The density of expected persistence diagrams and its kernel based estimation

Frédéric Chazal, Vincent Divol

The density of expected persistence diagrams and its kernel based estimation

Frédéric Chazal
Inria Saclay
Palaiseau, France
frederic.chazal@inria.fr

Vincent Divol
École Normale Supérieure,
Paris, France
vincent.divol@ens.fr

Abstract

Persistence diagrams play a fundamental role in Topological Data Analysis where they are used as topological descriptors of filtrations built on top of data. They consist in discrete multisets of points in the plane $\mathbb{R}^2$ that can equivalently be seen as discrete measures in $\mathbb{R}^2$. When the data come as a random point cloud, these discrete measures become random measures whose expectation is studied in this paper. First, we show that for a wide class of filtrations, including the Čech and Rips-Vietoris filtrations, the expected persistence diagram, that is a deterministic measure on $\mathbb{R}^2$, has a density with respect to the Lebesgue measure. Second, building on the previous result we show that the persistence surface recently introduced in [1] can be seen as a kernel estimator of this density. We propose a cross-validation scheme for selecting an optimal bandwidth, which is proven to be a consistent procedure to estimate the density.

2012 ACM Subject Classification Theory of computation → Randomness, geometry and discrete structures → Computational geometry

Keywords and phrases topological data analysis, persistence diagrams, subanalytic geometry

Funding This work was partially supported by the Advanced Grant of the European Research Council GUDHI (Geometric Understanding in Higher Dimensions) and a collaborative research agreement between Inria and Fujitsu.

1 Introduction

Persistent homology [16], a popular approach in Topological Data Analysis (TDA), provides efficient mathematical and algorithmic tools to understand the topology of a point cloud by tracking the evolution of its homology at different scales. Specifically, given a scale (or time) parameter r and a point cloud $x = (x_1, \dots, x_n)$ of size n, a simplicial complex $\mathcal{K}(x, r)$ is built on $\{1, \dots, n\}$ thanks to some procedure, such as, e.g., the nerve of the union of balls of radius r centered on the point cloud or the Vietoris-Rips complex. Letting the scale r increase gives rise to an increasing sequence of simplicial complexes $\mathcal{K}(x) = (\mathcal{K}(x, r))_r$ called a *filtration*. When a simplex is added in the filtration at a time r, it either "creates" or "fills" some hole in the complex. Persistent homology keeps track of the birth and death of these holes and encodes them as a *persistence diagram* that can be seen as a relevant and stable [6, 7] multi-scale topological descriptor of the data. A persistence diagram D_s is thus a collection of pairs of numbers, each of those pairs corresponding to the birth time and the death time of a s-dimensional hole. A precise definition of persistence diagram can be found,

for example, in [16, 8]. Mathematically, a diagram is a multiset of points in

$$\Delta = \{\mathbf{r} = (r_1, r_2), 0 \leq r_1 < r_2 \leq \infty\}. \tag{1}$$

Note that in a general setting, points $\mathbf{r} = (r_1, r_2)$ in diagrams can be "at infinity" on the line $\{r_2 = \infty\}$ (e.g. a hole may never disappear). However, in the cases considered in this paper, this will be the case for a single point for 0-dimensional homology, and this point will simply be discarded in the following.

In statistical settings, one is often given a (i.i.d.) sample of (random) point clouds $\mathbb{X}_1, \ldots, \mathbb{X}_N$ and filtrations $\mathcal{K}(\mathbb{X}_1), \ldots, \mathcal{K}(\mathbb{X}_N)$ built on top of them. We consider the set of persistence diagrams $D_s[\mathcal{K}(\mathbb{X}_1)], \ldots, D_s[\mathcal{K}(\mathbb{X}_N)]$, which are thought to contain relevant topological information about the geometry of the underlying phenomenon generating the point clouds. The space of persistence diagrams is naturally endowed with the so-called *bottleneck distance* [12] or some variants. However, the resulting metric space turns out to be highly non linear, making the statistical analysis of distributions of persistence diagrams rather awkward, despite several interesting results such as, e.g., [28, 15, 10]. A common scheme to overcome this difficulty is to create easier to handle statistics by mapping the diagrams to a vector space thanks to a feature map Ψ, also called a representation (see, e.g., [1, 2, 4, 9, 11, 20, 25]). A classical idea to get information about the typical shape of a random point cloud is then to estimate the expectation $E[\Psi(D_s[\mathcal{K}(\mathbb{X}_i)])]$ of the distribution of representations using the mean representation

$$\overline{\Psi}_N := \frac{\sum_{i=1}^N \Psi(D_s[\mathcal{K}(\mathbb{X}_i)])}{N}. \tag{2}$$

In this direction, [4] introduces a representation called persistence landscape, and shows that it satisfies law of large numbers and central limit theorems. Similar theorems can be shown for a wide variety of representations: it is known that $\overline{\Psi}_N$ is a consistent estimator of $E[\Psi(D_s[\mathcal{K}(\mathbb{X}_i)])]$. Although it may be useful for a classification task, this mean representation is still somewhat disappointing from a theoretical point of view. Indeed, what exactly $E[\Psi(D_s[\mathcal{K}(\mathbb{X}_i)])]$ is, has been scarcely studied in a non-asymptotic setting, i.e. when the cardinality of the random point cloud $\mathbb{X}_i$ is fixed or bounded.

Asymptotic results, when the size of the considered point clouds goes to infinity, are well understood for some non-persistent descriptors of the data, such as the Betti numbers: a natural question in geometric probability is to study the asymptotics of the s-dimensional Betti numbers $\beta_s(\mathcal{K}(\mathbb{X}_n, r_n))$ where $\mathbb{X}_n$ is a point cloud of size n and under different asymptotics for r_n. Notable results on the topic include [17, 30, 31]. Considerably less results are known about the asymptotic properties of fundamentally persistent descriptors of the data: [3] finds the right order of magnitude of maximally persistent cycles and [14] shows the convergence of persistence diagrams on stationary process in a weak sense.

Contributions of the paper.

In this paper, representing persistence diagrams as discrete measures, i.e. as element of the space of measures on $\mathbb{R}^2$, we establish non-asymptotic global properties of various representations and persistence-based descriptors. A multiset of points is naturally in bijection with the discrete measure defined on $\mathbb{R}^2$ created by putting Dirac measures on each point of the multiset, with mass equal to the multiplicity of the point. In this paper a persistence diagram D_s is thus represented as a discrete measure on Δ and with a slight abuse of notation, we will write

$$D_s = \sum_{\mathbf{r} \in D_s} \delta_{\mathbf{r}}, \tag{3}$$

where $\delta_{\mathbf{r}}$ denotes the Dirac measure in $\mathbf{r}$ and where, as mentioned above, points with infinite persistence are simply discarded. A wide class of representations, including the persistence surface [1] (variants of this object have been also introduced [11, 20, 25]), the accumulated persistence function [2] or persistence silhouette [9] are conveniently expressed as $\Psi(D_s) = D_s(f) := \sum_{\mathbf{r}\in D_s} f(\mathbf{r})$ for some function f on Δ. Given a random set of points $\mathbb{X}$, the expected behavior of the representations $E[D_s[\mathcal{K}(\mathbb{X})](f)]$ is well understood if the expectation $E[D_s[\mathcal{K}(\mathbb{X})]]$ of the distribution of persistence diagrams is understood, where the expectation $E[\mu]$ of a random discrete measure μ is defined by the equation $E[\mu](B) = E[\mu(B)]$ for all Borel sets B (see [23] for a precise definition of $E[\mu]$ in a more general setting). Our main contribution (Theorem 7) consists in showing that for a large class of situations the expected persistence diagram $E[D_s[\mathcal{K}(\mathbb{X})]]$, which is a measure on $\Delta \subset \mathbb{R}^2$, has a density p with respect to the Lebesgue measure on $\mathbb{R}^2$. Therefore, $E[\Psi(D_s[\mathcal{K}(\mathbb{X})])]$ is equal to $\int pf$, and if properties of the density p are shown (such as smoothness), those properties will also apply to the expectation of the representation Ψ.

The main argument of the proof of Theorem 7 relies on the basic observation that for point clouds $\mathbb{X}$ of given size n, the filtration $\mathcal{K}(\mathbb{X})$ can induce a finite number of ordering configurations of the simplices. The core of the proof consists in showing that, under suitable assumptions, this ordering is locally constant for almost all $\mathbb{X}$. As one needs to use geometric arguments, having properties only satisfied almost everywhere is not sufficient for our purpose. One needs to show that properties hold in a stronger sense, namely that the set on which it is satisfied is a dense open set. Hence, a convenient framework to obtain such properties is given by subanalytic geometry [26]. Subanalytic sets are a class of subsets of $\mathbb{R}^d$ that are locally defined as linear projections of sets defined by analytic equations and inequations. As most considered filtrations in Topological Data Analysis result from real algebraic constructions, such sets naturally appear in practice. On open sets where the combinatorial structure of the filtration is constant, the way the points in the diagrams are matched to pairs of simplices is fixed: only the times/scales at which those simplices appear change. Under an assumption of smoothness of those times, and using the coarea formula [24], a classical result of geometric measure theory generalizing the change of variables formula in integrals, one then deduces the existence of a density for $E[D_s[\mathcal{K}(\mathbb{X})]]$.

Among the different representations of the form $\Psi(D) = D(f)$, persistence surface is of particular interest. It is defined as the convolution of a diagram with a gaussian kernel. Hence, the mean persistence surface can be seen as a kernel density estimator of the density p of Theorem 7. As a consequence, the general theory of kernel density estimation applies and gives theoretical guarantees about various statistical procedures. As an illustration, we consider the bandwidth selection problem for persistence surfaces. Whereas Adams et al. [1] states that any reasonable bandwidth is sufficient for a classification task, we give arguments for the opposite when no "obvious" shapes appear in the diagrams. We then propose a cross-validation scheme to select the bandwidth matrix. The consistency of the procedure is shown using Stone's theorem [27]. This procedure is implemented on a set of toy examples illustrating its relevance.

The paper is organized as follow: section 2 is dedicated to the necessary background in geometric measure theory and subanalytic geometry. Results are stated in section 3, and the main theorem is proved in section 4. It is shown in section 5 that the main result applies to the Čech and Rips-Vietoris filtrations. Section 6 is dedicated to the statistical study of persistence surface, and numerical illustrations are found in section 7. All the technical proofs that are not essential to the understanding of the idea and results of the paper have been moved to the Appendix.

2 Preliminaries

2.1 The coarea formula

The proof of the existence of the density of the expected persistence diagram depends heavily on a classical result in geometric measure theory, the so-called coarea formula (see [24] for a gentle introduction to the subject). It consists in a more general version of the change of variables formula in integrals. Let (M, ρ) be a metric space. The diameter of a set $A \subset (M, \rho)$ is defined by $\sup_{x,y \in A} \rho(x, y)$.

▶ **Definition 1.** Let k be a non-negative integer. For $A \subset M$, and $\delta > 0$, consider

$$\mathcal{H}_k^\delta(A) := \inf \left\{ \sum_i \operatorname{diam}(U_i)^k, A \subset \bigcup_i U_i \text{ and } \operatorname{diam}(U_i) < \delta \right\}. \tag{4}$$

The *k-dimensional Haussdorf measure* on M of A is defined by $\mathcal{H}_k(A) := \lim_{\delta \to 0} \mathcal{H}_k^\delta(A)$.

If M is a d-dimensional submanifold of $\mathbb{R}^D$, the d-dimensional Haussdorf measure coincides with the volume form associated to the ambient metric restricted to M. For instance, if M is an open set of $\mathbb{R}^D$, the Haussdorf measure is the D-dimensional Lebesgue measure.

▶ **Theorem 2** (Coarea formula [24])**.** *Let M (resp. N) be a smooth Riemannian manifold of dimension m (resp n). Assume that $m \geq n$ and let $\Phi : M \to N$ be a differentiable map. Denote by $D\Phi$ the differential of Φ. The Jacobian of Φ is defined by $J\Phi = \sqrt{\det((D\Phi) \times (D\Phi)^t)}$. For $f : M \to \mathbb{R}_+$ a positive measurable function, the following equality holds:*

$$\int_M f(x) J\Phi(x) d\mathcal{H}_m(x) = \int_N \left(\int_{x \in \Phi^{-1}(\{y\})} f(x) d\mathcal{H}_{m-n}(x) \right) d\mathcal{H}_n(y). \tag{5}$$

In particular, if $J\Phi > 0$ almost everywhere, one can apply the coarea formula to $f \times (J\Phi)^{-1}$ to compute $\int_M f$. Having $J\Phi > 0$ is equivalent to have $D\Phi$ of full rank: most of the proof of our main theorem consists in showing that this property holds for certain functions Φ of interest.

2.2 Background on subanalytic sets

We now give basic results on subanalytic geometry, whose proofs are given in Appendix. See [26] for a thorough review of the subject. Let $M \subset \mathbb{R}^D$ be a connected real analytic submanifold possibly with boundary, whose dimension is denoted by d.

▶ **Definition 3.** A subset X of M is *semianalytic* if each point of M has a neighbourhood $U \subset M$ such that $X \cap U$ is of the form

$$\bigcup_{i=1}^{p} \bigcap_{j=1}^{q} X_{ij}, \tag{6}$$

where X_{ij} is either $f_{ij}^{-1}(\{0\})$ or $f_{ij}^{-1}((0, \infty))$ for some analytic functions $f_{ij} : U \to \mathbb{R}$.

▶ **Definition 4.** A subset X of M is *subanalytic* if for each point of M, there exists a neighborhood U of this point, a real analytic manifold N and A, a relatively compact semianalytic set of $N \times M$, such that $X \cap U$ is the projection of A on M. A function $f : X \to \mathbb{R}$ is subanalytic if its graph is subanalytic in $M \times \mathbb{R}$. The set of real-valued subanalytic functions on X is denoted by $\mathcal{S}(X)$.

A point x in a subanalytic subset X of M is smooth (of dimension k) if, in some neighbourhood of x in M, X is an analytic submanifold (of dimension k). The maximal dimension of a smooth point of X is called the dimension of X. The smooth points of X of dimension d are called regular, and the other points are called singular. The set $\mathrm{Reg}(X)$ of regular points of X is an open subset of M, possibly empty; the set of singular points is denoted by $\mathrm{Sing}(X)$.

▶ **Lemma 5.** *(i) For $f \in \mathcal{S}(M)$, the set $A(f)$ on which f is analytic is an open subanalytic set of M. Its complement is a subanalytic set of dimension smaller than d.*

Fix X a subanalytic subset of M. Assume that $f, g : X \to \mathbb{R}$ are subanalytic functions such that the image of a bounded set is bounded. Then,

(ii) The functions fg and $f+g$ are subanalytic.

(iii) The sets $f^{-1}(\{0\})$ and $f^{-1}((0,\infty))$ are subanalytic in M.

As a consequence of point (i), for $f \in \mathcal{S}(M)$, one can define its gradient ∇f everywhere but on some subanalytic set of dimension smaller than d.

▶ **Lemma 6.** *Let X be a subanalytic subset of M. If the dimension of X is smaller than d, then $\mathcal{H}_d(X) = 0$.*

As a direct corollary, we always have

$$\mathcal{H}_d(X) = \mathcal{H}_d(\mathrm{Reg}(X)). \tag{7}$$

Write $\mathcal{N}(M)$ the class of subanalytic subsets X of M with $\mathrm{Reg}(X) = \emptyset$. We have just shown that $\mathcal{H}_d \equiv 0$ on $\mathcal{N}(M)$. They form a special class of negligeable sets. We say that a property is verified *almost subanalytically everywhere* (a.s.e.) if the set on which it is not verified is included in a set of $\mathcal{N}(M)$. For example, Lemma 5 implies that ∇f is defined a.s.e..

3 The density of expected persistence diagrams

Let $n > 0$ be an integer. Write $\mathcal{F}_n$ the collection of non-empty subsets of $\{1, \dots, n\}$. Let $\varphi = (\varphi[J])_{J \in \mathcal{F}_n} : M^n \to \mathbb{R}^{\mathcal{F}_n}$ be a continuous function. The function φ will be used to construct the persistence diagram and is called a *filtering function*: a simplex J is added in the filtration at the time $\varphi[J]$. Write for $x = (x_1, \dots, x_n) \in M^n$ and for J a simplex, $x(J) := (x_j)_{j \in J}$. We make the following assumptions on φ:

(K1) *Absence of interaction:* For $J \in \mathcal{F}_n$, $\varphi[J](x)$ only depends on $x(J)$.

(K2) *Invariance by permutation:* For $J \in \mathcal{F}_n$ and for $(x_1, \dots, x_n) \in M^n$, if τ is a permutation of $\{1, \dots, n\}$, then $\varphi[J](x_{\tau(1)}, \dots, x_{\tau(n)}) = \varphi[J](x_1, \dots, x_n)$.

(K3) *Monotony:* For $J \subset J' \in \mathcal{F}_n$, $\varphi[J] \leq \varphi[J']$.

(K4) *Compatibility:* For a simplex $J \in \mathcal{F}_n$ and for $j \in J$, if $\varphi[J](x_1, \dots, x_n)$ is not a function of x_j on some open set U of M^n, then $\varphi[J] \equiv \varphi[J \backslash \{j\}]$ on U.

(K5) *Smoothness:* The function φ is subanalytic and the gradient of each of its entries (which is defined a.s.e.) is non vanishing a.s.e..

Assumptions (K2) and (K3) ensure that a filtration $\mathcal{K}(x)$ can be defined thanks to φ by:

$$\forall J \in \mathcal{F}_n,\ J \in \mathcal{K}(x, r) \iff \varphi[J](x) \leq r. \tag{8}$$

Assumption (K1) means that the moment a simplex is added in the filtration only depends on the position of its vertices, but not on their relative position in the point cloud. For $J \in \mathcal{F}_n$, the gradient of $\varphi[J]$ is a vector field in TM^n. Its projection on the jth coordinate is denoted by $\nabla^j \varphi[J]$: it is a vector field in TM defined a.s.e.. The persistence diagram of the filtration $\mathcal{K}(x)$ for s-dimensional homology is denoted by $D_s[\mathcal{K}(x)]$.

▶ **Theorem 7.** *Fix $n \geq 1$. Assume that M is a real analytic compact d-dimensional connected submanifold possibly with boundary and that $\mathbb{X}$ is a random variable on M^n having a density with respect to the Haussdorf measure $\mathcal{H}_{dn}$. Assume that $\mathcal{K}$ satisfies the assumptions (K1)-(K5). Then, for $s \geq 0$, the expected measure $E[D_s[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on Δ.*

▶ Remark. The condition that M is compact can be relaxed in most cases: it is only used to ensure that the subanalytic functions appearing in the proof satisfy the boundedness condition of Lemma 5. For the Čech and Rips-Vietoris filtrations, one can directly verify that the function φ (and therefore the functions appearing in the proofs) satisfies it when $M = \mathbb{R}^d$. Indeed, in this case, the filtering functions are semi-algebraic.

Classical filtrations such as the Rips-Vietoris and Čech filtrations do not satisfy the full set of assumptions (K1)-(K5). Specifically, they do not satisfy the second part of assumption (K5): all singletons $\{j\}$ are included at time 0 in those filtrations so that $\varphi[\{j\}] \equiv 0$, and the gradient $\nabla\varphi[\{j\}]$ is therefore null everywhere. This leads to a well-known phenomenon on Rips-Vietoris and Čech diagrams: all the non-infinite points of the diagram for 0-dimensional homology are included in the vertical line $\{0\} \times [0, \infty)$. A theorem similar to Theorem 7 still holds in this case:

▶ **Theorem 8.** *Fix $n \geq 1$. Assume that M is a real analytic compact d-dimensional connected submanifold and that $\mathbb{X}$ is a random variable on M^n having a density with respect to the Haussdorf measure $\mathcal{H}_{dn}$. Define assumption (K5'):*

(K5') The function φ is subanalytic and the gradient of its entries J of size greater than 1 is non vanishing a.s.e.. Moreover, for $\{j\}$ a singleton, $\varphi[\{j\}] \equiv 0$.

Assume that $\mathcal{K}$ satisfies the assumptions (K1)-(K4) and (K5'). Then, for $s \geq 1$, $E[D_s[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on Δ. Moreover, $E[D_0[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on the vertical line $\{0\} \times [0, \infty)$.

The proof of Theorem 8 is very similar to the proof of Theorem 7. It is therefore relegated to the appendix.

One can easily generalize Theorem 7 and assume that the size of the point process $\mathbb{X}$ is itself random. For $n \in \mathbb{N}$, define a function $\varphi^{(n)} : M^n \to \mathbb{R}^{\mathcal{F}_n}$ satisfying the assumption (K1)-(K5). If x is a finite subset of M, define $\mathcal{K}(x)$ by the filtration associated to $\varphi^{(|x|)}$ where $|x|$ is the size of x. We obtain the following corollary, proven in the appendix.

▶ **Corollary 9.** *Assume that $\mathbb{X}$ has some density with respect to the law of a Poisson process on M of intensity $\mathcal{H}_d$, such that $E\left[2^{|\mathbb{X}|}\right] < \infty$. Assume that $\mathcal{K}$ satisfies the assumptions (K1)-(K5). Then, for $s \geq 0$, $E[D_s[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on Δ.*

The condition $E\left[2^{|\mathbb{X}|}\right] < \infty$ ensures the existence of the expected diagram and is for example satisfied when $\mathbb{X}$ is a Poisson process with finite intensity.

As the way the filtration is created is smooth, one may actually wonder whether the density of $E[D_s[\mathcal{K}(\mathbb{X})]]$ is smooth as well: it is the case as long as the way the points are sampled is smooth. Recalling that a function is said to be of class C^k if it is k times differentiable, with a continuous kth derivative, we have the following result.

▶ **Theorem 10.** *Fix $0 \leq k \leq \infty$ and assume that $\mathbb{X} \in M^n$ has some density of class C^k with respect to $\mathcal{H}_{nd}$. Then, for $s \geq 0$, the density of $E[D_s[\mathcal{K}(\mathbb{X})]]$ is of class C^k.*

The proof is based on classical results of continuity under the integral sign as well as an use of the implicit function theorem: it can be found in the appendix.

As a corollary of Theorem 10, we obtain the smoothness of various expected descriptors computed on persistence diagrams. For instance, the expected birth distribution and the expected death distribution have smooth densities under the same hypothesis, as they are obtained by projection of the expected diagram on some axis. Another example is the smoothness of the expected Betti curves. The sth Betti number $\beta_s^r(\mathcal{K}(x))$ of a filtration $\mathcal{K}(x)$ is defined as the dimension of the sth homology group of $\mathcal{K}(x,r)$. The Betti curves $r \mapsto \beta_s^r(\mathcal{K}(x))$ are step functions which can be used as statistics, as in [29] where they are used for a classification task on time series. With few additional work (see proof in Appendix), the expected Betti curves are shown to be smooth.

▶ **Corollary 11.** *Under the same hypothesis than Theorem 10, for $s \geq 0$, the expected Betti curve $r \mapsto E[\beta_s^r(\mathcal{K}(\mathbb{X}))]$ is a C^k function.*

4 Proof of Theorem 7

First, one can always replace M^n by $A(\varphi) = \bigcap_{J \in \mathcal{F}_n} A(\varphi[J])$, as Lemma 5 implies that it is an open set whose complement is in $\mathcal{N}(M^n)$. We will therefore assume that φ is analytic on M^n.

Given $x \in M^n$, the different values taken by $\varphi(x)$ on the filtration can be written $r_1 < \cdots < r_L$. Define $E_l(x)$ the set of simplices J such that $\varphi[J](x) = r_l$. The sets $E_1(x), \ldots, E_L(x)$ form a partition of $\mathcal{F}_n$ denoted by $\mathcal{A}(x)$.

▶ **Lemma 12.** *For a.s.e. $x \in M^n$, for $l \geq 1$, $E_l(x)$ has a minimal element J_l (for the partial order induced by inclusion).*

Proof. Fix $J, J' \subset \{1, \ldots, n\}$ with $J \neq J'$ and $J \cap J' \neq \emptyset$. consider the subanalytic functions $f : x \in M^n \mapsto \varphi[J](x) - \varphi[J'](x)$ and $g : x \in M^n \mapsto \varphi[J](x) - \varphi[J \cap J'](x)$. The set

$$C(J, J') := \{f = 0\} \cap \{g > 0\}. \tag{9}$$

is a subanalytic subset of M^n. Assume that it contains some open set U. On U, $\varphi[J](x)$ is equal to $\varphi[J'](x)$. Therefore, it does not depend on the entries x_j for $j \in J \backslash J'$. Hence, by assumption (K4), $\varphi[J](x)$ is actually equal to $\varphi[J \cap J'](x)$ on U. This is a contradiction with having $g > 0$ on U. Therefore, $C(J, J')$ does not contain any open set, and all its points are singular: $C(J, J')$ is in $\mathcal{N}(M^n)$. If $J \cap J' = \emptyset$, similar arguments show that $C(J, J') = \{f = 0\}$ cannot contain any open set: it would contradict assumption (K5). On the complement of

$$C := \bigcup_{J \neq J' \subset \{1, \ldots, n\}} C(J, J'), \tag{10}$$

having $\varphi[J](x) = \varphi[J'](x)$ implies that this quantity is equal to $\varphi[J \cap J'](x)$. This show the existence of a minimal element J_l to $E_l(x)$ on the complement of C. This property is therefore a.s.e. satisfied. ◀

▶ **Lemma 13.** *A.s.e., $x \mapsto \mathcal{A}(x)$ is locally constant.*

Proof. Fix $\mathcal{A}_0 = \{E_1, \ldots, E_l\}$ a partition of $\mathcal{F}_n$ induced by some filtration, with minimal elements $J_1, \ldots, J_l$. Consider the subanalytic functions F, G defined, for $x \in M^n$, by

$$F(x) = \sum_{l=1}^{L} \sum_{J \in E_l} (\varphi[J](x) - \varphi[J_l](x)) \text{ and } G(x) = \sum_{l \neq l'} (\varphi[J_l](x)) - \varphi[J_{l'}](x))^2 .$$

The set $\{x \in M^n, \mathcal{A}(x) = \mathcal{A}_0\}$ is exactly the set $C(\mathcal{A}_0) = \{F = 0\} \cap \{G > 0\}$, which is subanalytic. The sets $C(\mathcal{A}_0)$ for all partitions $\mathcal{A}_0$ of $\mathcal{F}_n$ define a finite partition of the space M^n. On each open set $\mathrm{Reg}(C(\mathcal{A}_0)))$, the application $x \mapsto \mathcal{A}(x)$ is constant. Therefore, $x \mapsto \mathcal{A}(x)$ is locally constant everywhere but on $\bigcup_{\mathcal{A}_0} \mathrm{Sing}(C(\mathcal{A}_0)) \in \mathcal{N}(M^n)$. ◀

Therefore, the space M^n is partitioned into a negligeable set of $\mathcal{N}(M^n)$ and some open subanalytic sets $U_1, \dots, U_R$ on which $\mathcal{A}$ is constant.

▶ **Lemma 14.** *Fix $1 \leq r \leq R$ and assume that $J_1, \dots, J_L$ are the minimal elements of $\mathcal{A}$ on U_r. Then, for $1 \leq l \leq L$ and $j \in J_l$, $\nabla^j \varphi[J_l] \neq 0$ a.s.e. on U_r.*

Proof. By minimality of J_l, for $j \in J_l$, the subanalytic set $\{\nabla^j \varphi[J_l] = 0\} \cap U_r$ cannot contain an open set. It is therefore in $\mathcal{N}(M^n)$. ◀

Fix $1 \leq r \leq R$ and write

$$V_r = U_r \setminus \left(\bigcup_{l=1}^{L} \bigcup_{j=1}^{|J_l|} \{\nabla^j \varphi[J_l] = 0\} \right).$$

The complement of V_r in U_r is still in $\mathcal{N}(M^n)$. For $x \in V_r$, $D_s[\mathcal{K}(x)]$ is written $\sum_{i=1}^{N} \delta_{\mathbf{r}_i}$, where $\mathbf{r}_i = (\varphi[J_{l_1}](x), \varphi[J_{l_2}](x)) =: (b_i, d_i)$. The integer N and the simplices J_{l_1}, J_{l_2} depend only on V_r. Note that d_i is always greater than b_i, so that J_{l_2} cannot be included in J_{l_1}. The map $x \mapsto \mathbf{r}_i$ has it differential of rank 2. Indeed, take $j \in J_{l_2} \backslash J_{l_1}$. By Lemma 14, $\nabla^j \varphi[J_{l_2}](x) \neq 0$. Also, as $\varphi[J_{l_1}]$ only depends on the entries of x indexed by J_{l_1} (assumption (K1)), $\nabla^j \varphi[J_{l_1}](x) = 0$. Furthermore, take j' in J_{l_1}. By Lemma 14, $\nabla^{j'} \varphi[J_{l_1}](x) \neq 0$. This implies that the differential is of rank 2.

We now compute the sth persistence diagram for $s \geq 0$. Write κ the density of $\mathbb{X}$ with respect to the measure $\mathcal{H}_{nd}$ on M^n. Then,

$$\begin{aligned} E[D_s[\mathcal{K}(\mathbb{X})]] &= \sum_{r=1}^{R} E\left[\mathbb{1}\{\mathbb{X} \in V_r\} D_s[\mathcal{K}(\mathbb{X})]\right] = \sum_{r=1}^{R} E\left[\mathbb{1}\{\mathbb{X} \in V_r\} \sum_{i=1}^{N_r} \delta_{\mathbf{r}_i}\right] \\ &= \sum_{r=1}^{R} \sum_{i=1}^{N_r} E\left[\mathbb{1}\{\mathbb{X} \in V_r\} \delta_{\mathbf{r}_i}\right] \end{aligned}$$

Write μ_{ir} the measure $E[\mathbb{1}\{\mathbb{X} \in V_r\}\delta_{\mathbf{r}_i}]$. To conclude, it suffices to show that this measure has a density with respect to the Lebesgue measure on Δ. This is a consequence of the coarea formula. Define the function $\Phi_{ir} : x \in V_r \mapsto \mathbf{r}_i = (\varphi[J_{l_1}](x), \varphi[J_{l_2}](x))$. We have already seen that Φ_{ir} is of rank 2 on V_r, so that $J\Phi_{ir} > 0$. By the coarea formula (see Lemma 2), for a Borel set B in Δ,

$$\begin{aligned} \mu_{ir}(B) = P(\Phi_{ir}(\mathbb{X}) \in B, \mathbb{X} \in V_r) &= \int_{V_r} \mathbb{1}\{\Phi_{ir}(x) \in B\} \kappa(x) d\mathcal{H}_{nd}(x) \\ &= \int_{u \in B} \int_{x \in \Phi_{ir}^{-1}(\{u\})} (J\Phi_{ir}(x))^{-1} \kappa(x) d\mathcal{H}_{nd-2}(x) du. \end{aligned}$$

Therefore, μ_{ir} has a density with respect to the Lebesgue measure on Δ equal to

$$p_{ir}(u) = \int_{x \in \Phi_{ir}^{-1}(\{u\})} (J\Phi_{ir}(x))^{-1} \kappa(x) d\mathcal{H}_{nd-2}(x). \tag{11}$$

Finally, $E[D_s[\mathcal{K}(\mathbb{X})]]$ has a density equal to

$$p(u) = \sum_{r=1}^{R} \sum_{i=1}^{N_r} \int_{x \in \Phi_{ir}^{-1}(\{u\})} (J\Phi_{ir}(x))^{-1} \kappa(x) d\mathcal{H}_{nd-2}(x). \tag{12}$$

▶ Remark. Notice that, for n fixed, the above proof, and thus the conclusion, of Theorem 7 also works if the diagrams are represented by normalized discrete measures, i.e. probability measures defined by

$$D_s = \frac{1}{|D_s|} \sum_{\mathbf{r} \in D_s} \delta_{\mathbf{r}}. \tag{13}$$

5 Examples

We now note that the Rips-Vietoris and the Čech filtrations satisfy the assumptions (K1)-(K4) and (K5') when $M = \mathbb{R}^d$ is an Euclidean space. Note that the similar arguments show that weighted versions of those filtrations (see [5]) satisfy assumptions (K1)-(K5).

5.1 Rips-Vietoris filtration

For the Rips-Vietoris filtration, $\varphi[J](x) = \max_{i,j \in J} \|x_i - x_j\|$. The function φ clearly satisfies (K1), (K2) and (K3). It is also subanalytic, as it is the maximum of semi-algebraic functions.

For $x \in M^n$ and $J \in \mathcal{F}_n$ a simplex of size greater than one, $\varphi[J](x) = \|x_i - x_j\|$ for some indices i, j. Those indices are locally stable, and $\varphi[J](x) = \varphi[\{i,j\}](x)$: hypothesis (K4) is satisfied. Furthermore, on this set,

$$\nabla \varphi[\{i,j\}](x) = \left(\frac{x_i - x_j}{\|x_i - x_j\|}, \frac{x_j - x_i}{\|x_i - x_j\|} \right) \neq 0. \tag{14}$$

Hence, (K5') is also satisfied: both Theorem 8 and Theorem 10 are satisfied for the Rips-Vietoris filtration.

5.2 Čech filtration

The ball centered at x of radius r is denoted by $B(x,r)$. For the Čech filtration,

$$\varphi[J](x) = \inf_{r>0} \left\{ \bigcap_{j \in J} B(x_j, r) \neq \emptyset \right\}. \tag{15}$$

First, it is clear that (K1), (K2) and (K3) are satisfied by φ.

We give without proof a characterization of the Čech complex.

▶ Proposition 15. Let x be in M^n and fix $J \in \mathcal{F}_n$. If the circumcenter of $x(J)$ is in the convex hull of $x(J)$, then $\varphi[J](x)$ is the radius of the circumsphere of $x(J)$. Otherwise, its projection on the convex hull belongs to the convex hull of some subsimplex $x(J')$ of $x(J)$ and $\varphi[J](x) = \varphi[J'](x)$.

▶ **Definition 16.** The Cayley-Menger matrix of a k-simplex $x = (x_1, \dots, x_k) \in M^k$ is the symmetric matrix $(M(x)_{i,j})_{i,j}$ of size $k+1$, with zeros on the diagonal, such that $M(x)_{1,j} = 1$ for $j > 1$ and $M(x)_{i+1,j+1} = \|x_i - x_j\|^2$ for $i, j \leq k$.

▶ Proposition 17 (see [13]). Let $x \in M^k$ be a point in general position. Then, the Cayley-Menger matrix $M(x)$ is invertible with $(M(x))^{-1}_{1,1} = -2r^2$, where r is the radius of the circumsphere of x. The kth other entries of the first line of $M(x)^{-1}$ are the barycentric coordinates of the circumcenter.

Therefore, the application which maps a simplex to its circumcenter is analytic, and the set on which the circumcenter of a simplex belongs in the interior of its convex hull is a subanalytic set. On such a set, the function φ is also analytic, as it is the square root of the inverse a matrix which is polynomial in x. Furthermore, on the open set on which the circumcenter is outside the convex hull, we have shown that $\varphi[J](x) = \varphi[J'](x)$ for some subsimplex J': assumption (K4) is satisfied.

Finally, let us show that assumption (K5') is satisfied. The previous paragraph shows the subanalyticity of φ. For $J \in \mathcal{F}_n$ a simplex of size greater than one, there exists some subsimplex J' such that $\varphi[J](x)$ is the radius of the circumsphere of $x(J')$. It is clear that there cannot be an open set on which this radius is constant. Thus, $\nabla\varphi[J]$ is a.s.e. non null.

6 Persistence surface as a kernel density estimator

Persistence surface is a representation of persistence diagrams introduced by Adams & al. in [1]. It consists in a convolution of a diagram with a kernel, a general idea that has been repeatedly and fruitfully exploited, with slight variations, for instance in [11, 20, 25]. For $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel and H a bandwidth matrix (e.g. a symmetric positive definite matrix), let for $u \in \mathbb{R}^2$,

$$K_H(u) = \det(H)^{-1/2} K(H^{-1/2} \cdot u). \tag{16}$$

For D a diagram, $K : \mathbb{R}^2 \to \mathbb{R}$ a kernel, H a bandwidth matrix and $w : \mathbb{R}^2 \to \mathbb{R}_+$ a weight function, one defines the persistence surface of D with kernel K and weight function w by:

$$\forall u \in \mathbb{R}^2,\ \rho(D)(u) := \sum_{\mathbf{r} \in D} w(\mathbf{r}) K_H(u - \mathbf{r}) = D(wK_H(u - \cdot)) \tag{17}$$

Assume that $\mathbb{X}$ is some point process satisfying the assumptions of Theorem 7. Then, for $s \geq 1$, $\mu := E[D_s[\mathcal{K}(\mathbb{X})]]$ has some density p with respect to the Lebesgue measure on Δ. Therefore, μ_w, the measure having density w with respect to μ, has a density equal to $w \times p$ with respect to the Lebesgue measure. The mean persistence surface $E[\rho(D_s[\mathcal{K}(\mathbb{X})])]$ is exactly the convolution of μ_w by some kernel function: the persistence surface $\rho(D_s[\mathcal{K}(\mathbb{X})])$ is actually a kernel density estimator of $w \times p$.

If a point cloud approximates a shape, then its persistence diagram (for the Čech filtration for instance) is made of numerous points with small persistences and a few meaningful points of high persistences which corresponds to the persistence diagram of the "true" shape. As one is interested in the latter points, a weight function w, which is typically an increasing function of the persistence, is used to suppress the importance of the topological noise in the persistence surface. Adams & al. [1] argue that in this setting, the choice of the bandwidth matrix H has few effects for statistical purposes (e.g. classification), a claim supported by numerical experiments on simple sets of synthetic data, e.g. torus, sphere, three clusters, etc.

However, in the setting where the datasets are more complicated and contain no obvious "real" shapes, one may expect the choice of the bandwidth parameter H to become more critical: there are no highly persistent, easily distinguishable points in the diagrams anymore and the precise structure of the density functions of the processes becomes of interest. We now show that a cross validation approach allows the bandwidth selection task to be done in an asymptotically consistent way. This is a consequence of a generalization of Stone's theorem [27] when observations are not random vectors but random measures.

Assume that $\mu_1, \ldots, \mu_N$ are i.i.d. random measures on $\mathbb{R}^2$, such that there exists a deterministic constant C with $|\mu_1| \leq C$. Assume that the expected measure $E[\mu_1]$ has a

bounded density p with respect to the Lebesgue measure on $\mathbb{R}^2$. Given a kernel $K : \mathbb{R}^2 \to \mathbb{R}$ and a bandwidth matrix H, one defines the kernel density estimator

$$\hat{p}_H(x) := \frac{1}{N}\sum_{i=1}^{N}\int K_H(x-y)\mu_i(dy). \tag{18}$$

The optimal bandwidth H_{opt} minimizes the Mean Integrated Square Error (MISE)

$$MISE(H) := E\left[\|p-\hat{p}_H\|^2\right] = E\left[\int (p(x)-\hat{p}_H(x))^2\,dx\right]. \tag{19}$$

Of course, as p is unknown, $MISE(H)$ cannot be computed. Minimizing $MISE(H)$ is equivalent to minimize $J(H) := MISE(H) - \|p\|^2$. Define

$$\hat{p}_{iH}(x) := \frac{1}{N-1}\sum_{j\neq i}\int K_H(x-y)\mu_j(dy) \tag{20}$$

and

$$\hat{J}(H) := \frac{1}{N^2}\sum_{i,j}\iint K_H^{(2)}(x-y)\mu_i(dx)\mu_j(dy) - \frac{2}{N}\sum_i\int \hat{p}_{iH}(x)\mu_i(dx), \tag{21}$$

where $K^{(2)} : x \mapsto \int K(x-y)K(y)dy$ denotes the convolution of K with itself. The quantity $\hat{J}(H)$ is an unbiased estimator of $J(H)$. The selected bandwidth $\hat{H}$ is then chosen to be equal to $\arg\min_H \hat{J}(H)$.

▶ **Theorem 18** (Stone's theorem [27]). *Assume that the kernel K is nonnegative, Hölder continuous and has a maximum attained in 0. Also assume that the density p is bounded. Then, $\hat{H}$ is asymptotically optimal in the sense that*

$$\frac{\|p-\hat{p}_{\hat{H}}\|}{\|p-\hat{p}_{H_{opt}}\|} \xrightarrow[N\to\infty]{} 1 \ \ a.s.. \tag{22}$$

Note that the gaussian kernel $K(x) = \exp(-\|x\|^2/2)$ satisfies the assumptions of Theorem 18.

Let $\mathbb{X}_1, \ldots, \mathbb{X}_N$ be i.i.d. processes on M having a density with respect to the law of a Poisson process of intensity $\mathcal{H}_d$. Assume that there exists a deterministic constant C with $|\mathbb{X}_i| \leq C$. Then, Theorem 18 can be applied to $\mu_i = D_s[\mathcal{K}(\mathbb{X}_i)]$. Therefore, *the cross validation procedure* (21) *to select H the bandwidth matrix in the persistence surface ensures that the mean persistence surface*

$$\overline{\rho}_N := \frac{1}{N}\sum_{i=1}^{N}\rho(D_s[\mathcal{K}(\mathbb{X}_i)]) \tag{23}$$

is a good estimator of p the density of $E[D_s[\mathcal{K}(\mathbb{X}_1)]]$.

7 Numerical illustration

Three sets of synthetic data are considered (see Figure 1). The first one (a) is made of $N = 40$ sets of $n = 300$ i.i.d. points uniformly sampled in the square $[0,1]^2$. The second one (b) is made of N samples of a clustered process: $n/3$ cluster's centers are uniformly sampled in the square. Each center is then replaced with 3 i.i.d. points following a normal

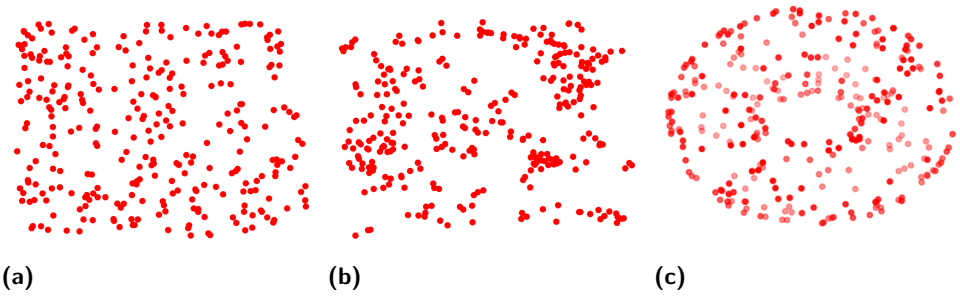


Figure 1 Realization of the processes (a), (b) and (c) described in Section 7.

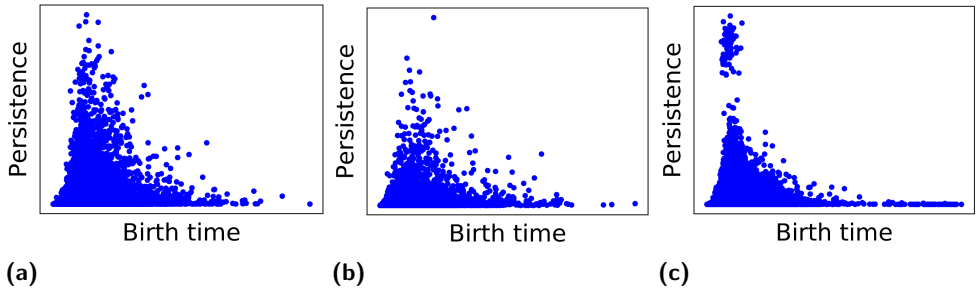
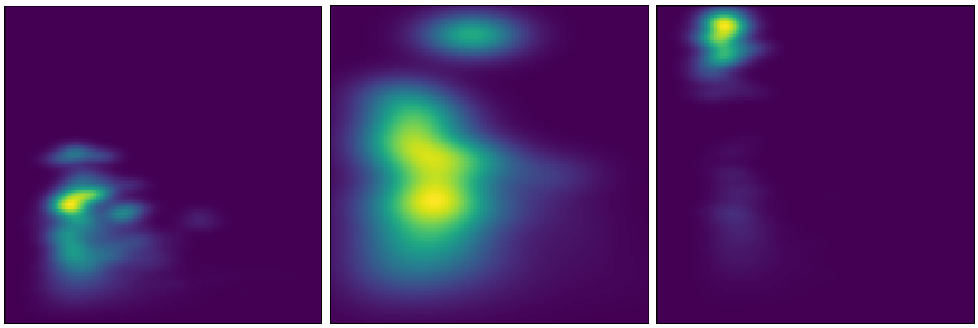


Figure 2 Superposition of the $N = 40$ diagrams of class (a), (b) and (c), transformed under the map $\mathbf{r} \to (r_1, r_2 - r_1)$.

distribution of standard deviation $0.01 \times n^{-1/2}$. The third dataset (c) is made of N samples of n uniform points on a torus of inner radius 1 and outer radius 2. For each set, a Čech persistence diagram for 1-dimensional homology is computed. Persistence diagrams are then transformed under the map $(r_1, r_2) \mapsto (r_1, r_2 - r_1)$, so that they now live in the upper-left quadrant of the plane. Figure 2 shows the superposition of the diagrams in each class. One may observe the slight differences in the structure of the topological noise over the classes (a) and (b). The cluster of most persistent points in the diagrams of class (c) correspond to the two holes of a torus and are distinguishable from the rest of the points in the diagrams of the class, which form topological noise. The persistence diagrams are weighted by the weight function $w(\mathbf{r}) = (r_2 - r_1)^3$, as advised in [19] for two-dimensional point clouds. The bandwidth selection procedure will be applied to the measures having density w with respect to the diagrams, e.g. a measure is a sum of weighted Dirac measures.

For each class of dataset, the score $\hat{J}(H)$ is computed for a set of bandwidth matrices of the form $h^2 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, for 50 values h evenly spaced on a log-scale between 10^{-5} and 1. Note that the computation of $\hat{J}(H)$ only involves the computations of $K_H(\mathbf{r}_1 - \mathbf{r}_2)$ for points $\mathbf{r}_1$, $\mathbf{r}_2$ in different diagrams. Hence, the complexity of the computation of $\hat{J}(H)$ is in $O(T^2)$, where T is the sum of the number of points in the diagrams of a given class. If this is too costly, one may use a subsampling approach to estimate the integrals. The selected bandwidth were respectively $h = 0.22, 0.60, 0.17$. Persistence surfaces for the selected bandwidth are displayed in Figure 3. The persistence of the "true" points of the torus are sufficient to suppress the topological noise: only two yellow areas are seen in the persistence surface of the torus. Note that the two areas can be separated, whereas it is not obvious when looking at the superposition of the diagrams, and would not have been obvious with an



(a) (b) (c)

Figure 3 Persistence surfaces for each class (a), (b) and (c), computed with the weight function $w(\mathbf{r}) = (r_2 - r_1)^3$ and with the bandwidth matrix selected by the cross-validation procedure.

arbitrary choice of bandwidth. The bandwidth for class (b) may look to have been chosen too big. However, there is much more variability in class (b) than in the other classes: this phenomenon explains that the density is less peaked around a few selected areas than in class (a).

Illustrations on non-synthetic data are shown in the appendix: similar behaviors are observed.

8 Conclusion and further works

Taking a measure point of view to represent persistence diagrams, we have shown that the expected behavior of persistence diagrams built on top of random point sets reveals to have a simple and interesting structure: a measure on $\mathbb{R}^2$ with density with respect to Lebesgue measure that is as smooth as the random process generating the data points! This opens the door to the use of effective kernel density estimation techniques for the estimation of the expectation of topological features of data. Our approach and results also seem to be particularly well-suited to the use of recent results on the Lepski method for parameter selection [22] in statistics, a research direction that deserves further exploration. As many persistence-based features considered among the literature - persistence images, birth and death distributions, Betti curves,... - can be expressed as linear functional of the discrete measure representation of diagrams, our results immediately extend to them. The ability to select the parameters on which these features are dependent in a well-founded statistical way also opens the door to a well-justified usage of persistence-based features in further supervised and un-supervised learning tasks.

References

1 Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: a stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

2 Christophe Biscio and Jesper Møller. The accumulated persistence function, a new useful functional summary statistic for topological data analysis, with a view to brain artery trees and spatial point process applications. *arXiv preprint arXiv:1611.00630*, 2016.

3 Omer Bobrowski, Matthew Kahle, Primoz Skraba, et al. Maximally persistent cycles in random geometric complexes. *The Annals of Applied Probability*, 27(4):2032–2060, 2017.

4 Peter Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015.

5 Mickaël Buchet, Frédéric Chazal, Steve Y Oudot, and Donald R Sheehy. Efficient and robust persistent homology for measures. *Computational Geometry*, 58:70–96, 2016.

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

7 F. Chazal, V. de Silva, and S. Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, 2014.

8 Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The structure and stability of persistence modules*. SpringerBriefs in Mathematics. Springer, 2016.

9 Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Proceedings of the thirtieth annual symposium on Computational geometry*, page 474. ACM, 2014.

10 Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *Journal of Machine Learning Research*, 16:3603–3635, 2015. URL: `http://jmlr.org/papers/v16/chazal15a.html`.

11 Yen-Chi Chen, Daren Wang, Alessandro Rinaldo, and Larry Wasserman. Statistical analysis of persistence intensity functions. *arXiv preprint arXiv:1510.02502*, 2015.

12 David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

13 HSM Coxeter. The circumradius of the general simplex. *The Mathematical Gazette*, pages 229–231, 1930.

14 Trinh Khanh Duy, Yasuaki Hiraoka, and Tomoyuki Shirai. Limit theorems for persistence diagrams. *arXiv preprint arXiv:1612.08371*, 2016.

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

16 D. Morozov H. Edelsbrunner. Persistent homology. In *Handbook of Discrete and Computational Geometry (3rd Ed - To appear)*. CRC Press (to appear), 2017.

17 Matthew Kahle, Elizabeth Meckes, et al. Limit the theorems for betti numbers of random simplicial complexes. *Homology, Homotopy and Applications*, 15(1):343–374, 2013.

18 Ludger Kaup and Burchard Kaup. *Holomorphic functions of several variables: an introduction to the fundamental theory*, volume 3. Walter de Gruyter, 1983.

19 Genki Kusano, Kenji Fukumizu, and Yasuaki Hiraoka. Kernel method for persistence diagrams via kernel embedding and weight factor. *arXiv preprint arXiv:1706.03472*, 2017.

20 Genki Kusano, Yasuaki Hiraoka, and Kenji Fukumizu. Persistence weighted gaussian kernel for topological data analysis. In *International Conference on Machine Learning*, pages 2004–2013, 2016.

21 J.H. Kwak and S. Hong. *Linear Algebra*. Birkhäuser Boston, 2004.

22 Claire Lacour, Pascal Massart, and Vincent Rivoirard. Estimator selection: a new method with applications to kernel density estimation. *arXiv preprint arXiv:1607.05091*, 2016.

23 Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

24 F. Morgan. *Geometric Measure Theory: A Beginner's Guide*. Elsevier Science, 2016.

25 Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4741–4748, 2015.

26 M. Shiota. *Geometry of Subanalytic and Semialgebraic Sets*. Progress in mathematics. Springer, 1997.

27 Charles J Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, pages 1285–1297, 1984.

28 Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014.

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

30 D Yogeshwaran, Robert J Adler, et al. On the topology of random complexes built over stationary point processes. *The Annals of Applied Probability*, 25(6):3338–3380, 2015.

31 D. Yogeshwaran, Eliran Subag, and Robert J. Adler. Random geometric complexes in the thermodynamic regime. *Probability Theory and Related Fields*, 167(1):107–142, Feb 2017. doi:10.1007/s00440-015-0678-9.

A Proofs of the subanalytic elementary lemmas

▶ **Lemma 19.** *(i) For $f \in \mathcal{S}(M)$, the set $A(f)$ on which f is analytic is an open subanalytic set of M. Its complement is a subanalytic set of dimension smaller than d.*

Fix X a subanalytic subset of M. Assume that $f, g : X \to \mathbb{R}$ are subanalytic functions such that the image of a bounded set is bounded. Then,

(ii) The functions fg and $f+g$ are subanalytic.

(iii) The sets $f^{-1}(\{0\})$ and $f^{-1}((0,\infty))$ are subanalytic in M.

Proof. (i) Section I.2.1 in [26] states that $A(f)$ is subanalytic. Therefore, its complement E is also subanalytic: it is enough to show that E is of empty interior to conclude.

Claim: The set F of points x where f is not analytic but G_f is locally a real analytic manifold in $(x, f(x))$ is a subanalytic set of empty interior.

Proof: Assume F contains an open set U. Replacing U by a smaller open set if necessary, there exists some local parametrization of $U_f = \{(x, f(x)),\ x \in U\}$ by some analytic function $\Phi : V \to \mathbb{R}$, V being a neighborhood of U_f in $M \times \mathbb{R}$. Denote by $\nabla^u \Phi \in \mathbb{R}$ the gradient of Φ with respect to the real variable $u \in \mathbb{R}$. The set Z on which $\nabla^u \Phi = 0$ is an analytic subset of V. As G_f is the graph of a function, $Z \cap G_f$ is made of isolated points: one can always assume that those points are not in U_f. Therefore, there exists some neighborhood V' of U_f which does not intersect Z. One can now apply the analytic implicit function theorem (see for instance [18, Section 8]) anywhere on U_f: for $(x_0, u_0) \in U_f$, there exists some neighborhood $W \subset V'$ and an analytic function $g : \Omega \to \mathbb{R}$, Ω being a neighborhood of x_0, such that, on W

$$\Phi(x, u) = 0 \iff u = g(x).$$

As we also have $\Phi(x, u) = 0$ if and only if $u = f(x)$, $f \equiv g$ on Ω and f is analytic on Ω. This is a contradiction with having f not analytic in every point of U. ■

Now, the set E is the union of F and of $E \cap G$ where G is the projection on M of $\mathrm{Reg}(G_f)$. As, by definition, $\mathrm{Reg}(G_f)$ is of empty interior, G is also of empty interior. Therefore, E is of empty interior, which is equivalent to say that its dimension is smaller than d.

(i) See [26, Section II.1.1].

(ii) See [26, Section II.1.6].

◀

▶ **Lemma 20.** *Let X be a subanalytic subset of M. If the dimension of X is smaller than d, then $\mathcal{H}_d(X) = 0$.*

Proof. Write k the dimension of X. First, one can always assume that X is closed, as $\mathcal{H}_d(\overline{X}) \geq \mathcal{H}_d(X)$. Therefore, there exists some real analytic manifold N of dimension k and a proper real analytic mapping $\Psi : N \to M$ such that $\Psi(N) = X$ (see [26, Section I.2.1]). The set X can be written as the union of some compact sets X_K for $K \geq 0$. It is enough to show that $\mathcal{H}_d(X_K) = 0$. The set X_K can be written $\Psi(\Psi^{-1}(X_K))$, where $\Psi^{-1}(X_K)$ is some compact subset of N. We have $\mathcal{H}_d(\Psi^{-1}(X_K)) = 0$ because N is of dimension $k < d$. Furthermore, as Ψ is analytic on Y, it is Lipschitz on $\Psi^{-1}(X_K)$. Therefore, $\mathcal{H}_d(\Psi(\Psi^{-1}(X_K))) = \mathcal{H}_d(X_K)$ is also null. ◀

B Proof of Theorem 8

▶ **Theorem 8.** *Fix $n \geq 1$. Assume that M is a real analytic compact d-dimensional connected submanifold and that $\mathbb{X}$ is a random variable on M^n having a density with respect to the Haussdorf measure $\mathcal{H}_{dn}$. Define assumption (K5'):*

(K5') The function φ is subanalytic and the gradient of its entries J of size greater than 1 is non vanishing a.s.e.. Moreover, for $\{j\}$ a singleton, $\varphi[\{j\}] \equiv 0$.

Assume that $\mathcal{K}$ satisfies the assumptions (K1)-(K4) and (K5'). Then, for $s \geq 1$, $E[D_s[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on Δ. Moreover, $E[D_0[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on the vertical line $\{0\} \times [0, \infty)$.

We indicate how to change the proof of Theorem 7 when assumption (K5') is satisfied instead of assumption (K5). In the partition $E_1(x), \ldots, E_L(x)$ of $\mathcal{F}_n$, the set $E_1(x)$ plays a special role: it corresponds to the value $r_1 = 0$ and contains all the singletons, which satisfy $\varphi[\{j\}] \equiv 0$ by assumption. Lemma 12 holds for $l > 1$ and one can always define $J_1 = \{1\}$ to be a minimal element of $E_1(x)$. With this convention in mind, it is straightforward to check that Lemma 13 still holds and that Lemma 14 is satisfied as well for $l > 1$. Now, one can define in a likewise manner the sets V_r. For $x \in V_r$, the diagram $D_s[\mathcal{K}(x)]$ is still decomposed $\sum_{i=1}^{N} \delta_{\mathbf{r}_i}$, with $\mathbf{r}_i = (\varphi[J_{l_1}](x), \varphi[J_{l_2}](x))$. If $s > 0$, the end of the proof is similar. However, for $s = 0$, the pairs of simplices (J_{l_1}, J_{l_2}) are made of one singleton J_{l_1} and of one 2-simplex J_{l_2}. As φ is null on singletons, the points in this diagram are all included in the vertical line $L_0 := \{0\} \times [0, \infty)$. The map $\Phi_{ir} : x \in V_r \mapsto \mathbf{r}_i \in L_0$ has a differential of rank 1, as Lemma 14 ensures that $\nabla^j \varphi[J_{l_2}](x) \neq 0$ for $j \in J_{l_2}$. One can apply the coarea formula to Φ_{ir} to conclude to the existence of a density with respect to the Lebesgue measure on L_0.

C Proof of Corollary 9

▶ **Corollary 9.** *Assume that $\mathbb{X}$ has some density with respect to the law of a Poisson process on M of intensity $\mathcal{H}_d$, such that $E\left[2^{|\mathbb{X}|}\right] < \infty$. Assume that $\mathcal{K}$ satisfies the assumptions (K1)-(K5). Then, for $s \geq 0$, $E[D_s[\mathcal{K}(\mathbb{X})]]$ has a density with respect to the Lebesgue measure on Δ.*

The diagram $D_s[\mathcal{K}(\mathbb{X})]$ can be written

$$D_s[\mathcal{K}(\mathbb{X})] = \sum_{n \geq 0} \mathbb{1}\{|\mathbb{X}| = n\} D_s[\mathcal{K}(\mathbb{X})], \tag{24}$$

and Theorem 7 states that $\mathbb{1}\{|\mathbb{X}| = n\}D_s[\mathcal{K}(\mathbb{X})]$ has a density p_n with respect to the Lebesgue measure on Δ. Take B a Borel set in Δ:

$$\begin{aligned} E[D_s[\mathcal{K}(\mathbb{X})]](B) &= \sum_{n\geq 0} E[\mathbb{1}\{|\mathbb{X}| = n\}D_s[\mathcal{K}(\mathbb{X})]](B) \\ &= \sum_{n\geq 0} \int_B p_n = \int_B \sum_{n\geq 0} p_n \text{ by Fubini-Torelli's theorem.} \end{aligned}$$

It is possible to use Fubini-Torelli's theorem because $E[D_s[\mathcal{K}(\mathbb{X})]](B)$ is finite. Indeed, as $D_s[\mathbb{X}]$ is always made of less than $2^{|\mathbb{X}|}$ points, and as we have supposed that $E\left[2^{|\mathbb{X}|}\right] < \infty$, the measure $E[D_s[\mathcal{K}(\mathbb{X})]]$ is finite as well.

D Proof of Theorem 10

▶ **Theorem 10.** *Fix $0 \leq k \leq \infty$ and assume that $\mathbb{X} \in M^n$ has some density of class C^k with respect to $\mathcal{H}_{nd}$. Then, for $s \geq 0$, the density of $E[D_s[\mathcal{K}(\mathbb{X})]]$ is of class C^k.*

Given the expression (11), it is sufficient to show that integrating a function along the fibers is a smooth operation in the fibers. We only show that the density is continuous. Continuity of the higher orders derivatives is obtained in a similar fashion. The proof is a standard application of the implicit function theorem.

Using the same notations than in the proof of Theorem 7, fix $1 \leq r \leq R$ and $1 \leq i \leq N_r$. We will show that p_{ir} is continuous. As the indices r and i are now fixed, we drop the dependency in the notation: $V := V_r$ and $\Phi := \Phi_{ir}$. By using a partition of unity and taking local diffeomorphisms, one can always assume that $V \subset \mathbb{R}^d$. Define the function $f : (x, u) \in V \times \Delta \mapsto \Phi(x) - u \in \mathbb{R}^2$. We have already shown in the proof of Theorem 7 that for $x_0 \in V$, there exists two indices a_1 and a_2 (depending on x_0) such that the minor $M(x_0) = (D\Phi(x_0))_{a_{1,2}}$ is invertible. Rewrite $x \in V$ in (y, z) where $z = (x_{a_1}, x_{a_2}) \in \mathbb{R}^2$. By the implicit function theorem, for (x_0, u_0) such that $f(x_0, u_0) = 0$, there exists a neighborhood $\Omega_{x_0} \subset V \times \Delta$ of (x_0, u_0) and an analytic function $g_{x_0} : W_{y_0} \times Y_{u_0} \to \mathbb{R}^2$ defined on a neighborhood of (y_0, u_0) such that for $(x, u) \in \Omega_{x_0}$

$$f(x, u) = 0 \iff z = g_{x_0}(y, u).$$

The sets $(\Omega_{x_0})_{x_0 \in V}$ constitutes an open cover of the fiber $f^{-1}(0)$. Consider a smooth partition of unity $(\rho_{x_0})_{x_0 \in V}$ subordinate to this cover. Then, for all $(x, u) \in f^{-1}(0)$

$$(J\Phi(x))^{-1}\kappa(x) = \sum_{x_0 \in V} \rho_{x_0}(y, u, g_{x_0}(y, u))(J\Phi(y, g_{x_0}(y, u)))^{-1}\kappa(y, g_{x_0}(y, u))$$

Therefore,

$$\begin{aligned} p_{ir}(u) &= \int_{x \in \Phi^{-1}(u)} (J\Phi(x))^{-1}\kappa(x) d\mathcal{H}_{nd-2}(x) \\ &= \sum_{x_0 \in V} \int_{y \in W_{y_0}} \rho_{x_0}(y, u, g_{x_0}(y, u))(J\Phi(y, g_{x_0}(y, u)))^{-1}\kappa(y, g_{x_0}(y, u)) dy. \end{aligned} \tag{25}$$

We are now faced with a classical continuity under the integral sign problem. First, the Cauchy-Binet formula (see [21, Example 2.15]) states that $J\Phi$ is equal to the square root of the sum of the squares of the determinants of all 2×2 minors of $D\Phi$. Therefore, $J\Phi(x)$ is greater than the determinant of $M(x)$, the minor of f of indices a_1 and a_2. The implicit

function theorem gives the exact value of $M(x)$. Indeed, for $X = (x, u) \in \Omega_{x_0}$, and for any index k,

$$\frac{\partial g}{\partial X_k}(y,u) = -\left(M^{-1} \cdot \frac{\partial f}{\partial X_k}\right)(y,u,g(y,u)) \tag{26}$$

Take $X_k = u_{1,2}$. Then, $\partial f/\partial X_k = (-1,0)$, resp. $(0,-1)$. Therefore,

$$M^{-1}(y,u,g(y,u)) = \frac{\partial g}{\partial u}(y,u,g(y,u)) \tag{27}$$

As ρ_{x_0} has a compact support, it suffices to show that the integrand is bounded by a constant independent of u. The only issue is that $(J\Phi)^{-1}$ may diverge. Equation (27) shows that it is bounded by $\det \partial g/\partial u$. This is bounded, as g is analytic on the compact support of ρ_{x_0}: each term in the sum (25) is continuous. By the compactness of M, all the partitions of unity can be taken finite, and a finite sum of continuous functions is continuous. This proves the continuity of p.

E Proof of Corollary 11

▶ **Corollary 11.** *Under the same hypothesis than Theorem 10, for $s \geq 0$, the expected Betti curve $r \mapsto E[\beta_s^r(\mathcal{K}(\mathbb{X}))]$ is a C^k function.*

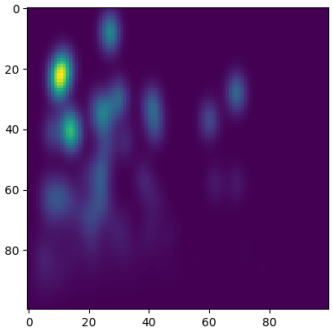
Define $f(r,u)$ to be equal to 1 if $u_1 \leq r \leq u_2$ and 0 otherwise. Then, $\beta_s^r(\mathcal{K}(\mathbb{X}))$ is equal to $D_s[\mathcal{K}(\mathbb{X})](f(r,\cdot))$. Therefore, the expectation $E[\beta_s^r(\mathcal{K}(\mathbb{X}))]$ is equal to

$$\int p(u) f(r,u) du. \tag{28}$$

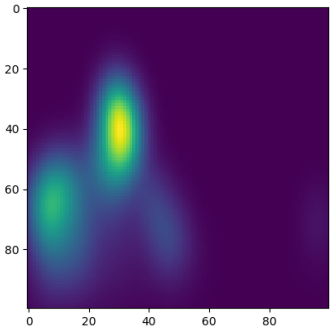
As we assumed that the hypothesis of Theorem 10 were satisfied, the density p is smooth. Moreover, $p(u)f(r,u)$ is smaller than $p(u)$. The function p being integrable, one can apply the continuity under the integral sign theorem to conclude that $r \mapsto E[\beta_s^r(\mathcal{K}(\mathbb{X}))]$ is continuous. Higher-order derivatives are obtained in a similar fashion.

F Bandwidth selection on accelerometer data

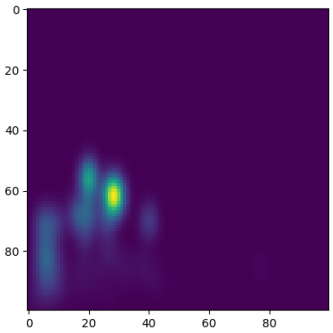
The walk of 3 persons A, B and C, has been recorded using the accelerometer sensor of a smartphone in their pocket, giving rise to 3 multivariate time series in $\mathbb{R}^3$. Using a sliding window, each serie have been splitted in a list of 10 times series made of 200 consecutive points. Using a time-delay embedding technique, those new time series are embedded into $\mathbb{R}^9$: these are the point clouds on which we build the Rips filtration. For each person, the set of 10 persistence diagrams is transformed under the map $(r_1, r_2) \mapsto (r_1, r_2 - r_1)$. The persistence diagrams are weighted by the weight function $w(\mathbf{r}) = (r_2 - r_1)^3$. For each person, the scores $\hat{J}(H)$ are computed for a set of bandwidth matrix of the form $h^2 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, for 20 values h evenly spaced on a log-scale between 10^{-3} and 10^{-1}. The selected bandwidths are $0.0089, 0.01833$ and 0.0089 and the corresponding persistence images are displayed in figure 4. The three images show very distinct patterns: a reasonable machine learning algorithm will easily make the distinction between the three classes using the images as input.



(a)



(b)



(c)

Figure 4 Persistence surfaces for each person A,B and C, computed with the weight function $w(\mathbf{r}) = (r_2 - r_1)^3$ and with the bandwidth matrix selected by the cross-validation procedure.