HAL
open science

High-Dimensional Topological Data Analysis

Frédéric Chazal

To cite this version:

Frédéric Chazal. High-Dimensional Topological Data Analysis. 3rd Handbook of Discrete and Computational Geometry, CRC Press, 2016. hal-01316989

HAL Id: hal-01316989

https://inria.hal.science/hal-01316989v1

Submitted on 17 May 2016

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

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

27 HIGH-DIMENSIONAL TOPOLOGICAL DATA ANALYSIS

Frédéric Chazal

INTRODUCTION

Modern data often come as point clouds embedded in high dimensional Euclidean spaces, or possibly more general metric spaces. They are usually not distributed uniformly, but lie around some highly nonlinear geometric structures with nontrivial topology. *Topological data analysis* (TDA) is an emerging field whose goal is to provide mathematical and algorithmic tools to understand the topological and geometric structure of data. This chapter provides a short introduction to this new field through a few selected topics. The focus is deliberately put on the mathematical foundations rather than specific applications, with a particular attention to stability results asserting the relevance of the topological information inferred from data.

The chapter is organized in four sections. Section 27.1 is dedicated to distance-based approaches that establishes the link between TDA and curve and surface reconstruction in computational geometry. Section 27.2 considers homology inference problems and introduces the idea of interleaving of spaces and filtrations, a fundamental notion in TDA. Section 27.3 is dedicated to the use of persistent homology and its stability properties to design robust topological estimators in TDA. Section 27.4 briefly presents a few other settings and applications of TDA, including dimensionality reduction, visualization and simplification of data.

27.1 GEOMETRIC INFERENCE AND RECONSTRUCTION

Topologically correct reconstruction of geometric shapes from point clouds is a classical problem in computational geometry. The case of smooth curve and surface reconstruction in $\mathbb{R}^3$ has been widely studied over the last two decades and has given rise to a wide range of efficient tools and results that are specific to dimension 2 and 3; see Chapter 36. Geometric structures underlying data often appear to be of higher dimension and much more complex than smooth manifolds. This section presents a set of techniques based on the study of distance-like functions leading to general reconstruction and geometric inference results in any dimension.

GLOSSARY

Homotopy equivalence: Given two topological spaces X and Y, two maps $f_0, f_1 : X \to Y$ are *homotopic* if there exists a continuous map $H : [0,1] \times X \to Y$ such that for all $x \in X$, $H(0,x) = f_0(x)$ and $H(1,x) = f_1(x)$. The two spaces X

and Y are said to be *homotopy equivalent*, or to *have the same homotopy type* if there exist two continuous maps $f : X \to Y$ and $g : Y \to X$ such that $g \circ f$ is homotopic to the identity map in X and $f \circ g$ is homotopic to the identity map in Y.

Isotopy: Given $X, Y \subseteq \mathbb{R}^d$, an *(ambient) isotopy* between X and Y is a continuous map $F : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ such that $F(.,0)$ is the identity map on $\mathbb{R}^d$, $F(X,1) = Y$ and for any $t \in [0,1]$, $F(.,t)$ is an homeomorphism of $\mathbb{R}^d$.

Probability measure: A *probability measure* μ on $\mathbb{R}^d$ is a function mapping every (Borel) subset B of $\mathbb{R}^d$ to a nonnegative number $\mu(B)$ such that whenever $(B_i)_{i\in I}$ is a countable family of disjoint (Borel) subsets, then $\mu(\cup_{i\in I} B_i) = \sum_{i\in I} \mu(B_i)$, and $\mu(\mathbb{R}^d) = 1$. The *support* of μ is the smallest closed set S such that $\mu(\mathbb{R}^d \backslash S) = 0$. Probability measures are similarly defined on metric spaces.

Hausdorff distance: Given a compact subset $K \subset \mathbb{R}^d$, the distance function from K, $d_K : \mathbb{R}^d \to [0, +\infty)$, is defined by $d_K(x) = \inf_{y\in K} d(x,y)$. The Hausdorff distance between two compact subsets $K, K' \subset \mathbb{R}^d$ is defined by $d_H(K,K') = \|d_K - d_{K'}\|_\infty = \sup_{x\in\mathbb{R}^d} |d_K(x) - d_{K'}(x)|$.

DISTANCE-BASED APPROACHES AND GEOMETRIC INFERENCE

The general problem of geometric inference can be stated in the following way: given an *approximation* P (e.g., a point cloud) of a *geometric object* K in $\mathbb{R}^d$, is it possible to reliably and efficiently estimate the topological and geometric properties of K? Obviously, it needs to be instantiated in precise frameworks by defining the class of geometric objects that are considered and the notion of distance between these objects. The idea of distance-based inference is to associate to each object a real valued function defined on $\mathbb{R}^d$ such that the sublevel sets of this function carry some geometric information about the object. Then, proving geometric inference results boils down to the study of the stability of the sublevel sets of these functions under perturbations of the objects.

Distance to compact sets and distance-like functions. A natural and classical example is to consider the set of compact subsets of $\mathbb{R}^d$, which includes both continuous shapes and point clouds. The space of compact sets is endowed with the Hausdorff distance and to each compact set $K \subset \mathbb{R}^d$ is associated its distance function $d_K : \mathbb{R}^d \to [0, +\infty)$ defined by $d_K(x) = \inf_{y\in K} d(x,y)$. The properties of the ***r-offsets*** $K^r = d_K^{-1}([0,r])$ of K (i.e., the union of the balls of radius r centered on K) can then be used to compare and relate the topology of the offsets of compact sets that are close to each other with respect to the Hausdorff distance. When the compact K is a smooth submanifold, this leads to basic methods for the estimation of the homology and homotopy type of K from an approximate point cloud P, under mild sampling conditions [NSW08, CL08]. This approach extends to a larger class of nonsmooth compact sets K and leads to stronger results on the inference of the isotopy type of the offsets of K [CCSL09a]. It also leads to results on the estimation of other geometric and differential quantities such as normals [CCSL09b], curvatures [CCSLT09] or boundary measures [CCSM10] from shapes sampled with a moderate amount of noise (with respect to Hausdorff distance).

These results mainly rely on the ***stability*** of the map associating to a compact set K its distance function d_K (i.e., $\|d_K - d_{K'}\|_\infty = d_H(K,K')$ for any compact sets $K, K' \subset \mathbb{R}^d$) and on the *1-semiconcavity* of the squared distance function d_K^2 (i.e.,

the convexity of the map $x \to \|x\|^2 - d_K^2(x))$ motivating the following definition.

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

The 1-semiconcavity property of a distance-like function φ allows to define its gradient vector field $\nabla\varphi : \mathbb{R}^d \to \mathbb{R}^d$. Although not continuous, this gradient vector field can be integrated [Pet06] into a continuous flow that is used to compare the geometry of the sublevel sets of two close distance functions. In particular, the topology of the sublevel sets of a distance-like function φ can only change at levels corresponding to critical points, i.e. points x such that $\|\nabla_x\varphi\| = 0$:

LEMMA 27.1.1 Isotopy Lemma [Gro93, Proposition 1.8]
Let φ be a distance-like function and $r_1 < r_2$ be two positive numbers such that φ has no critical point in the subset $\varphi^{-1}([r_1, r_2])$. Then all the sublevel sets $\varphi^{-1}([0, r])$ are isotopic for $r \in [r_1, r_2]$.

This result suggests the following definitions.

Definition: Let φ be a distance-like function. We denote by $\varphi^r = \varphi^{-1}([0, r])$ the r sublevel set of φ.

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

Note that the isotopy lemma implies that all the sublevel sets of φ between r and $r + \mathrm{wfs}_\varphi(r)$ have the same topology. Comparing two close distance-like functions, if φ and ψ are two distance-like functions, such that $\|\varphi - \psi_\infty\| \leq \varepsilon$ and $\mathrm{wfs}_\varphi(r) > 2\varepsilon$, $\mathrm{wfs}_\psi(r) > 2\varepsilon$, then for every $0 < \eta \leq 2\varepsilon$, $\varphi^{r+\eta}$ and $\psi^{r+\eta}$ have the same homotopy type. An improvement of this result leads to the following reconstruction theorem from [CCSM11].

THEOREM 27.1.2 Reconstruction Theorem
Let φ, ψ be two distance-like functions such that $\|\varphi - \psi\|_\infty < \varepsilon$, with $\mathrm{reach}_\alpha(\varphi) \geq R$ 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, and that their normals and curvatures can be compared [CCSL09b, CCSLT09].

As an example, distance functions from compact sets are obviously distance-like and the above reconstruction result gives the following result.

THEOREM 27.1.3 *Let $K \subset \mathbb{R}^d$ be a compact set and let $\alpha \in (0, 1]$ be such that $r_\alpha = \mathrm{reach}_\alpha(d_K) > 0$. If $P \subset \mathbb{R}^d$ such that $d_H(K, P) \leq \kappa\alpha$ with*

$\kappa < \alpha^2/(5\alpha^2+12)$, then the offsets K^r and $P^{r'}$ are homotopy equivalent when

$$0 < r < r_\alpha \quad \text{and} \quad \frac{4\mathrm{d_H(P,K)}}{\alpha^2} \leq \mathrm{r}' \leq \mathrm{r}_\alpha - 3\mathrm{d_H(P,K)}.$$

In particular, if K is a smooth submanifold of $\mathbb{R}^d$, then $r_1 > 0$ and $P^{r'}$ is homotopy equivalent to K.

It is interesting to notice that indeed, distance-like functions are closely related to distance functions from compact sets: any distance-like function $\varphi : \mathbb{R}^d \to \mathbb{R}^+$ is the restriction to a hyperplane of the distance function from a compact set in $\mathbb{R}^{d+1}$ [CCSM11, Prop.3.1].

DTM AND KERNEL DISTANCES: THE MEASURE POINT OF VIEW

The major drawback of the geometric inference framework derived from the Hausdorff distance and distances between compact sets is its instability in the presence of outliers in the approximate data (i.e., points that are not close to the underlying geometric object). One way to circumvent this problem is to consider the approximate data as an empirical measure (i.e., a weighted sum of Dirac measures centered on the data points) rather than a point cloud, and to consider the probability measures on $\mathbb{R}^d$ instead of the compact subsets of $\mathbb{R}^d$ as the new class of geometric objects.

As the distance between a point $x \in \mathbb{R}^d$ and a compact set K is defined as the radius of the smallest ball centered at x and containing a point of K, a basic and natural idea to associate a distance-like function to a probability measure is to mimic this definition in the following way: given a probability measure μ and a parameter $0 \leq l < 1$, define the function $\delta_{\mu,l} : \mathbb{R}^d \to \mathbb{R}_+$ by

$$\delta_{\mu,l} : x \in \mathbb{R}^d \mapsto \inf\{r > 0 : \mu(\bar{B}(x,r)) > l\}$$

where $\bar{B}(x,r)$ is the closed Euclidean ball of radius r with center x. Unfortunately, the map $\mu \to \delta_{\mu,l}$ turns out to be, in general, not continuous for standard metrics on the space of probability measures. This continuity issue is fixed by averaging over the parameter l.

Definition: Let μ be a probability measure on $\mathbb{R}^d$, and $m \in (0,1]$ be a positive parameter. The function defined by

$$d^2_{\mu,m} : \mathbb{R}^d \to \mathbb{R}_+,\ x \mapsto \frac{1}{m}\int_0^m \delta_{\mu,l}(x)^2 dl$$

is called the *distance-to-measure (DTM) function to μ with parameter m.*

From a practical point of view, if $P \subset \mathbb{R}^d$ is a finite point cloud and $\mu = \frac{1}{|P|}\sum_{x\in P}\delta_x$ is the uniform measure on P then for any x the function $l \to \delta_{\mu,l}(x)$ is constant on the intervals $(k/|P|, (k+1)/|P|)$ and equal to the distance between x and its kth nearest neighbor in P. As an immediate consequence for $m = k/|P|$,

$$d^2_{\mu,m}(x) = \frac{1}{k}\sum_{i=1}^{k} \|x - X_{(i)}(x)\|^2$$

where $X_{(i)}(x)$ is the ith nearest neighbor of x in P. In other word, $d^2_{\mu,m}(x)$ is just the average of the squared distances from x to its first k nearest neighbors.

Distance-to-measure functions turn out to be distance-like; see Theorem 27.1.4 for distance-to-measures below. The application of Theorem 27.1.2 of the previous section to DTM functions require stability properties relying on a well-chosen metric on the space of measures. For this reason, the space of probability measures is equipped with a so-called *Wasserstein distance* W_p ($p \geq 1$) whose definition relies on the notion of transport plan between measures which is strongly related to the theory of optimal transport [Vil03].

A *transport plan* between two probability measures μ and ν on $\mathbb{R}^d$ is a probability measure π on $\mathbb{R}^d \times \mathbb{R}^d$ such that for every $A, B \subseteq \mathbb{R}^d$ $\pi(A \times \mathbb{R}^d) = \mu(A)$ and $\pi(\mathbb{R}^d \times B) = \nu(B)$. Intuitively $\pi(A \times B)$ corresponds to the amount of mass of μ contained in A that will be transported to B by the transport plan. Given $p \geq 1$, the p-cost of such a transport plan π is given by

$$\mathcal{C}_p(\pi) = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, d\pi(x, y) \right)^{1/p}$$

This cost is finite when the measures μ and ν both have finite *p-moments*, i.e. $\int_{\mathbb{R}^d} \|x\|^p \, d\mu(x) < +\infty$ and $\int_{\mathbb{R}^d} \|x\|^p \, d\nu(x) < +\infty$. The set of probability measures on $\mathbb{R}^d$ with finite p-moment includes all probability measures with compact support, such as, e.g., empirical measures. The *Wasserstein distance* of order p between two probability measures μ and ν on $\mathbb{R}^d$ with finite p-moment is the minimum p-cost $\mathcal{C}_p(\pi)$ of a transport plan π between μ and ν. It is denoted by $W_p(\mu, \nu)$.

For geometric inference, the interest of Wasserstein distance comes from its weak sensibility to the presence of a small number of outliers. For example, consider a reference point cloud P with N points, and define a noisy version P' by replacing n points in P by points $o_1, \ldots, o_n$ such that $d_P(o_i) \geq R$ for some $R > 0$. Considering the cost of the transport plan between P' and P that moves the outliers back to their original position, and keeps the other points fixed we get $W_p(\mu_P, \mu_{P'}) \leq \frac{n}{N}(R + \mathrm{diam}(P))$ while the Hausdorff distance between P and P' is at least R. Hence, if the number of outliers is small, i.e. $n \ll N$, the Wasserstein distance between μ_P and $\mu_{P'}$ remains small. Moreover, if the N points of P are independently drawn from a common measure μ then μ_P converges almost surely to μ in the Wasserstein metric W_p (see [BGV07] for precise statements).

THEOREM 27.1.4 Stability of distance-to-measures [CCSM11]

For any probability measure μ in $\mathbb{R}^d$ and $m \in (0, 1)$ the function $d_{\mu,m}$ is distance-like. Moreover, if ν is another probability measures on $\mathbb{R}^d$ and $m > 0$, then

$$\|d_{\mu,m} - d_{\nu,m}\|_\infty \leq \frac{1}{\sqrt{m}} \, \mathrm{W}_2(\mu, \nu).$$

This theorem allows to apply the reconstruction theorem (Theorem 27.1.2) to recover topological and geometric information of compact shapes from noisy data containing outliers [CCSM11, Cor. 4.11].

More recently, a new family of distance-like functions associated to probability measures, called *kernel distances*, has been introduced in [PWZ15] that are closely related to classical kernel-based density estimators. They offer similar, but complementary, properties as the DTM functions and come with stability properties ensuring the same topological guarantees for topological and geometric inference.

Probabilistic and statistical considerations. The distance-based approach is well-suited to explore reconstruction and geometric inference from a statistical perspective, in particular when data are assumed to be randomly sampled. The problem of approximation of smooth manifolds with respect to the Hausdorff distance from random samples under different models of noise has been studied in [GPP+12a, GPP+12b]. The statistical analysis of DTM and kernel distances remains largely unexplored despite a few recent preliminary results [CMM15, CFL+14]; see also the open problems below.

Some open problems. Here are a few general problems related to distance-based approach that remain open or partly open.

1. The computation of the DTM at a given point only require to compute nearest neighbors but the efficient global computation of the DTM, e.g. to obtain its sublevel sets or its persistent homology, turns out to have prohibitive complexity as it is closely related to the computation of higher order Voronoi diagrams. The difficulty of efficiently approximating the DTM function is still rather badly understood despite a few results in this direction [BCOS15, GMM13, Mér13]; see also Chapter 29.

2. The dependence of DTM functions on the parameter m raises the problem of the choice of this parameter. The same problem also occurs with the kernel distances that depend on a bandwith parameter. Very little is known about the dependency of DTM on m (the situation is slightly better for the kernel distances) and data driven methods to chose these parameters still need to be developed. Preliminary results in this direction have recently been obtained in [CMM15, CFL+14].

RECONSTRUCTION IN HIGH DIMENSION

Although the above mentioned approaches provide general frameworks for geometric inference in any dimension, they do not directly lead to efficient reconstruction algorithms. Here, a reconstruction algorithm is meant to be an algorithm that:
- takes as input a finite set of points P sampled from an unknow shape K,
- outputs a triangulation or a simplicial complex that approximates K, and
- provides a topologically correct reconstruction (i.e., homeomorphic or isotopic to K) when certain sampling conditions quantifying the quality of the approximation of K by P are satisfied.

Efficient algorithms with such guarantees is possible if we restrict to specific classes of shapes to reconstruct.

- **Low dimensional smooth manifolds in high dimension:** except for the case of curve and surface reconstruction in $\mathbb{R}^2$ and $\mathbb{R}^3$; see Chapter 36, the attempts to develop effective reconstruction algorithms for smooth manifolds in arbitrary dimension remain quite limited. Extending smooth manifold reconstruction algorithms in $\mathbb{R}^3$ to $\mathbb{R}^d$, $d > 3$, raises several major difficulties. In particular, important topological properties of restricted Delaunay triangulations used for curve and surface reconstruction are no longer hold in higher dimensions preventing direct generalization of the existing low dimensional algorithms. Moreover classical data structures involved in reconstruction algorithms, such as the Delaunay triangulation, are global and have their complexity depending exponentially on the ambient dimension that make them

almost intractable in dimension larger than 3. However, a few attempts have been made to overcome these issues. In [BGO09], using the so-called ***witness complex*** [SC04], the authors design a reconstruction algorithm whose complexity scales up with the intinsic dimension of the submanifold. More recently, a new data structure, the *tangential Delaunay complex*, has been introduced and used to design effective reconstruction algorithms for smooth low dimensional submanifolds of $\mathbb{R}^d$ [BG13].

- **Filamentary structures and stratified spaces:** 1-dimensional filamentary structures appear in many domains (road networks, network of blood vessels, astronomy, etc.) and can be modeled as 1-dimensional stratified sets, or (geometric) graphs. Various methods, motivated and driven by specific applications have been developed to reconstruct such structures from point cloud data. From a general perspective the (relatively) simple structure of graphs allows to propose new approaches to design metric graph reconstruction algorithms coming with various topological guarantees, e.g. homeomorphy or homotopy type and closeness in the Gromov-Hausdorf metric [GSBW11, ACC+12, CHS15]. Despite a few attempts [BCSE+07, BWM12], reconstruction of stratified sets of higher dimension turns out to be a much more difficult problem that remains largely open.

27.2 HOMOLOGY INFERENCE

The results on geometric inference from the previous section provide a general theoretical framework to "reconstruct" unknown shapes from approximate data. However, it is not always desirable to fully reconstruct a geometric object to infer some relevant topological properties from data. This is illustrated in this section by two examples. First, we consider a weaker version of the reconstruction paradigm where the goal is to infer topological invariants, more precisely homology and Betti numbers. Second, we consider coverage problems in sensor networks that can be answered using homology computations. Both examples rely on the idea that relevant topological information cannot always be directly inferred from the data at a given scale, but by considering how topological features relate to each other across different scales. This fundamental idea raises the notion of interleaving between spaces and filtrations and leads to persistence-based methods in TDA that are considered in the next section.

GLOSSARY

Abstract simplicial complex: Given a set X, an abstract simplicial complex C with vertex set X is a set of finite subsets of X, the simplices, such that the elements of X belong to C and if $\sigma \in C$ and $\tau \subset \sigma$ then $\tau \in C$.

Homology: Intuitively, homology (with coefficient in a field) associates to any topological space X, a family of vector spaces, the so-called homology groups $H_k(X)$, $k = 0, 1, \ldots$, each of them encoding k-dimensional topological features of X. A fundamental property of homology is that any continuous function $f : X \to Y$ induces a linear map $f_* : H_k(X) \to H_k(Y)$ between homology groups that encodes the way the topological features of X are mapped to the

topological features of Y by f. This linear map is an isomorphism when f is an homeomorphism or an homotopy equivalence (homology is thus a homotopy invariant). See [Hat01] or Chapter 23 for a formal definition.

Betti numbers: The kth Betti number of X, denoted $\beta_k(X)$, is the rank of $H_k(X)$ and represents the number of "independent" k-dimensional features of X: for example, $\beta_0(X)$ is the number of connected components of M, $\beta_1(X)$ the number of independent cycles or tunnels, $\beta_2(X)$ the number of cavities, etc.

Čech complex: Given P a subset of a metric space X and $r > 0$, the Čech complex $\check{\text{C}}\text{ech}(P, r)$ built on top of P, with parameter r is the abstract simplicial complex defined as follows: (i) the vertices of $\check{\text{C}}\text{ech}(P, r)$ are the points of P and (ii) $\sigma = [p_0, \ldots, p_k] \in \check{\text{C}}\text{ech}(P, r)$ if and only if the intersection of balls of radius r and centered at the p_i's have nonempty intersection.

Vietoris-Rips complex: Given a metric space (X, d_X) and $r \geq 0$, the Vietoris-Rips complex $\text{Rips}(X, r)$ is the (abstract) simplicial complex defined by i) the vertices of $\text{Rips}(X, r)$ are the points of X and, ii) $\sigma = [x_0, \ldots, x_k] \in \text{Rips}(X, r)$ if and only if $d_X(x_i, x_j) \leq r$ for any $i, j \in \{0, \ldots, k\}$.

ČECH COMPLEX, VIETORIS-RIPS COMPLEX AND HOMOLOGY INFERENCE

An important advantage of simplicial complexes is that they are not only combinatorial objects but they can also be seen as topological spaces. Let C be a finite simplicial complex with vertex set $X = \{x_1, \ldots, x_n\}$. Identifying each x_i with the point e_i of $\mathbb{R}^n$ whose all coordinates are 0 except the ith which is equal to 1, one can identify each simplex $\sigma = [x_{i_0}, \cdot x_{i_k}] \in C$ with the convex hull of the points $e_{i_0}, \cdot e_{i_k}$. The union of these sets inherits a topology as a subset of $\mathbb{R}^n$ and is called the geometric realisation of C in $\mathbb{R}^n$. In the following, the topology or the homotopy type of a simplicial complex refers to the ones of its geometric realisation.

Thanks to this double nature, simplicial complexes play a fundamental role to bridge the gap between continuous shapes and their discrete representations. In particular, the classical nerve theorem [Hat01][Corollary 4G3] is fundamental in TDA to relate continuous representation of shapes to discrete description of their topology through simplicial complexes.

Definition: Let X be a topological space and let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of X, i.e. a family of open subset such that $X = \cup_{i \in I} U_i$. The *nerve of* $\mathcal{U}$, denoted $N(\mathcal{U})$ is the (abstract) simplicial complex defined by:
(i) the vertices of $N(\mathcal{U})$ are the U_i's,
(ii) $\sigma = [U_{i_0}, \ldots, U_{i_k}] \in N(\mathcal{U})$ if and only if $\bigcap_{j=0}^{k} U_{i_j} \neq \emptyset$.

THEOREM 27.2.1 Nerve Theorem

Let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of a paracompact topological space X. If any nonempty intersection of finitely many sets in $\mathcal{U}$ is contractible, then X and $N(\mathcal{U})$ are homotopy equivalent. In particular, their homology groups are isomorphic.

An immediate consequence of the Nerve theorem is that under the assumption of Theorem 27.1.2, the computation of the homology of a smooth submanifold $K \subset \mathbb{R}^d$ approximated by a finite point cloud P boils down to the computation of the homology of the Čech complex $\check{\text{C}}\text{ech}(P, r)$ for some well-chosen radius r.

However this direct approach suffers from several drawbacks: first, computing the nerve of a union of balls requires the extensive use of an awkward predicate testing the nonemptiness of the intersection of finite sets of balls; second, the suitable choice of the radius r relies on the knowledge of the reach of K and of the Hausdorff distance between K and P that are usually not available. Moreover, the assumption that the underlying shape K is a smooth manifold is often too restrictive in practical applications. To overcome this latter restriction, [CL07, CSEH07] consider the linear map $H_k(P^\varepsilon) \to H_k(P^{3\varepsilon})$ induced by the inclusion $P^\varepsilon \hookrightarrow P^{3\varepsilon}$ of small offsets of P and prove that its rank is equal to the kth Betti number of K^δ when $d_H(K,P) < \varepsilon < \mathrm{wfs}(K)/4$ and $0 < \delta < \mathrm{wfs}(K)$, where $\mathrm{wfs}(K) = \mathrm{wfs}_{d_k}(0)$ is the infimum of the positive critical values of d_K. The idea of using nested pairs of offsets to infer the homology of compact sets was initially introduced in [Rob99] for the study of attractors in dynamical systems. Beyond homology, the inclusion $P^\varepsilon \hookrightarrow P^{3\varepsilon}$ also induces group morphisms between the homotopy groups of theses offsets whose images are isomorphic to the homotopy groups of K^δ [CL07]. Homotopy inference and the use of homotopy information in TDA raise deep theoretical and algorithmic problems and remains rather unexplored despite a few attempts such as, e.g. [BM13].

The homotopy equivalences between $P^\varepsilon, P^{3\varepsilon}$ and $\check{\mathrm{C}}\mathrm{ech}(P^\varepsilon), \check{\mathrm{C}}\mathrm{ech}(P^{3\varepsilon})$ respectively given by the nerve theorem can be chosen in such a way that they commute with the inclusion $P^\varepsilon \hookrightarrow P^{3\varepsilon}$, leading to an algorithm for homology inference based upon the Čech complex. To overcome the difficulty raised by the computation of the Čech complex, [CO08] proposes to replace it by the Vietoris-Rips complex. Using the elementary interleaving relation

$$\check{\mathrm{C}}\mathrm{ech}(P, r/2) \subseteq \mathrm{Rips}(P, r/2) \subset \check{\mathrm{C}}\mathrm{ech}(P, r)$$

one easily obtains that, for any integer $k = 0, 1 \ldots$, the rank of the linear map $H_k(\mathrm{Rips}(P,\varepsilon)) \to H_k(\mathrm{Rips}(P, 4\varepsilon))$ is equal to that of $H_k(K^\delta)$ when $2d_H(P,K) < \varepsilon < (\mathrm{wfs}(K) - d_H(P,K))/4$ and $0 < \delta < \mathrm{wfs}(K)$. A similar result also holds for witness complexes built on top of the input data P. To overcome the problem of the choice of the Vietoris-Rips parameter ε, a greedy algorithm is proposed in [CO08] that maintains a nested sequence of Vietoris-Rips complexes and eventually compute the Betti numbers of the offsets K^δ for various relevant scales δ. When K is an m-dimensional smooth submanifold of $\mathbb{R}^d$ this algorithm recovers the Betti numbers of K in times at most $c(m)n^5$ where $n = |P|$ and $c(m)$ is a constant depending exponentially on m and linearly on d. Precise information about the complexity of the existing homology inference algorithms is available in [Oud15, Chapter 4].

From a statistical perspective, when K is a smooth submanifold and P is a random sample, the estimation of the homology has been considered in [NSW08, NSW11] while [BRS$^+$12] provide minimax rates of convergence.

COVERAGE PROBLEMS IN SENSOR NETWORKS

Given sensors located at a set of nodes $P = \{p_1, \ldots, p_n\} \subset \mathbb{R}^d$ spread out in a bounded region $D \subset \mathbb{R}^d$, assume that each sensor can sense its environment within a disc of fixed *covering radius* $r_c > 0$. Basic coverage problems in sensor network address the question of the full coverage of D by the sensing areas covered by the sensor. When the exact position of the nodes is not known but only the

graph connecting sensors within distance less than some *communication radius* $r_c > 0$ from each other, Vietoris-Rips complexes appear as a natural tool to infer topological information about the covered domain. Following this idea, [SG07a, SG07b] propose to use the homology of nested pairs of such simplicial complexes to certify that the domain D is covered by the union of the covering discs in various settings.

More precisely, assume that each node can detect and communicate with other nodes via a strong signal within radius $r_s > 0$ and via a weak signal within a radius $r_s > 0$, respectively, such that $r_c \geq r_s/\sqrt{2}$ and $r_w \geq r_s\sqrt{10}$. Assume moreover that the nodes can detect the presence of the boundary ∂D within a *fence detection radius* r_f and denote by $F \subset P$ the set of nodes that are at distance at most r_f from ∂D. Regarding the domain D, assume that $D \setminus (\partial D)^{r_f + r_s/\sqrt{2}}$ is connected and the injectivity radius of the hypersurface $d_{\partial D}^{-1}(r_f)$ is larger than r_s. Then [SG07a] introduces the following criterion involving the relative homology of the pairs of Vietoris-Rips complexes built on top of F and P.

THEOREM 27.2.2 Coverage criterion

if the morphism between relative homology groups

$$i_\star : H_d(\mathrm{Rips}(P, F, r_s)) \to H_d(\mathrm{Rips}(P, F, r_w))$$

induced by the inclusion of the pairs of complexes

$$i : (\mathrm{Rips}(P, r_s), \mathrm{Rips}(F, r_s)) \hookrightarrow (\mathrm{Rips}(P, r_w), \mathrm{Rips}(F, r_w))$$

is nonzero, then $D \setminus (\partial D)^{r_f + r_s/\sqrt{2}}$ *is contained in the union of the balls of radius* r_c *with centers the points of* P.

This result has given rise to a large literature on topological methods in sensor networks. In particular, regarding the robustness of this criterion, its stability under perturbations of the networks is studied in [HK14]. Similar ideas, combined with zigzag persistent homology, have also been used to address other problems such as, e.g. the detection of evasion paths in mobile sensor networks [AC15].

27.3 PERSISTENCE-BASED INFERENCE

Beyond homology, persistent homology (see Chapter 23) plays a central role in topological data analysis. It is usually used in two different ways. It may be applied to functions defined on data in order to estimate topological features of these functions (number and relevance of local extrema, homology of sublevel sets, etc.). Persistent homology may also be applied to geometric filtrations built on top of the data in order to infer topological information about the global structure of data. These two ways give rise to two main persistence-based pipelines that are presented in the two next sections and illustrated in Figure 27.3.1. The resulting persistence diagrams are then used to reveal and characterize topological features for further data analysis tasks (classification, clustering, learning, etc.). From a theoretical perspective, the stability properties of persistent homology allow to establish the stability and thus the relevance of these features.

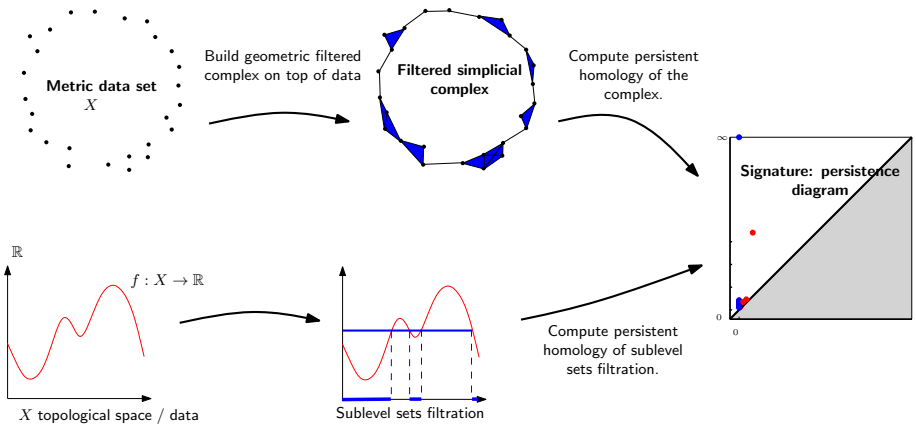

FIGURE 27.3.1
The classical pipelines for persistence in TDA.

GLOSSARY

Filtered simplicial complex: given a simplicial complex C and a finite or infinite subset $A \subset \mathbb{R}$, a *filtration of* C is a family $(C_\alpha)_{\alpha \in A}$ of subcomplexes of C such that for any $\alpha \leq \alpha'$, $C_\alpha \subseteq C_{\alpha'}$ and $C = \cup_{\alpha \in A} C_\alpha$.

Sublevel set filtration: Given a topological space X and a function $f : X \to \mathbb{R}$, the *sublevel set filtration of* f is the nested family of sublevel sets of f: $(f^{-1}((-\infty, \alpha]))_{\alpha \in \mathbb{R}}$.

Metric space: A metric space is a pair (X, d_X) where X is a set and $d_X : X \times X \to \mathbb{R}_+$ is a nonnegative map such that for any $x, y, z \in X$, $d_X(x, y) = d_X(y, x)$, $d_X(x, y) = 0$ if and only if $x = y$ and $d_X(x, z) \leq d_X(x, y) + d_X(y, z)$.

Gromov-Hausdorff distance: The Gromov-Hausdorff distance extends the notion of Hausdorff distance between compact subsets of a same metric spaces to general spaces. More precisely, given two compact metric spaces (X, d_X) and (Y, d_Y) and a third metric space (Z, d_Z), a map $\varphi : X \to Z$ (resp., $\psi : Y \to Z$) is an isometric embedding if for any $x, x' \in X$, $d_Z(\varphi(x), \varphi(x')) = d_X(x, x')$ (resp., any $y, y' \in Y$, $d_Z(\psi(y), \psi(y')) = d_Y(y, y')$). The *Gromov-Hausdorff distance* $d_{GH}(X, Y)$ between X and Y is defined as the infimum of the Hausdorff distances $d_H(\varphi(X), \psi(Y))$ where the infimum is taken over all the metric spaces (Z, d_Z) and all the isometric embeddings $\varphi : X \to Z$ and $\psi : Y \to Z$.

Persistent homology: Persistent homology provides a framework and efficient algorithms to encode the evolution of the homology of families of nested topological spaces (filtrations) indexed by a set of real numbers, such as the sublevel sets filtration of a function, a filtered complex, etc. These indices may often be seen as scales, as for example in the case of the Vietoris-Rips filtration where the index is the radius of the balls used to build the complex. Given a filtration $(F_\alpha)_{\alpha \in A}$, its homology changes as α increases: new connected components can appear, existing connected components can merge, cycles and cavities can appear or be filled, etc. Persistent homology tracks these changes, identifies features and associates, to each of them, an interval or lifetime from α_{birth} to α_{death}. For instance, a connected component is a feature that is born at the smallest α such that the component is present in F_α, and dies when it merges with an older connected component. Intuitively, the longer a feature persists, the more rele-

vant it is. The set of intervals representing the lifetime of the identified features is called the barcode of the filtration. As an interval can also be represented as a point in the plane with coordinates $(\alpha_{birth}, \alpha_{death})$, the set of points (with multiplicity) representing the intervals is called the persistence diagram of the filtration. See Chapter 23 for formal definitions.

Bottleneck distance: Given two persistence diagrams, D and D', the *bottleneck distance* $d_B(D, D')$ is defined as the infimum of $\delta \geq 0$ for which there exists a matching between the diagrams, such that two points can only be matched if their distance is less than δ and all points at distance more than δ from the diagonal must be matched. See Chapter 23, for more details.

PERSISTENCE OF SUBLEVEL SET FILTRATIONS

Persistent homology of sublevel set filtration of functions may be used from two different perspectives in TDA.

Collections of complex objects. When data are collections in which each element is already a "complex" geometric object such as, e.g., image or shape, functions defined on each data element may be used to highlight some of their features. The persistence diagrams of the sublevel set filtrations of such functions can be used for comparison and classification of the data elements. The bottleneck distance between the diagrams is then used as a measure of dissimilarity between the elements. The idea of using persistence of functions defined on images and shapes was first introduced in the setting of *size theory* where it was used for shape analysis [VUFF93]; see also [FL99] for a survey. These ideas do not restrict to images and shapes and can also be applied to other "geometric" data such as, for example, textures or hand gesture data [LOC14, RHBK15]. In practical applications the main difficulty of this approach is in the design of functions whose persistent homology provides sufficiently informative and discriminative features for further classification or learning tasks.

Scalar field analysis. Another problem arising in TDA is the estimation of the persistent homology of a function defined on a possibly unknown manifold, from a finite approximation. As an example, assume that we are given a collection of sensors spread out in some region and that these sensors measure some physical quantity, such as temperature or humidity. Assuming that the nodes do not know their geographic location but that they can detect which other nodes lie in their vicinity, the problem is then to recover global topological information about the physical quantity through the estimation of its persistence diagrams. Another example is the estimation of the persistence diagrams of a probability density function f defined on some domain from a finite set of points sampled according to f. The persistence diagram of f may be used to provide information about the modes (peaks) of f and their shape and prominence. More formally, the problem can be stated in the following way: given an unknown metric space X and a function $f : X \to \mathbb{R}$ whose values are known only at a finite set of sample points $P \subset X$, can we reliably estimate the persistence diagrams of the sublevel set filtration of f?

When X is a compact Riemannian manifold and f is a Lipschitz function, [CGOS11] provide an algorithm computing a persistence diagram whose bottleneck distance to the diagram of f is upperbounded by a function depending on the Lipschitz constant of f and on $d_H(P, X)$ when this latter quantity is smaller than

some geometric quantity, namely the so-called *convexity radius* of the manifold X. Applied to the case where f is a density estimate, this result has led to new clustering algorithms on Riemannian manifold where persistence is used to identify and characterize relevant clusters [CGOS13]. Applied to curvature-like functions on surfaces, it has also been used for shape segmentation [SOCG10]. As already noted in Section 27.1, the dependence of the quality of the estimated persistence diagrams on the Hausdorff distance $d_H(P, X)$ makes this approach very sensitive to data corrupted by outliers. Some recent attempts have been made to overcome this issue [BCD+15] but the existing results apply in only very restrictive settings and the problem remains largely open.

PERSISTENCE-BASED SIGNATURES

Relevant multiscale topological signatures of data can be defined using the persistent homology of filtered simplicial complexes built on top of the data. Formally, given a metric space (Y, d_Y)—the data—approximate a (possibly unknown) metric space (X, d_X) the idea is to build a filtered simplicial complex on top of Y whose homotopy type, homology or persistent homology is related to the one of X. Considering the Vietoris-Rips filtration $\mathbb{R}\mathrm{ips}(X)$, it was proven in [Hau95] that if X is a closed Riemannian manifold, then for any sufficiently small $\alpha > 0$, $\mathrm{Rips}(X, \alpha)$ is homotopy equivalent to X. This result was later generalized to prove that if (Y, d_Y) is close enough to (X, d_X) with respect to the Gromov-Hausdorff distance, then there exists $\alpha > 0$ such that $\mathrm{Rips}(Y, \alpha)$ is homotopy equivalent to X [Lat01]. Quantitative variants of this result were obtained in [ALS13] for a class of compact subsets of $\mathbb{R}^d$. Considering the whole filtration and its persistent homology allows to relax the assumtions made on X. For the Čech and Vietoris-Rips complexes, the following stability result holds in any compact metric space [CSO14].

THEOREM 27.3.1 Stability of persistence-based signatures
Let (X, d_X) and (Y, d_Y) be two compact metric spaces. Then

$$\mathrm{d_b}(\mathsf{dgm}(\mathrm{H}(\check{\mathbb{C}}\mathrm{ech}(X))), \mathsf{dgm}(\mathrm{H}(\check{\mathbb{C}}\mathrm{ech}(Y)))) \leq 2\mathrm{d_{GH}}(X, Y),$$
$$\mathrm{d_b}(\mathsf{dgm}(\mathrm{H}(\mathbb{R}\mathrm{ips}(X))), \mathsf{dgm}(\mathrm{H}(\mathbb{R}\mathrm{ips}(Y)))) \leq 2\mathrm{d_{GH}}(X, Y)$$

where $\mathsf{dgm}(\mathrm{H}(\check{\mathbb{C}}\mathrm{ech}(X)))$ (resp., $\mathsf{dgm}(\mathrm{H}(\mathbb{R}\mathrm{ips}(X)))$) denotes the persistence diagrams of the Čech (resp., Vietoris-Rips) filtrations built on top of X and $\mathrm{d_b}(.,.)$ is the bottleneck distance.

This result indeed holds for larger families of geometric complexes built on top of metric spaces, in particular for the so-called witness complexes [SC04], and also extends to spaces endowed with a dissimilarity measure (no need of the triangle inequality). Computing persistent homology of geometric filtrations built on top of data is a classical strategy in TDA; see for example [CISZ08] for an "historical" application.

A first version of Theorem 27.3.1, restricted to the case of finite metric spaces, is given in [CCSG+09] where it is applied to shape comparison and classification. From a practical perspective, the computation of the Gromov-Hausdorff distance between two metric spaces is in general out of reach, even for finite metric spaces with relatively small cardinality. The computation of persistence diagrams of geometric filtrations built on top of metric spaces thus provides a tractable way to

compare them. It is however important to notice that the size of the k-dimensional skeleton of geometric filtrations, such as the Rips-Vietoris or Čech complexes, built on top of n data points is $O(n^k)$ leading to severe practical restriction for their use. Various approaches have been proposed to circumvent this problem. From an algorithmic point of view, new data structures have been proposed to efficiently represent geometric filtrations [BM14] and compute their persistence; see Chapter 23. Other lighter filtrations have also been proposed, such as the graph induced complex [DFW15] or the sparse Rips complex [She13]. From a statistical point of view, subsampling and bootstrap methods have been proposed to avoid the prohibitive computation of the persistent homology on filtrations built on the whole data; see next paragraph. Despite these recent attempts, the practical computation of persistent homology of geometric filtrations built on top of large data set remains a severe issue.

STATISTICAL ANALYSIS OF PERSISTENCE-BASED SIGNATURES

In the context of data analysis, where data usually carries some noise and outliers, the study of persistent homology from a statistical perspective has recently attracted some interest. Assuming that the data $X_n = \{x_1, \ldots, x_n\}$ is an i.i.d. sample from some probability measure μ supported on a compact metric space (M, d_M), the persistence diagram of a geometric filtrations built on top of X_n becomes a random variable distributed according a probability measure in the space of persistence diagrams endowed with the bottleneck distance. Recent efforts have been made to understand and exploit the statistical properties of these distributions of diagrams. For example, building on the stability result for persistence-based signatures, [CGLM15] establish convergence rates for the diagrams built on top of X_n to the diagrams built on top of M as $n \to +\infty$. In the same direction, considering subsamples of fixed size m, [BGMP14, CFL$^+$15a] prove stability results for the associated distributions of diagrams under perturbations of the probability measure μ in the Gromov-Prohorov and Wasserstein metrics respectively. The latter results provide new promising methods for inferring persistence-based topological information that are resilient to the presence of noise and outliers in the data and that turn out to be practically efficient (persistent homology being computed on filtrations built on top of small fixed size subsamples).

More generally, a main difficulty in the use of persistent homology in statistical settings hinges on the fact that the space of persistence diagram is highly nonlinear. This makes the definition and computation of basic statistical quantities such as, e.g. means, nonobvious. Despite this difficulty it has been shown that several standard statistical notions and tools can still be defined and used with persistent diagrams, such as Fréchet means [MMH11], confidence sets [FLR$^+$14], or bootstrap techniques [CFL$^+$15b], etc. Attempts have also been made to find new representations of persistence diagrams as elements of linear spaces in which statistical tools are easier to handle. A particularly interesting contribution in this direction is the introduction of the notion of *persistence landscape*, a representation of persistence diagrams as a family of piecewise linear functions on the real line [Bub15].

27.4 OTHER APPLICATIONS OF TOPOLOGICAL METHODS IN DATA ANALYSIS

Topological Data Analysis has known an important development during the last decade and it now includes a broad spectrum of tools, methods and applications that go beyond the mathematical results presented in the first three sections of this paper. In this section, we present other directions in which TDA has been developed or applied.

VISUALIZATION AND DIMENSIONALITY REDUCTION

Beyond mathematical and statistical relevance, the efficient and easy-to-understand visualization of the topological and geometric structure of data is an important task in data analysis. The TDA toolbox proposes a few methods to represent and visualize some topological features of data.

Data visualization using Mapper. *Mapper* is a method to visualize high dimensional and complex data using simplicial complexes. Introduced in [SMC07], it relies on the idea that local and partial clustering of the data leads to a cover of the whole data whose nerve provides a simplified representation of the global structure. Given a data set X, a function $f : X \to \mathbb{R}$ and a finite cover $(I_i)_{i=1,\dots,n}$ of $f(X) \subset \mathbb{R}$ by a family of intervals, the Mapper method first clusterizes each preimage $f^{-1}(I_i)$, of the interval I_i to obtain a (finite) cover $U_1, \dots, U_{k_i}$ of $f^{-1}(I_i)$. The union of the obtained clusters for all the intervals I_i's is a cover of X and Mapper outputs a graph, the 1-skeleton of this cover. The method is very flexible as it leaves the choice of the function f, the cover $(I_i)_{i=1,\dots,n}$ and the clustering methods to the user. The output graph provides an easy to visualize representation of the structure of the data driven by the function f. The Mapper algorithm has been popularized and is widely used as a visualization tool to explore and discover hidden insights in high dimensional data sets; see, e.g. [Car09, LSL$^+$13] for a precise description and a discussion on the Mapper algorithm. When the length of the intervals I_i's is small, the output of Mapper can be seen as a discrete version of the Reeb graph of the function f. However, despite a few recent results, the theoretical analysis of the Mapper method and its formal connection with Reeb graph remain an open research area.

Morse theory. Other topological methods, including in particular Morse Theory, are also successfully used for data visualization, but in a rather different perspective than Mapper. The interested reader is referred to the following collection of books providing a good survey on the topic: [PTHT11, PHCF12, BHPP14].

Circular coordinates and dimensionality reduction. Non linear dimensionality reduction (NLDR) includes a set of techniques whose aim is to represent high dimensional data in low dimensional spaces while preserving the intrinsic structure of the data. Classical NLDR methods map the data in a low dimensional Euclidean space $\mathbb{R}^k$ assuming that real-valued coordinates are sufficient to correctly and efficiently parametrize the underlying structure M (which is assumed to be a manifold) of the data. More precisely, NLDR methods intend to infer a set of functions $f_1, \dots, f_k : M \to \mathbb{R}$ such the map $F = (f_1, \dots, f_k) : M \to \mathbb{R}^k$ is an embedding

preserving the geometric structure of M. As a consequence, the theoretical guarantees of NLDR methods require M to have a very simple geometry. For example, ISOMAP [TSL00] assumes M to be isometric to a convex open subset of $\mathbb{R}^k$. To enrich the class of functions used to parametrize the data, [SMVJ11] introduces a persistence-based method to detect and construct circular coordinates, i.e., functions $f : M \to \mathbb{S}^1$ where $\mathbb{S}^1$ is the unit circle. The approach relies on the classical property that $\mathbb{S}^1$ is the classifying space of the first cohomology group (with integer coefficients) $H^1(M, \mathbb{Z})$, i.e. $H^1(M, \mathbb{Z})$ is equal to the set of equivalence classes of maps from M to $\mathbb{S}^1$, where two maps are equivalent if they are homotopic [Hat01]. The method consists first in building a filtered simplicial complex on top of the data and using persistent cohomology to identify relevant, i.e. persistent, cohomology classes. Then a smooth (harmonic) cocycle is chosen in each of these classes and integrated to give a circular function on the data.

This approach opens the door to new NLDR methods combining real-valued and circle-valued coordinates. Using time-delay embedding of times series and time dependent data [Tak81], the circular coordinates approach also opens the door to new topological approaches in time series analysis [PH13, Rob14].

TOPOLOGICAL DATA ANALYSIS IN SCIENCES

Despite its youth, TDA has already led to promising applications and results in various domains of science and the TDA toolbox is now used with many different kind of data. The following list provides a short and nonexhaustive selection of domains where topological approaches appear to be particularly promising.

- Biology: biology is currently probably the largest field of application of TDA. There already exists a vast litterature using persistent homology and the Mapper algorithm to analyze various types of biological data; see, e.g. [DCCW+10, NLC11] for an application to breast cancer data.
- Networks analysis: Beyond sensor networks problems, the use topological data analysis tools to understand and analyze the structure of networks has recently attracted some interest. A basic idea is to build filtered simplicial complexes on top of weighted networks and to compute their persistent homology. Despite a few existing preliminary experimental results, this remains a widely unexplored research direction.
- Material science: Persistent homology recently found some promising applications in the study of structure of materials, such as for example granular media [KGKM13] or amorphous materials [NHH+15].
- Shape analysis: The geometric nature of 2D and 3D shapes makes topological methods particularly relevant to design shape descriptors for various tasks such as classification, segmentation of registration; see, for example, [CZCG05, dFL12, dFL11, COO15].

27.5 FURTHER READINGS

[Car09, Ghr08]: two survey papers that present various aspects of TDA addressing a large audience.

[Oud15]: a recent book that offers a very good introduction.

Although not discussed in this chapter, (discrete) Morse theory, Reeb graphs [DW13] and, more recently, category and sheaf theory are among the mathematical tools used in TDA. An introduction to these topics from a computational and applied perspective can be found in the recent books [EH10, Ghr14].

RELATED CHAPTERS

Chapter 23: Random simplicial complexes
Chapter 26: Persistent Homology
Chapter 36: Curve and Surface Reconstruction
Chapter 44: Nearest Neighbors in High-Dimensional Spaces

REFERENCES

[AC15] H. Adams and G. Carlsson. Evasion paths in mobile sensor networks. *Internat. J. Robotics Research*, 34:90–104, 2015.

[ACC+12] M. Aanjaneya, F. Chazal, D. Chen, M. Glisse, L. Guibas, and D. Morozov. Metric graph reconstruction from noisy data. *Internat. J. Comput. Geom. Appl.*, 22:305–325, 2012.

[ALS13] D. Attali, A. Lieutier, and D. Salinas. Vietorisrips complexes also provide topologically correct reconstructions of sampled shapes. *Comput. Geom.*, 46:448–465, 2013.

[BCD+15] M. Buchet, F. Chazal, T.K. Dey, F. Fan, S. Oudot, and Y. Wang. Topological analysis of scalar fields with outliers. In *31st Sympos. Comput. Geom.*, pages 827–841, ACM Press, 2015.

[BCOS15] M. Buchet, F. Chazal, S. Oudot, and D. Sheehy. Efficient and robust persistent homology for measures. In *Proc. 26th ACM-SIAM Sympos. Discrete Algorithms*, pages 168–180, 2015.

[BCSE+07] P. Bendich, D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and D. Morozov. Inferring local homology from sampled stratified spaces. In *48th IEEE Sympos. Found. Comp. Sci.*, pages 536–546, 2007.

[BG13] J.-D. Boissonnat and A. Ghosh. Manifold reconstruction using tangential Delaunay complexes. *Discrete Comput. Geom.*, 51, 2013.

[BGMP14] A.J. Blumberg, I. Gal, M.A. Mandell, and M. Pancia. Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces. *Found. Comput. Math.*, 14:745–789, 2014.

[BGO09] J.-D. Boissonnat, L.J. Guibas, and S.Y. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. *Discrete Comput. Geom.*, 42:37–70, 2009.

[BGV07] F. Bolley, A. Guillin, and C. Villani. Quantitative Concentration Inequalities for Empirical Measures on Non-compact Spaces. *Probab. Theory Rel.*, 137:541–593, 2007.

[BHPP14] P.-T. Bremer, I. Hotz, V. Pascucci, and R. Peikert, editors. *Topological Methods in Data Analysis and Visualization III: Theory, Algorithms, and Applications.* Mathematics and Visualization, Springer, Heidelberg,, Heidelberg, 2014.

[BM13] A.J. Blumberg and M.A. Mandell. Quantitative homotopy theory in topological data analysis. *Found. Comput. Math.*, 13:885–911, 2013.

[BM14] J.-D. Boissonnat and C. Maria. The simplex tree: An efficient data structure for general simplicial complexes. *Algorithmica*, 70:406–427, 2014.

[BRS+12] S. Balakrishna, A. Rinaldo, D. Sheehy, A. Singh, and L.A. Wasserman. Minimax rates for homology inference. In *Proc. 15th Conf. Artificial Intelligence and Statistics*, pages 64–72. JMLR W&CP, 2012.

[Bub15] P. Bubenik. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16:77–102, 2015.

[BWM12] P. Bendich, B. Wang, and S. Mukherjee. Local homology transfer and stratification learning. In *Proc. 23rd ACM-SIAM Sympos. Discrete Algorithms*, pages 1355–1370, 2012.

[Car09] G. Carlsson. Topology and data. *Bull. Amer. Math. Soc.*, 46:255–308, 2009.

[CCSG+09] F. Chazal, D. Cohen-Steiner, L.J. Guibas, F. Mémoli, and S.Y. Oudot. Gromov-Hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum*, 28:1393–1403, 2009.

[CCSL09a] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in Euclidean space. *Discete Comput. Geom.*, 41:461–479, 2009.

[CCSL09b] F. Chazal, D. Cohen-Steiner, and A. Lieutier. Normal cone approximation and offset shape isotopy. *Comput. Geom.*, 42:566–581, 2009.

[CCSLT09] F. Chazal, D. Cohen-Steiner, A. Lieutier, and B. Thibert. Stability of curvature measures. *Computer Graphics Forum*, 28:1485–1496, 2009.

[CCSM10] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Boundary measures for geometric inference. *Found. Comput. Math.*, 10:221–240, 2010.

[CCSM11] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Found. Comput. Math.*, 11:733–751, 2011.

[CFL+14] F. Chazal, B.T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Robust topological inference: Distance to a measure and kernel distance. Preprint, `arXiv:1412.7197`, 2014.

[CFL+15a] F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Subsampling methods for persistent homology. In *Proc. 32nd Internat. Conf. Machine Learning (ICML)*, pages 2143–2151, JMLR W&CP, 2015.

[CFL+15b] F. Chazal, B.T. Fasy, F. Lecci, A. Rinaldo, and L. Wasserman. Stochastic convergence of persistence landscapes and silhouettes. *J. Comp. Geom.*, 6:140–161, 2015.

[CGLM15] F. Chazal, M. Glisse, C. Labruère, and B. Michel. Convergence rates for persistence diagram estimation in topological data analysis. *J. Machine Learning Research*, 16:3603–3635, 2015.

[CGOS11] F. Chazal, L.J. Guibas, S.Y. Oudot, and P. Skraba. Scalar field analysis over point cloud data. *Discrete Comput. Geom.*, 46:743–775, 2011.

[CGOS13] F. Chazal, L.J. Guibas, S.Y. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. *J. ACM*, 60, 2013.

[CHS15] F. Chazal, R. Huang, and J. Sun. Gromov-hausdorff approximation of filamentary structures using reeb-type graphs. *Discrete Comput. Geom.*, 53:621–649, 2015.

[CISZ08] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *Internat. J. Computer Vision*, 76:1–12, 2008.

[CL07] F. Chazal and A. Lieutier. Stability and computation of topological invariants of solids in $\mathbb{R}^n$. *Discrete Comput. Geom.*, 37:601–617, 2007.

[CL08] F. Chazal and A. Lieutier. Smooth manifold reconstruction from noisy and non-uniform approximation with guarantees. *Comput. Geom.*, 40:156–170, 2008.

[CMM15] F. Chazal, P. Massart, and B. Michel. Rates of convergence for robust geometric inference. Preprint, *arXiv:1505.07602*, 2015.

[CO08] F. Chazal and S.Y. Oudot. Towards persistence-based reconstruction in euclidean spaces. In *Proc. 24th Sympos. Comput. Geom.*, pages 232–241, ACM Press, 2008.

[COO15] M. Carrière, S.Y Oudot, and M. Ovsjanikov. Stable topological signatures for points on 3d shapes. *Computer Graphics Forum*, 34:1–12, 2015.

[CSEH07] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37:103–120, 2007.

[CSO14] F. Chazal, V. de Silva, and S. Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173:193–214, 2014.

[CZCG05] G. Carlsson, A. Zomorodian, A. Collins, and L.J. Guibas. Persistence barcodes for shapes. *Internat. J. Shape Model*, 11, 2005.

[DCCW+10] D. DeWoskin, J. Climent, I. Cruz-White, M. Vazquez, C. Park, and J. Arsuaga. Applications of computational homology to the analysis of treatment response in breast cancer patients. *Topology and its Applications*, 157:157–164, 2010.

[dFL11] B. di Fabio and C. Landi. A mayer-vietoris formula for persistent homology with an application to shape recognition in the presence of occlusions. *Found. Comput. Math.*, 11:499–527, 2011.

[dFL12] B. di Fabio and C. Landi. Persistent homology and partial similarity of shapes. *Pattern Recognit. Lett.*, 33:1445–1450, 2012.

[DFW15] T.K. Dey, F. Fan, and Y. Wang. Graph induced complex on point data. *Comput. Geom.*, 48:575–588, 2015.

[DW13] T.K. Dey and Y. Wang. Reeb graphs: approximation and persistence. *Discrete Comput. Geom.*, 49:46–73, 2013.

[EH10] H. Edelsbrunner and J.L. Harer. *Computational Topology: An Introduction.* AMS, Providence, 2010.

[FL99] P. Frosini and C. Landi. Size theory as a topological tool for computer vision. *Pattern Recognit. Image Anal.*, 9:596–603, 1999.

[FLR+14] B.T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, A. Singh, et al. Confidence sets for persistence diagrams. *Ann. Statist.*, 42:2301–2339, 2014.

[Ghr08] R. Ghrist. Barcodes: The persistent topology of data. *Bulletin AMS*, 45:61–75, 2008.

[Ghr14] R. Ghrist. *Elementary Applied Topology.* CreateSpace, 2014.

[GMM13] L. Guibas, D. Morozov, and Q. Mérigot. Witnessed k-distance. *Discrete Comput. Geom.*, 49:22–45, 2013.

[GPP+12a] C.R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.*, 40:941–963, 2012.

[GPP+12b] C.R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *J. Machine Learning Research*, 13:1263–1291, 2012.

[Gro93] K. Grove. Critical point theory for distance functions. In *Proc. Symposia in Pure Mathematics*, vol. 54, 1993.

[GSBW11] X. Ge, I. Safa, M. Belkin, and Y. Wang. Data skeletonization via reeb graphs. In *Advances in Neural Information Processing Systems*, pages 837–845, 2011.

[Hat01] A. Hatcher. *Algebraic Topology.* Cambridge Univ. Press, 2001.

[Hau95] J.-C. Hausmann. On the vietoris-rips complexes and a cohomology theory for metric spaces. *Ann. Math. Stud.*, 138:175–188, 1995.

[HK14] Y. Hiraoka and G. Kusano. Coverage criterion in sensor networks stable under perturbation. Preprint, `arXiv:1409.7483`, 2014.

[KGKM13] M. Kramar, A. Goullet, L. Kondic, and K. Mischaikow. Persistence of force networks in compressed granular media. *Physical Review E*, 87, 2013.

[Lat01] J. Latschev. Vietoris-rips complexes of metric spaces near a closed riemannian manifold. *Arch. Math.*, 77:522–528, 2001.

[LOC14] C. Li, M. Ovsjanikov, and F. Chazal. Persistence-based structural recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2003–2010, 2014.

[LSL+13] P.Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson. Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3, 2013.

[Mér13] Q. Mérigot. Lower bounds for k-distance approximation. In *Proc. 29th Sympos. Comput. Geom.*, pages 435–440, ACM Press, 2013.

[MMH11] Y. Mileyko, S. Mukherjee, and J. Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27, 2011.

[NHH+15] T. Nakamura, Y. Hiraoka, A. Hirata, E.G. Escolar, and Y. Nishiura. Persistent homology and many-body atomic structure for medium-range order in the glass. *Nanotechnology*, 26, 2015.

[NLC11] M. Nicolau, A.J. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. USA*, 108:7265–7270, 2011.

[NSW08] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39:419–441, 2008.

[NSW11] P. Niyogi, S. Smale, and S. Weinberger. A topological view of unsupervised learning from noisy data. *SIAM J. Comput.*, 40:646–663, 2011.

[Oud15] S.Y. Oudot. *Persistence Theory: From Quiver Representations to Data Analysis.* Vol. 209 of *AMS Math. Surveys Monographs*, AMS, Providence, 2015.

[Pet06] A. Petrunin. Semiconcave functions in Alexandrov's geometry. In *Surveys in Differential Geometry*, vol. 11, pages 137–201. International Press, Somerville, 2006.

[PH13] J.A. Perea and J. Harer. Sliding windows and persistence: An application of topological methods to signal analysis. *Found. Comput. Math.*, 15:799–838, 2013.

[PHCF12] R. Peikert, H. Hauser, H. Carr, and R. Fuchs, editors. *Topological Methods in Data Analysis and Visualization II: Theory, Algorithms, and Applications*. Mathematics and Visualization, Springer, Heidelberg, 2012.

[PTHT11] V. Pascucci, X. Tricoche, H. Hagen, and J. Tierny. *Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications*, 1st edition. Springer, Heidelberg, 2011.

[PWZ15] J. Phillips, B. Wang, and Y. Zheng. Geometric inference on kernel density estimates. In *Proc. 31st Sympos. Comput. Geom.*, pages 857–871, ACM Press, 2015.

[RHBK15] J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt. A stable multi-scale kernel for topological machine learning. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[Rob99] V. Robins. Towards computing homology from finite approximations. In *Topology Proceedings*, vol. 24, pages 503–532, 1999.

[Rob14] M. Robinson. *Topological Signal Processing*. Springer, Heidelberg, 2014.

[SC04] V. de Silva and G. Carlsson. Topological estimation using witness complexes. In *Proc. 1st Eurographics Conf. on Point-Based Graphics*, pages 157–166, 2004.

[SG07a] V. de Silva and R. Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7:339–358, 2007.

[SG07b] V. de Silva and R. Ghrist. Homological sensor networks. *Notices Amer. Math. Soc.*, 54, 2007.

[She13] D.R. Sheehy. Linear-size approximations to the Vietoris-Rips filtration. *Discrete Comput. Geom.*, 49(4):778–796, 2013.

[SMC07] G. Singh, F. Mémoli, and G. Carlsson. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *Proc. Eurographics Sympos. on Point-Based Graphics (SPBG)*, pages 91–100. Eurographics, 2007.

[SMVJ11] V. de Silva, D. Morozov, and M. Vejdemo-Johansson. Persistent cohomology and circular coordinates. *Discrete Comput. Geom.*, 45:737–759, 2011.

[SOCG10] P. Skraba, M. Ovsjanikov, F. Chazal, and L. Guibas. Persistence-based segmentation of deformable shapes. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 45–52, 2010.

[Tak81] F. Takens. *Detecting Strange Attractors in Turbulence*. Springer, Heidelberg, 1981.

[TSL00] J.B. Tenenbaum, V. De Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[Vil03] C. Villani. *Topics in Optimal Transportation*, AMS, Providence, 2003.

[VUFF93] A. Verri, C. Uras, P. Frosini, and M. Ferri. On the use of size functions for shape analysis. *Biological Cybernetics*, 70:99–107, 1993.