Analysis of Scalar Fields over Point Cloud Data

Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, Primoz Skraba

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Analysis of Scalar Fields over Point Cloud Data

Frédéric Chazal — Leonidas J. Guibas — Steve Y. Oudot — Primoz Skraba

N° 6576

July 2008

Thème COM

ISSN 0249-6399 ISRN INRIA/RR--6576--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

Analysis of Scalar Fields over Point Cloud Data

Frédéric Chazal *, Leonidas J. Guibas †, Steve Y. Oudot ‡, Primoz Skraba §

Thème COM — Systèmes communicants
Équipe-Projet Geometrica

Rapport de recherche n° 6576 — July 2008 — 24 pages

Abstract: Given a real-valued function f defined over some metric space $\mathbb{X}$, is it possible to recover some structural information about f from the sole information of its values at a finite set $L \subseteq \mathbb{X}$ of sample points, whose pairwise distances in $\mathbb{X}$ are given? We provide a positive answer to this question. More precisely, taking advantage of recent advances on the front of stability for persistance diagrams, we introduce a novel algebraic construction, based on a pair of nested families of simplicial complexes built on top of the point cloud L, from which the persistance diagram of f can be faithfully approximated. We derive from this construction a series of algorithms for the analysis of scalar fields from point cloud data. These algorithms are simple and easy to implement, they have reasonable complexities, and they come with theoretical guarantees. To illustrate the genericity and practicality of the approach, we also present some experimental results obtained in various applications, ranging from clustering to sensor networks.

Key-words: Persistent homology, Persistence modules, Sampling theory, Vietoris-Rips complexes, Morse theory

* frederic.chazal@inria.fr
† guibas@cs.stanford.edu
‡ steve.oudot@inria.fr
§ primoz@stanford.edu

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Analyse de champs scalaires sur des nuages de points

Résumé : Étant donné une fonction scalaire f définie sur un espace métrique $\mathbb{X}$, est-il possible d'extraire de l'information sur la structure du graphe de f à partir de la seule donnée de ses valeurs sur un ensemble fini L d'échantillons de $\mathbb{X}$, ainsi que des distances géodésiques entre les points de L ? Cet article répond positivement à cette question. Plus précisément, en nous appuyant sur des résultats récents sur la stabilité des diagrammes de persistance, nous introduisons une nouvelle construction algébrique utilisant une paire de familles de complexes simpliciaux imbriqués, à partir de laquelle le diagramme de persistance de f peut être calculé de manière approchée. Nous déduisons de cette construction algébrique une famille d'algorithmes pour l'analyse des champs scalaires à partir de nuages de points. Ces algorithmes sont simples et faciles à implanter, ils ont des complexités raisonnables, ainsi que des garanties théoriques. Afin d'illustrer la généricité de notre approche, nous présentons des résultats expérimentaux obtenus dans diverses applications, comme la classification ou les réseaux de capteurs sans fils.

Mots-clés : Homologie persistante, modules de persistance, théorie de l'échantillonnage, complexes de Rips-Vietoris, théorie de Morse.

1 Introduction

Suppose we are given a collection of sensors spread out in some planar region, and suppose that these sensors measure some intensive 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, is it possible to recover some high-level information about the measured quantity, such as the number of its peaks or valleys, as well as a sense of their prominence? Consider now the case were we are given a finite set of sample points in Euclidean space, drawn from some unknown probability density f. Suppose that we can compute at each of these points a rough estimate of the local density. Can we then infer the number of prominent peaks of f, which we could later use as the input parameter to a clustering algorithm? Can we tell how to merge the basins of attraction of the maxima, in order to guide the clustering? Consider finally the case where a movie database is provided together with a similarity measure between movies and a measure of popularity for each movie. Can we extract the prominent peaks of the popularity measure, so as to provide information on the general trends of the public's tastes?

These three scenarios are just special instances of a same generic problem: given an unknown domain $\mathbb{X}$ and a scalar field $f : \mathbb{X} \to \mathbb{R}$ whose values are known only at a finite set L of sample points, the goal is to extract some structural information about f from the sole information of the pairwise distances between the data points and of their function values. The nature of the sought-for information is highly application-dependent. In the above scenarios one is mainly interested in finding the peaks and valleys of the function, together with their respective basins of attraction[1]. In addition, it is desirable to have a mechanism for distinguishing between significant and insignificant peaks or valleys of f, which requires to introduce some notion of prominance for the critical points of a function. This is where *topological persistence* comes into play: inspired from Morse theory, this framework describes the evolution of the topology of the sublevel-sets of f, *i.e.* the sets of type $f^{-1}((-\infty, a])$, as parameter a ranges from $-\infty$ to $+\infty$. Topological changes occur only at critical points of f, which can be paired in some natural way. For instance, a new connected component appears in $f^{-1}((-\infty, a])$ when a reaches the f-value of a local minimum, and this component gets connected to the rest of the sublevel-set as a reaches the f-value of a saddle. The outcome of this process is a set of intervals, called a *persistence barcode*, each of which corresponds to a pair of critical points and gives the birth and death times of a homological feature of the sublevel-sets of f — see Figure 1. An equivalent representation is by a multiset of points in the plane, called a *persistence diagram*, where the coordinates of each point correspond to the endpoints of some interval in the barcode. Such barcodes or diagrams can be used to guide the simplification of the graphs of real-valued functions by iterative cancellations of critical pairs [14, 15]. As such, they provide the desired information for evaluating the prominence of the peaks and valleys (and in fact of all the critical points) of a scalar field.

Thus, our problem becomes the following: given $\mathbb{X}, f, L$ as above, is it possible to approximate the persistence diagram of f from the pairwise distances between the points of L and from the values of f at these points? The main contribution of the paper is a positive answer to this question. More precisely, in Section 3 we exhibit a novel algebraic construction, based on a pair of nested families of simplicial complexes — derived from the so-called *Rips complexes* of L, defined below — from which the persistence diagram of f can be approximated (Theorem 3.1). We also show the robustness of our construction with respect to noise in the pairwise distances or function values (Theorems 3.7 and 3.9). From these structural results we derive algorithms (Section 4) for approximating the persistence diagram of f from its values at a finite set of samples, both in static (fixed f) and in dynamic (time-varying f) settings. We also give a procedure for finding the basins of attraction of the peaks of f inside the point cloud L, and for

[1] In the context of clustering, this approach to the problem is reminiscent of Mean Shift [11].

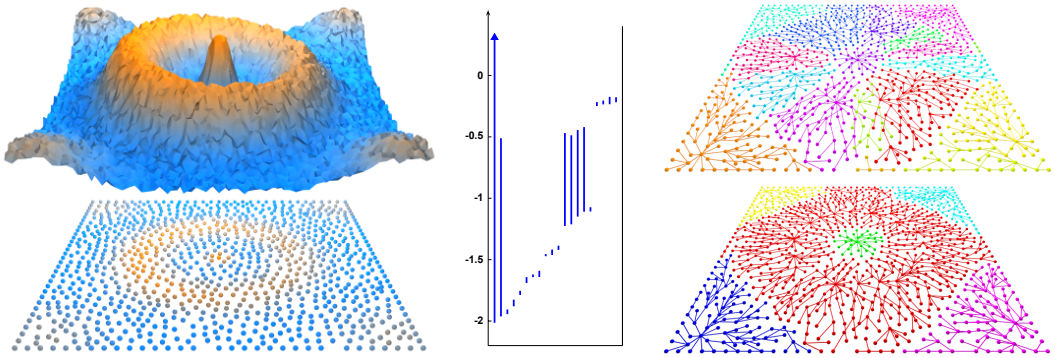

Figure 1: *Left: a noisy scalar field f defined over a planar square domain $\mathbb{X}$. Center: approximation of the 0-dimensional persistence barcode of $(-f)$ from a finite sampling of $\mathbb{X}$: the long intervals correspond to the six prominent peaks (including the top of the crater) of f. Right: approximate basins of attraction of the peaks of f in L, before (top) and after (bottom) merging non-persistent clusters, thus revealing the intuitive structure of f.*

merging these basins according to the persistence information, as illustrated in Figure 1 (right). Our algorithms are based on variants [9, 10] of the celebrated persistence algorithm. They can be easily implemented, they have reasonable complexities, and they are provably correct for the most part. To illustrate the versatility of the approach, we provide experimental results obtained in a variety of applications (Section 5): without pretending in any way to give definitive solutions to the considered problems, we aim at showing the potential of our method and its possible interest for the community.

Related work. Topological persistence and its applications have been an extensively studied topic since the introduction of the persistence algorithm by Edelsbrunner *et al.* [14]. First designed for simplicial complexes in $\mathbb{R}^3$, this algorithm was later extended to compute the persistent homology of discrete functions over arbitrary finite simplicial complexes [27]. A number of variants were also proposed, for instance to cope with changes in the function over time [10] or to handle pairs of functions defined over nested pairs of spaces [9]. All these methods deal with functions defined over simplicial complexes, and in some sense our work suggests a way of extending the approach to a more general class of spaces *via* finite sampling and modulo some (controlled) errors in the output.

Topological persistence has already been used in the past for the analysis and simplification of scalar fields. The original persistence paper [14] showed how to simplify the graph of a piecewise-linear (PL) real-valued function f defined over a simplicial complex $\mathbb{X}$ in $\mathbb{R}^3$, by iteratively cancelling the pairs of critical points provided by the persistence barcode of f. This approach was later refined, in the special case where $\mathbb{X}$ is a triangulated 2-manifold, to only cancel the pairs corresponding to short intervals in the barcode, thus removing all topological noise up to a certain prescribed amplitude [15]. In parallel, people have considered computing accurate or simplified representations of Morse-Smale complexes, which capture important information about the structure of scalar fields. Indeed, the Morse-Smale complex of a function $f : \mathbb{X} \to \mathbb{R}$ is a partition of the space $\mathbb{X}$ into regions where the flow induced by the gradient vector field of f is uniform. Building upon the idea of iterative cancellations of pairs of critical points, it is possible to construct hierarchies of increasingly coarse Morse-Smale complexes from PL functions defined over triangulated 2- or 3-manifolds [1, 3, 13, 19, 20]. Although our question of finding the basins

of attraction of the peaks of a scalar field may seem a simplistic variant of the above problems, we claim that it is in fact not, as our context is much more general and our knowledge of the function f is much weaker. In particular, the resort to a PL approximation of f in our potentially high-dimensional or even non-Euclidean setting would be prohibitively costly, if not impossible. Note also that, in applications such as the scenarios described at the beginning of the section, the knowledge of the basins of attraction of the (significant) peaks of the scalar field is sufficient for further processing.

A last trend of work in which persistence has played a prominent role is homology inference from point cloud data, where the goal is to recover the homological type of an unknown space $\mathbb{X}$ from a finite set L of sample points. The idea is to consider the function *distance to* L, either inside $\mathbb{X}$ or in some ambient space $\mathbb{Y}$ where $\mathbb{X}$ is embedded. Under sufficient sampling density, this function approximates the distance to $\mathbb{X}$, and therefore their persistence diagrams are close, by a stability result due to Cohen-Steiner *et al.* [8]. Thus, the sole knowledge of the sample points is enough to approximate the persistence diagram of the distance to $\mathbb{X}$, from which the homology of $\mathbb{X}$ is easily inferred [6, 8]. In practice however, the cost of estimating the distance to L at every point of $\mathbb{X}$ or of some ambient space $\mathbb{Y}$ is prohibitive, thus requiring the resort to auxiliary algebraic constructions. Among the most popular ones is the Rips complex $R_\alpha(L)$, which is the abstract simplicial complex whose simplices correspond to non-empty subsets of L of diameter less than α. The building of this complex only involves comparisons of distances, which makes it a good candidate data structure in practice. Furthermore, as proved in [7], a pair of nested Rips complexes $R_\alpha(L) \subseteq R_\beta(L)$ can provably-well capture the homology of the underlying space $\mathbb{X}$, even though none of the individual complexes does. Our algebraic construction (see Section 3) is directly inspired from this property, and in fact our theoretical analysis is articulated in the same way as in [7], namely: we first work out structural properties of unions of geodesic balls, which we prove to also hold for their nerves (also called *Čech complexes*); then, using strong relationships between families of Čech and of Rips complexes, we derive structural properties for the latter. Note however that the core of the analysis differs significantly from [7], because our families of complexes are built differently. In particular, the classical notion of stability for persistence diagrams, as introduced in [8], is not broad enough for our setting, where it is replaced by a generalized notion recently proposed by Chazal *et al.* [4].

2 Background

Our analysis uses singular homology with coefficients in a commutative ring R, assumed to be a field throughout the paper and omitted in the notations. We also use some elements of Riemannian geometry, as well as of Morse theory (mainly in Section 4.3). We refer the reader to [2, 21, 22] for comprehensive introductions to these topics.

2.1 Geodesic ε-samples on Riemannian manifolds

Throughout the paper, and unless otherwise stated, $\mathbb{X}$ denotes a compact Riemannian manifold possibly with boundary, and $\mathrm{d}_\mathbb{X}$ denotes its geodesic distance. Our analysis turns out to hold for a larger class of length spaces, however for simplicity we restrict the focus of the paper to the Riemannian setting. Given a point $x \in \mathbb{X}$ and a real value $r \geq 0$, let $B_\mathbb{X}(x, r)$ denote the open geodesic ball of center x and radius r, namely: $B_\mathbb{X}(x, r) = \{y \in \mathbb{X},\ \mathrm{d}_\mathbb{X}(x, y) < r\}$. For all sufficiently small values $r \geq 0$, the ball $B_\mathbb{X}(x, r)$ is known to be *strongly convex*, that is: for every pair of points y, y' in the closure of $B_\mathbb{X}(x, r)$, there exists a unique shortest path in $\mathbb{X}$ between y and y', and the interior of this path is included in $B_\mathbb{X}(x, r)$. Let $\varrho_c(x) > 0$ be the supremum of the radii such that this property holds. Since $\mathbb{X}$ is compact, the infimum of $\varrho_c(x)$ over the points

of $\mathbb{X}$ is positive, and known as the *strong convexity radius* of $\mathbb{X}$, noted $\varrho_c(\mathbb{X})$. This quantity plays an important role in the paper because strongly convex sets are contractible[2], and because intersections of strongly convex sets are also strongly convex.

In the sequel, L denotes a finite set of points of $\mathbb{X}$ that form a *geodesic ε-sample* of $\mathbb{X}$, namely: $\forall x \in \mathbb{X}$, $\mathrm{d}_{\mathbb{X}}(x, L) < \varepsilon$. Here, parameter ε is homogeneous to a length, and it controls the density of the point cloud L. Our theoretical claims will assume this density to be high enough, *via* a condition on ε stipulating that the latter is at most a fraction of the strong convexity radius of $\mathbb{X}$.

2.2 Persistence modules and filtrations

The main algebraic objects under study here are *persistence modules*. A persistence module is a family $\{\Phi_\alpha\}_{\alpha\in\mathbb{R}}$ of R-modules together with a family $\{\phi_\alpha^\beta : \Phi_\alpha \to \Phi_\beta\}_{\alpha\leq\beta\in\mathbb{R}}$ of homomorphisms such that $\forall \alpha \leq \beta \leq \gamma$, $\phi_\alpha^\gamma = \phi_\beta^\gamma \circ \phi_\alpha^\beta$ and $\phi_\alpha^\alpha = \mathrm{id}_{\Phi_\alpha}$. Persistence modules are often derived from *filtrations*, which are families $\{F_\alpha\}_{\alpha\in\mathbb{R}}$ of topological spaces that are nested with respect to inclusion. For all $\alpha \leq \beta$, the canonical inclusion map $F_\alpha \hookrightarrow F_\beta$ induces homomorphisms between the homology groups $H_k(F_\alpha) \to H_k(F_\beta)$ of all dimensions $k \in \mathbb{N}$. Thus, for any fixed k the family $\{H_k(F_\alpha)\}_{\alpha\in\mathbb{R}}$ forms a persistence module, called *kth persistent homology module of* $\{F_\alpha\}_{\alpha\in\mathbb{R}}$, where the homomorphisms between R-modules are understood to be those induced by inclusions.

An important class of filtrations are the ones formed by the sublevel-sets of real-valued functions. Given a topological space $\mathbb{X}$ and a function $f : \mathbb{X} \to \mathbb{R}$, the *sublevel-sets filtration* of f is the family $\{F_\alpha\}_{\alpha\in\mathbb{R}}$ of subspaces of $\mathbb{X}$ of type $F_\alpha = f^{-1}((-\infty, \alpha])$. This family forms a filtration because $f^{-1}((-\infty, \alpha]) \subseteq f^{-1}((-\infty, \beta])$ whenever $\alpha \leq \beta$. A real-valued function that has played a prominent role in homology inference is the geodesic distance to a finite point cloud L. The 0-sublevel set of this function is L itself, while for any $\alpha > 0$ its α-sublevel set is the closure of the so-called *α-offset* L^α, defined as the union of the open geodesic balls of radius α about the points of L, namely: $L^\alpha = \bigcup_{p\in L} B_{\mathbb{X}}(p, \alpha)$. Important structural properties of growing families of open balls, some of which will be exploited in Section 3.1 of this paper, follow from the properties of the sublevel-sets of the distance function [5, 23].

Since the offsets of a point cloud L can be difficult to manipulate, they are often replaced by purely combinatorial constructions in practice. A natural choice is to use the *nerve* of the family of open geodesic balls used in the definition of the α-offset of L. Specifically, the nerve of the family $\{B_{\mathbb{X}}(p, \alpha)\}_{p\in L}$ is the abstract simplicial complex of vertex set L whose simplices correspond to non-empty subsets of the family whose elements have a non-empty common intersection. This complex is also known as the *Čech complex*, and noted $C_\alpha(L)$. Thanks to the duality that exists between unions of open balls and their nerves (see Lemma 3.4 below), Čech complexes $C_\alpha(L)$ enjoy many interesting properties that will be exploited as well in Section 3.1.

An even simpler combinatorial construction is the so-called *(Vietoris-)Rips complex* $R_\alpha(L)$, which is the abstract simplicial complex of vertex set L whose simplices correspond to non-empty subsets of L of geodesic diameter less than α. The building of this complex only involves comparisons of distances, which makes it a good candidate data structure in practice. Furthermore, Rips complexes are known to be closely related to Čech complexes through the following sequence of inclusions, which holds in any arbitrary metric space (see *e.g.* [7]):

$$\forall \alpha > 0,\ C_{\frac{\alpha}{2}}(L) \subseteq R_\alpha(L) \subseteq C_\alpha(L) \tag{1}$$

[2] A topological space is contractible if it can be continuously deformed to a point within itself.

Several other combinatorial constructions, such as the α-shape or the witness complex, have been proven to be useful in the context of homology inference. These will not be considered in the paper.

Finally, let us mention that the above constructions are parametrized by a unique quantity α, which one usually lets vary from 0 to $+\infty$ to get a filtration. In contrast, our filtrations will be obtained by fixing α to some constant value and by letting the vertex set grow from $\emptyset$ to L.

2.3 Persistence diagrams and stability

Persistence diagrams have been introduced as a succint way of describing the algebraic structure of a persistence module [27]. There is a restriction though: without any further assumptions, the algebraic structure of a persistence module can be arbitrarily complicated, thereby making it impossible to find a descriptor that is both succint and complete. This is where the concept of *tameness*[3] comes into play:

Definition 2.1 *A persistence module* $(\{\Phi_\alpha\}_{\alpha\in\mathbb{R}}, \{\phi_\alpha^\beta\}_{\alpha\leq\beta\in\mathbb{R}})$ *is* tame *if* $\forall \alpha < \beta$, $\text{rank}\ \phi_\alpha^\beta < +\infty$.

This condition is restrictive enough for persistence diagrams to be well-defined, yet the concept of tameness remains sufficiently wide to encompass a large class of persistence modules. In particular, all persistent homology modules of nested families of finite simplicial complexes are tame. As a consequence, all persistence modules introduced in Sections 3 and 4 of this paper will be tame.

Following [4], the persistence diagram of a tame persistence module $(\{\Phi_\alpha\}_{\alpha\in\mathbb{R}}, \{\phi_\alpha^\beta\}_{\alpha\leq\beta\in\mathbb{R}})$ is defined as a multiset of points in the extended plane $\bar{\mathbb{R}}^2$, where $\bar{\mathbb{R}} = \mathbb{R}\cup\{-\infty, +\infty\}$. This multiset is obtained as the limit of the following iterative process: given arbitrary values $a, \varepsilon > 0$, we discretize the persistence module over the integer scale $a+\varepsilon\mathbb{Z}$, considering the subfamily $\{\Phi_{a+k\varepsilon}\}_{k\in\mathbb{Z}}$ of vector spaces together with the subfamily $\{\phi_{a+k\varepsilon}^{a+l\varepsilon}\}_{k\leq l\in\mathbb{Z}}$ of linear maps. Its persistence diagram is defined naturally[4] as the set of vertices of the regular grid $(a+\varepsilon\mathbb{Z})\times(a+\varepsilon\mathbb{Z})$ in $\bar{\mathbb{R}}^2$, plus the diagonal $\Delta = \{(x,x),\ x\in\bar{\mathbb{R}}\}$, where each grid vertex $(a+k\varepsilon, a+l\varepsilon)$ is given the (finite) multiplicity $\text{mult}(a+k\varepsilon, a+l\varepsilon) = \text{rank}\ \phi_{a+k\varepsilon}^{a+(l-1)\varepsilon} - \text{rank}\ \phi_{a+k\varepsilon}^{a+l\varepsilon} + \text{rank}\ \phi_{a+(k-1)\varepsilon}^{a+l\varepsilon} - \text{rank}\ \phi_{a+(k-1)\varepsilon}^{a+(l-1)\varepsilon}$, while each point of Δ is given infinite multiplicity. Then, the persistence diagram of $(\{\Phi_\alpha\}_{\alpha\in\mathbb{R}}, \{\phi_\alpha^\beta\}_{\alpha\leq\beta\in\mathbb{R}})$ is the limit multiset obtained as $\varepsilon\to 0$, which is known to be independent of the choice of a [4].

An important property of persistence diagrams is their stability under small perturbations. Cohen-Steiner *et al.* [8] proposed the first result in this vein: given two tame continuous real-valued functions f, g defined over a same triangulable space $\mathbb{X}$, for all $k\in\mathbb{N}$ the bottleneck distance between the persistence diagrams of the kth persistent homology modules of their sublevel-sets filtrations is at most $\sup_{x\in\mathbb{X}}|f(x)-g(x)|$. Recall that the bottleneck distance $\text{d}_B^\infty(A,B)$ between two multisets in $\bar{\mathbb{R}}^2$ endowed with the l^∞ norm is the quantity $\min_\gamma \max_{p\in A}\|p-\gamma(p)\|_\infty$, where γ ranges over all bijections from A to B. Recently, Chazal *et al.* [4] extended this result by dropping the continuity and triangulability conditions, as well as the functional setting. To do so, they had to introduce a new notion of proximity between persistence modules:

Definition 2.2 *Two persistence modules* $(\{\Phi_\alpha\}_{\alpha\in\mathbb{R}},\ \{\phi_\alpha^\beta\}_{\alpha\leq\beta\in\mathbb{R}})$ *and* $(\{\Psi_\alpha\}_{\alpha\in\mathbb{R}},\ \{\psi_\alpha^\beta\}_{\alpha\leq\beta\in\mathbb{R}})$ *are (strongly)* ε-interleaved *if there exist two families of homomorphisms,* $\{\mu_\alpha : \Phi_\alpha\to\Psi_{\alpha+\varepsilon}\}_{\alpha\in\mathbb{R}}$ *and* $\{\nu_\alpha : \Psi_\alpha\to\Phi_{\alpha+\varepsilon}\}_{\alpha\in\mathbb{R}}$*, such that* $\phi_\alpha^\beta = \nu_{\beta-\varepsilon}\circ\psi_{\alpha+\varepsilon}^{\beta-\varepsilon}\circ\mu_\alpha$ *and* $\psi_\alpha^\beta = \mu_{\beta-\varepsilon}\circ\phi_{\alpha+\varepsilon}^{\beta-\varepsilon}\circ\nu_\alpha$ *for all* $\beta\geq\alpha+2\varepsilon$.

Under these conditions, Chazal *et al.* proved the following generalized stability result [4]:

[3] We borrow this concept from [4], where it is called 0-*tameness* and made weaker than in [8].

[4] In the particular case of a discretized persistence module, this definition does coincide with the one of [8].

Theorem 2.3 *If two tame persistence modules are ε-interleaved, then, in the extended plane $\bar{\mathbb{R}}^2$ endowed with the l^∞ norm, the bottleneck distance between their persistence diagrams is at most ε.*

An important special case is that of the kth persistent homology modules of two filtrations $\{F_\alpha\}_{\alpha\in\mathbb{R}}$ and $\{G_\alpha\}_{\alpha\in\mathbb{R}}$ such that $F_\alpha \subseteq G_{\alpha+\varepsilon}$ and $G_\alpha \subseteq F_{\alpha+\varepsilon}$ for all $\alpha \in \mathbb{R}$. In this case, the maps μ_α and ν_α induced at homology level by the inclusions $F_\alpha \hookrightarrow G_{\alpha+\varepsilon}$ and $G_\alpha \hookrightarrow F_{\alpha+\varepsilon}$ make the two persistence modules ε-interleaved, and Theorem 2.3 guarantees that their persistence diagrams are ε-close.

Note about the exposition. In order to simplify the exposition in the sequel, we allow ourselves some degree of sloppiness in the notations. Specifically, we omit the ranges of the indices when these are obvious, thus designating a filtration of parameter $\alpha \in \mathbb{R}$ by $\{F_\alpha\}$, and a persistence module of parameters $\alpha \leq \beta \in \mathbb{R}$ by $(\{\Phi_\alpha\}, \{\phi_\alpha^\beta\})$. In addition, we use the following *shortcuts*: the persistence diagram of the kth persistent homology module of a filtration $\{F_\alpha\}$ is simply called the kth persistence diagram of $\{F_\alpha\}$. Furthermore, the filtration itself is said to be tame if its kth persistent homology module is tame for all values $k \in \mathbb{N}$. At a higher level, the kth persistent homology module of a real-valued function f refers by default to the kth persistent homology module of its sublevel-sets filtration, and f is said to be tame if its sublevel-sets filtration is tame. Finally, the kth persistence diagram of f is the kth persistence diagram of its sublevel-sets filtration.

3 Structural properties

Let $\mathbb{X}$ be a Riemannian manifold, possibly with boundary, and let $f : \mathbb{X} \to \mathbb{R}$ be a tame c-Lipschitz function. Assuming $\mathbb{X}$ and f to be unknown, we want to approximate the kth persistence diagram of f from the values of the function at a finite set L of sample points that form a geodesic ε-sample of $\mathbb{X}$. The main result of the section (Theorem 3.1 below) claims that this is possible using an algebraic construction based on Rips complexes. The main advantage of this construction is that it leads to an easy-to-compute data structure, which will be described in the algorithms Section 4. From now on, L_α denotes the set $L \cap f^{-1}((-\infty, \alpha])$.

Our construction is inspired from [7], where it is shown that a pair of nested Rips complexes can provably-well capture the homology of a domain even though none of the individual Rips complexes does. Given a fixed parameter $\delta > 0$, we use two Rips-based filtrations simulatneously, $\{R_\delta(L_\alpha)\}_{\alpha\in\mathbb{R}}$ and $\{R_{2\delta}(L_\alpha)\}_{\alpha\in\mathbb{R}}$, and we consider the persistence modules formed at homology level by the images of the homomorphisms induced by the inclusions $R_\delta(L_\alpha) \hookrightarrow R_{2\delta}(L_\alpha)$. Specifically, for all $k \in \mathbb{N}$ and all $\alpha \leq \beta$ we have the following induced commutative diagram at homology level:

$$\begin{array}{ccc} H_k(R_\delta(L_\beta)) & \to & H_k(R_{2\delta}(L_\beta)) \\ \uparrow & & \uparrow \\ H_k(R_\delta(L_\alpha)) & \to & H_k(R_{2\delta}(L_\alpha)) \end{array}$$

Letting Γ_α^k be the image of $H_k(R_\delta(L_\alpha)) \to H_k(R_{2\delta}(L_\alpha))$, we get that the above commutative diagram induces a map $\gamma_\alpha^\beta : \Gamma_\alpha^k \to \Gamma_\beta^k$. Since this is true for all $\alpha \leq \beta$, the family $\{\Gamma_\alpha^k\}_{\alpha\in\mathbb{R}}$ of vector spaces, together with the family $\{\gamma_\alpha^\beta\}_{\alpha\leq\beta}$ of linear maps, forms a persistence module. By analogy with the terminology of Section 2, we call it the kth persistent homology module of the nested pair of filtrations $\{R_\delta(L_\alpha) \hookrightarrow R_{2\delta}(L_\alpha)\}_{\alpha\in\mathbb{R}}$, and its persistence diagram the kth persistence diagram of the nested pair. This construction is in fact not specific to families of Rips complexes, and it allows to define a persistence module $(\{\Gamma_\alpha^k\}, \{\gamma_\alpha^\beta\})$ from the k-dimensional

homology groups of any pair of filtrations $\{G_\alpha\}$ and $\{G'_\alpha\}$ that is nested with respect to inclusion: $\forall \alpha \in \mathbb{R}$, $G_\alpha \subseteq G'_\alpha$.

Theorem 3.1 *Let $\mathbb{X}$ be a compact Riemannian manifold, possibly with boundary, and $f : \mathbb{X} \to \mathbb{R}$ a tame c-Lipschitz function. Let also L be a geodesic ε-sample of $\mathbb{X}$. If $\varepsilon < \frac{1}{4}\varrho_c(\mathbb{X})$, then for any $\delta \in [2\varepsilon,\ \frac{1}{2}\varrho_c(\mathbb{X}))$ and any $k \in \mathbb{N}$, the kth persistent homology modules of f and of the nested pair of filtrations $\{R_\delta(L_\alpha) \hookrightarrow R_{2\delta}(L_\alpha)\}_{\alpha\in\mathbb{R}}$ are $2c\delta$-intreleaved. Therefore, the bottleneck distance between their persistence diagrams is at most $2c\delta$, by Theorem 2.3.*

In practice, the kth persistent homology module of the pair of filtrations $\{R_\delta(L_\alpha) \hookrightarrow R_{2\delta}(L_\alpha)\}_{\alpha\in\mathbb{R}}$ does not have to be built explicitly since its persistence diagram can be computed directly from the filtrations $\{R_\delta(L_\alpha)\}_{\alpha\in\mathbb{R}}$ and $\{R_{2\delta}(L_\alpha)\}_{\alpha\in\mathbb{R}}$ [9]. The next two sections are devoted to the proof of Theorem 3.1. The core argument, based on a technique of algebraic topology called *diagram chasing*, is presented in Section 3.2. It makes use of preliminary results on unions of balls and their nerves, introduced in Section 3.1. Finally, Section 3.3 addresses the robustness of our main result with respect to small perturbations of the geodesic distances or function values.

3.1 Preliminaries: unions of geodesic balls and their nerves

Let $\delta > 0$ be a fixed parameter. Consider the filtration $\{L^\delta_\alpha\}_{\alpha\in\mathbb{R}}$ formed by the δ-offsets of the subsets L_α. Recall that the δ-offset of L_α is defined by $L^\delta_\alpha = \bigcup_{p\in L_\alpha} B_{\mathbb{X}}(p,\delta)$.

Lemma 3.2 *Let $\mathbb{X}, f, L$ be as in Theorem 3.1. Then, for any $\delta \geq \varepsilon$, the sublevel-sets filtration $\{F_\alpha\}$ of f is $c\delta$-interleaved with $\{L^\delta_\alpha\}_{\alpha\in\mathbb{R}}$. Hence, $\forall k \in \mathbb{N}$, the bottleneck distance between their kth persistence diagrams is at most $c\delta$, by Theorem 2.3.*

Proof. Consider an arbitrary value $\alpha \in \mathbb{R}$ and take a point $p \in F_\alpha$. Since L is a geodesic ε-sample of $\mathbb{X}$, there exists some point $q \in L$ such that $\mathrm{d}_{\mathbb{X}}(p,q) < \varepsilon \leq \delta$. Since f is c-Lipschitz, we have $f(q) \leq f(p) + c\delta \leq \alpha + c\delta$, which implies that $q \in L \cap F_{\alpha+c\delta}$. Hence, p belongs to $L^\delta_{\alpha+c\delta}$. Reciprocally, take a point $p \in L^\delta_\alpha$. By definition, there exists some point $q \in L_\alpha$ such that $\mathrm{d}_{\mathbb{X}}(p,q) < \delta$. Since f is c-Lipschitz, we have $f(p) \leq f(q) + c\delta \leq \alpha + c\delta$. Therefore, p belongs to $F_{\alpha+c\delta}$. This proves that $\{F_\alpha\}$ and $\{L^\delta_\alpha\}$ are $c\delta$-interleaved. □

We now turn our focus to the nerves $C_\delta(L_\alpha)$ of the offsets L^δ_α, where δ remains a fixed parameter:

Lemma 3.3 *Let $\mathbb{X}, f, L$ be as in Theorem 3.1. If $\varepsilon < \varrho_c(\mathbb{X})$, then, for any $\delta \in [\varepsilon,\ \varrho_c(\mathbb{X}))$ and any $k \in \mathbb{N}$, the bottleneck distance between the kth persistence diagrams of f and of the filtration $\{C_\delta(L_\alpha)\}_{\alpha\in\mathbb{R}}$ is at most $c\delta$.*

The proof of the lemma, detailed below, relies on the following technical result[5] from [7], which relates *good* open covers to their nerves. Given a topological space $\mathbb{X}$ and a family $\{U_a\}_{a\in A}$ of open subsets covering $\mathbb{X}$, the family defines a *good* cover if, for every finite subset S of A, the common intersection $\bigcap_{a\in S} U_a$ is either empty or contractible.

Lemma 3.4 (Lemma 3.4 of [7]) *Let $\mathbb{X} \subseteq \mathbb{X}'$ be two paracompact spaces, and let $\mathcal{U} = \{U_a\}_{a\in A}$ and $\mathcal{U}' = \{U'_{a'}\}_{a'\in A'}$ be good open covers of $\mathbb{X}$ and $\mathbb{X}'$ respectively, based on finite parameter sets $A \subseteq A'$, such that $U_a \subseteq U'_a$ for all $a \in A$. Then, the homotopy equivalences $\mathcal{NU} \to \mathbb{X}$ and $\mathcal{NU}' \to \mathbb{X}'$ provided by the Nerve Theorem [21, §4G] commute with the canonical inclusions $\mathbb{X} \hookrightarrow \mathbb{X}'$ and $\mathcal{NU} \hookrightarrow \mathcal{NU}'$ at homology level.*

[5]Note that the statement of Lemma 3.4 of [7] assumes the parameter sets A, A' to be equal. However, the proof of the lemma only uses the facts that $A \subseteq A'$ and that the cover $\mathcal{U}$ is subordinate to the cover $\mathcal{U}'$ on its own index set A. In addition, the statement of the lemma does not specify that the homotopy equivalences considered are the ones provided by the Nerve Theorem [21, §4G], but this appears clearly in the proof of the lemma.

Proof of Lemma 3.3. We claim that, for all $\alpha \in \mathbb{R}$, the family of open balls $\{B_{\mathbb{X}}(p,\delta)\}_{p\in L_\alpha}$ forms a good open cover of the set L_α^δ, that is: $\forall l \in \mathbb{N}$, $\forall p_1, \cdots, p_l \in L_\alpha$, the intersection $I = B_{\mathbb{X}}(p_1,\delta) \cap \cdots \cap B_{\mathbb{X}}(p_l,\delta)$ is either empty or contractible. Indeed, assuming that I is non-empty, we have that each ball $B(p_i,\delta)$ is strongly convex because $\delta < \varrho_c(\mathbb{X})$. As a consequence, I itself is strongly convex and therefore contractible, as mentioned in Section 2.1. Thus, $\{B_{\mathbb{X}}(p,\delta)\}_{p\in L_\alpha}$ forms a good open cover of L_α^δ. Since this is true for all $\alpha \in \mathbb{R}$, Lemma 3.4 guarantees that, for all $k \in \mathbb{N}$ and all $\alpha \leq \beta$, the rank of the homomorphism $H_k(C_\delta(L_\alpha)) \to H_k(C_\delta(L_\beta))$ induced by inclusion is the same as the rank of $H_k(L_\alpha^\delta) \to H_k(L_\beta^\delta)$. Therefore, the filtrations $\{C_\delta(L_\alpha)\}_{\alpha\in\mathbb{R}}$ and $\{L_\alpha^\delta\}_{\alpha\in\mathbb{R}}$ have identical kth persistence diagrams. The result follows then from Lemma 3.2. □

With these preliminary results at hand, we can now proceed to the proof of our main result.

3.2 Proof of Theorem 3.1

We will in fact prove the following more general (yet technical) result:

Lemma 3.5 *Let $\mathbb{X}, f, L$ be as in Theorem 3.1. Suppose there exist $\varepsilon' \leq \varepsilon'' \in [\varepsilon,\ \varrho_c(\mathbb{X}))$ and two filtrations, $\{G_\alpha\}$ and $\{G'_\alpha\}$, such that: $\forall \alpha \in \mathbb{R}$, $C_\varepsilon(L_\alpha) \subseteq G_\alpha \subseteq C_{\varepsilon'}(L_\alpha) \subseteq G'_\alpha \subseteq C_{\varepsilon''}(L_\alpha)$. Then, $\forall k \in \mathbb{N}$, the kth persistent homology modules of f and of the nested pair of filtrations $\{G_\alpha \hookrightarrow G'_\alpha\}_{\alpha\in\mathbb{R}}$ are $c\varepsilon''$-intreleaved.*

Applying Lemma 3.5 with $\varepsilon' = \delta$, $\varepsilon'' = 2\delta$, $G_\alpha = R_\delta(L_\alpha)$ and $G'_\alpha = R_{2\delta}(L_\alpha)$ gives Theorem 3.1, the sequence of inclusions assumed in the statement of Lemma 3.5 being ensured by Eq. (1) in this case. Lemma 3.5 itself will be instrumental in Section 3.3, in proving the robustness of our main result with respect to small perturbations of geodesic distances or function values.

Proof of Lemma 3.5. Let $k \in \mathbb{N}$, and let $(\{\Gamma_\alpha^k\}, \{\gamma_\alpha^\beta\})$ denote the kth persistent homology module of the nested pair of filtrations $\{G_\alpha \hookrightarrow G'_\alpha\}_{\alpha\in\mathbb{R}}$. For all $\alpha \leq \beta$, the sequence of inclusions assumed in the statement of the lemma induces the following commutative diagram at homology level:

$$\begin{array}{ccccccccc}
H_k(C_\varepsilon(L_\beta)) & \stackrel{a_\beta}{\to} & H_k(G_\beta) & \stackrel{b_\beta}{\to} & H_k(C_{\varepsilon'}(L_\beta)) & \stackrel{d_\beta}{\to} & H_k(G'_\beta) & \stackrel{e_\beta}{\to} & H_k(C_{\varepsilon''}(L_\beta)) \\
\uparrow i_\alpha^\beta & & \uparrow j_\alpha^\beta & & \uparrow l_\alpha^\beta & & \uparrow m_\alpha^\beta & & \uparrow n_\alpha^\beta \\
H_k(C_\varepsilon(L_\alpha)) & \stackrel{a_\alpha}{\to} & H_k(G_\alpha) & \stackrel{b_\alpha}{\to} & H_k(C_{\varepsilon'}(L_\alpha)) & \stackrel{d_\alpha}{\to} & H_k(G'_\alpha) & \stackrel{e_\alpha}{\to} & H_k(C_{\varepsilon''}(L_\alpha))
\end{array} \tag{2}$$

This diagram encodes important relations between the persistence module $(\{\Gamma_\alpha^k\}, \{\gamma_\alpha^\beta\})$ and the homology groups of the Čech complexes. It implies for instance that the rank of γ_α^β is at most the rank of $H_k(C_{\varepsilon'}(L_\alpha)) \to H_k(C_{\varepsilon'}(L_\beta))$. Indeed, by definition, $\gamma_\alpha^\beta : \Gamma_\alpha^k \to \Gamma_\beta^k$ is the restriction of m_α^β to $\operatorname{im} d_\alpha \circ b_\alpha$, therefore we have $\operatorname{im} \gamma_\alpha^\beta = \operatorname{im} m_\alpha^\beta \circ d_\alpha \circ b_\alpha = \operatorname{im} d_\beta \circ l_\alpha^\beta \circ b_\alpha$, which implies that $\operatorname{rank} \gamma_\alpha^\beta = \operatorname{rank} d_\beta \circ l_\alpha^\beta \circ b_\alpha \leq \operatorname{rank} l_\alpha^\beta$. Similarly, the rank of $H_k(C_\varepsilon(L_\alpha)) \to H_k(C_{\varepsilon''}(L_\beta))$ is equal to $\operatorname{rank} e_\beta \circ (m_\alpha^\beta \circ d_\alpha \circ b_\alpha) \circ a_\alpha \leq \operatorname{rank} m_\alpha^\beta \circ d_\alpha \circ b_\alpha = \operatorname{rank} \gamma_\alpha^\beta$. Thus, for all $\alpha \leq \beta$, the rank of the homomorphism γ_α^β is sandwiched between the ranks of $H_k(C_\varepsilon(L_\alpha)) \to H_k(C_{\varepsilon''}(L_\beta))$ and $H_k(C_{\varepsilon'}(L_\alpha)) \to H_k(C_{\varepsilon'}(L_\beta))$. If ever these lower and upper bounds were to be equal for all $\alpha \leq \beta$, then we could conclude that $(\{\Gamma_\alpha^k\}, \{\gamma_\alpha^\beta\})$ has the same persistence diagram as the kth persistent homology module of the filtration $\{C_{\varepsilon'}(L_\alpha)\}_{\alpha\in\mathbb{R}}$, which by Lemma 3.3 is close to the kth persistence diagram of f. However, in full generality the lower and upper bounds may differ, therefore we have to work out additional relations involving the homology of the sublevel-sets filtration $\{F_\alpha\}$ of f.

Since $\varepsilon \leq \varepsilon' \leq \varepsilon'' < \varrho_c(\mathbb{X})$, the proof of Lemma 3.3 enables us to consider the isomorphisms $h_\alpha : H_k(C_\varepsilon(L_\alpha)) \to H_k(L_\alpha^\varepsilon)$, $h'_\alpha : H_k(C_{\varepsilon'}(L_\alpha)) \to H_k(L_\alpha^{\varepsilon'})$ and $h''_\alpha : H_k(C_{\varepsilon''}(L_\alpha)) \to H_k(L_\alpha^{\varepsilon''})$

induced at homology level by the homotopy equivalences provided by the Nerve Theorem [21, §4G]. According to Lemma 3.2, their images are related to the kth persistent homology module of $\{F_\alpha\}$ through the following sequence of homomorphisms induced by inclusions: $\forall \alpha, \beta$ s.t. $\beta - \alpha \geq c(\varepsilon + \varepsilon'')$,

$$\begin{array}{ccccccccc} H_k(F_{\alpha-c\varepsilon}) & \stackrel{t_\alpha}{\rightarrow} & H_k(L_\alpha^\varepsilon) & \stackrel{u_\alpha}{\rightarrow} & H_k(L_\alpha^{\varepsilon'}) & \stackrel{v_\alpha}{\rightarrow} & H_k(L_\alpha^{\varepsilon''}) & \stackrel{w_\alpha}{\rightarrow} & H_k(F_{\alpha+c\varepsilon''}) \\ & & & & & & & & \downarrow s_\alpha^\beta \\ H_k(F_{\beta+c\varepsilon''}) & \stackrel{w_\beta}{\leftarrow} & H_k(L_\beta^{\varepsilon''}) & \stackrel{v_\beta}{\leftarrow} & H_k(L_\beta^{\varepsilon'}) & \stackrel{u_\beta}{\leftarrow} & H_k(L_\beta^\varepsilon) & \stackrel{t_\beta}{\leftarrow} & H_k(F_{\beta-c\varepsilon}) \end{array} \tag{3}$$

Combining (2) and (3) with isomorphisms h_α, h'_α, h''_α, h_β, h'_β, h''_β, we get a full diagram relating $(\{\Gamma_\alpha^k\}, \{\gamma_\alpha^\beta\})$ to the kth persistent homology module of $\{F_\alpha\}$. Note that this diagram may not fully commute: for instance, there is no particular reason why the linear map m_α^β should be identical to $d_\beta \circ b_\beta \circ a_\beta \circ h_\beta^{-1} \circ t_\beta \circ s_\alpha^\beta \circ w_\alpha \circ h''_\alpha \circ e_\alpha$. Nevertheless, the subdiagram of Eq. (2) commutes for all $\alpha \leq \beta$, because it is induced by inclusions. Furthermore, Lemma 3.4 ensures that the following subdiagrams (where the new homomorphisms ${l'}_\alpha^\beta$ and ${n'}_\alpha^\beta$ are induced by inclusions) also commute for all $\alpha \leq \beta$:

$$\begin{array}{ccccc} H_k(C_\varepsilon(L_\alpha)) & \stackrel{b_\alpha \circ a_\alpha}{\longrightarrow} & H_k(C_{\varepsilon'}(L_\alpha)) & \stackrel{e_\alpha \circ d_\alpha}{\longrightarrow} & H_k(C_{\varepsilon''}(L_\alpha)) \\ \downarrow h_\alpha & & \downarrow h'_\alpha & & \downarrow h''_\alpha \\ H_k(L_\alpha^\varepsilon) & \stackrel{u_\alpha}{\longrightarrow} & H_k(L_\alpha^{\varepsilon'}) & \stackrel{v_\alpha}{\longrightarrow} & H_k(L_\alpha^{\varepsilon''}) \end{array} \tag{4}$$

$$\begin{array}{ccccc} H_k(C_\varepsilon(L_\beta)) & \stackrel{b_\beta \circ a_\beta}{\longrightarrow} & H_k(C_{\varepsilon'}(L_\beta)) & \stackrel{e_\beta \circ d_\beta}{\longrightarrow} & H_k(C_{\varepsilon''}(L_\beta)) \\ \downarrow h_\beta & & \downarrow h'_\beta & & \downarrow h''_\beta \\ H_k(L_\beta^\varepsilon) & \stackrel{u_\beta}{\longrightarrow} & H_k(L_\beta^{\varepsilon'}) & \stackrel{v_\beta}{\longrightarrow} & H_k(L_\beta^{\varepsilon''}) \end{array} \tag{5}$$

$$\begin{array}{ccccc} H_k(C_{\varepsilon'}(L_\alpha)) & \stackrel{l_\alpha^\beta}{\longrightarrow} & H_k(C_{\varepsilon'}(L_\beta)) & \qquad H_k(C_{\varepsilon''}(L_\alpha)) & \stackrel{n_\alpha^\beta}{\longrightarrow} & H_k(C_{\varepsilon''}(L_\beta)) \\ \downarrow h'_\alpha & & \downarrow h'_\beta & \qquad \downarrow h''_\alpha & & \downarrow h''_\beta \\ H_k(L_\alpha^{\varepsilon'}) & \stackrel{{l'}_\alpha^\beta}{\longrightarrow} & H_k(L_\beta^{\varepsilon'}) & \qquad H_k(L_\alpha^{\varepsilon''}) & \stackrel{{n'}_\alpha^\beta}{\longrightarrow} & H_k(L_\beta^{\varepsilon''}) \end{array} \tag{6}$$

For all $\alpha \in \mathbb{R}$, let $\phi_\alpha : \Gamma_\alpha^k \to H_k(F_{\alpha+c\varepsilon''})$ be the restriction of the map $w_\alpha \circ h''_\alpha \circ e_\alpha$ to the subspace $\Gamma_\alpha^k = \operatorname{im} d_\alpha \circ b_\alpha \subseteq H_k(G'_\alpha)$. Symmetrically, let $\psi_{\alpha-c\varepsilon} : H_k(F_{\alpha-c\varepsilon}) \to \Gamma_\alpha^k$ be the map $d_\alpha \circ b_\alpha \circ a_\alpha \circ h_\alpha^{-1} \circ t_\alpha$. Its image is indeed included in the subspace $\Gamma_\alpha^k = \operatorname{im} d_\alpha \circ b_\alpha \subseteq H_k(G'_\alpha)$. To prove that the persistence module $(\{\Gamma_\alpha^k\}, \{\gamma_\alpha^\beta\})$ is $c\varepsilon''$-interleaved with the kth persistent homology module of $\{F_\alpha\}$, it suffices to show that (a.) the map $\psi_{\beta-c\varepsilon} \circ s_\alpha^\beta \circ \phi_\alpha$ is equal to m_α^β over the subspace $\Gamma_\alpha^k \subseteq H_k(G_\alpha)$ for all $\beta \geq \alpha + c(\varepsilon + \varepsilon'')$, and (b.) the map $\phi_\beta \circ m_\alpha^\beta \circ \psi_{\alpha-c\varepsilon}$ is equal to the homomorphism $s_{\alpha-c\varepsilon}^{\beta+c\varepsilon''} : H_k(F_{\alpha-c\varepsilon}) \to H_k(F_{\beta+c\varepsilon''})$ induced by inclusion for all $\beta \geq \alpha$. Our proof uses *diagram chasing* arguments involving diagrams (2) to (6):

Consider first the map $\phi_\beta \circ m_\alpha^\beta \circ \psi_{\alpha-c\varepsilon}$. Replacing ϕ_β and $\psi_{\alpha-c\varepsilon}$ by their definitions, we get $w_\beta \circ h''_\beta \circ (e_\beta \circ m_\alpha^\beta) \circ d_\alpha \circ b_\alpha \circ a_\alpha \circ h_\alpha^{-1} \circ t_\alpha$, which by commutativity of (2) is equal to $w_\beta \circ h''_\beta \circ (n_\alpha^\beta \circ e_\alpha) \circ d_\alpha \circ b_\alpha \circ a_\alpha \circ h_\alpha^{-1} \circ t_\alpha$. Now, by commutativity of (4), we have $e_\alpha \circ d_\alpha \circ b_\alpha \circ a_\alpha \circ h_\alpha^{-1} = {h''}_\alpha^{-1} \circ v_\alpha \circ u_\alpha$, therefore $\phi_\beta \circ m_\alpha^\beta \circ \psi_{\alpha-c\varepsilon}$ is equal to $w_\beta \circ (h''_\beta \circ n_\alpha^\beta \circ {h''}_\alpha^{-1}) \circ v_\alpha \circ u_\alpha \circ t_\alpha$, which by commutativity of the rightmost diagram of (6) is equal to $w_\beta \circ {n'}_\alpha^\beta \circ v_\alpha \circ u_\alpha \circ t_\alpha$, which is precisely $s_{\alpha-c\varepsilon}^{\beta+c\varepsilon''}$.

Consider now the map $\psi_{\beta-c\varepsilon} \circ s_\alpha^\beta \circ \phi_\alpha$. Since by definition we have $\Gamma_\alpha^k = \operatorname{im} d_\alpha \circ b_\alpha \subseteq \operatorname{im} d_\alpha$, the fact that $\psi_{\beta-c\varepsilon} \circ s_\alpha^\beta \circ \phi_\alpha$ coincides with m_α^β over Γ_α^k is a direct consequence of the fact that the map $\psi_{\beta-c\varepsilon} \circ s_\alpha^\beta \circ \phi_\alpha \circ d_\alpha$ equals $m_\alpha^\beta \circ d_\alpha$ over $H_k(C_{\varepsilon'}(L_\alpha))$, which we will now prove. Replacing ϕ_α and $\psi_{\beta-c\varepsilon}$ by their definitions, we get $d_\beta \circ (b_\beta \circ a_\beta \circ h_\beta^{-1}) \circ t_\beta \circ s_\alpha^\beta \circ w_\alpha \circ ({h''}_\alpha \circ e_\alpha \circ d_\alpha)$,

which by commutativity of (4) and (5) is equal to $d_\beta \circ {h'}_\beta^{-1} \circ u_\beta \circ t_\beta \circ s_\alpha^\beta \circ w_\alpha \circ v_\alpha \circ {h'}_\alpha$. Now, observe that $u_\beta \circ t_\beta \circ s_\alpha^\beta \circ w_\alpha \circ v_\alpha$ is nothing but the homomorphism ${l'}_\alpha^\beta$ induced by the inclusion $L_\alpha^{\varepsilon'} \hookrightarrow L_\beta^{\varepsilon'}$. Therefore, we have $\psi_{\beta-c\varepsilon} \circ s_\alpha^\beta \circ \phi_\alpha \circ d_\alpha = d_\beta \circ ({h'}_\beta^{-1} \circ {l'}_\alpha^\beta \circ {h'}_\alpha)$, which is equal to $d_\beta \circ l_\alpha^\beta$ by commutativity of the leftmost diagram of (6). Finally, we have $d_\beta \circ l_\alpha^\beta = m_\alpha^\beta \circ d_\alpha$ by commutativity of (2). Thus, $\psi_{\beta-c\varepsilon} \circ s_\alpha^\beta \circ \phi_\alpha$ coincides with m_α^β over Γ_α^k.

It follows from the last two paragraphs that the two families of homomorphisms $\{\phi_\alpha\}$ and $\{\psi_\alpha\}$ make the persistence module $(\{\Gamma_\alpha^k\}, \{\gamma_\alpha^\beta\})$ $c\varepsilon''$-interleaved with the kth persistent homology module of f. This concludes the proof of Lemma 3.5. □

3.3 Stability with respect to noise

The guarantees provided by Theorem 3.1 hold as far as exact geodesic distances and function values are used in the construction of the Rips complexes. In practice however, function values are often obtained from physical measurements with inherent noise, while geodesic distances are not known in advance and have to be estimated through some neighborhood graph distance. We claim that our analysis is generic enough to handle these practical situations.

Consider first the case where function values are noisy. More precisely, given a geodesic ε-sample L of some Riemannian manifold $\mathbb{X}$, and a c-Lipschitz tame function $f : \mathbb{X} \to \mathbb{R}$, assume that the data points $p \in L$ are assigned values $\tilde{f}(p)$ that are different from $f(p)$, and let $\zeta = \max_{p \in L} |\tilde{f}(p) - f(p)|$. For convenience, for all $\alpha \in \mathbb{R}$ we introduce the set $\tilde{L}_\alpha$ of points of L whose $\tilde{f}$-values are at most α. Note that $\tilde{L}_\alpha$ may neither contain nor be contained in L_α in general. However, we have $\tilde{L}_\alpha \subseteq L_{\alpha+\zeta}$, which, plugged into the proof of Lemma 3.2, yields the following variant of that result:

Lemma 3.6 *$\forall \delta \geq \varepsilon$, the sublevel-sets filtration of f is $(c\delta + \zeta)$-interleaved with $\{\tilde{L}_\alpha^\delta\}_{\alpha \in \mathbb{R}}$.*

The rest of the analysis of Sections 3.1 and 3.2 carries through, with L_α replaced by $\tilde{L}_\alpha$ for all $\alpha \in \mathbb{R}$ and $c\varepsilon$ and $c\varepsilon''$ replaced respectively by $c\varepsilon + \zeta$ and $c\varepsilon'' + \zeta$ in Eq. (3) and in the rest of the proof of Lemma 3.5. We thus obtain the following new bounds:

Theorem 3.7 *Let $\mathbb{X}, f, L$ be as in Theorem 3.1. Assume that the values of f at the points of L are known within a precision of ζ. Then, for any $\delta \in [2\varepsilon, \frac{1}{2}\varrho_c(\mathbb{X}))$ and any $k \in \mathbb{N}$, the kth persistent homology modules of f and of the nested pair of filtrations $\{R_\delta(\tilde{L}_\alpha) \hookrightarrow R_{2\delta}(\tilde{L}_\alpha)\}_{\alpha \in \mathbb{R}}$ are $(2c\delta + \zeta)$-intreleaved. Therefore, the bottleneck distance between their persistence diagrams is at most $2c\delta + \zeta$, by Theorem 2.3.*

Consider now the case where geodesic distances are noisy. Specifically, assume that the geodesic distance $\mathrm{d}_\mathbb{X}$ is replaced by the distance d_G in some neighborhood graph G built on top of the point cloud L. This graph can be either weighted of unweighted, depending on the application. For instance, in unsupervised learning the edges of G are often weighted by the Euclidean distances between their vertices [24], while in sensor networks edges are usually unweighted because retrieving the exact geographic locations of the sensor nodes can be difficult — see *e.g.* [25, §4.4]. Generally speaking, weighted graphs provide better approximations of geodesic distances, but their construction requires to have additional information at hand, such as extrinsic distances between the data points. We therefore focus on unweighted graphs, which correspond to the most general case. In order to make theoretical claims, we assume that G is a *μ-disk graph*[6], that is: a pair of data points form an edge in G if and only if their geodesic

[6]Our analysis also handles quasi μ-disk graphs, modulo a degradation of the bounds on the approximation error.

distance is less than μ. Assuming that the input point cloud L is a geodesic ε-sample of some Riemannian manifold $\mathbb{X}$, we will use the following bounds on the graph distance [16, Lemma 6.1]:

$$\forall i,j \in \{1,\cdots,n\},\ \frac{\mathrm{d}_{\mathbb{X}}(x_i,x_j)}{\mu} \leq \mathrm{d}_G(x_i,x_j) \leq 1+\lambda\frac{\mathrm{d}_{\mathbb{X}}(x_i,x_j)}{\mu},\ \text{where } \lambda = 1+4\frac{\varepsilon}{\mu}. \tag{7}$$

The two Rips-based filtrations introduced at the beginning of Section 3 are now defined with respect to d_G, and no longer $\mathrm{d}_{\mathbb{X}}$. To emphasize this aspect, we denote them respectively by $\{R^G_\delta(L_\alpha)\}_{\alpha\in\mathbb{R}}$ and $\{R^G_{\delta'}(L_\alpha)\}_{\alpha\in\mathbb{R}}$. In Theorem 3.1 we set $\delta' = 2\delta$ because geodesic distances were exact. We will now show that noise in geodesic distances can be handled by taking a slightly larger δ'. We first relate $\{R^G_\delta(L_\alpha)\}_{\alpha\in\mathbb{R}}$ and $\{R^G_{\delta'}(L_\alpha)\}_{\alpha\in\mathbb{R}}$ to Čech filtrations defined with respect to $\mathrm{d}_{\mathbb{X}}$:

Lemma 3.8 *Let $\lambda = 1+4\frac{\varepsilon}{\mu}$ be as in Eq. (7). Assume that $\delta \geq 1+2\lambda\frac{\varepsilon}{\mu}$, $\varepsilon' \geq \mu\delta$, $\delta' \geq 1+2\lambda\frac{\varepsilon'}{\mu}$ and $\varepsilon'' \geq \mu\delta'$. Then, $\forall\alpha\in\mathbb{R}$, $C_\varepsilon(L_\alpha) \subseteq R^G_\delta(L_\alpha) \subseteq C_{\varepsilon'}(L_\alpha) \subseteq R^G_{\delta'}(L_\alpha) \subseteq C_{\varepsilon''}(L_\alpha)$.*

Proof. Let us prove that $R^G_\xi(L_\alpha) \subseteq C_{\mu\xi}(L_\alpha)$ and $C_\xi(L_\alpha) \subseteq R^G_{1+2\lambda\xi/\mu}(L_\alpha)$ for any arbitrary value $\xi \geq 0$. The lemma will then follow by letting ξ be consecutively equal to ε, δ, ε', and δ'.

Consider first a simplex $\{x_1,\cdots,x_l\}$ of $R^G_\xi(L_\alpha)$. Eq. (7) implies that $\mathrm{d}_{\mathbb{X}}(x_1,x_i) \leq \mu\mathrm{d}_G(x_1,x_i) < \mu\xi$ for all $i\in\{1,\cdots,l\}$. This means that the open geodesic balls of same radius $\mu\xi$ about the points x_i have x_1 in their common intersection, which is therefore non-empty. As a consequence, the simplex belongs to $C_{\mu\xi}(L_\alpha)$. Consider now a simplex $\{x_1,\cdots,x_l\}$ of $C_\xi(L_\alpha)$. The open geodesic balls of same radius ξ about the points x_i have a non-empty common intersection, therefore the pairwise geodesic distances between the points are less than 2ξ. It follows then from Eq. (7) that the diameter of the simplex in the graph distance is at most $1+2\lambda\frac{\xi}{\mu}$. Thus, the simplex belongs to $R^G_{1+2\lambda\xi/\mu}(L_\alpha)$. □

Letting $G_\alpha = R^G_\delta(L_\alpha)$ and $G'_\alpha = R^G_{\delta'}(L_\alpha)$, where δ,δ' and $\varepsilon \leq \varepsilon' \leq \varepsilon''$ satisfy the conditions of Lemma 3.8, we can now apply Lemma 3.5 to get the following guarantee:

Theorem 3.9 *Let $\mathbb{X}, f, L$ be as in Theorem 3.1. Assume that the geodesic distance $\mathrm{d}_{\mathbb{X}}$ is replaced by the graph distance d_G in the unweighted μ-disk graph G built on top of L. Let $\lambda = 1+4\frac{\varepsilon}{\mu}$. Then, for any $\delta \geq 1+2\lambda\frac{\varepsilon}{\mu}$, any $\delta' \in [1+2\lambda\delta,\ \frac{1}{\mu}\varrho_c(\mathbb{X}))$, and any $k\in\mathbb{N}$, the kth persistent homology modules of f and of the nested pair of filtrations $\{R^G_\delta(L_\alpha) \hookrightarrow R^G_{\delta'}(L_\alpha)\}_{\alpha\in\mathbb{R}}$ are $c\mu\delta'$-intreleaved. Therefore, the bottleneck distance between their persistence diagrams is at most $c\mu\delta'$, by Theorem 2.3.*

This result provides sufficient conditions on parameters δ,δ' for the analysis of the previous sections to hold in the case where geodesic distances are not exact. Note that simple expressions can be derived for δ and δ', which can be later used in our algorithms. For instance, if we assume that $\mu \geq 4\varepsilon$, then $\lambda \leq 2$ and therefore we can choose $\delta = 2$ and $\delta' = 9$. Then, the conclusion of Theorem 3.9 holds provided that δ' is less than $\frac{1}{\mu}\varrho_c(\mathbb{X})$, from which derives the following condition on the sampling density ε and communication radius μ: $4\varepsilon \leq \mu < \frac{1}{9}\varrho_c(\mathbb{X})$.

4 Algorithms

Section 4.1 presents the core algorithm, which derives from the structural results of Section 3. The subsequent sections introduce two improvements to the algorithm: the first one deals with time-varying functions (Section 4.2), the second one extracts additional spatial information (Section 4.3).

4.1 Core algorithm

The algorithm takes as input a n-dimensional vector v, a $n \times n$ distance matrix D, and a parameter $\delta \geq 0$. The ith entry of v stands for the function value at the ith point of the data set, while the entries $D_{i,j} = D_{j,i}$ give the distance between points i and j. No geographic coordinates are to be provided, so that the algorithm can virtually be applied in any arbitrary metric space. For clarity of exposition, we assume that the entries of v are sorted, that is: $v_1 \leq v_2 \leq \cdots \leq v_n$. They are not in our implementation. The algorithm proceeds in two steps:

1. it builds two families of Rips complexes: $R_\delta(\{1\}) \subseteq R_\delta(\{1,2\}) \subseteq \cdots \subseteq R_\delta(\{1,2,\cdots,n\})$ and $R_{2\delta}(\{1\}) \subseteq R_{2\delta}(\{1,2\}) \subseteq \cdots \subseteq R_{2\delta}(\{1,2,\cdots,n\})$. The ith complex in each family is computed from the sub-matrix of D spanned by the rows and columns of indices $1,\cdots,i$. The time of appearance of its simplices that are not in the $(i-1)$th complex is set to v_i.
2. for k ranging from zero to the dimension of the complexes, it computes the kth persistence diagram of the nested pair of filtrations $\{R_\delta(\{1,\cdots,i\}) \hookrightarrow R_{2\delta}(\{1,\cdots,i\})\}_{1 \leq i \leq n}$.

Upon termination, the algorithm returns the persistence diagrams computed at step 2. The quality of this output is guaranteed by the structural results of Section 3, under sufficient sampling density and in the absence of noise. Observe indeed that the filtrations built at step 1. are the same as the ones considered in Theorem 3.1, which therefore provides the following theoretical guarantee:

Theorem 4.1 *If the data points form a geodesic ε-sample of some Riemannian manifold $\mathbb{X}$, with $\varepsilon < \frac{1}{4}\,\varrho_c(\mathbb{X})$, and if the input distance matrix D gives the exact geodesic distances between the data points, then, for any input $\delta \in [2\varepsilon, \frac{1}{2}\varrho_c(\mathbb{X}))$ and any tame c-Lipschitz function $f : \mathbb{X} \to \mathbb{R}$ whose values at the data points are given exactly by the input vector v, the kth persistence diagram output by the algorithm lies at bottleneck distance at most $2c\delta$ of the kth persistence diagram of f.*

Note that the output of the algorithm also gives the homology groups of the underlying space $\mathbb{X}$. Indeed, $H_k(\mathbb{X})$ is isomorphic to the linear span of the k-dimensional homological features that are infinitely persistent in the kth persistence diagram of f. Now, by Theorem 4.1, the bottleneck distance between the diagram of f and the one computed at step 2. of the algorithm is finite, therefore the infinitely-persistent homological features in both diagrams are in bijection.

One drawback of our approach is that it is not parameter-free, which makes its behavior dependent on the choice of the input parameter δ. In some sense, this parameter controls the scale at which the algorithm will process the data. The issue of finding the *right* scale is ubiquitous in geometric data analysis, and several solutions based on the idea of persistence have been proposed. We suggest to consider a whole range of values of δ, between zero and infinity (or any sufficiently large value). For each value in this range[7], we apply the algorithm and report the infinitely-persistent homological features in the output persistence diagrams, which supposedly coincide with the ones of the underlying space $\mathbb{X}$, according to Theorem 4.1. Then, following [17] and subsequent work, we claim that relevant ranges of scales can be identified as ranges of values of δ over which the numbers of infinitely-persistent homological features in all the diagrams are stable.

Finally, note that Theorems 3.7 and 3.9 provide theoretical guarantees similar to the ones of Theorem 4.1 in cases where the input vector v of function values or the input distance matrix D is noisy. As explained in Section 3.3, this latter case requires to set $\delta = 2$ and to replace 2δ by $\delta' = 9$ in the construction of the two families of Rips complexes at step 1. of the algorithm.

[7] In fact, we only have to consider the finitely many values of δ at which the combinatorial structures of the Rips complexes change.

Implementation and complexity. The running time of the algorithm can be bounded in terms of the size of the data structure, provided that a careful implementation is built. In our case, the two families of complexes introduced at step 1. are built simultaneously as filtrations of the largest of the Rips complexes, $R_{2\delta}(\{1,\cdots,n\})$, which by definition contains all the other complexes of the two families. As emphasized in [7], the simplices of $R_{2\delta}(\{1,\cdots,n\})$ are in bijection with the cliques of its 1-skeleton graph. Therefore, we first build this graph in $O(n^2)$ time by comparing the entries of the matrix D with the threshold 2δ. Then, we construct the simplices of $R_{2\delta}(\{1,\cdots,n\})$ iteratively, by increasing dimension. First, all vertices are created. Then, for each simplex $\{i_1,\cdots,i_k\}$ created, we look at its 1-ring neighborhood in the graph, and for each vertex i_l in this neighborhood, we check whether $\{i_1,\cdots,i_k,i_l\}$ forms a clique. If so, then this new simplex is created, and its diameter $\max_{1\leq r<s\leq l} D_{r,s}$ and appearance time $\max_{1\leq r\leq l} v_{j_r}$ are stored. The time spent checking whether we have a clique and computing the new diameter and appearance time from the ones of the original simplex $\{i_1,\cdots,i_k\}$ is $O(k)$, while the size of the 1-ring neighborhood is $O(n)$. Thus, the total time spent building the complex is $O(ndN)$, where d is the dimension of the complex and N is its total number of simplices. Then, within $O(N\log N)$ time, we order the created simplices according to their appearance times, to build the filtration of parameter 2δ. As for the filtration of parameter δ, observe that each of its simplices must appear in both filtrations at the same time. Therefore, we can build the filtration of parameter δ in $O(N)$ time by scanning through the sorted list of simplices in the filtration of parameter 2δ and reporting the simplices that have diameter at most δ. Finally, we perform step 2. by running the algorithm of [9] on our two filtrations. This variant of the standard persistence algorithm has the same worst-case running time of $O(N^3)$.

Theorem 4.2 *The total running time of the algorithm is $O(ndN+N^3)$, where d is the dimension of $R_{2\delta}(\{1,\cdots,n\})$ and N is its total number of simplices.*

Step 2. is clearly the pacing phase of our method. However, it is reported in [27] that, although the worst-case running time of the persistence algorithm is $O(N^3)$, in most practical cases it has an almost-linear behavior. Thus, the running time of our method is likely to be $O(ndN)$ in practice.

Note also that $R_{2\delta}(\{1,\cdots,n\})$ could potentially span the full $(n-1)$-simplex and therefore have as many as 2^n simplices. However, there are important cases where the size of the complex remains bounded. For instance, when the data points are uniformly sampled along a m-dimensional Riemannian manifold, a packing argument detailed in [7] shows that the size of the complex is at most $2^{2^m}n$, and that it even reduces to $2^{O(m^2)}n$ if a reasonable upper bound on m is known. This reduces the running time of the algorithm to $2^{O(m^2)}n^3$ and thereby makes the approach tractable when the data points sample uniformly some low-dimensional manifold, possibly embedded in high-dimensional space. Sampling uniformity can be achieved in practice by a landmarking strategy [17].

4.2 Time-varying functions

It is commonplace in sensor networks and related areas that the functions under study vary with time. In monitoring applications for instance, one wants to get a high-level description of the distribution of some intensive quantity like temperature or humidity over a fixed domain. Such quantities vary typically on a day scale, and a natural goal is to be able to maintain accurate approximations to their persistence diagrams under such variations.

We model the problem as follows: given a finite point cloud $L=\{x_1,\cdots,x_n\}$ that is a geodesic ε-sample of some fixed Riemannian manifold $\mathbb{X}$, we want to maintain accurate approximations to the persistence diagrams of some time-varying function $f_t:\mathbb{X}\to\mathbb{R}$ whose values

are known only at the points of L and at a finite number of instants $t_0 \leq t_1 \leq \cdots \leq t_k$. We assume f_{t_i} to be tame and c-Lipschitz for all i, for some fixed constant c. Thanks to this assumption, Theorem 4.1 provides us with theoretical guarantees regarding the quality of the output persistence diagrams at every instant t_i. The dynamic version of the algorithm works as follows:

It performs an initialization step at time t_0, where it simply applies the core algorithm as in the static setting. The filtrations of parameters δ and 2δ are stored as two arrays of simplices, sorted according to their times of appearance, which are derived from the values of f_{t_0} at the vertices.

At every subsequent instant t_j we need to update the two filtrations, and then to recompute their kth persistence diagram for all values k between zero and their dimension. In fact, since $\mathbb{X}$ and L remain fixed throughout the process, the distance matrix D does not change and therefore the Rips complexes $R_\delta(L)$ and $R_{2\delta}(L)$ remain the same. Thus, updating the filtrations boils down to re-sorting their simplices according to the new appearance times induced by f_{t_j}. Computing the new appearance times is done by scanning through the filtrations, and for each simplex, finding the vertex of maximal f_{t_j}-value. Then, re-sorting the simplices of each filtration is done in-place in the array of the filtration using *insertion sort*. The reason for using this particular sorting algorithm is that it decomposes the permutation on the simplices into a sequence of *inversions*[8]. This sequence is then provided as input to the *vineyards*[9] variant of the persistence algorithm [10], which uses this information to update the kth persistence diagram for all values k at once.

The time complexity of the initialization stage is the same as the one of the static algorithm, namely $O(N^3)$, where N is the total number of simplices of $R_{2\delta}(L)$. Then, at every subsequent instant t_j, the time spent updating the appearance times is $O(dN)$, where d is the dimension of $R_{2\delta}(L)$. Consider now the permutation π_j on the simplices induced by the change from function $f_{t_{j-1}}$ to function f_{t_j}. A key feature of insertion sort is that it decomposes π_j into a minimal sequence of inversions, of size $|\pi_j|$. Its time complexity is thus $O(N+|\pi_j|)$. Finally, the *vineyards* algorithm updates the persistence diagrams in $O(N)$ time per inversion. Hence, the total time spent by our method at instant t_j is $O((1+d+|\pi_j|)N)$. Although d is bounded by $\log N$, in the worst case $|\pi_j|$ can be up to $\Theta(N^2)$, thereby raising the complexity to $\Theta(N^3)$, which is no better than if the filtrations and persistence diagrams were re-computed entirely at time t_j. However, this is a worst-case analysis, and in many practical situations $|\pi_j|$ is likely to be small. If for instance the values $f_t(x_i)$ at the data points follow polynomial trajectories in time[10], such that only a constant number (say two) of such trajectories meet at any given time, then between two instants t_{j-1} and t_j that are close enough only two function values $f_t(x_i), f_t(x_j)$ are permuted. As a consquence, only the stars of x_i, x_j in $R_{2\delta}(L)$ are affected by π_j, and therefore we have $|\pi_j| = O(d_v^2)$, where d_v denotes the size of the largest possible star of a vertex of $R_{2\delta}(L)$. The update time of our method at t_j becomes then $O((d+d_v^2)N) = O((d+d_v^2)d_v n)$. If the input point cloud uniformly samples some Riemannian manifold of dimension m (known within a constant factor), then we have $d = O(m)$ and $d_v = 2^{O(m^2)}$, which reduces the update time to $2^{O(m^2)}n$ — a quantity that is linear in the size of the input, modulo a constant factor that depends on the intrinsic dimensionality of the data.

Finally, let us mention that, similarly to the standard persistence algorithm, the *vineyards* algorithm has been observed to run much faster in practice than expected in theory [10]. Typically, the observed running time is constant per simplex inversion. This reduces the update time of our method to $O(d+|\pi_j|)$ in the general case, and even to a constant $2^{O(m^2)}$ in the practical setting described above.

[8] An inversion is a transposition between two simplices that are adjacent in the array.

[9] Originally designed for a single filtration, this algorithm was adapted to our context in Appendix A of [9].

[10] This is the usual assumption in the *kinetic data structures* framework [18].

4.3 Extracting additional spatial information

Assume the domain $\mathbb{X}$ underlying the data to be a m-dimensional Riemannian manifold, and the unknown function $f : \mathbb{X} \to \mathbb{R}$ to be a Morse function, *i.e.* a smooth function with only nondegenerate critical points. We want to recover the ascending regions (a.k.a. stable manifolds) of the maxima of f, as well as the descending regions (a.k.a. unstable manifolds) of its minima. The ascending region of a maximum p is the set of points of $\mathbb{X}$ that eventually reach p by moving along the flow induced by the gradient vector field of f. Symmetrically, the descending region of a minimum q is the set of points that eventually reach q by moving against the gradient flow. These regions share many interesting properties, among which the following ones are of particular interest to us: ascending (resp. descending) regions form pairwise disjoint open cells homeomorphic to $\mathbb{R}^m$ that cover $\mathbb{X}$ up to a subset of measure zero. In other words, they can be used as a tool for segmenting the domain $\mathbb{X}$ according to the basins of attraction of the maxima (resp. minima) of f. Furthermore, they can be used as the main building block of the Morse-Smale decomposition of $\mathbb{X}$ induced by f, since the faces of the complex are obtained as intersections of ascending and descending regions. Note that the ascending regions of f are the descending regions of $-f$, so the problem reduces to finding the descending regions of f from its values at a finite sampling L of $\mathbb{X}$.

As in the previous sections, the geographic locations of the data points are not assumed to be known, and the algorithm uses only the connectivity between the data points in the 1-skeleton graph of the Rips complex $R_{2\delta}(L)$, called the *Rips graph* G from now on. For simplicity, we assume that the values of f at the data points are all different. This genericity condition is easily ensured by an infinitesimal perturbation of f. At a high level, our method is composed of two phases: first, it approximates the gradient vector field of f at the vertices of G and clusters them according to the (approximate) basins of attraction in the graph G; second, it uses the 0th persistence diagram of f to merge the clusters of short lifespans with longer-lasting clusters. The clustering technique used in the first phase is in fact not new, and it has been shown to be quite unstable under small perturbations of the function, both in theory [12] and in practice [26]. The novelty of our approach lies in the way it uses persistence to merge clusters and regain some stability.

In the first phase, we iterate over the vertices of G, in the order of their f-values. At each vertex v, the direction of $-\nabla f$ is approximated by the edge e of G that connects v to a neighbor u minimizing the quantity $\frac{f(u)-f(v)}{|e|}$, where $|e|$ is the length of the edge — computed during the construction of the Rips graph G. If no neighbor of v has a lower f-value than v, then v is a local minimum of f in G and is therefore kept disconnected. Such a vertex v is called a *sink*. Note that every non-sink vertex w is connected to a proper neighbor in G, and by following the approximate direction of $-\nabla f$ in the graph we eventually reach a sink because the value of f decreases strictly along the path followed. We declare this sink as the center of the cluster to which w belongs.

Recall now that the core algorithm (Section 4.1) approximates the kth persistence diagram of f *via* the kth persistent homology module of the nested pair of filtrations $\{R_\delta(L_\alpha) \hookrightarrow R_{2\delta}(L_\alpha)\}_{\alpha\in\mathbb{R}}$, where by definition $L_\alpha = L \cap f^{-1}((-\infty, \alpha])$. In the special case of zero-dimensional homology however, we know that each vertex appears both in $R_\delta(L_\alpha)$ and in $R_{2\delta}(L_\alpha)$ at the same time, therefore im $H_0(R_\delta(L_\alpha)) \to H_0(R_{2\delta}(L_\alpha))$ is isomorphic to $H_0(R_{2\delta}(L_\alpha))$. As a result, the 0th persistence diagram of f is in fact approximated by the 0th persistence diagram of the filtration $\{R_{2\delta}(L)\}_{\alpha\in\mathbb{R}}$, which can be computed easily from its 1-skeleton graph G using a variant of the standard union-find data structure, described in [14]. This is what phase two of our algorithm does. The outcome is a set of pairs (v, e), where v is a local minimum of f in the graph G, and e is an edge of G that connects the connected component created by v in G to some

older connected component. Stated differently, v is a sink and e is the first edge that connects its cluster to some other cluster of center u. If the lifespan[11] of the cluster of v is shorter than some user-defined threshold λ, then the algorithm merges the cluster of v into the cluster of u.

Our implementation uses only one pass through the graph G, during which the approximate gradients at the vertices are computed and the clusters are formed and merged on the fly using the union-find data structure of [14]. Thus, once the Rips graph G is built, the remaining running time is $O(|G|A^{-1}(|G|))$, where $|G|$ is the size of G and A is the Ackermann function. In addition, Theorem 4.1 provides the following theoretical guarantee on the output of the algorithm:

Theorem 4.3 *Assume L to be a geodesic ε-sample of $\mathbb{X}$, with $\varepsilon < \frac{1}{4}\,\varrho_c(\mathbb{X})$, and f to be tame and c-Lipschitz. Assume further that there exist two non-negative values $\mathrm{d}_2 > \mathrm{d}_1 + 16c\varepsilon$ such that the 0th persistence diagram of f has the following well-separated structure: $\mathrm{D}_0 f = \mathrm{D}_1 \cup \mathrm{D}_2$, with $\max\{p_y - p_x,\ p \in \mathrm{D}_1\} \leq \mathrm{d}_1$ and $\min\{q_y - q_x,\ q \in \mathrm{D}_2\} \geq \mathrm{d}_2$. Then, for any Rips parameter $\delta \in [2\varepsilon, \min\{\frac{1}{2}\varrho_c(\mathbb{X}),\ \frac{\mathrm{d}_2 - \mathrm{d}_1}{8c}\})$ and any threshold $\lambda \in (\mathrm{d}_1 + 4c\delta,\ \mathrm{d}_2 - 4c\delta)$, the number of clusters computed by our algorithm is equal to the number of basins of attraction of minima of f on $\mathbb{X}$ whose lifespans are at least λ. Furthermore, there is a pairing between clusters and basins of attraction that modifies the birth times by at most $2c\delta$.*

The well-separatedness of the 0th persistence diagram of f can be interpreted as a signal-to-noise ratio condition: the relevant peaks or valleys of f must be significantly more persistent than the non-relevant ones, as measured by the difference between their lifespans. Under such a condition, it is possible to threshold the diagrams of f and of the Rips complex $R_{2\delta}(L)$ so that the remaining finite point sets in both diagrams are in bijection and lie at small bottleneck distance of each other.

In addition to the above stability guarantee, it would be desirable to have an approximation result that bounds the distance between the set of data points falling into the cluster of a given sink and the basin of attraction of the corresponding minimum of f in $\mathbb{X}$. To the best of our knowledge, this question remains open.

5 Applications & Discussion

We now illustrate the relevance and generality of our approach through three specific applications. For each application, we describe the context and show some experimentation validation. We also provide timings information in Table 1.

data set	dimension	# vertices	# edges	Rips graph (sec.)	clustering (sec.)	total (sec.)
crater	2	1,048	7,095	0.01	0.00	0.01
torus	3	2,034	7,650	0.01	0.00	0.01
four Gaussians	2	6,354	51,946	0.07	0.02	0.09
hand	2	19,470	158,395	0.27	0.05	0.32
double spiral	2	114,563	2,116,035	2.43	0.61	3.04
octopus	3	770,196	9,540,143	14.56	7.11	21.67

Table 1: *Timings on an Intel Core 2 Duo T7500 @ 2.20GHz with 2GB of RAM. We used the C++ library ANN [28] for the proximity queries involved in the construction of the Rips graph. The* clustering *phase comprises both steps of the algorithm of Section 4.3, which are performed simultaneously.*

[11] Defined as the difference between the times at which e and v appear in the Rips graph G.

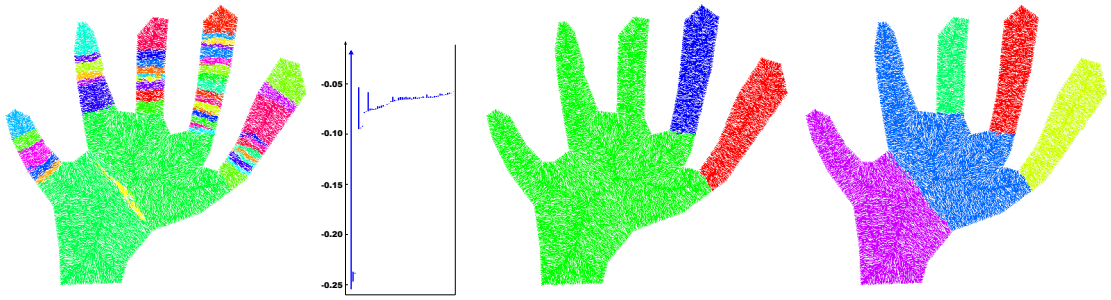

Figure 2: *Segmentation result on a sampled hand-shaped 2-D domain. The segmentation function is the Euclidean distance to the subset of the data points lying on the boundary of the domain. The barcode shows only three long intervals, corresponding to the palm of the hand and to the two rightmost fingers (center-right image), which have significant bottlenecks at their base. This suggests that the above function is not well-suited for segmenting this type of shape. Indeed, when a finger (such as the index in our example) has no bottleneck, the exact distance to the boundary has no local maximum inside this finger, therefore no ascending region separates it from the rest of the hand. In practice, the inaccuracy of our gradient estimation creates artificial local maxima which, by chance, cover the fingers (left). However, our barcode reveals that their ascending regions are actually not persistent. The rightmost image shows the result obtained with a smaller persistence threshold τ, which divides the palm of the hand before separating it from the index finger.*

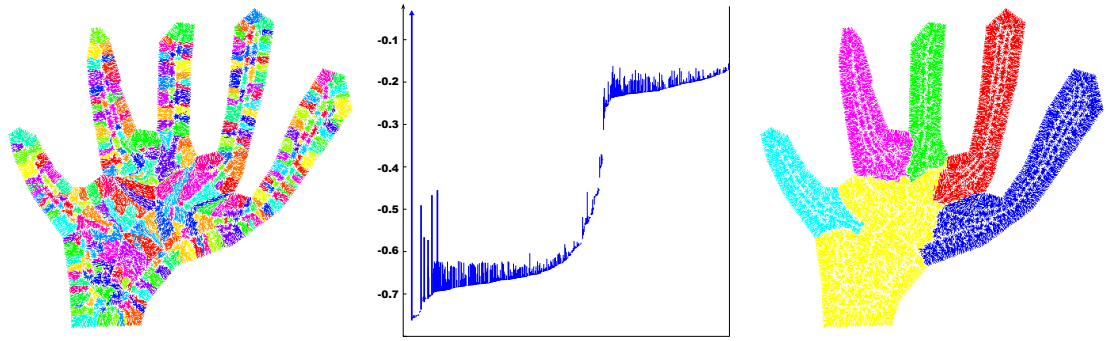

Figure 3: *Result obtained on the same data set as in Figure 2, using the normalized diameter of the set of nearest boundary points as the segmentation function. The barcode shows six long intervals corresponding to the palm of the hand and to the five fingers. The results before and after merging non-persistence clusters are shown respectively to the left and to the right of the barcode.*

Clustering. Clustering attempts to group points by assuming they are drawn from some unknown probability distribution. Our approach is inspired by Mean-Shift clustering [11]. Given an input point cloud L, we use a simple density estimator to approximate the local density at the points of L. As Figure 6 shows, our estimator can be quite noisy. However, our emphasis is not on accurate density estimation, but rather on clustering with noisy density estimates. Our estimator is provided together with L as input to the algorithm of Section 4.3, which clusters the points of L according to the basins of attraction of the local maxima of the estimator in the

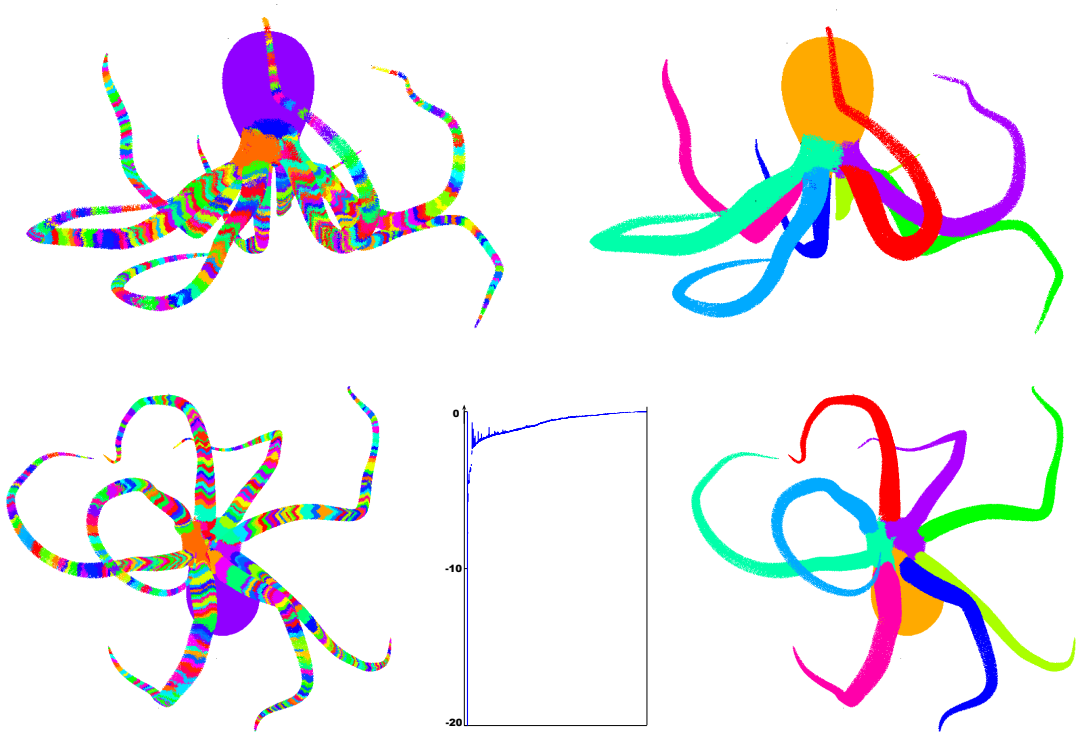

Figure 4: *Segmentation result obtained from a sampling of the interior of an octopus in 3-D. The segmentation function is the squared distance to the boundary of the body, whose barcode (center) somewhat emphasizes the bottlenecks at the base of the legs. With this segmentation function, there is a small range of values of the persistence threshold τ (easily computed from the barcode) that allow to recover the eight legs and the head. Pictures on the left show the result before merging clusters, while pictures on the right show the result after merging.*

Rips graph $G_{2\delta}$ built over L. Due to the noisy nature of the estimator, we get a myriad of small clusters before the merging phase. The novelty of our approach is to provide visual feedback to the user in the form of an approximate persistence barcode of the estimator, from which the user can choose a relevant merging parameter τ. For instance, the example of Figure 6 is highly non-linear and noisy, yet the barcode clearly shows two long intervals, suggesting that there are two main clusters.

Another important feature of our approach is to make a clear distinction between the merging criterion, governed by τ and based solely on persistence information, and the approximation accuracy of the basins of attraction of the maxima, governed by the Rips parameter δ and based solely on spatial information. In the example of Figure 6, reducing δ while keeping τ fixed enabled us to separate the two spirals from the background while keeping them separate and integral.

Shape Segmentation. The goal of shape segmentation is to partition a given shape into *meaningful* segments, such as fingers on a hand. This problem is ill-posed by nature, as meaningfulness is a subjective notion. Given a sampled shape $\mathbb{X}$, our approach is to apply the algorithm

of Section 4.3 on some *segmentation function* $f : \mathbb{X} \to \mathbb{R}$ derived from the geometric features of $\mathbb{X}$. The output is a partition of the point cloud into clusters corresponding roughly to the basins of attraction of the significant peaks of f. Thus, we cast the segmentation problem into another problem, namely the one of finding a relevant segmentation function for a given class of data.

We investigated two functions in our experiments: Distance from a point to the set of samples on the boundary, as proposed in [12, 26]; Diameter of the set of nearest samples on the boundary, normalized by the previous distance. We chose these two functions as a demonstration, but our method can be applied virtually with any segmentation function. The approximate barcodes computed by the algorithm provide information on the stability of the different segments. This information can be viewed as an indicator of the relevance of a given segmentation function on a particular class of data. In Figures 2 and 3, the barcodes suggest that the second function is superior to the first one at separating the fingers from the palm of a hand. Yet, the second function turned out to be too noisy on the octopus data set of Figures 4 and 5.

Sensor networks. Our approach was originally designed with the sensor network framework in mind, where physical quantities such as temperature or humidity are measured by a collection of communicating sensors, and where the goal is to answer qualitative queries such as how many significant *hot spots* are being sensed. Purely geometric approaches cannot be applied in this setting, since geographic location is usually unavailable. Rough pairwise geodesic distances however are available, in the form of graph distances in the communication network. With this data at hand, the algorithms of Section 4 can find the number of hot spots, provide an estimation of their prominance and of their size in the network, and track them as the quantity being measured changes. The computations are done in a centralized way, after a data aggregation step.

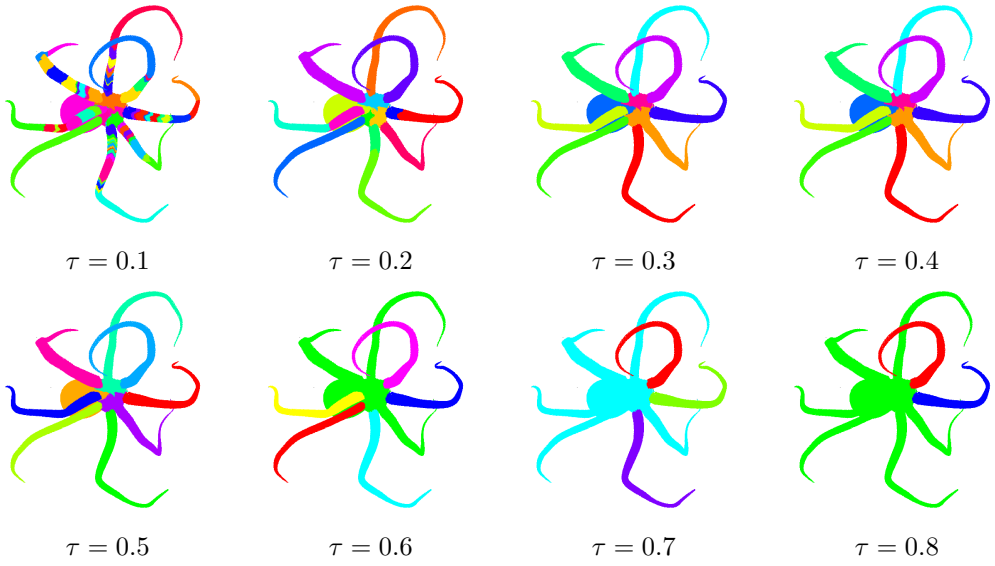

Figure 5: *Influence of the persistence threshold τ on the data set of Figure 4.*

Final remarks

The potential of our approach stems from the observation that many problems can be reduced to the analysis of some scalar field defined over a given point cloud data. With the theoretical and algorithmic tools developed in this paper at hand, the users can cast these problems into the one of finding the scalar field that is most suitable for their particular purposes. Thus, clustering is turned into a density estimation problem, while shape segmentation is turned into finding a relevant segmentation function for a given class of shapes. Many application-specific questions arise from this paradigm, which we do not pretend to solve in the paper. Some of them, related to the above scenarios, will be addressed in subsequent work.

References

[1] P.-T. Bremer, H. Edelsbrunner, B. Hamann, and V. Pascucci. Topological hierarchy for functions on triangulated surfaces. *IEEE Trans. Vis. Comput. Graphics*, 10:385–396, 2004.

[2] M. Do Carmo. *Riemannian Geometry*. Birkhäuser, Boston, Basel, Berlin, 1992.

[3] F. Cazals, F. Chazal, and T. Lewiner. Molecular shape analysis based upon the Morse-Smale complex and the Connolly function. In *Proc. 19th Annu. ACM Sympos. Comput. Geom.*, pages 237–246, 2003.

[4] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Y. Oudot. Proximity of persistence modules and their diagrams. Research Report 6568, INRIA, November 2008.

[5] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in Euclidean space. In *Proc. 22nd Annu. Sympos. Comput. Geom.*, pages 319–326, 2006.

[6] F. Chazal and A. Lieutier. Stability and computation of topological invariants of solids in $\mathbb{R}^n$. *Discrete Comput. Geom.*, 37(4):601–617, 2007.

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

[8] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *Proc. 21st ACM Sympos. Comput. Geom.*, pages 263–271, 2005.

[9] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and D. Morozov. Persistent homology for kernels and images. Preprint, 2008.

[10] D. Cohen-Steiner, H. Edelsbrunner, and D. Morozov. Vines and vineyards by updating persistence in linear time. In *Proc. 22nd Sympos. on Comput. Geom.*, pages 119–126, 2006.

[11] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[12] T. K. Dey and R. Wenger. Stability of critical points with interval persistence. *Discrete Comput. Geom.*, 38:479–512, 2007.

[13] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical Morse complexes for piecewise linear 2-manifolds. In *Proc. 17th Annu. Sympos. Comput. Geom.*, pages 70–79, 2001.

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

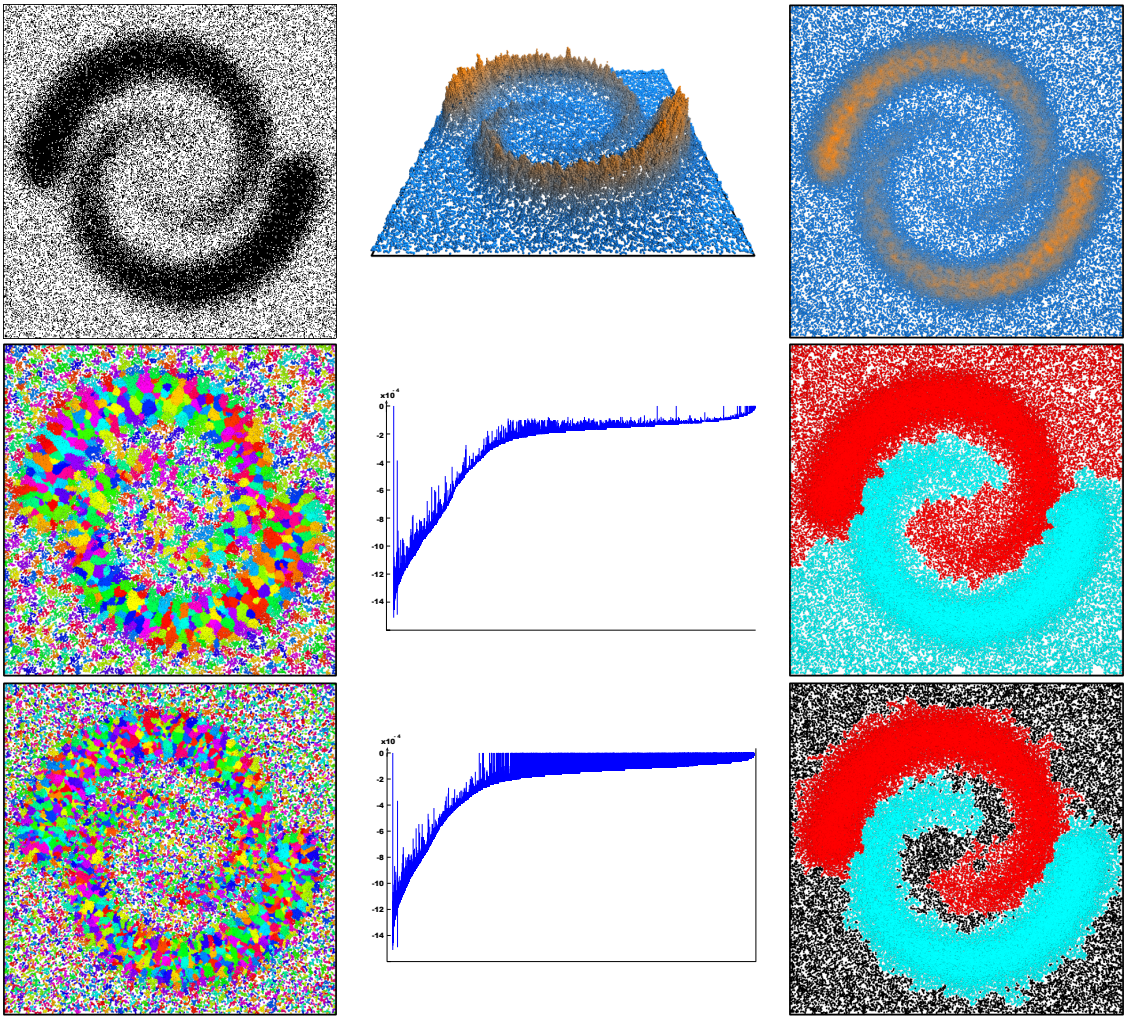

Figure 6: *A result in clustering. The top row shows the input provided to the algorithm of Section 4.3: the data points (left), or rather their pairwise Euclidean distances, and the estimated density function f (center and right). The 3-d view of f illustrates how noisy this function can be in practice, thereby emphasizing the importance of our robustness results (Theorems 3.7 and 3.9). The two bottom rows show the results of the algorithm when applied to $-f$ to get the ascending regions of the maxima of f. Two different Rips parameters have been used: $\delta = 15$ (center row) and $\delta = 10$ (bottom row). Each row shows the result of the clustering before (left) and after (right) merging non-persistent clusters. The persistence barcodes, shown in the center column, contain two prominent intervals (to the left) corresponding to the two main clusters. Since the estimated density is everywhere positive, the barcodes have been thresholded at 0. Thus, intervals reaching 0 correspond to independent connected components in the Rips graph. The ones to the right of the barcodes are treated as noise and their corresponding clusters shown in black: this is because they appear lately, meaning that their corresponding peaks of f are low.*

[15] H. Edelsbrunner, D. Morozov, and V. Pascucci. Persistence-sensitive simplification of functions on 2-manifolds. In *Proc. 22nd Sympos. on Comput. Geom.*, pages 127–134, 2006.

[16] J. Gao, L. Guibas, S. Oudot, and Y. Wang. Geodesic Delaunay triangulation and witness complex in the plane. Full version, partially published in *Proc. 18th ACM-SIAM Sympos. on Discrete Algorithms*, pages 571–580, 2008. Full draft available at: `http://graphics.stanford.edu/projects/lgl/papers/ggow-gtwcp-08/ggow-gdtwcp-08-full.pdf`.

[17] L. G. Guibas and S. Y. Oudot. Reconstruction using witness complexes. In *Proc. 18th Sympos. on Discrete Algorithms*, pages 1076–1085, 2007.

[18] L. J. Guibas. Kinetic data structures — a state of the art report. In P. K. Agarwal, L. E. Kavraki, and M. Mason, editors, *Proc. Workshop Algorithmic Found. Robot.*, pages 191–209. A. K. Peters, Wellesley, MA, 1998.

[19] A. Gyulassy, V. Natarajan, V. Pascucci, P.-T. Bremer, and B. Hamann. Topology-based simplification for feature extraction from 3d scalar fields. In *Proc. IEEE Conf. Visualization*, pages 275–280, 2005.

[20] A. Gyulassy, V. Natarajan, V. Pascucci, P.-T. Bremer, and B. Hamann. A topological approach to simplification of three-dimensional scalar fields. *IEEE Trans. Vis. Comput. Graphics*, 12(4):474–484, 2006.

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

[22] John W. Milnor. *Morse Theory*. Princeton University Press, Princeton, NJ, 1963.

[23] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, to appear.

[24] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[25] F. Zhao and L. J. Guibas. *Wireless Sensor Networks*. Morgan Kaufmann, 2004.

[26] X. Zhu, R. Sarkar, and J. Gao. Shape segmentation and applications in sensor networks. In *Proc. INFOCOM*, pages 1838–1846, 2007.

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

[28] `http://www.cs.umd.edu/~mount/ANN/`.

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399