Topological Data Analysis

Jean-Daniel Boissonnat and Frédéric Chazal and Bertrand Michel

Abstract It has been observed since a long time that data are often carrying interesting topological and geometric structures. Characterizing such structures and providing efficient tools to infer and exploit them is a challenging problem that asks for new mathematics and that is motivated by a real need from applications. This paper is an introduction to Topological Data Analysis (TDA), a new field that emerged during the last two decades with the objective of understanding and exploiting the topological structure of modern and complex data. The paper surveys some important mathematical and algorithmic developments in TDA as well as software solutions that are currently used to address various applied and industrial problems.

1 Introduction

The recent years have seen all domains of science, economy and even everyday life overwhelmed with massive amounts of data. Bringing scientists, industrials and citizens to the most relevant, often unexpected, features and giving them the tools to discover and extract the best knowledge out of their data are fundamental challenges for our modern society.

During the last decades, the wide availability of measurement devices and simulation tools has not only led to an explosion in the amount of available data, but also in a spectacular increase in their complexity, making their analysis more and more challenging: data are often represented as points in a high dimensional space, or as

J.-D. Boissonnat
Université Côte d'Azur Inria e-mail: jean-daniel.boissonnat@inria.fr

F. Chazal
Inria Saclay e-mail: frederic.chazal@inria.fr

B. Michel
Ecole Centrale de Nantes e-mail: Bertrand.Michel@ec-nantes.fr

complex objects like a 2D or 3D image, a meshed shape, a multivariate time serie, a graph...

This challenge has led to the development of a wide variety of new mathematical theories and tools among which topology and geometry have recently shown to be particularly relevant. Indeed, a closer look at data often shows that they carry geometric and topological patterns and structures that are immensely helpful in analyzing the systems and phenomena from which they have been generated and that can be used for further machine learning tasks - see Figure 1 for a concrete illustrative example. From this observation, made by practitioners in many academic and industrial domains, emerged the need to develop new mathematical and effective tools to capture and exploit topological information from data. This gave rise to a new research domain known as Topological Data Analysis (TDA). Although one can trace back geometric approaches in data analysis quite far in the past, TDA emerged from various works in applied (algebraic) topology and computational geometry during the first decade of the century. It really started as a field with the pioneering works of [52] and [79] in persistent homology and was popularized in a landmark paper in 2009 [21]. It is interesting to mention that persistent homology was already introduced and developed earlier, in some restricted setting, in the 90's by [55, 74] under the name of size theory and, in a pure mathematics context, by [4] under the name of framed Morse complex.

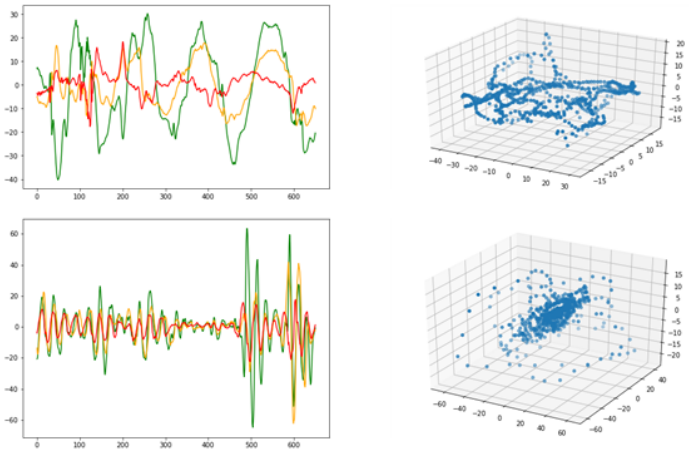

Fig. 1 The left plots represent the 3D acceleration of the arm of a subject making specific disordered movements, measured by an inertial device developed by Sysnav, a French SME, during a short interval of time. The 3 colored curves correspond to the 3 coordinates of the acceleration vector measured in the device coordinate system. The right point clouds are plots of the trajectory of the acceleration vector in $\mathbb{R}^3$. It shows that the "chaotic" behavior of the considered multivariate time-series carries interesting topological patterns that are typical of the movements. The aim of TDA is to provide relevant and efficient tools to characterize qualitatively and quantitatively such patterns in order to exploit them -in combination with other features - for further analysis of the movement.

TDA aims at designing and providing relevant and efficient 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, considerable efforts have been made to settle the mathematical and algorithmic foundations of TDA and to provide robust, efficient and easy to use software such as the Gudhi library (C++ and Python) [64] and its R software interface [53]. Although this new mathematical field is still young and evolves rapidly, TDA already provides a set of mature and efficient tools that are complementary to other tools in data science. Adding them to the toolbox of the data scientist turned out to be remarkably useful in several applications and industrial problems.

About this paper

This paper is not a survey on TDA and does not aim at presenting an exhaustive overview of the field and its numerous applications. Its goal is to explain and illustrate, through a few selected topics reflecting the experience of the authors in the field, and without technical details, how the need to understand and exploit the topological structure of data has led to new mathematical ideas. The resulting developments are playing an essential role to bring TDA from a set of ad-hoc or heuristic methods to a well-founded field. They also contribute to bring experimental algorithms and prototype codes to efficient software for industrial applications.

To avoid the introduction of too technical notions and digressions, and make the paper as easy-to-read as possible for non experts, a brief glossary recalling some classical but important definitions has been included in the last section of the paper[1]. References to specific aspects of TDA are given all along the paper. Additionally, the reader can find several surveys [47, 77] and textbooks [68, 56, 51, 12], covering different aspects of the field.

2 The need of new mathematical and algorithmic tools

Formalizing the notion of topological and geometric structure of data is a tedious question. Inferring relevant topological and geometric information from data raises difficult problems requiring new mathematical tools. The reasons for these difficulties are many but are closely related to the three fundamental following facts.

- First of all, data are discrete, i.e. finite sets of observations, while topological and geometric quantities are usually associated to continuous shapes. It is thus necessary to introduce intermediate geometric models that both faithfully approximate or summarize the data and carry information about the underlying shapes around which they have been sampled. The highly non linear nature of the space of such

[1] The first occurences, in the paper, of the notions defined in the glossary are put in italic in the text

models makes the approximation theory for topological and geometric invariants much more difficult than the classical approximation theories for functions. This problem has been addressed from different perspectives among which the distance-based approaches, presented in the next section, have given rise to new mathematical developments. This is still an active research area.

- Second, although they appear to be concentrated around geometric shapes, real data are usually corrupted by noise and outliers. It is also often observed that the topological and geometric features underlying data are strongly dependent on the scale at which they are considered. Quantifying and distinguishing topological/geometric noise from topological/geometric signal to infer relevant scale-dependent or multi-scale information is a subtle problem that does not benefit from the standard signal processing and statistical tools. It requires the development of new tools and approaches. Persistent homology, presented in Section 3.2, emerged at the beginning of the century as a new approach to address such problems. Since then it has grown as a new mathematical theory with impressive developments going from fundamental mathematics to algorithms and concrete applications.
- Third, even when the data are concentrated around low dimensional shapes, the possibly high dimensionality of the spaces in which they are embedded raises severe algorithmic and practical issues. Classical data structures and algorithms from computational topology and geometry quickly become inefficient in practice when the dimensionality of the data increases. Although these questions, at the crossing of computer science and mathematics, are not discussed in details in this paper, they are of fundamental importance to provide efficient and implemented software tools. They are subject to intense research activities.

3 The emergence of geometric inference and persistent homology

Many approaches have been proposed, to address the difficulties mentioned above. However, since these difficulties are often driven by the constraints and needs of specific applications, the solutions are mostly ad-hoc and are not easy to generalize or exploit in other settings. During the last two decades many efforts have been made to address them in general mathematical frameworks that have led to new mathematical developments and significant progress on the practical side.

The general problem underlying TDA can be roughly summarized in the following ill-posed question: given a finite set of points $\mathbb{X}_n = \{x_0, \cdots, x_n\}$ in $\mathbb{R}^d$, or in a more general metric space, is it possible to reliably and efficiently estimate topological and geometric properties reflecting the global structure of $\mathbb{X}_n$?

This question can be more precisely formalized when the points of $\mathbb{X}_n$ are assumed to be sampled around a given compact subset K. In that case, it boils down to the estimation of topological or geometric invariants of K. However, assuming the existence of a well-defined shape underlying the data to which the estimated topological invariants could be compared, is in many cases a too restrictive assumption.

Then, the estimation of geometric and topological invariants is relevant only if one can establish that the invariants remain stable under perturbations of the input data.

3.1 Distance-based geometric inference

An intuitive way to associate a continuous geometric structure to a discrete data set is to consider union of balls centered on the data points. 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, $K^r = \cup_{x\in K} B(x,r) = d_K^{-1}([0,r])$, 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 $d_K(x) = \inf_{y\in K} \|x-y\|$. The idea underlying distance-based approaches is, instead of directly comparing $\mathbb{X}_n$ and K, to compare the topology of the foliations defined by the level sets of d_K and $d_{\mathbb{X}_n}$.

This is made possible thanks to a fundamental property of the squared distance function: for any compact set K, d_K is semi-concave, i.e. $x \to \|x^2\| - d_K^2(x)$ is convex. Distance functions inherit interesting differential properties from semi-concavity that allow to relate the topology of the offsets of compact sets that are close to each other with respect to the *Hausdorff distance* $d_H(.,.)$ between compact sets. For example, when K is a smooth and compact submanifold of $\mathbb{R}^d$, this leads to a basic method to reliably estimate the topology (homotopy type) and the homology groups of K from well-chosen offsets of $\mathbb{X}_n$ under explicit mild conditions on $d_H(\mathbb{X}_n, K)$ [67, 45]. As geometric structures underlying data are not always as regular as smooth manifolds, the result has been extended to a larger class of non smooth compact sets K and led to stronger results on the inference of the topology of the offsets of K [32]. This approach, based on the study of distance functions, also led to results on the estimation of other geometric and differential quantities of K such as normal cones [31], curvature measures [34, 33] and boundary measures [35]. Some of these estimators were adapted to find applications for feature detection - such as sharp edges or corners - on 3D shapes [65].

Covers and nerves to compute the topology of union of balls

From an algorithmic perspective, the advantage of estimating topological features of data from union of balls is that, thanks to the so-called *Nerve Theorem* in algebraic topology, their topology is fully described by an *abstract simplicial complex* (a purely combinatorial structure) encoding the intersection patterns of the balls. More precisely, given a union of balls in Euclidean space, $\bigcup_{i=1}^n B(x_i, r_i)$, connecting the data points x_i, x_j by edges whenever the two corresponding balls $B(x_i, r_i), B(x_j, r_j)$ intersect gives rise two a graph with the same connectivity as the union of balls. To go beyond connectivity, one can also connect $(k+1)$-uple of points $x_{i_0}, \cdots x_{i_k}$ whenever $\bigcap_{j=0}^k B(x_{i_j}, r_{i_j}) \neq \emptyset$ - see Figure 3. The resulting object, called the *nerve* of the union of balls is a *simplicial complex* and allows to identify higher dimensional topological features such as cycles, voids and their higher dimensional counterparts. It follows

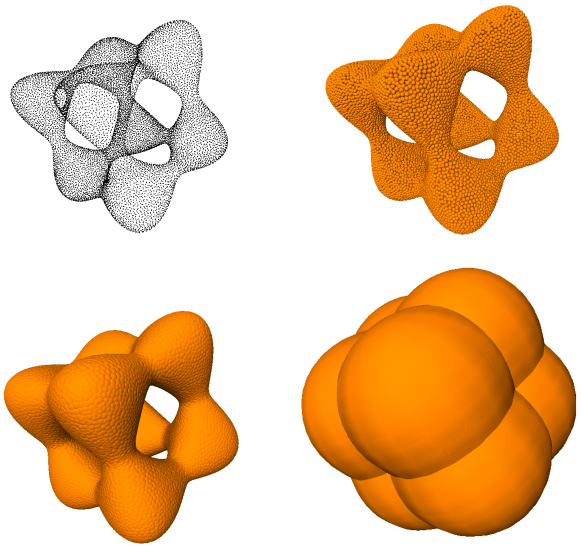

Fig. 2 The level sets of the distance function to a point cloud sampled on a smooth surface (upper left picture), the tangle cube defined by the implicit equation $x^4 - 5x^2 + y^4 - 5y^2 + z^4 - 5z^2 + 11.8 = 0$, for different values. The offset on the upper left picture has the same topology as the sampled surface.

from the Nerve Theorem that the homotopy type of $\bigcup_{i=1}^{n} B(x_i, r_i)$ is determined by its nerve. As a consequence, many topological invariants such as the *homology groups*, the *Betti numbers* and the Euler characteristic of a union of ball can be efficiently computed from its nerve.

Indeed, simplical complexes play a fundamental role in TDA where they are widely used to infer topological features from data. On one hand, they are classical well-studied topological objects that are well adapted to model geometric structures from data. One the other hand, they are combinatorial objects that are building blocks of most TDA algorithms. As a consequence, they are perfectly well-suited to bridge the gap between continuous spaces and discrete data structures that can be processed by computer programs.

The previous approach faces two issues. First, computing a union of balls in high dimensions is very difficult and takes time that is exponential in the ambient dimension even if the object of interest is low dimensional. Second, only the homotopy type of the geometric shape can be recovered, not a triangulation homeomorphic to the shape. These restrictions can be removed when the shape is a manifold and the amount of noise in the data is limited. In particular, the tangential Delaunay complex [13, 9] can construct a triangulation of a manifold from a dense sample in time only linear in the dimension of the ambient space. This result has been further studied in a statistical context, proving that the reconstruction is minimax optimal [1].

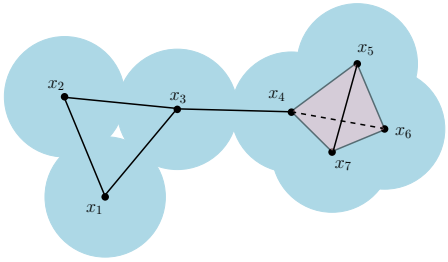

Fig. 3 The nerve of a union of balls in $\mathbb{R}^2$. Notice that, although the balls are included in $\mathbb{R}^2$, their nerve contains the tetrahedron $[x_4, x_5, x_6, x_7]$ and cannot be embedded in $\mathbb{R}^2$: this is an abstract simplicial complex.

Another use of covers and nerves: the Mapper algorithm

Nerves are more generally defined for any family of subsets covering a data set. They provide, in many cases, an interesting way to summarize, visualize and explore the topological structure of data. This natural idea was first proposed for TDA in [72], giving rise to the so-called Mapper algorithm 1. Despite its simplicity, this algorithm has been successfully used in a large variety of problems as an exploratory tool for clustering and identify important features of data - see, for example, [78, 63]. It has also given birth to a successfull company, Ayasdi (https://www.ayasdi.com/) that have made it one of its core technology. Despite its successes, the Mapper algorithm remains an exploratory tool that is very sensitive to various parameters requiring a tedious involvement of the user to be correctly tuned. Overcoming this problem to make Mapper a fully automated tool is an active mathematical research topic. Based on preliminary results on the stability of Mapper proposed in [27], advances towards a statistically well-founded version of Mapper have been obtained recently in [26], but new mathematical ideas are necessary to go beyond these preliminary results and make significant progress.

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 $C_{U,i} \cap C_{U',j} \neq \emptyset$

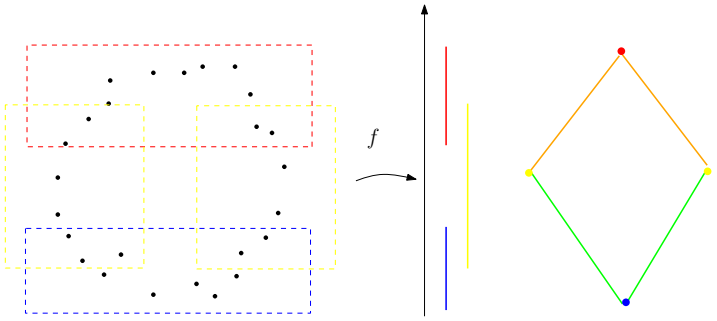

Fig. 4 The mapper algorithm applied to a point cloud sampled around a circle and the height function.

Distance-based inference with noisy data

Real world data are often corrupted by noise and the distance-based approach may fail completely in the presence of outliers. Indeed, adding even a single outlier to the point cloud can dramatically change the distance function and make the above methods fail. To circumvent this problem, we adopted a measure theoretic approach and introduced an alternative distance-like function, the distance-to-a-measure (DTM) [36]. DTM is robust to noise and its sublevel sets still carry relevant topological information.

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. Intuitively, $\delta_{P,u}(x)$ is the smallest ball centered at x that contains a fraction u of the total mass of the probability distribution P. To avoid issues due to discontinuities of the map $P \to \delta_{P,u}$, the distance-to-measure function (DTM) with parameter $m \in [0,1]$ is then defined by averaging the functions $\delta_{P,u}$ for $u \in [0,m]$:

$$d_{P,m}(x) : x \in \mathbb{R}^d \mapsto \left(\frac{1}{m}\int_0^m \delta_{P,u}^2(x)du\right)^{1/2}. \tag{1}$$

A nice property of the DTM proved in [36] 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

$$\|d_{P,m} - d_{\tilde{P},m}\|_\infty \leq m^{-\frac{1}{2}} W_2(P,\tilde{P}) \tag{2}$$

where W_2 is the Wasserstein distance [2] for the Euclidean metric on $\mathbb{R}^d$, with exponent 2. This property implies that the DTM associated to close distributions in the

[2] See [76] for a definition of the Wasserstein distance between probability distributions.

Wasserstein metric have close sublevel sets. Moreover, the function $d^2_{P,m}$ is semiconcave ensuring strong regularity properties on the geometry of its sublevel sets. Using these observations, [36] 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, the DTM, for $m = k/n$, can be approximated using the empirical measure P_n instead of P which is defined by the simple following formula,

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

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, making it well-adapted for algorithmic and practical use. Moreover, its convergence properties and, more generally approximations of the DTM, have been carefully studied [40, 46, 57, 69, 20]. The practical use of the DTM raises the question of the choice of the parameter m that has been partly addressed in [46, 40] but still remains largely open. New mathematical ideas will probably be necessary to provide a complete and practically rigorous answer to this problem. The study and results on DTM have led to further works and applications in various directions such as topological data analysis [19], GPS traces analysis [28], density estimation [6], hypothesis testing [17] or clustering [44]. The use of the DTM in the setting of an industrial research project, with the Japanese company Fujitsu, whose goal was to address an anomaly detection problem from inertial sensor data in bridge and building monitoring [60] has also recently led to new mathematical developments on the inference of relevant topological features from data with noise and outliers [3].

3.2 Persistent homology

Distance-based inference methods presented in the previous section offer a set of efficient methods to estimate topological features (homology, homotopy type) of data, or even reconstruct an approximating shape, at a given scale r, by considering union of balls of radius r or r-sublevel sets of the DTM. Their validity relies on specific sampling assumptions and on the regularity of some fixed shape underlying the data. However, in many real cases, such assumptions are difficult to assert, or are not even satisfied. Moreover, the underlying structure of the data may strongly depend on the scale r at which it is considered as illustrated on Figure 5.

In some cases, it also happens that none of the offsets of the data have the correct topology as illustrated on Figure 6.

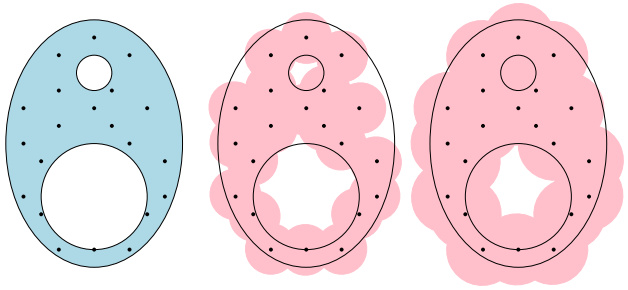

Fig. 5 A sampled 2D domain in $\mathbb{R}^2$ with a small and big holes. Depending on the choice of the radius r, the union of balls centered on the data points may have the topology of a disk with one or two holes. In such a case, selecting the appropriate radius from the data to infer the correct topology is an ill-posed problem.

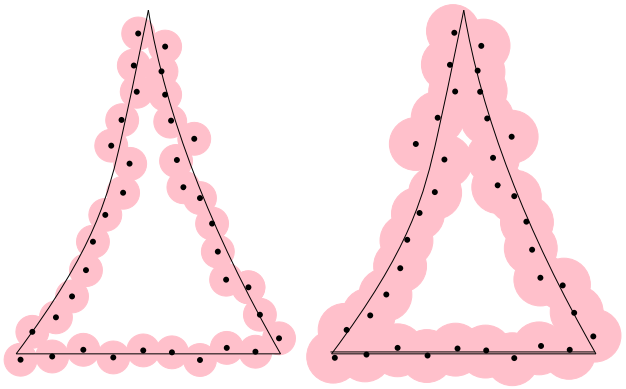

Fig. 6 When the sampling and regularity conditions provided by distance-based inference results are not satisfied, it may happen that there is no correct choice of radius to infer the topology of the underlying shape. Here, none of the union of balls centered on the data point succeeds to recover the topology of the sampled simple curve.

To overcome these problems and give a well-defined and formal meaning to multiscale topological structure of a wide range of data, it was thus necessary to introduce new mathematical ideas and tools. This is where persistent homology, a central concept in TDA, comes into play! Preliminary ideas related to persistent homology can be traced back quite far in the past in pure mathematics [66, 4]. The theory emerged in its present form at the begining of the century [75, 52, 52] as a tool to compute, study and efficiently encode multiscale topological features of data. Persistent homology provides a framework and efficient algorithms to encode the evolution of the *homology* of families of nested topological spaces or simplicial complexes indexed by a set of real numbers, called *filtrations*.

In TDA, two different kinds of filtrations are classically considered. When data come as a point cloud in metric space (not necessarily Euclidean), one can construct various filtered abstract simplicial complexes whose vertex set is the set of data

points. An example of such a filtration is the Čech complex filtration, which consists of the nerves of a growing family of balls centered on the data points (see the previous section). In that case, the persistent homology of such a filtration reflects the global topological structure of the data at different scales corresponding to the radii of the considered balls. When data come as a collection of already complex objects such as images, 3D shapes or graphs, functions defined on each of them may be used to highlight some of their features. The persistent homology of the sublevel or upperlevel set filtrations of these functions then provide topological information. This new type of information can be fruitfully used to compare and classify the data elements, and can be combined with other non topological features to enhance the results.

Given a filtration $\text{Filt} = (F_r)_{r \in T}$ of a simplicial complex or a topological space, the homology of F_r changes as r increases: new connected components can appear, existing components 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, called a *persistence diagram* in $\mathbb{R}^2$ where the coordinates of each point are the starting and end points of the corresponding intervals, as illustrated on a simple example on Figure 7. The length of the persistence intervals - or equivalently the vertical distance of the persistence point to the diagonal in $\mathbb{R}^2$ - reflects the life span of the topological features along the filtration. Intuitively, the longer an interval is, the more relevant is the corresponding topological feature.

A simple but fundamental observation, at the root of important mathematical developments in the theory of persistent homology, is that persistence diagrams can be defined in a purely algebraic way. Given a filtration $\text{Filt} = (F_r)_{r \in T}$, 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*.

More generally, a persistence module $\mathbb{V}$ over a subset T of the real numbers $\mathbb{R}$ is an indexed family of vector spaces over a field $\mathbb{K}$, $(V_r \mid r \in T)$ 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 favorable cases, a persistence module can be uniquely decomposed, up to a re-ordering of the terms, in a direct sum of irreducible modules of the form $\mathbb{I}_{(b,d)} = (I_r \mid r \in T)$ where $I_r = \mathbb{K}$ if $r \in (b, d)$ and $I_r = 0$ otherwise, and where all the linear maps from $\mathbb{K}$ to $\mathbb{K}$ are the identity. The persistence barcode/diagram of $\mathbb{V}$ is then defined as the set of intervals (b, d) in this decomposition.

The study of general algebraic persistence modules opened a new research direction in mathematics that is knowing various and important developments. In [48, 29], it is proven that the notion of persistence diagram can be defined for a large class of abstract persistence modules, that are not necessarily decomposable. This generalization, which may appear of purely theoretical interest, reveals extremely useful to study the convergence and statistical properties of persistence diagrams and

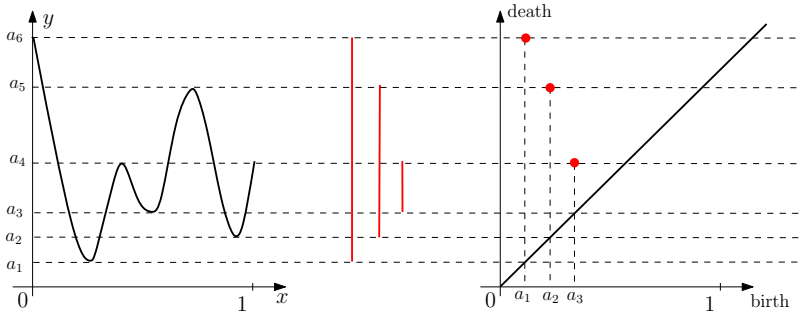

Fig. 7 The persistence barcode and the persistence diagram of the sublevel set filtration $(F_r = f^{-1}((-\infty, r]))_{r\in\mathbb{R}}$ of a function $f : [0, 1] \to \mathbb{R}$. All the sublevel sets of f are either empty or a union of interval, so the only non trivial topological information they carry is their 0-dimensional homology, i.e. their number of connected components. At $r = a_1$, persistent homology registers the birth of a first connected component in F_{a_1} and start a first interval. Then, F_r remains connected until r reaches the value a_2 where a second connected component appears and a second interval is created. Similarly, when r reaches a_3, a new connected component appears and a new interval is created. 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. When r reaches a_5, 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). 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.

related topological features under various models - see for example [42, 39]. It is interesting to notice that these mathematical developments were initially motivated by an applied problem of topological inference [43]. This problem involvesromov persistence modules that were not the homology groups of a filtration and to which the already existing theory did not applied.

A fundamental property of persistent homology in TDA is that persistence diagrams are stable, ensuring that the topological features of data obtained from persistence diagrams are robust to various types of perturbations. In the case of filtrations defined by the sublevel sets of functions, this property can be stated in the following way.

Theorem 1 (Stability of persistence diagrams for functions [49])

Let $f, g : X \to \mathbb{R}$ be two real-valued continuous functions defined on a compact triangulable topological space X. Then for any integer k,

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

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.* $(H_k(g^{-1}(-\infty, r]))|r \in \mathbb{R})$*) and* d_B *the* bottleneck distance *between diagrams.*

The above result appeared to be a particular case of a much more general algebraic stability result in persistence theory. This later revealed to be a powerful tool to derive stability results in various settings [48]. In particular, it led to establish the stability of the persistence diagrams of classical filtrations built on top of data sets for perturbations with respect to the Gromov-Hausdorff metric [37].

Persistent homology for machine learning

Thanks to the stability results and to natural invariance properties, persistence diagrams can be used as robust and relevant multiscale topological signatures of data. Moreover, in many cases, persistence diagrams can be easily interpreted by the user since persistence intervals represent the life span of some topological features. This makes them particularly attractive to be used as discriminative features for classification and other machine learning tasks. See [30, 62] for some examples in classification, or [50] for an industrial application to arrhythmia classification that led to a patent application.

However, persistence diagrams cannot be directly used for such tasks: indeed persistence diagrams come as unordered finite sets of points in the plane while most of classical data analysis and machine learning algorithms require vectors as input. This linearization problem has been the subject of an intense research activity during the last few years, motivated in part by concrete applied and industrial problems. Some approaches intend to design representations that preserve some of the stability properties of persistence diagrams, like for example persistence landscapes [18] and images [2], while some others follow more heuristic approaches driven by specific needs, like for example [73]. The construction of kernels on the space of diagrams have also been proposed in order to use persistence diagrams in combination with kernel-based methods such as SVM. See, e.g. [70, 25]. More recent works have also proposed data-driven methods to automatically learn an adapted representation of persistence diagrams for a given task [59]. Despite various applied and industrial successes, these works remain very dependent on a specific considered data or problem and usually require a deep understanding of persistence theory to be applied. This prevents non expert users and engineers to really benefit from the resulting tools. To facilitate the work of non expert users to chose and correctly use the appropriate TDA tools for their problems, there is a need to better understand and analyze, from a mathematical perspective, the existing linearization methods and to provide global frameworks in which they can be compared. This problem is at the core of an industrial research partnership between the authors and the Japanese company Fujitsu, that goes from the most mathematical aspects to concrete industrial problems. It already led to several new tools to overcome this issue [24, 71].

The algorithmic and software challenges of TDA

TDA tools and persistent homology computations raise challenging algorithmic problems due to the size of the combinatorial objects (simplicial complexes and filtrations) involved in computations. Many efforts and progress have been made in computational topology to address these problems and improve the whole pipeline of TDA. On one end of the pipeline, new efficient data structures have been proposed to represent simplicial complexes and filtrations compactly [11, 16, 14]. New algorithms have also been designed to approximate filtrations in a controlled way [15]. Algorithms for computing persistent homology or cohomology have also made remarkable progress [5, 8, 10]. Lastly, efficient algorithms are now available to process persistent diagrams and, in particular, to compute distances between persistent diagrams and to compare and cluster such diagrams [61].

On the practical side, reliable and efficient state-of-the-art software have been developed to support industrial applications of TDA. These software also help mathematical research by providing tools to experiment and explore theoretical ideas and new research directions. Building on the mathematical and algorithmic progress of TDA during the last decade, several individuals and teams working on computational topology have developed their own software - e.g. DIONYSUS [3], PHAT [4], RIPSER [5] - implementing specific methods and functionalities. Taking a more global approach, funded by the ERC project GUDHI, a high quality open source software platform called GUDHI[6] has been developed to provide a unified framework for the central data structures and algorithms in TDA. The library is written in C++ with a Python interface that makes the functionalities of the library easily accessible to data scientists [64]. An interface with R has also been developed [53]. The development of the GUDHI library is a long-term project, supervised by an editorial board, and submitted to a rigorous review process. The library is already used by industrial partners of the authors (Sysnav, Fujitsu, IFPEN) and serves as a basis for joint research and concrete developments.

4 New research directions

New research directions in TDA are motivated by the increasing variety of applied and industrial problems where TDA is relevant. Most of them require innovative mathematical ideas and approaches. Providing a global picture of these new research directions is beyond the scope of this paper. We only list a few of them that are of particular importance, both from the mathematical and applied side.

[3] http://www.mrzv.org/software/dionysus/

[4] https://bitbucket.org/phat-code/phat

[5] https://github.com/Ripser/ripser

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

- *Statistical aspects of geometric inference and* TDA*:* the study of inference and estimation problems in TDA from a statistical perspective has started to attract some attention since a few years ago and is now an active research theme. It gave rise to a significant literature that has led down the mathematical foundations of Statistical Topological Data Analysis. Beyond its mathematical interest, this research direction also intend to provide effective new approaches to address concrete problems. As an example, one can mention the study of bootstraping and subsampling strategies for persistent homology. They led to practical tools able to infer relevant topological features from data sets that are too large to be handled by classical TDA algorithms [41, 39]. Another example is the design of confidence intervals and statistical tests [7, 54, 17].
- *Persistent homology in an algebraic framework:* persistent homology, considered in an algebraic and category theory perspective, is knowing impressive developments and generalizations in pure mathematics. They raise deep mathematical questions, reveal unexpected connections with other areas of mathematics like, e.g. symplectic geometry, but they also find their roots in very concrete motivations. As an example, one can mention the study of multidimensional persistence [23] that, instead of considering filtrations indexed by real numbers, consider filtrations with multidimensional indices in $\mathbb{R}^k$, $k > 1$, or zig-zag persistence that consider non increasing sequences of spaces [22].
- TDA *and machine learning:* as already mentioned in the previous section, combining TDA with other machine learning approaches and algorithms is a research direction motivated by applied and industrial problems that currently widely contributes to stimulate TDA. Many applied and experimental works and results demonstrate the interest and the potential of this research direction and there is a real need, to go further, to address it from a mathematical perspective. For example, understanding the structure and the properties of all the existing representations of persistent homology - persistence landscapes [18], persistence images [2], Betti curves [73],... - in a general framework is an important question for practical purposes. Despite recent works in this direction [38], this kind of question remains rather unexplored.
- Software tools: Software tools have made tremendous progress in the recent past. It is expected that this will continue in near future both in terms of new functionalities associated to new theoretical advances and in terms of efficiency. As the field of TDA expand and is becoming an essential tool in modern data analysis, feedback from applied fields and industry will push further experimental research and software development.

5 Conclusion

Giving sense to the notion of topological and geometric structures, that are concepts usually associated to continuous spaces, to discrete data is a natural but important general challenge. Although it has been largely explored for a long time in very

particular cases[7],, there was a real need to propose new mathematical and algorithmic approaches to address the problem in the more general settings provided by data science and its numerous applied problems. During the last decade, TDA has largely contributed to serve this need by developing a new mathematical research area at the crossing of mathematics, statistics and computer science. An important specificity of TDA is that if it has known applied and industrial successes, this is mainly because it has established its roots in pure and applied mathematics. This enabled to address the challenges at the right level of generality and to develop concepts and models that turned out to be relevant and useful in a large variety of settings. It is also important to underline that part of TDA, still motivated by concrete problems, is also actively expanding towards more theoretical domains resulting in many publications in pure mathematics journals and books.

A brief glossary

- **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 $d_H(K, K') = \|d_K - d_{K'}\|_\infty = \sup_{x \in \mathbb{R}^d} |d_K(x) - d_{K'}(x)|$.
- **Homotopy type.** 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. Spaces with the same homotopy type have isomorphic homology groups.
- **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

[7] for example in CAD industry, the construction of numerical geometric models (meshes) of real world shapes from scans or other measurements has been widely studied

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 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, 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.

- **Covers and nerves.** A *cover* $\mathcal{U} = (U_i)_{i\in I}$ of a topological space X is a family of sets U_i such that $X = \cup_{i\in I} U_i$. The *nerve of* $\mathcal{U}$ is the abstract simplicial complex $C(\mathcal{U})$ whose vertices are the U_i's and such that

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

The nerve of a union of balls of given radius r centered on a set of points in Euclidean space, or in a more general metric space, is also known as the *Čech complex*. The nested family of Čech complexes, for $r \geq 0$ is called the Čech filtration of the set of points.

Theorem 2 (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, i.e having the homotopy type of a point. Then, X and the nerve $C(\mathcal{U})$ are homotopy equivalent.

- **Homology and Betti numbers.** Homology (with coefficient in a given field) is a classical object from algebraic topology that 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 - see [58] for a formal definition. The k^{th} Betti number of X, denoted $\beta_k(X)$, is the rank of $H_k(X)$. It corresponds to the number of "independent" k-dimensional features of X: for example, $\beta_0(X)$ is the number of connected components of M, $\beta_1(X)$ is the number of independent cycles or tunnels, $\beta_2(X)$ the number of cavities, etc... 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.
- **Filtrations.** 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 $C = \cup_{\alpha \in A} C_\alpha$. Given a topological space X and a function $f : X \to \mathbb{R}$, the *sub level set filtration of* f is the nested family of sublevel sets of f: $(f^{-1}((-\infty, \alpha]))_{\alpha \in \mathbb{R}}$.

- **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.

References

1. Aamari, E., Levrard, C.: Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction. Discrete and Computational Geometry (2018). URL https://hal.archives-ouvertes.fr/hal-01245479
2. Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., Ziegelmeier, L.: Persistence images: a stable vector representation of persistent homology. Journal of Machine Learning Research **18**(8), 1–35 (2017)
3. Anai, H., Chazal, F., Glisse, M., Ike, Y., Inakoshi, H., Tinarrage, R., Umeda, Y.: Dtm-based filtrations. arXiv preprint arXiv:1811.04757. To appear in the Proc. of the Abel Symposium 2018. (2019)
4. Barannikov, S.A.: The framed Morse complex and its invariants. Adv. Soviet Math. **21**, 93–115 (1994)
5. Bauer, U., Kerber, M., Reininghaus, J., Wagner, H.: PHAT – persistent homology algorithms toolbox. Journal of Symbolic Computation **78**, 76–90 (2017)
6. Biau, G., Chazal, F., Cohen-Steiner, D., Devroye, L., Rodriguez, C.: A weighted k-nearest neighbor density estimate for geometric inference. Electronic Journal of Statistics **5**, 204–237 (2011)
7. Blumberg, A., Gal, I., Mandell, M., Pancia, M.: Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces. Foundations of Computational Mathematics **14**(4), 745–789 (2014). DOI 10.1007/s10208-014-9201-4. URL http://dx.doi.org/10.1007/s10208-014-9201-4
8. Boissonnat, J., Dey, T.K., Maria, C.: The compressed annotation matrix: An efficient data structure for computing persistent cohomology. Algorithmica **73**(3), 607–619 (2015). DOI 10.1007/s00453-015-9999-4. URL https://doi.org/10.1007/s00453-015-9999-4
9. Boissonnat, J., Dyer, R., Ghosh, A., Oudot, S.Y.: Only distances are required to reconstruct submanifolds. Comput. Geom. **66**, 32–67 (2017). DOI 10.1016/j.comgeo.2017.08.001. URL https://doi.org/10.1016/j.comgeo.2017.08.001
10. Boissonnat, J., Maria, C.: Computing persistent homology with various coefficient fields in a single pass. In: Algorithms - ESA 2014 - 22th Annual European Symposium, Wroclaw, Poland, September 8-10, 2014. Proceedings, pp. 185–196 (2014). DOI 10.1007/978-3-662-44777-2_16. URL https://doi.org/10.1007/978-3-662-44777-2_16
11. Boissonnat, J., Maria, C.: The simplex tree: An efficient data structure for general simplicial complexes. Algorithmica **70**(3), 406–427 (2014). DOI 10.1007/s00453-014-9887-3. URL https://doi.org/10.1007/s00453-014-9887-3
12. Boissonnat, J.D., Chazal, F., Yvinec, M.: Geometric and Topological Inference, vol. 57. Cambridge University Press (2018)
13. Boissonnat, J.D., Ghosh, A.: Manifold reconstruction using tangential Delaunay complexes. Discrete and computational Geometry (November) (2103)
14. Boissonnat, J.D., Karthik, C.: An Efficient Representation for Filtrations of Simplicial Complexes. ACM Transactions on Algorithms **14** (2018). URL https://hal.inria.fr/hal-01883836

15. Boissonnat, J.D., Pritam, S., Pareek, D.: Strong Collapse for Persistence. In: ESA 2018 - 26th Annual European Symposium on Algorithms, pp. 67:1–67:13. Helsinki, Finland (2018). DOI 10.4230/LIPIcs. URL https://hal.inria.fr/hal-01886165
16. Boissonnat, J.D., Srikanta, K.C., Tavenas, S.: Building Efficient and Compact Data Structures for Simplicial Complexes. Algorithmica (2016). DOI 10.1007/s00453-016-0207-y. URL https://hal.inria.fr/hal-01364648
17. Brécheteau, C.: The DTM-signature for a geometric comparison of metric-measure spaces from samples (2017)
18. Bubenik, P.: Statistical topological data analysis using persistence landscapes. Journal of Machine Learning Research **16**, 77–102 (2015)
19. Buchet, M., Chazal, F., Dey, T.K., Fan, F., Oudot, S.Y., Wang, Y.: Topological analysis of scalar fields with outliers. In: Proc. Sympos. on Computational Geometry (2015)
20. Buchet, M., Chazal, F., Oudot, S., Sheehy, D.R.: Efficient and robust persistent homology for measures. In: Proceedings of the 26th ACM-SIAM symposium on Discrete algorithms. SIAM. SIAM (2015)
21. Carlsson, G.: Topology and data. AMS Bulletin **46**(2), 255–308 (2009)
22. Carlsson, G., De Silva, V.: Zigzag persistence. Foundations of computational mathematics **10**(4), 367–405 (2010)
23. Carlsson, G., Zomorodian, A.: The theory of multidimensional persistence. Discrete & Computational Geometry **42**(1), 71–93 (2009)
24. Carrière, M., Chazal, F., Ike, Y., Lacombe, T., Royer, M., Umeda, Y.: A general neural network architecture for persistence diagrams and graph classification. arXiv preprint arXiv:1904.09378 (2019)
25. Carriere, M., Cuturi, M., Oudot, S.: Sliced wasserstein kernel for persistence diagrams (2017). To appear in ICML-17
26. Carrière, M., Michel, B., Oudot, S.: Statistical analysis and parameter selection for mapper (2017)
27. Carrière, M., Oudot, S.: Structure and stability of the 1-dimensional mapper. arXiv preprint arXiv:1511.05823 (2015)
28. Chazal, F., Chen, D., Guibas, L., Jiang, X., Sommer, C.: Data-driven trajectory smoothing. In: Proc. ACM SIGSPATIAL GIS (2011)
29. Chazal, F., Cohen-Steiner, D., Glisse, M., Guibas, L., Oudot, S.: Proximity of persistence modules and their diagrams. In: SCG, pp. 237–246 (2009)
30. Chazal, F., Cohen-Steiner, D., Guibas, L.J., Mémoli, F., Oudot, S.Y.: Gromov-hausdorff stable signatures for shapes using persistence. Computer Graphics Forum (proc. SGP 2009) pp. 1393–1403 (2009)
31. Chazal, F., Cohen-Steiner, D., Lieutier, A.: Normal cone approximation and offset shape isotopy. Computational Geometry **42**(6), 566–581 (2009)
32. Chazal, F., Cohen-Steiner, D., Lieutier, A.: A sampling theory for compact sets in euclidean space. Discrete & Computational Geometry **41**(3), 461–479 (2009)
33. Chazal, F., Cohen-Steiner, D., Lieutier, A., Mérigot, Q., Thibert, B.: Inference of curvature using tubular neighborhoods. In: Modern Approaches to Discrete Curvature, pp. 133–158. Springer (2017)
34. Chazal, F., Cohen-Steiner, D., Lieutier, A., Thibert, B.: Stability of Curvature Measures. Computer Graphics Forum (proc. SGP 2009) pp. 1485–1496 (2008)
35. Chazal, F., Cohen-Steiner, D., Mérigot, Q.: Boundary measures for geometric inference. Found. Comp. Math. **10**, 221–240 (2010)
36. Chazal, F., Cohen-Steiner, D., Mérigot, Q.: Geometric inference for probability measures. Foundations of Computational Mathematics **11**(6), 733–751 (2011)
37. Chazal, F., De Silva, V., Oudot, S.: Persistence stability for geometric complexes. Geometriae Dedicata **173**(1), 193–214 (2014)
38. Chazal, F., Divol, V.: The density of expected persistence diagrams and its kernel based estimation. In: 34th International Symposium on Computational Geometry (SoCG 2018) - LIPIcs-Leibniz International Proceedings in Informatics, vol. 99. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2018)

39. Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., Wasserman, L.: Subsampling methods for persistent homology. In: D. Blei, F. Bach (eds.) Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pp. 2143–2151. JMLR Workshop and Conference Proceedings (2015). URL http://jmlr.org/proceedings/papers/v37/chazal15.pdf
40. Chazal, F., Fasy, B.T., Lecci, F., Michel, B., Rinaldo, A., Wasserman, L.: Robust topological inference: Distance to a measure and kernel distance. Journal of Machine Learning Research (2014)
41. Chazal, F., Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L.: Stochastic convergence of persistence landscapes and silhouettes. Journal of Computational Geometry **6**(2), 140–161 (2015)
42. Chazal, F., Glisse, M., Labruère, C., Michel, B.: Convergence rates for persistence diagram estimation in topological data analysis. Journal of Machine Learning Research **16**, 3603–3635 (2015). URL http://jmlr.org/papers/v16/chazal15a.html
43. Chazal, F., Guibas, L., Oudot, S., Skraba, P.: Scalar field analysis over point cloud data. Discrete & Computational Geometry **46**(4), 743–775 (2011). DOI 10.1007/s00454-011-9360-x. URL http://dx.doi.org/10.1007/s00454-011-9360-x
44. Chazal, F., Guibas, L.J., Oudot, S.Y., Skraba, P.: Persistence-based clustering in riemannian manifolds. Journal of the ACM (JACM) **60**(6), 41 (2013)
45. Chazal, F., Lieutier, A.: Smooth manifold reconstruction from noisy and non uniform approximation with guarantees. Computational Geometry Theory and Applications **40**, 156–170 (2008)
46. Chazal, F., Massart, P., Michel, B.: Rates of convergence for robust geometric inference. Electron. J. Statist **10**, 2243–2286 (2016)
47. Chazal, F., Michel, B.: An introduction to topological data analysis: fundamental and practical aspects for data scientists. arXiv preprint arXiv:1710.04019 (2017)
48. Chazal, F., de Silva, V., Glisse, M., Oudot, S.: The structure and stability of persistence modules. SpringerBriefs in Mathematics. Springer (2016)
49. Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. Discrete & Computational Geometry **37**(1), 103–120 (2007)
50. Dindin, M., Umeda, Y., Chazal, F.: Topological data analysis for arrhythmia detection through modular neural networks. arXiv preprint arXiv:1906.05795 (2019)
51. Edelsbrunner, H., Harer, J.: Computational Topology: An Introduction. AMS (2010)
52. Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological persistence and simplification. Discrete Comput. Geom. **28**, 511–533 (2002)
53. Fasy, B.T., Kim, J., Lecci, F., Maria, C.: Introduction to the R package TDA. arXiv preprint arXiv:1411.1830 (2014)
54. Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., Singh, A.: Confidence sets for persistence diagrams. The Annals of Statistics **42**(6), 2301–2339 (2014)
55. Frosini, P.: Measuring shapes by size functions. In: Intelligent Robots and Computer Vision X: Algorithms and Techniques, vol. 1607, pp. 122–133. International Society for Optics and Photonics (1992)
56. Ghrist, R.W.: Elementary applied topology, vol. 1. CreateSpace Independent Publishing Platform
57. Guibas, L., Morozov, D., Mérigot, Q.: Witnessed k-distance. Discrete Comput. Geom. **49**, 22–45 (2013)
58. Hatcher, A.: Algebraic Topology. Cambridge Univ. Press (2001)
59. Hofer, C., Kwitt, R., Niethammer, M., Uhl, A.: Deep learning with topological signatures. In: Advances in Neural Information Processing Systems, pp. 1634–1644 (2017)
60. Laboratories, F.: Estimating the degradation state of old bridges-fijutsu supports ever-increasing bridge inspection tasks with AI technology. Fujitsu Journal (2018). URL https://journal.jp.fujitsu.com/en/2018/03/01/01/
61. Lacombe, T., Cuturi, M., Oudot, S.: Large Scale computation of Means and Clusters for Persistence Diagrams using Optimal Transport. In: NIPS. Montreal, Canada (2018). URL https://hal.inria.fr/hal-01966674

62. Li, C., Ovsjanikov, M., Chazal, F.: Persistence-based structural recognition. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pp. 2003–2010 (2014). DOI 10.1109/CVPR.2014.257
63. Lum, P., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., Carlsson, G.: Extracting insights from the shape of complex data using topology. Scientific reports **3** (2013)
64. Maria, C., Boissonnat, J.D., Glisse, M., Yvinec, M.: The GUDHI library: Simplicial complexes and persistent homology. In: International Congress on Mathematical Software, pp. 167–174. Springer (2014)
65. Mérigot, Q., Ovsjanikov, M., Guibas, L.: Robust Voronoi-based Curvature and Feature Estimation. In: Proc. SIAM/ACM Joint Conference on Geometric and Physical Modeling, pp. 1–12 (2009)
66. Morse, M.: Rank and span in functional topology. Annals of Mathematics **41**(2), 419–454 (1940)
67. Niyogi, P., Smale, S., Weinberger, S.: Finding the homology of submanifolds with high confidence from random samples. Discrete & Computational Geometry **39**(1-3), 419–441 (2008)
68. Oudot, S.Y.: Persistence Theory: From Quiver Representations to Data Analysis, *AMS Mathematical Surveys and Monographs*, vol. 209. American Mathematical Society (2015)
69. Phillips, J.M., Wang, B., Zheng, Y.: Geometric inference on kernel density estimates. arXiv preprint 1307.7760 (2014)
70. Reininghaus, J., Huber, S., Bauer, U., Kwitt, R.: A stable multi-scale kernel for topological machine learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4741–4748 (2015)
71. Royer, M., Chazal, F., Ike, Y., Umeda, Y.: Atol: Automatic topologically-oriented learning. arXiv preprint arXiv:1909.13472 (2019)
72. Singh, G., Mémoli, F., Carlsson, G.E.: Topological methods for the analysis of high dimensional data sets and 3d object recognition. In: SPBG, pp. 91–100. Citeseer (2007)
73. Umeda, Y.: Time series classification via topological data analysis. Transactions of the Japanese Society for Artificial Intelligence **32**(3), D–G72_1 (2017)
74. Verri, A., Uras, C., Frosini, P., Ferri, M.: On the use of size functions for shape analysis. Biological cybernetics **70**(2), 99–107 (1993)
75. Verri, A., Uras, C., Frosini, P., Ferri, M.: On the use of size functions for shape analysis. Biological Cybernetics **70**(2), 99–107 (1993). DOI 10.1007/BF00200823. URL http://dx.doi.org/10.1007/BF00200823
76. Villani, C.: Topics in Optimal Transportation. American Mathematical Society (2003)
77. Wasserman, L.: Topological data analysis. Annual Review of Statistics and Its Application **5**, 501–532 (2018)
78. Yao, Y., Sun, J., Huang, X., Bowman, G.R., Singh, G., Lesnick, M., Guibas, L.J., Pande, V.S., Carlsson, G.: Topological methods for exploring low-density states in biomolecular folding pathways. The Journal of chemical physics **130**(14), 144115 (2009)
79. Zomorodian, A., Carlsson, G.: Computing persistent homology. Discrete Comput. Geom. **33**(2), 249–274 (2005)