 1, \cdots, n$. The use of the Gromov-Hausdorff distance allows us to consider this set of observations as an abstract metric space of cardinality n, independently of the way it is embedded in M. This general framework includes the more standard approach consisting in estimating the support with respect to the Hausdorff distance by restraining the values of $\widehat{\mathbb{X}}$ to the compact sets included in M.

The finite set $\mathbb{X}_n := \{X_1, \ldots, X_n\}$ is a natural estimator of the support $\mathbb{X}_\mu$. In several contexts discussed in the following, $\mathbb{X}_n$ shows optimal rates of convergence to $\mathbb{X}_\mu$ with respect to the Hausdorff distance. For some constants $a, b > 0$, we say that μ satisfies the (a, b)-standard assumption if for any $x \in \mathbb{X}_\mu$ and any $r > 0$,

$$\mu(B(x, r)) \geq \min(ar^b, 1). \tag{4}$$

This assumption has been widely used in the literature of set estimation under Hausdorff distance (Cuevas and Rodríguez-Casal, 2004; Singh et al., 2009).

Theorem 10. [Chazal et al. (2014b)] *Assume that the probability measure μ on M satisfies the (a,b)-standard assumption, then for any $\varepsilon > 0$:*

$$\mathbb{P}\left(\mathrm{d_b}\left(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n))\right) > \varepsilon\right) \leq \min\left(\frac{2^b}{a\varepsilon^b}\exp(-na\varepsilon^b), 1\right). \tag{5}$$

Moreover,

$$\limsup_{n\to\infty}\left(\frac{n}{\log n}\right)^{1/b}\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n))) \leq C_1$$

almost surely, and

$$\mathbb{P}\left(\mathrm{d_b}\left(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n))\right) \leq C_2\left(\frac{\log n}{n}\right)^{1/b}\right)$$

converges to 1 *when $n \to \infty$, where C_1 and C_2 only depend on a and b.*

Let $\mathcal{P} = \mathcal{P}(a,b,M)$ be the set of all the probability measures on the metric space (M,ρ) satisfying the (a,b)-standard assumption on M:

$$\mathcal{P} := \left\{\mu \text{ on } M \mid \mathbb{X}_\mu \text{ is compact and } \forall x \in \mathbb{X}_\mu, \forall r > 0,\ \mu\left(B(x,r)\right) \geq \min\left(1, ar^b\right)\right\}. \tag{6}$$

The next theorem gives upper and lower bounds for the rate of convergence of persistence diagrams. The upper bound is a consequence of Theorem 10, while the lower bound is established using Le Cam's lemma.

Theorem 11. [Chazal et al. (2014b)] *For some positive constants a and b,*

$$\sup_{\mu\in\mathcal{P}}\mathbb{E}\left[\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n)))\right] \leq C\left(\frac{\log n}{n}\right)^{1/b}$$

where the constant C only depends on a and b (not on M). Assume moreover that there exists a non isolated point x in M and consider any sequence $(x_n) \in (M \setminus \{x\})^{\mathbb{N}}$ such that $\rho(x,x_n) \leq (an)^{-1/b}$. Then for any estimator $\widehat{\mathrm{dgm}}_n$ of $\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu))$:

$$\liminf_{n\to\infty}\rho(x,x_n)^{-1}\sup_{\mu\in\mathcal{P}}\mathbb{E}\left[\mathrm{d_b}(\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \widehat{\mathrm{dgm}}_n)\right] \geq C'$$

where C' is an absolute constant.

Consequently, the estimator $\mathrm{dgm}(\mathrm{Filt}(\mathbb{X}_n))$ is minimax optimal on the space $\mathcal{P}(a,b,M)$ up to a logarithmic term as soon as we can find a non-isolated point in M and a sequence (x_n) in M such that $\rho(x_n,x) \sim (an)^{-1/b}$. This is obviously the case for the Euclidean space $\mathbb{R}^d$.

Additive noise. Consider the convolution model where the observations satisfy $Y_i = X_i + \varepsilon_i$ where $X_1, \ldots X_n$ are sampled according to a measure μ as in the previous paragraph and where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. standard Gaussian random variables. It can be deduced from the results of Genovese et al. (2012) that the minimax convergence rates for the persistence diagram estimation in this context is upper bounded by some rate of the order of $(\log n)^{-1/2}$. However, giving a tight lower bound for this problem appears to be more difficult than for the support estimation problem.

5.7.2 Estimation of the persistent homology of functions

Theorem 7 opens the door to the estimation of the persistent homology of functions defined on $\mathbb{R}^d$, on a submanifold of $\mathbb{R}^d$ or more generally on a metric space. One important direction of research on this topic concerns various versions of robust TDA. One option is to study the persistent homology of the upper level sets of density estimators Fasy et al. (2014b). A different approach, more closely related to the distance function, but robust to noise, consists in studying the persistent homology of the sub level sets of the distance to measure defined in Section 4.2 Chazal et al. (2014a). The persistent homology of regression functions has also been studied in Bubenik et al. (2010). The alternative approach of Bobrowski et al. (2014) which is based on the inclusion map between nested pairs of estimated level sets can be applied with kernel density and regression kernel estimators to estimate persistence homology of density functions and regression functions.

5.7.3 Statistics for other signatures

Convergence and confidence regions (see next paragraph) can be proposed for persistence landscapes using similar stability results. However, a complete minimax description of the problem would also require to prove the corresponding lower bounds. Functional convergence for persistence landscapes and silhouettes have been studied in Chazal et al. (2015b).

5.7.4 Confidence regions for persistent homology

For many applications, in particular when the point cloud does not come from a geometric shape, persistence diagrams can be quite complex to analyze. In particular, many topological features are closed to the diagonal. Since they correspond to topological structures that die very soon after they appear in the filtration, these points are generally considered as noise, see Figure 16. Confidence regions of persistence diagram are rigorous answers to the problem of distinguishing between signal and noise in these representations.

The stability results given in Section 5.6 motivate the use of the bottleneck distance to define confidence regions. However alternative distances in the spirit of Wasserstein distances can proposed too. When estimating a persistence diagram dgm with an estimator $\hat{\text{dgm}}$, we typically look for some value η_α such that

$$P(\mathrm{d_b}(\hat{\text{dgm}}, \text{dgm}) \geq \eta_\alpha) \leq \alpha,$$

for $\alpha \in (0,1)$. Let B_α be the closed ball of radius α for the bottleneck distance and centered at $\hat{\text{dgm}}$ in the space of persistence diagrams. Following Fasy et al. (2014b), we can visualize the signatures of the points belonging to this ball in various ways. One first option is to center a box of side length 2α at each point of the persistence diagram $\hat{\text{dgm}}$. An alternative solution is to visualize the confidence set by adding a band at (vertical) distance $\eta_\alpha/2$ from the diagonal (the bottleneck distance being defined for the ℓ_∞ norm), see Figure 18 for an illustration. The points outside the band are then considered as significant topological features, see Fasy et al. (2014b) for more details.

Several methods have been proposed in Fasy et al. (2014b) to estimate η_α in the definition of the confidence region for the persistent homology of the measure support and for the sub-level sets of a density function. Except for the bottleneck bootstrap (see further), all the methods proposed in these papers rely on the stability results for persistence diagrams: confidence sets for diagrams can be derived from confidence sets in the sample space.

Subsampling approach. This method is based on a confidence region for the support K of the distribution of the sample in Hausdorff distance. Let $\tilde{\mathbb{X}}_b$ be a subsample of size b drawn from the sample $\tilde{\mathbb{X}}_n$, where $b = o(n/logn)$. Let $q_b(1-\alpha)$ be the quantile of of the distribution

of Haus $\left(\tilde{\mathbb{X}}_b, \mathbb{X}_n\right)$. Take $\hat{\eta}_\alpha := 2\hat{q}_b(1-\alpha)$ where $\hat{q}_b$ is an estimation $q_b(1-\alpha)$ using a standard Monte Carlo procedure. Under an (a,b) standard assumption, and for n large enough, Fasy et al. (2014b) show that

$$\begin{aligned} P\left(\mathrm{d_b}\left(\mathrm{dgm}\left(\mathrm{Filt}(K)\right), \mathrm{dgm}\left(\mathrm{Filt}(\mathbb{X}_n)\right)\right) > \hat{\eta}_\alpha\right) &\leq P\big(\mathrm{Haus}\left(K, \mathbb{X}_n\right) > \hat{\eta}_\alpha\big) \\ &\leq \alpha + O\left(\frac{b}{n}\right)^{1/4}. \end{aligned}$$

Bottleneck Bootstrap. The stability results often leads to conservative confidence sets. An alternative strategy is the bottleneck bootstrap introduced in Chazal et al. (2016b). We consider the general setting where a persistence diagram $\hat{\mathrm{dgm}}$ is defined from the observation $(X_1, \ldots, X_n)$ in a metric space . This persistence diagram corresponds to the estimation of an underlying persistence diagram dgm, which can be related for instance to the support of the measure, or to the sublevel sets of a function related to this distribution (for instance a density function when the X_i's are in $\mathbb{R}^d$). Let $(X_1^*, \ldots, X_n^*)$ be a sample from the empirical measure defined from the observations $(X_1, \ldots, X_n)$. Let also $\hat{\mathrm{dgm}}^*$ be the persistence diagram derived from this sample. We then can take for η_α the quantity $\hat{\eta}_\alpha$ defined by

$$P(\mathrm{d_b}(\hat{\mathrm{dgm}}^*, \hat{\mathrm{dgm}}) > \hat{\eta}_\alpha \,|\, X_1, \ldots, X_n) = \alpha. \tag{7}$$

Note that $\hat{\eta}_\alpha$ can be easily estimated with Monte Carlo procedures. It has been shown in Chazal et al. (2016b) that the bottleneck bootstrap is valid when computing the sublevel sets of a density estimator.

Confidence bands for landscapes. A bootstrap algorithm can be used to construct confidence bands for landscapes (Chazal et al., 2015b). However the setting of this paper is slightly different than before since it is now assumed that we observe several landscapes $\lambda_1, \ldots, \lambda_N$ drawn i.i.d. from a random distribution in the space of landscapes. In this context the multiplier bootstrap strategy can be applied to construct a confidence band for $\mathbb{E}(\lambda_1)$.

5.8 Central tendency for persistent homology

The space of persistence diagrams being not an Hilbert space, the definition of a *mean persistence diagram* is not obvious and unique. One first approach to define a central tendency in this context is to define a Fréchet mean in this context. Indeed it has been proved in Mileyko et al. (2011) that the space of persistence diagrams is a Polish space. Fréchet means have also been characterized in Turner et al. (2014a). However they are may not be unique and there are very difficult to compute in practice. To overcome the problem of computational costs, sampling strategies can be proposed to compute topological signatures based on persistence landscapes. Given a large point cloud, the idea is to extract many subsamples, to compute the landscape for each subsample and then to combine the information.

We assume that the diameter of M is finite and upper bounded by $\frac{T}{2}$, where T is the same constant as in the definition of persistence landscapes in Section 5.4. For ease of exposition, we focus on the case $k=1$, and set $\lambda(t) = \lambda(1,t)$. However, the results we present in this section hold for $k > 1$.

For any positive integer m, let $X = \{x_1, \cdots, x_m\} \subset \mathbb{X}_\mu$ be a sample of m points from μ. The corresponding persistence landscape is λ_X and we denote by Ψ_μ^m the measure induced by $\mu^{\otimes m}$ on the space of persistence landscapes. Note that the persistence landscape λ_X can be seen as a single draw from the measure Ψ_μ^m. The point-wise expectations of the (random) persistence landscape under this measure is defined by $\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)], t \in [0,T]$. The average landscape $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$ has a natural empirical counterpart, which can be used as its unbiased

estimator. Let $S_1^m, \ldots, S_\ell^m$ be ℓ independent samples of size m from $\mu^{\otimes m}$. We define the empirical average landscape as

$$\overline{\lambda_\ell^m}(t) = \frac{1}{b}\sum_{i=1}^{b} \lambda_{S_i^m}(t), \quad \text{for all } t \in [0,T], \tag{8}$$

and propose to use $\overline{\lambda_\ell^m}$ to estimate $\lambda_{\mathbb{X}_\mu}$. Note that computing the persistent homology of $\mathbb{X}_n$ is $O(\exp(n))$, whereas computing the average landscape is $O(b\exp(m))$.

Another motivation for this subsampling approach is that it can be also applied when μ is a discrete measure with support $\mathbb{X}_N = \{x_1, \ldots, x_N\} \subset M$. This framework can be very common in practice, when a continuous (but unknown measure) is approximated by a discrete uniform measure μ_N on $\mathbb{X}_N$.

The average landscape $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$ is an interesting quantity on its own, since it carries some stable topological information about the underlying measure μ, from which the data are generated. In particular,

Theorem 12. [Chazal et al. (2015a)] *Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where μ and ν are two probability measures on M. For any $p \geq 1$ we have*

$$\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq 2\, m^{\frac{1}{p}} W_p(\mu,\nu),$$

where W_p is the pth Wasserstein distance on M.

The result of Theorem 12 is useful for two reasons. First, it tells us that for a fixed m, the expected "topological behavior" of a set of m points carries some stable information about the underlying measure from which the data are generated. Second, it provides a lower bound for the Wasserstein distance between two measures, based on the topological signature of samples of m points.

5.9 Persistent homology and machine learning

In some domains persistence diagrams obtained from data can be directly interpreted and exploited for better understanding of the phenomena from which the data have been generated. This, for example, the case in the study of force fields in granular media Kramar et al. (2013) or of atomic structures in glass Nakamura et al. (2015) in material science, in the study of the evolution of convection patterns in fluid dynamics Kramár et al. (2016) or in the analysis of nanoporous structures in chemistry Lee et al. (2017) where topological features can be rather clearly related to specific geometric structures and patterns in the considered data.

There are many other cases where persistence features cannot be easily or directly interpreted but present valuable information for further processing. However, the highly non linear nature of diagrams prevents them to be immediately used as standard features in machine learning algorithms. Persistence landscapes and their variants, introduced in Section 5.4 offer a first option to convert persistence diagrams into elements of a vector space and have been used, for example, for protein binding Kovacev-Nikolic et al. (2016) or object recognition Li et al. (2014). In the same vein, the construction of kernels for persistence diagrams that preserve their stability properties has recently attracted some attention. Most of them have been obtained by considering diagrams as discrete measures in $\mathbb{R}^2$. Convolving a symmetrized (with respect to the diagonal) version of persistence diagrams with a 2D Gaussian distribution, Reininghaus et al. (2015) introduce a multi-scale kernel and apply it to shape classification and texture recognition problems. Considering Wasserstein distance between projections of persistence diagrams on lines, Carriere and Oudot (2017) build another kernel and test its performance on several benchmarks. Other kernels, still obtain by considering persistence diagrams as measures, have also been proposed in Kusano et al. (2017).

Various other vector summaries of persistence diagrams have been proposed and then used as features for different problems. For example, basic summaries are considered in Bonis et al. (2016) and combined with quantization and pooling methods to address non rigid shape analysis problems; *Betti curves* extracted from persistence diagrams are used with 1-dimensional Convolutional Neural Networks (CNN) to analyze time dependent data and recognize human activities from inertial sensors in Umeda (2017); *persistence images* are introduced in Adams et al. (2017) and are considered to address some inverse problems using linear machine learnig models in Obayashi and Hiraoka (2017).

Connections between persistence homology and deep learning have also very recently started to be explored. For example, as already mentioned above, Umeda (2017) combine persistent homology with CNNs to analyze multivariate time-dependent data. Approaches combining persistence and deep learning have also been proposed in molecular biology - see, e.g., Cang and Wei (2017).

The above mentioned kernels and vector summaries of persistence diagrams are built independently of the considered data analysis or learning task. Moreover, it appears that in many cases the relevant topological information is not carried by the whole persistence diagrams but is concentrated in some localized regions that may not be obvious to identify. This usually makes the choice of a relevant kernel or vector summary very difficult for the user. To overcome this issue, Hofer et al. (2017) proposes a deep learning approach that allows to learn the relevant topological features for a given task.

As illustrated in this section, combining TDA and more specifically persistent homology, with machine learning has recently became an active research direction with already promising results but still many theoretical and practical open questions and problems.

6 TDA for data sciences with the GUDHI library

In this section we illustrate TDA methods with the Python library Gudhi[6] (Maria et al., 2014) together with popular libraries as numpy (Walt et al., 2011), scikit-learn (Pedregosa et al., 2011), pandas (McKinney et al., 2010).

6.1 Bootstrap and comparison of protein binding configurations

This example is borrowed from Kovacev-Nikolic et al. (2016). In this paper, persistent homology is used to analyze protein binding and more precisely it compares closed and open forms of the maltose-binding protein (MBP), a large biomolecule consisting of 370 amino acid residues. The analysis is not based on geometric distances in $\mathbb{R}^3$ but on a metric of *dynamical distances* defined by

$$D_{ij} = 1 - |C_{ij}|,$$

where C is the correlation matrices between residues. The data can be download at this link[7].

```
import numpy as np
import gudhi as gd
import pandas as pd
import seaborn as sns

corr_protein = pd.read_csv("mypath/1anf.corr_1.txt",
                           header=None,
                           delim_whitespace=True)
dist_protein_1 = 1- np.abs(corr_protein_1.values)
rips_complex_1= gd.RipsComplex(distance_matrix=dist_protein_1,
                               max_edge_length=1.1)
simplex_tree_1 = rips_complex_1.create_simplex_tree(max_dimension=2)
```

[6] http://gudhi.gforge.inria.fr/python/latest/

[7] https://www.researchgate.net/publication/301543862_corr

```
diag_1 = simplex_tree_1.persistence()
gd.plot_persistence_diagram(diag_1)
```

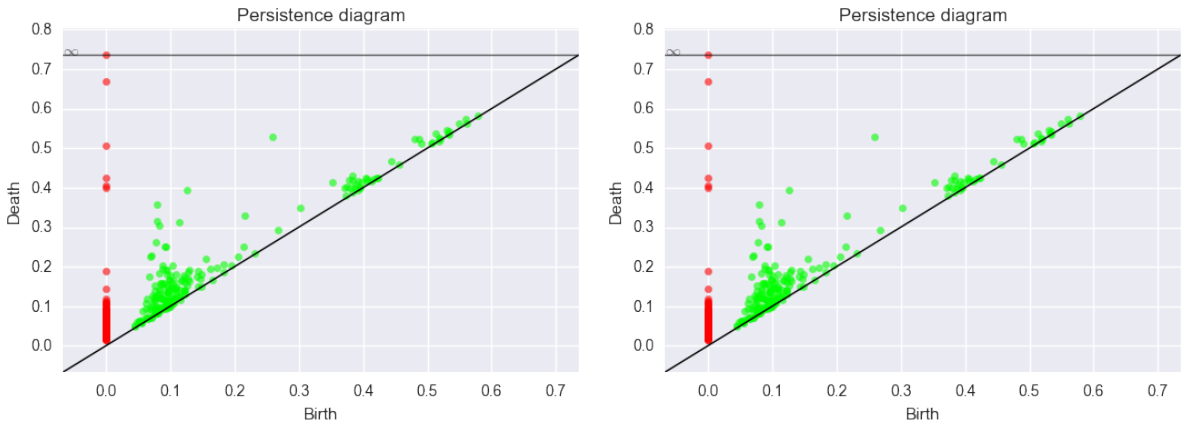


Figure 16: Persistence diagrams for two configurations of MBP.

For comparing persistence diagrams, we use the bottleneck distance. The block of statements given below computes persistence intervals and computes the bottleneck distance for 0-homology and 1-homology:

```
interv0_1 = simplex_tree_1.persistence_intervals_in_dimension(0)
interv0_2 = simplex_tree_2.persistence_intervals_in_dimension(0)
bot0  = gd.bottleneck_distance(interv0_1,interv0_2)

interv1_1 = simplex_tree_1.persistence_intervals_in_dimension(1)
interv1_2 = simplex_tree_2.persistence_intervals_in_dimension(1)
bot1  = gd.bottleneck_distance(interv1_1,interv1_2)
```

In this way, we can compute the matrix of bottleneck distances between the fourteen MPB. Finally, we apply a multidimensional scaling method to find a configuration in $\mathbb{R}^2$ which almost match with the bottleneck distances, see Figure 17. We use the scikit-learn library for the MDS:

```
import matplotlib.pyplot as plt
from sklearn import manifold

mds = manifold.MDS(n_components=2, dissimilarity="precomputed")
config = mds.fit(M).embedding_

plt.scatter(config [0:7,0], config [0:7, 1], color='red', label="closed")
plt.scatter(config [7:1,0], config [7:1, 1], color='blue', label="red")
plt.legend(loc=1)
```

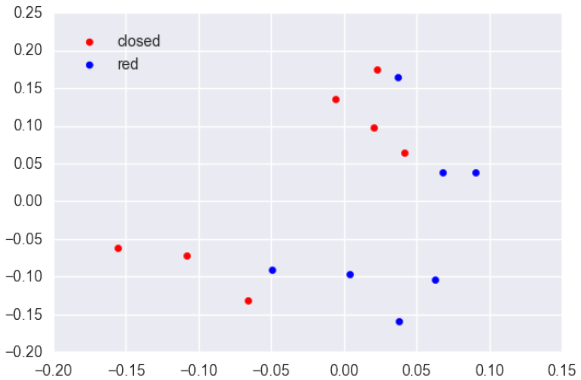


Figure 17: MDS configuration for the matrix of bottleneck distances.

We now define a band of confiance for a diagram using the bottleneck bootstrap approach. We resample over the lines (and columns) of the matrix of distances and we compute the bottleneck distance between the original persistence diagram and the bootstrapped persistence diagram. We repeat the procedure many times and finally we estimate the quantile 95% of this collection of bottleneck distances. We take the value of the quantile to define a confidence band on the original diagram (see Figure 18). However, such a procedure should be considered with caution because as far as we know the validity of the bottleneck bootstrap has not been proved in this framework.

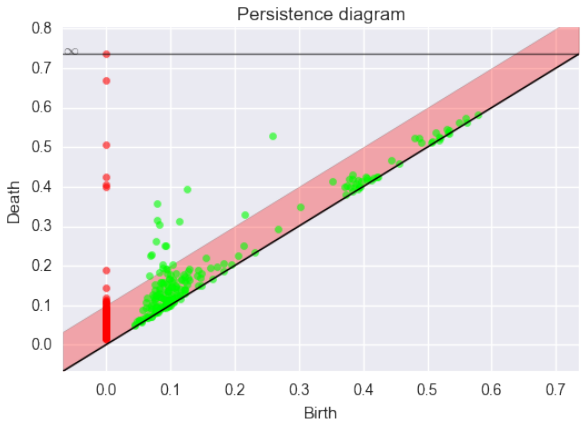


Figure 18: Persistence diagram and confidence region for the persistence diagram of a MBP.

6.2 Classification for sensor data

In this experiment, the 3d acceleration of 3 walkers (A, B and C) have been recorded from the sensor of a smart phone[8]. Persistence homology is not sensitive to the choice of axes and so no preprocessing is necessary to align the 3 times series according to the same axis. From these three times series, we have picked at random sequences of 8 seconds in the complete time series, that is 200 consecutive points of acceleration in $\mathbb{R}^3$. For each walker, we extract 100 time series in this way. The next block of statements computes the persistence for the alpha complex filtration for `data_A_sample`, one of the 100 times series of acceleration of Walker A.

```
alpha_complex_sample = gd.AlphaComplex(points = data_A_sample)
simplex_tree_sample = alpha_complex_sample.create_simplex_tree(max_alpha_square
    =0.3)
diag_Alpha = simplex_tree_sample.persistence()
```

From `diag_Alpha` we can then easily compute and plot the persistence landscapes, see Figure 19. For all the 300 times series, we compute the persistence landscapes for dimension 0 and 1 and we compute the three first landscapes for the 2 dimensions, see Figure 19. Moreover, each persistence landscape is discretized on 1000 points. Each time series is thus described by 6000 topological variables. To predict the walker from these features, we use a random forest (Breiman, 2001), which is known to be an efficient in such an high dimensional setting. We split the data into train and test tests at random several times. We finally obtain a averaged classification error around 0.95. We can also visualize les most important variables in the Random Forest, see Figure 20.

Acknowledgements This work was partly supported by the French ANR project TopData ANR-13-BS01-0008 and the ERC project GUDHI (Geometric Understanding in Higher Dimensions). We thank the authors of Kovacev-Nikolic et al. (2016) for making their data available.

[8]The dataset can be download at this link `http://bertrand.michel.perso.math.cnrs.fr/Enseignements/TDA/data_acc`

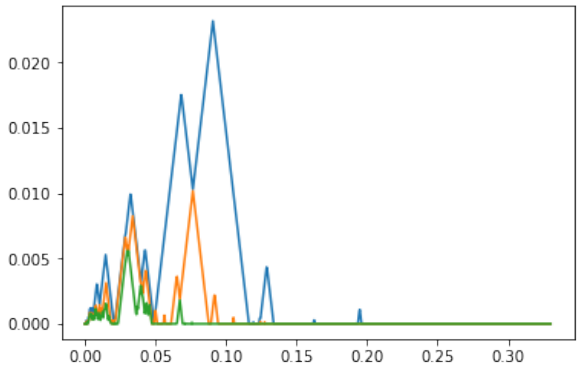


Figure 19: The three first landscapes for 0-homology of the alpha shape filtration defined for a time series of acceleration of Walker A.

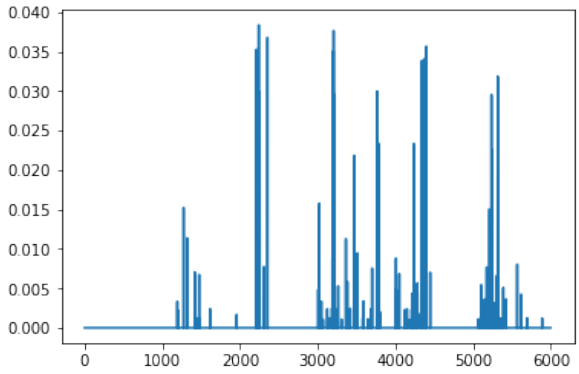


Figure 20: Variable importances of the landscapes coefficients for the classification of Walkers. The 3 000 first coefficients correspond to the three landscapes of dimension 0 and the 3 000 last coefficients to the three landscapes of dimension 1. There are 1000 coefficients per landscape. Note that the first landscape of dimension 0 is always the same using the Rips complex (a trivial landscape) and consequently the corresponding coefficients have a zero importance value.

References

Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. (2017). Persistence images: a stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35.

Balakrishna, S., Rinaldo, A., Sheehy, D., Singh, A., and Wasserman, L. A. (2012). Minimax rates for homology inference. *Journal of Machine Learning Research - Proceedings Track*, 22:64–72.

Biau, G., Chazal, F., Cohen-Steiner, D., Devroye, L., and Rodriguez, C. (2011). A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237.

Bobrowski, O., Mukherjee, S., and Taylor, J. (2014). Topological consistency via kernel estimation. *arXiv preprint arXiv:1407.5272*.

Bonis, T., Ovsjanikov, M., Oudot, S., and Chazal, F. (2016). Persistence-based pooling for shape pose recognition. In *Computational Topology in Image Context - 6th International Workshop, CTIC 2016, Marseille, France, June 15-17, 2016, Proceedings*, pages 19–29.

Brécheteau, C. (2017). The dtm-signature for a geometric comparison of metric-measure spaces from samples.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102.

Bubenik, P., Carlsson, G., Kim, P. T., and Luo, Z.-M. (2010). Statistical topology via morse theory persistence and nonparametric estimation. *Algebraic methods in statistics and probability II*, 516:75–92.

Buchet, M., Chazal, F., Dey, T. K., Fan, F., Oudot, S. Y., and Wang, Y. (2015a). Topological analysis of scalar fields with outliers. In *Proc. Sympos. on Computational Geometry*.

Buchet, M., Chazal, F., Oudot, S., and Sheehy, D. R. (2015b). Efficient and robust persistent homology for measures. In *Proceedings of the 26th ACM-SIAM symposium on Discrete algorithms. SIAM*. SIAM.

Cadre, B. (2006). Kernel estimation of density level sets. *Journal of multivariate analysis*, 97(4):999–1023.

Cang, Z. and Wei, G. (2017). Topologynet: Topology based deep convolutional and multitask neural networks for biomolecular property predictions. *PLoS Computational Biology*, 13(7):e1005690.

Carlsson, G. (2009). Topology and data. *AMS Bulletin*, 46(2):255–308.

Carrière, M., Michel, B., and Oudot, S. (2017). Statistical analysis and parameter selection for mapper.

Carrière, M. and Oudot, S. (2015). Structure and stability of the 1-dimensional mapper. *arXiv preprint arXiv:1511.05823*.

Carriere, M. and Oudot, S. (2017). Sliced wasserstein kernel for persistence diagrams. To appear in ICML-17.

Chazal, F. (2017). High-dimensional topological data analysis. In *Handbook of Discrete and Computational Geometry (3rd Ed - To appear)*, chapter 27. CRC Press.

Chazal, F., Chen, D., Guibas, L., Jiang, X., and Sommer, C. (2011a). Data-driven trajectory smoothing. In *Proc. ACM SIGSPATIAL GIS*.

Chazal, F., Cohen-Steiner, D., Glisse, M., Guibas, L., and Oudot, S. (2009a). Proximity of persistence modules and their diagrams. In *SCG*, pages 237–246.

Chazal, F., Cohen-Steiner, D., Guibas, L. J., M'emoli, F., and Oudot, S. Y. (2009b). Gromov-hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum (proc. SGP 2009)*, pages 1393–1403.

Chazal, F., Cohen-Steiner, D., and Lieutier, A. (2009c). Normal cone approximation and offset shape isotopy. *Comp. Geom. Theor. Appl.*, 42(6-7):566–581.

Chazal, F., Cohen-Steiner, D., and Lieutier, A. (2009d). A sampling theory for compact sets in euclidean space. *Discrete & Computational Geometry*, 41(3):461–479.

Chazal, F., Cohen-Steiner, D., Lieutier, A., and Thibert, B. (2008). Stability of Curvature Measures. *Computer Graphics Forum (proc. SGP 2009)*, pages 1485–1496.

Chazal, F., Cohen-Steiner, D., and Mérigot, Q. (2010). Boundary measures for geometric inference. *Found. Comp. Math.*, 10:221–240.

Chazal, F., Cohen-Steiner, D., and Mérigot, Q. (2011b). Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751.

Chazal, F., de Silva, V., Glisse, M., and Oudot, S. (2016a). *The structure and stability of persistence modules*. SpringerBriefs in Mathematics. Springer.

Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2014a). Robust topological inference: Distance to a measure and kernel distance. *to appear in JMLR*.

Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2015a). Subsampling methods for persistent homology. To appear in *Proceedings of the 32 st International Conference on Machine Learning (ICML-15)*.

Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., and Wasserman, L. (2015b). Stochastic convergence of persistence landscapes and silhouettes. *Journal of Computational Geometry*, 6(2):140–161.

Chazal, F., Glisse, M., Labruère, C., and Michel, B. (2014b). Convergence rates for persistence diagram estimation in topological data analysis. To appear in *Journal of Machine Learning Research*.

Chazal, F., Guibas, L. J., Oudot, S. Y., and Skraba, P. (2013). Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):41.

Chazal, F., Huang, R., and Sun, J. (2015c). Gromov—hausdorff approximation of filamentary structures using reeb-type graphs. *Discrete Comput. Geom.*, 53(3):621–649.

Chazal, F. and Lieutier, A. (2008a). Smooth manifold reconstruction from noisy and non-uniform approximation with guarantees. *Comp. Geom. Theor. Appl.*, 40(2):156–170.

Chazal, F. and Lieutier, A. (2008b). Smooth manifold reconstruction from noisy and non uniform approximation with guarantees. *Computational Geometry Theory and Applications*, 40:156–170.

Chazal, F., Massart, P., and Michel, B. (2016b). Rates of convergence for robust geometric inference. *Electron. J. Statist*, 10:2243–2286.

Chazal, F. and Oudot, S. Y. (2008). Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, SCG '08, pages 232–241, New York, NY, USA. ACM.

Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2015). Density level sets: Asymptotics, inference, and visualization. *arXiv preprint arXiv:1504.05438*.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2005). Stability of persistence diagrams. In *SCG*, pages 263–271.

Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. (2010). Lipschitz functions have l p-stable persistence. *Foundations of computational mathematics*, 10(2):127–139.

Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Adv. in Appl. Probab.*, 36(2):340–354.

Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Advances in Applied Probability*, pages 340–354.

De Silva, V. and Carlsson, G. (2004). Topological estimation using witness complexes. In *Proceedings of the First Eurographics Conference on Point-Based Graphics*, SPBG'04, pages 157–166, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

De Silva, V. and Ghrist, R. (2007). Homological sensor networks. *Notices of the American mathematical society*, 54(1).

Devroye, L. and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, 38(3):480–488.

Dey, T. K., Mémoli, F., and Wang, Y. (2016). Multiscale mapper: topological summarization via codomain covers. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 997–1013. Society for Industrial and Applied Mathematics.

Dey, T. K., Mémoli, F., and Wang, Y. (2017). Topological analysis of nerves, reeb spaces, mappers, and multiscale mappers. In *Proc. Sympos. Comput. Geom. (SoCG)*.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533.

Fasy, B. T., Kim, J., Lecci, F., and Maria, C. (2014a). Introduction to the r package tda. *arXiv preprint arXiv:1411.1830*.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014b). Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339.

Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012). Manifold estimation and singular deconvolution under hausdorff loss. *Ann. Statist.*, 40:941–963.

Ghrist, R. (2017). Homological algebra and data. *preprint*.

Grove, K. (1993). Critical point theory for distance functions. In *Proc. of Symposia in Pure Mathematics*, volume 54.

Guibas, L., Morozov, D., and Mérigot, Q. (2013). Witnessed k-distance. *Discrete Comput. Geom.*, 49:22–45.

Hatcher, A. (2001). *Algebraic Topology*. Cambridge Univ. Press.

Hofer, C., Kwitt, R., Niethammer, M., and Uhl, A. (2017). Deep learning with topological signatures. *arXiv preprint arXiv:1707.04041*.

Kovacev-Nikolic, V., Bubenik, P., Nikolić, D., and Heo, G. (2016). Using persistent homology and dynamical distances to analyze protein binding. *Statistical applications in genetics and molecular biology*, 15(1):19–38.

Kramar, M., Goullet, A., Kondic, L., and Mischaikow, K. (2013). Persistence of force networks in compressed granular media. *Physical Review E*, 87(4):042207.

Kramár, M., Levanger, R., Tithof, J., Suri, B., Xu, M., Paul, M., Schatz, M. F., and Mischaikow, K. (2016). Analysis of kolmogorov flow and rayleigh–bénard convection using persistent homology. *Physica D: Nonlinear Phenomena*, 334:82–98.

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2017). Kernel method for persistence diagrams via kernel embedding and weight factor. *arXiv preprint arXiv:1706.03472*.

Lee, Y., Barthel, S. D., Dłotko, P., Moosavi, S. M., Hess, K., and Smit, B. (2017). Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8.

Li, C., Ovsjanikov, M., and Chazal, F. (2014). Persistence-based structural recognition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2003–2010.

Lum, P., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific reports*, 3.

Maria, C., Boissonnat, J.-D., Glisse, M., and Yvinec, M. (2014). The gudhi library: Simplicial complexes and persistent homology. In *International Congress on Mathematical Software*, pages 167–174. Springer.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. SciPy Austin, TX.

Mileyko, Y., Mukherjee, S., and Harer, J. (2011). Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007.

Nakamura, T., Hiraoka, Y., Hirata, A., Escolar, E. G., and Nishiura, Y. (2015). Persistent homology and many-body atomic structure for medium-range order in the glass. *Nanotechnology*, 26(30):304001.

Niyogi, P., Smale, S., and Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441.

Niyogi, P., Smale, S., and Weinberger, S. (2011). A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40(3):646–663.

Obayashi, I. and Hiraoka, Y. (2017). Persistence diagrams with linear machine learning models. *arXiv preprint arXiv:1706.10082*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Petrunin, A. (2007). Semiconcave functions in Alexandrov's geometry. In *Surveys in differential geometry. Vol. XI*, pages 137–201. Int. Press, Somerville, MA.

Phillips, J. M., Wang, B., and Zheng, Y. (2014). Geometric inference on kernel density estimates. *arXiv preprint 1307.7760*.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, pages 855–881.

Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. (2015). A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4741–4748.

Seversky, L. M., Davis, S., and Berger, M. (2016). On time-series topological data analysis: new data and opportunities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 59–67.

Singh, A., Scott, C., and Nowak, R. (2009). Adaptive Hausdorff estimation of density level sets. *Ann. Statist.*, 37(5B):2760–2782.

Singh, G., Mémoli, F., and Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100. Citeseer.

Skraba, P., Ovsjanikov, M., Chazal, F., and Guibas, L. (2010). Persistence-based segmentation of deformable shapes. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 45–52.

Tsybakov, A. B. et al. (1997). On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969.

Turner, K., Mileyko, Y., Mukherjee, S., and Harer, J. (2014a). Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70.

Turner, K., Mukherjee, S., and Boyer, D. M. (2014b). Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4):310–344.

Umeda, Y. (2017). Time series classification via topological data analysis. *Transactions of the Japanese Society for Artificial Intelligence*, 32(3):D–G72_1.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

Walt, S. v. d., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30.

Yao, Y., Sun, J., Huang, X., Bowman, G. R., Singh, G., Lesnick, M., Guibas, L. J., Pande, V. S., and Carlsson, G. (2009). Topological methods for exploring low-density states in biomolecular folding pathways. *The Journal of chemical physics*, 130(14):144115.

Zomorodian, A. and Carlsson, G. (2005). Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274.