Subsampling Methods for Persistent Homology

Frédéric Chazal FREDERIC.CHAZAL@INRIA.FR
INRIA Saclay, Palaiseau, 91120, France

Brittany Terese Fasy BRITTANY@FASY.US
Computer Science Department, Tulane University, New Orleans, LA 70118

Fabrizio Lecci LECCI@CMU.EDU
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213

Bertrand Michel BERTRAND.MICHEL@UPMC.FR
LSTA, Université Pierre et Marie Curie (UPMC), Paris, 75005, France

Alessandro Rinaldo ARINALDO@CMU.EDU
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213

Larry Wasserman LARRY@CMU.EDU
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

Persistent homology is a multiscale method for analyzing the shape of sets and functions from point cloud data arising from an unknown distribution supported on those sets. When the size of the sample is large, direct computation of the persistent homology is prohibitive due to the combinatorial nature of the existing algorithms. We propose to compute the persistent homology of several subsamples of the data and then combine the resulting estimates. We study the risk of two estimators and we prove that the subsampling approach carries stable topological information while achieving a great reduction in computational complexity.

1. Introduction

Topological Data Analysis (TDA) refers to a collection of methods for finding topological structure in data (Carlsson, 2009). The input is a dataset drawn from a probability measure supported on an unknown set $\mathbb{X}$. The output is a collection of data summaries that are used to describe the topological features of $\mathbb{X}$. Comparisons between the datasets can then be done on these data summaries.

Homology, or more precisely persistent homology, appears as a fundamental tool for TDA. *Homology* associates to any topological space $\mathbb{X}$, a family of vector spaces (the so-called homology groups) $H_k(\mathbb{X})$, $k = 0, 1, \ldots$, each of them encoding topological features of $\mathbb{X}$. The k^{th} *Betti number* of $\mathbb{X}$, denoted β_k, is the rank of $H_k(\mathbb{X})$ and represents the number of k-dimensional features of $\mathbb{X}$: for example, β_0 is the number of connected components of $\mathbb{M}$, β_1 the number of independent cycles or "tunnels", β_2 the number of "voids", etc. (see Hatcher, 2001). Persistent homology (Edelsbrunner et al., 2002; Zomorodian & Carlsson, 2005) provides a framework and efficient algorithms to encode the evolution of the homology of families of nested topological spaces indexed by a set of real numbers that may often be seen as scales, such as the the union of growing balls, or a nested family of simplicial complexes built on top of the data. The obtained multiscale topological information is then represented in a simple way as a barcode or persistence diagram, providing relevant information about the data (Cohen-Steiner et al., 2007; Chazal et al., 2009; 2012a;b). The persistence diagram can be converted into a summary function called a persistence landscape (Bubenik, 2012); see Section 2. These landscapes are the data summaries that we focus on in this paper.

Contribution and Related Work. The time and space complexity of persistent homology algorithms is one of the main obstacles in applying TDA techniques to high-dimensional problems. To overcome the problem of computational costs, we propose the following strategy: given

Proceedings of the 32^{nd} *International Conference on Machine Learning*, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

a large point cloud, take several subsamples, compute the landscape for each subsample, and then combine the information. More precisely, let λ be a random persistence landscape from Ψ_μ^m, a measure on the space of landscape functions induced by a sample of size m from a metric measure space $(\mathbb{X}, \rho, \mu)$. We show that the average landscape is stable with respect to perturbations of the underlying measure μ in the Wasserstein metric; see Theorem 5. The empirical counterpart of the average landscape is $\overline{\lambda_n^m} = \frac{1}{n}\sum_{i=1}^n \lambda_i$, where $\lambda_1, \ldots, \lambda_n \sim \Psi_\mu^m$. The empirical average landscape can be used as an unbiased estimator of $\mathbb{E}_{\Psi_\mu^m}[\lambda]$ and as a biased estimator of $\lambda_{\mathbb{X}_\mu}$, the computationally expensive persistence landscape associated to the support of the measure μ. Unlike $\lambda_{\mathbb{X}_\mu}$, the estimator $\overline{\lambda_n^m}$ is robust to the presence of outliers. In the same spirit, we propose a different estimator constructed by choosing a sample of m points of $\mathbb{X}$ as close as possible to $\mathbb{X}_\mu$, and then computing its persistent homology to approximate $\lambda_{\mathbb{X}_\mu}$. See Section 3 for more details.

Closely related to our approach, the distribution of persistence diagrams associated to subsamples of fixed size was studied in Blumberg et al. (2014). There, the authors show that the distribution of persistence diagrams associated to subsamples of fixed size is stable with respect to perturbations of the underlying measure in the Gromov-Prohorov metric. Though similar in spirit, our approach relies on different techniques and, in particular, leads to easily computable summaries of the persistent homology of a given space. These summaries are particularly useful when the exact computation of the persistent homology is infeasible, as in the case of large point clouds or when the data are noisy or contain outliers.

Software. The computations in this paper were done using the R package **TDA** (Fasy et al., 2014a). The package includes a series of tools for the statistical analysis of persistent homology, including the methods described in Fasy et al. (2014b), Chazal et al. (2014b), Chazal et al. (2014a), and this paper.

Outline. Background on persistent homology is presented in Section 2. Our approach is introduced in Section 3, with a formal definition of the estimators briefly described in this introduction. Section 4 contains the stability result of the average landscape. Section 5 is devoted to the risk analysis of the proposed estimators. In Section 6, we apply our methods to two examples. We conclude with some remarks in Section 7 and defer proofs and technical details to the appendices.

2. Background

In this section, we briefly introduce the basics of persistence used in this paper. We refer the reader to Edelsbrunner & Harer (2010); Chazal et al. (2012b); Bubenik (2012) for more details.

2.1. Geometric Complexes

To compute the persistent homology from a set of data, we need to construct a set of structures called simplicial complexes. A simplicial complex $\mathcal{C}$ is a set of simplices (points, segments, triangles, etc) such that any face from a simplex in $\mathcal{C}$ is also in $\mathcal{C}$ and the intersection of any two simplices of $\mathcal{C}$ is a (possibly empty) face of these simplices.

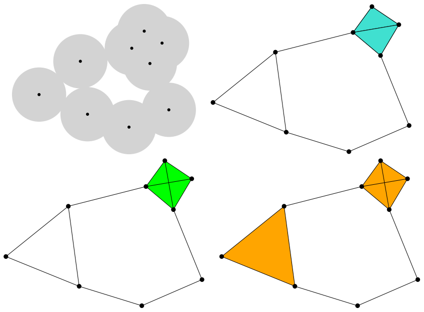

Figure 1. Top left: The α sublevel set of the distance function to a point set $\mathbb{X}$ in $\mathbb{R}^2$. Top right: the α-complex. Bottom left: $\mathrm{Cech}_\alpha(\mathbb{X})$. Bottom right: $\mathrm{Rips}_{2\alpha}(\mathbb{X})$. The last two complexes include a tetrahedron.

Given a metric space $\mathbb{X}$, we define three simplicial complexes whose vertex set is $\mathbb{X}$; see Figure 1 for illustrations. The *Vietoris-Rips complex* $\mathrm{Rips}_\alpha(\mathbb{X})$ is the set of simplices $[x_0, \ldots, x_k]$ such that $d_\mathbb{X}(x_i, x_j) \le \alpha$ for all (i, j). The *Čech complex* $\mathrm{Cech}_\alpha(\mathbb{X})$ is similarly defined as the set of simplices $[x_0, \ldots, x_k]$ such that there exists a point $x \in \mathbb{X}$ for which $d_\mathbb{X}(x, x_i) \le \alpha$ for all i. Note that these two complexes are related by $\mathrm{Rips}_\alpha(\mathbb{X}) \subseteq \mathrm{Cech}_\alpha(\mathbb{X}) \subseteq \mathrm{Rips}_{2\alpha}(\mathbb{X})$ and that their definition does not require $\mathbb{X}$ to be finite. When $\mathbb{X} \subset \mathbb{R}^d$, we also define the *$\alpha$-complex* as the set of simplices $[x_0, \ldots, x_k]$ such that there exists a ball of radius at most α containing $x_0, \ldots, x_k$ on its boundary and whose interior does not intersect $\mathbb{X}$.

Each family described above is non-decreasing with α: for any $\alpha \le \beta$, there is an inclusion of $\mathrm{Rips}_\alpha(\mathbb{X})$ in $\mathrm{Rips}_\beta(\mathbb{X})$, and similarly for the Čech and Alpha complexes. These sequences of inclusions are called *filtrations*. In the following, we let $\mathrm{Filt}(\mathbb{X}) := (\mathrm{Filt}_\alpha(\mathbb{X}))_{\alpha\in\mathcal{A}}$ denote a filtration corresponding to one of the parameterized complexes defined above.

2.2. Persistence Diagrams

The topology of $\mathrm{Filt}_\alpha(\mathbb{X})$ changes as α increases: new connected components can appear, existing connected components can merge, cycles and cavities can appear or be filled, etc. Persistent homology tracks these changes, identifies *features* and associates an *interval* or *lifetime* (from b to d) to them. For instance, a connected component is a feature that is born at the smallest α such that the component is present in $\mathrm{Filt}_\alpha(\mathbb{X})$, and dies when it merges with an older connected component. Intuitively, the longer a feature persists, the more relevant it is. The lifetime of a feature can be represented as a point in the plane with coordinates (b, d). The obtained set of points (with multiplicity) is called the *persistence diagram* $D(\mathrm{Filt}(\mathbb{X}))$ (and we will abuse terminology slightly by denoting it $D_{\mathbb{X}}$). Note that the diagram is entirely contained in the half-plane above the diagonal Δ defined by $y = x$, since death always occurs after birth. Chazal et al. (2012a) shows that this diagram is still well-defined under very weak hypotheses, and in particular $D(\mathrm{Filt}(\mathbb{X}))$ is well-defined for any compact metric space. The most persistent features (supposedly the most important) are those represented by the points furthest from the diagonal in the diagram; whereas, points close to the diagonal can be interpreted as (topological) noise, as they are indistinguishable from features that are born and die at the same value of α.

To avoid (minor) technical difficulties, we restrict our attention to diagrams D such that $(b, d) \in [0, T] \times [0, T]$ for all $(b, d) \in D$, for some fixed $T > 0$. Note that, in our setting, $D_{\mathbb{X}}$ satisfies this property as soon as T is larger than the diameter of $\mathbb{X}$. We denote by $\mathcal{D}_T$ the space of all such (restricted) persistence diagrams and we endow it with a metric called the *bottleneck distance* $\mathrm{d_b}$. Given two persistence diagrams, the bottleneck distance is defined as the infimum of the δ for which we can find a matching between the diagrams, such that two points can only be matched if their distance is less than δ and all points at distance more than δ from the diagonal must be matched.

A fundamental property of persistence diagrams, proven in Chazal et al. (2012a), is their *stability*. Recall that the Hausdorff distance between two compact subsets X, Y of a metric space $(\mathbb{X}, \rho)$ is $H(X, Y) = \max\left\{ \max_{x\in X} \min_{y \in Y} \rho(x, y),\ \max_{y\in Y} \min_{x\in X} \rho(x, y)\right\}$. If $\mathbb{X}$ and $\widetilde{\mathbb{X}}$ are two compact metric spaces, then one has

$$\mathrm{d_b}(D_{\mathbb{X}}, D_{\widetilde{\mathbb{X}}}) \leq 2\mathrm{d_{GH}}(\mathbb{X}, \widetilde{\mathbb{X}}), \tag{1}$$

where $\mathrm{d_{GH}}(\mathbb{X}, \widetilde{\mathbb{X}})$ denotes the Gromov-Hausdorff distance, i.e., the infimum Hausdorff distance between $\mathbb{X}$ and $\widetilde{\mathbb{X}}$ over all possible isometric embeddings into a common metric space. If $\mathbb{X}$ and $\widetilde{\mathbb{X}}$ are already embedded in the same metric space then (1) holds for $H(\cdot, \cdot)$ in place of $\mathrm{d_{GH}}(\cdot, \cdot)$.

2.3. Persistence Landscapes

The persistence landscape, introduced in Bubenik (2012), is a collection of continuous, piecewise linear functions $\lambda \colon \mathbb{Z}^+ \times \mathbb{R} \to \mathbb{R}$ that summarizes a persistence diagram. To define the landscape, consider the set of functions created by tenting each point $p = (x, y) = \left(\frac{b+d}{2}, \frac{d-b}{2}\right)$ representing a birth-death pair $(b, d) \in D$ as follows:

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

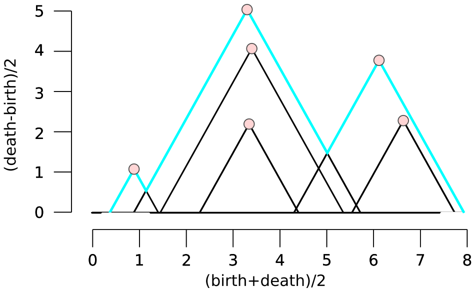

Figure 2. We use the rotated axes to represent a persistence diagram D. A feature $(b, d) \in D$ is represented by the point $(\frac{b+d}{2}, \frac{d-b}{2})$ (pink). In words, the x-coordinate is the average parameter value over which the feature exists, and the y-coordinate is the half-life of the feature. The cyan curve is the landscape $\lambda(1, \cdot)$.

We obtain an arrangement of piecewise linear curves by overlaying the graphs of the functions $\{\Lambda_p\}_p$; see Figure 2. The persistence landscape of D is a summary of this arrangement. Formally, the persistence landscape of D is the collection of functions

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

where kmax is the kth largest value in the set; in particular, 1max is the usual maximum function. We set $\lambda_D(k, t) = 0$ if the set $\{\Lambda_p(t)\}_p$ contains less than k points. For simplicity of exposition, if $D_{\mathbb{X}}$ is the persistence diagram of some metric space $\mathbb{X}$, then we use $\lambda_{\mathbb{X}}$ to denote $\lambda_{D_{\mathbb{X}}}$.

We denote by $\mathcal{L}_T$ the space of persistence landscapes corresponding to $\mathcal{D}_T$. From the definition of persistence landscape, we immediately observe that $\lambda_D(k, \cdot)$ is one-Lipschitz. The following additional properties are proven in Bubenik (2012).

Lemma 1. *Let D, D' be persistence diagrams. We have the following for any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$:*
(i) $\lambda_D(k, t) \geq \lambda_D(k + 1, t) \geq 0$.
(ii) $|\lambda_D(k, t) - \lambda_{D'}(k, t)| \leq \mathrm{d_b}(D, D')$.

For ease of exposition, we focus on the case $k = 1$, and set $\lambda_D(t) = \lambda_D(1, t)$. However, the results we present hold for $k > 1$. In fact, the results hold for more general summaries of persistence landscapes, including the silhouette defined in Chazal et al. (2014b).

3. The Multiple Samples Approach

Let $(\mathbb{X}, \rho)$ be a metric space of diameter at most $T/2$ and let $\mathcal{P}(\mathbb{X})$ be the space of probability measures on $\mathbb{X}$, such that, for any measure $\mu \in \mathcal{P}(\mathbb{X})$, its support $\mathbb{X}_\mu$ is a compact set. The space $\mathbb{X}_\mu$ is a natural object of interest in computational topology. Its persistent homology is usually approximated by the persistent homology of the distance function to a sample $X_N = \{x_1, \dots, x_N\} \subset \mathbb{X}_\mu$. Fasy et al. (2014b) propose several methods for the construction of confidence sets for the persistence diagram of $\mathbb{X}_\mu$, while Chazal et al. (2014c) establish optimal convergence rates for $d_b(D_{\mathbb{X}_\mu}, D_{X_N})$.

When N is too large, the computation of the persistent homology of X_N is prohibitive, due to the combinatorial complexity of the computation. Our aim is to study topological signatures of the data that can be efficiently computed in a reasonable time. We define such quantities by repeatedly sampling m points of $\mathbb{X}$ according to μ.

For any positive integer m, let $X = \{x_1, \cdots, x_m\} \subset \mathbb{X}$ be a sample of m points from the measure $\mu \in \mathcal{P}(\mathbb{X})$. The corresponding persistence landscape is λ_X and we denote by Ψ_μ^m the measure induced by $\mu^{\otimes m}$ on $\mathcal{L}_\mathcal{T}$. Note that the persistence landscape λ_X can be seen as a single draw from the measure Ψ_μ^m. We consider the point-wise expectations of the (random) persistence landscape under this measure: $\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)], t \in [0, T]$. The main result of this paper establishes the stability of this quantity under perturbation of μ, making it relevant from a topological point of view (see next section).

The average landscape $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$ has a natural empirical counterpart, which can be used as its unbiased estimator. Let $S_1^m, \dots, S_n^m$ be n independent samples of size m from μ. We define the empirical average landscape as

$$\overline{\lambda_n^m}(t) = \frac{1}{n} \sum_{i=1}^{n} \lambda_{S_i^m}(t), \qquad \text{for all } t \in [0, T], \tag{4}$$

and propose to use $\overline{\lambda_n^m}$ to estimate $\lambda_{\mathbb{X}_\mu}$. The variance of this estimator under the ℓ_∞-distance was studied in detail in Chazal et al. (2014b). Here instead we are concerned with the quantity $\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}_{\Psi_\mu^m}[\lambda_X]\|_\infty$, which can be seen as the bias component (see Section 5).

In addition to the average, we also consider using the *closest sample* to $\mathbb{X}_\mu$ in Hausdorff distance. The closest sample method consists in choosing a sample of m points of $\mathbb{X}$, as close as possible to $\mathbb{X}_\mu$, and then use this sample to build a landscape that approximates $\lambda_{\mathbb{X}_\mu}$. Let $S_1^m, \dots, S_n^m$ be n independent samples of size m from $\mu^{\otimes m}$. The closest sample is

$$\widehat{C_n^m} = \arg \min_{S \in \{S_1^m, \dots, S_n^m\}} H(S, X_\mu) \tag{5}$$

and the corresponding landscape function is $\widehat{\lambda_n^m} = \lambda_{\widehat{C_n^m}}$. Of course, the method requires the support of μ to be a known quantity.

Remark 2. *Computing the persistent homology of X_N is $O(\exp(N))$, whereas computing the average landscape is $O(n\exp(m))$ and the persistent homology of the closest sample is $O(nmN + \exp(m))$.*

Remark 3. *The general framework described above is valid for the case in which μ is a discrete measure with support $\mathbb{X}_\mu = \{x_1, \dots, x_N\} \subset \mathbb{R}^D$. For example, the following situation is very common in practice. Let $X_N = \{x_1, \dots, x_N\}$ be a given point cloud, for large but fixed $N \in \mathbb{N}$. When N is large, the computation of the persistent homology of X_N is infeasible. Instead, we consider the discrete uniform measure μ that puts mass $1/N$ on each point of X_N, and we propose to estimate $\lambda_{\mathbb{X}_u}$ by repeatedly subsampling $m \ll N$ points of X_N according to μ.*

We study the ℓ_∞-risk of the proposed estimators, $\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \overline{\lambda_n^m}\|_\infty\right]$ and $\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \widehat{\lambda_n^m}\|_\infty\right]$, under the following assumption on the underlying measure μ, which we refer to as the (a, b, r_0)-*standard assumption*: there exist positive constants a, b and $r_0 \geq 0$ such that

$$\forall r > r_0, \forall x \in \mathbb{X}_\mu \,, \; \mu(B(x, r)) \geq 1 \wedge ar^b. \tag{6}$$

For $r_0 = 0$, this is known as the (a, b)-standard assumption and has been widely used in the literature of set estimation under Hausdorff distance (Cuevas & Rodríguez-Casal, 2004; Cuevas, 2009; Singh et al., 2009) and more recently in the statistical analysis of persistence diagrams (Chazal et al., 2014c; Fasy et al., 2014b). We use the generalized version with $r_0 > 0$ to take into account the case in which μ is a discrete measure (in which case r_0 depends on N); see Appendix C for more details.

4. Stability of the Average Landscape

Consider the framework described in Section 3: m points are repeatedly sampled from the space $\mathbb{X}$ according to a measure $\mu \in \mathcal{P}(\mathbb{X})$. In this section, we show that the average landscape $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$ is an interesting quantity on its own, since it carries some stable topological information about the underlying measure μ, from which the data are generated.

Chazal et al. (2014b) provide a way to construct confidence bands for $\mathbb{E}_{\Psi_\mu^m}[\lambda_X]$. Here, we compare the average landscapes corresponding to two measures that are close to each other in the Wasserstein metric.

Definition 4. Given a metric space $(\mathbb{X}, \rho)$, the pth Wasserstein distance between two measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$ is $W_{\rho,p}(\mu,\nu) = \left(\inf_\Pi \int_{\mathbb{X}\times\mathbb{X}} [\rho(x,y)]^p d\Pi(x,y)\right)^{\frac{1}{p}}$, where the infimum is taken over all measures on $\mathbb{X} \times \mathbb{X}$ with marginals μ and ν.

Next, we show that the average behavior of the landscapes of sets of m points sampled according to any measure μ is stable with respect to the Wasserstein distance.

Theorem 5. *Let $(\mathbb{X}, \rho)$ be a metric space of diameter bounded by $T/2$. Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where $\mu, \nu \in \mathcal{P}(\mathbb{X})$ are two probability measures. For any $p \geq 1$ we have*

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

Remark 6. *For measures that are not defined on the same metric space, the inequality of Theorem 5 can be extended to Gromov-Wasserstein metric: $\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq 2 m^{\frac{1}{p}} GW_{\rho,p}(\mu,\nu)$.*

The result of Theorem 5 is useful for two reasons. First, it tells us that for a fixed m, the expected "topological behavior" of a set of m points carries some stable information about the underlying measure from which the data are generated. Second, it provides a lower bound for the Wasserstein distance between two measures, based on the topological signature of samples of m points.

The dependence on m of the upper bound of Theorem 5 seems to be necessary in this setting: intuitively, when m grows, the samples of m points converge to the support of μ and ν w.r.t. the Hausdorff distance. Therefore the expected landscapes should converge to the landscapes of the support of the measures. But, in general, two measures that are close in the Wasserstein metric can have support that have very different and unrelated topologies. Indeed, a similar dependence was also obtained in Blumberg et al. (2014) when considering the Gromov-Prohorov metric.

Note that in Theorem 5 we do not make any assumption on the measures μ and ν. If we assume that they both satisfy the (a, b, r_0)-standard assumption we can provide a different bound on the difference of the expected landscapes, based on the Hausdorff distance between the support of the two measures.

Theorem 7. *Let $(\mathbb{X}, \rho)$ be a metric space of diameter bounded by $T/2$. Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where $\mu, \nu \in \mathcal{P}(\mathbb{X})$ satisfy the (a, b, r_0)-standard assumption on $\mathbb{X}$. Define $r_m = 2\left(\frac{\log m}{am}\right)^{1/b}$. Then*

$$\begin{aligned} \|\mathbb{E}_{\Psi_\mu^m}(\lambda_X) - \mathbb{E}_{\Psi_\nu^m}(\lambda_Y)\|_\infty \leq 2H(\mathbb{X}_\mu, \mathbb{X}_\nu) + 4r_0 + \\ + 4r_m \mathbb{1}_{(r_0,\infty)}(r_m) + 4\, C_1(a,b)\, r_m \frac{1}{(\log m)^2}, \end{aligned}$$

where $C_1(a,b)$ is a constant depending on a and b.

The following result follows from Theorems 5 and 7.

Corollary 8. *Under the same assumptions of Theorem 7, we have that*

$$\begin{aligned} \|\mathbb{E}_{\Psi_\mu^m}(\lambda_X) - \mathbb{E}_{\Psi_\nu^m}(\lambda_Y)\|_\infty \leq 2 \min \Big\{ m^{\frac{1}{p}} W_p(\mu,\nu), \\ H(\mathbb{X}_\mu, \mathbb{X}_\nu) + 2r_0 + 2r_m \mathbb{1}_{(r_0,\infty)}(r_m) + \\ + 2\, C_1(a,b)\, r_m \frac{1}{(\log m)^2} \Big\}. \end{aligned}$$

5. Risk Analysis

In this section, we study the performance of the average landscape $\overline{\lambda_n^m}$ and of the landscape of the closest sample $\widehat{\lambda_n^m}$, as estimators of $\lambda_{\mathbb{X}_\mu}$. We start by decomposing the ℓ_∞-risk of the average landscape as follows. Set $\lambda_1 = \lambda_{S_1^m}$, with S_1^m a sample of size m from μ. Then,

$$\mathbb{E}\left\|\lambda_{\mathbb{X}_\mu} - \overline{\lambda_n^m}\right\|_\infty \leq \left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty + \mathbb{E}\left\|\overline{\lambda_n^m} - \mathbb{E}\lambda_1\right\|_\infty, \tag{7}$$

where the expectation of $\overline{\lambda_n^m}$ is w.r.t. $(\Psi_\mu^m)^{\otimes n}$ and the expectation of λ_1 is w.r.t. Ψ_μ^m.

For the bias term $\left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty$ we use the stability property to go back into $\mathbb{R}^d$:

$$\left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty \leq \mathbb{E}_{\Psi_\mu^m}\left\|\lambda_{\mathbb{X}_\mu} - \lambda_1\right\|_\infty \leq 2\mathbb{E}_{\mu^{\otimes m}} H(\mathbb{X}_\mu, X), \tag{8}$$

where X is a sample of size m from μ. Note that, if calculating $H(\mathbb{X}_\mu, X)$ is computationally feasible, then, in practice, $\mathbb{E}_{\mu^{\otimes m}} H(\mathbb{X}_\mu, X)$ can be approximated by the average of a large number B of values of $H(\mathbb{X}_\mu, X)$, for B different draws of subsamples $X \sim \mu^{\otimes m}$.

To give an explicit bound on the bias, we assume that μ satisfies the (a, b, r_0)-standard assumption.

Theorem 9. *Let $r_m = 2\left(\frac{\log m}{am}\right)^{1/b}$. If μ satisfies the (a, b, r_0)-standard assumption, then*

$$\begin{aligned} \left\|\lambda_{\mathbb{X}_\mu} - \mathbb{E}\lambda_1\right\|_\infty \leq & 2r_0 + 2r_m \mathbb{1}_{(r_0,\infty)}(r_m) + \\ & + 2C_1(a,b)\, r_m \frac{1}{(\log m)^2}, \end{aligned}$$

where $C_1(a,b)$ is a constant that depends on a and b.

Chazal et al. (2014b) control the variance term, which is of the order of $1/\sqrt{n}$. Therefore, if r_0 is negligible, we see that n should be taken of the order of $(m/\log m)^{2/b}$.

We now turn to the closest sample estimator $\widehat{\lambda}_n$ and investigate its ℓ_∞ risk $\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \widehat{\lambda_n^m}\|_\infty\right]$, where the expectation is with respect to $(\Psi_\mu^m)^{\otimes n}$. As before, in our analysis, we rely on the stability property $\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \widehat{\lambda_n^m}\|_\infty\right] \leq$

$2\mathbb{E}\left[H(\mathbb{X}_\mu, \widehat{C_n^m})\right]$, where the second expectation is with respect to $(\mu^{\otimes m})^{\otimes n}$.

Theorem 10. *Let $r_m = 2\Big(\frac{\log(2^b m)}{am}\Big)^{\frac{1}{b}}$. If $\mu \in \mathcal{P}(\mathbb{X})$ satisfies the (a, b, r_0)-standard assumption, then*

$$\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \widehat{\lambda_n^m}\|_\infty\right] \leq 2r_0 \;+\; 2r_m \mathbb{1}_{(r_0,\infty)}(r_m) + \\ + 2C_2(a,b)\, r_m \, \frac{1}{n\,[\log(2^b m)]^{n+1}},$$

where $C_2(a, b)$ is a constant that depends on a and b.

Remark 11. *The risk of the closest subsample method can in principle be smaller than the average landscape method. In Appendix C, we show that if μ is the discrete uniform measure on a point cloud of size N, sampled from a measure satisfying the $(a, b, 0)$-standard assumption, then r_0 is of the order of $(\frac{\log N}{N})^{1/b}$. When r_0 is negligible, the rates of theorems 9 and 10 are comparable, both of the order of $O(\frac{\log m}{m})^{1/b}$. However, the average method has another advantage: it is robust to outliers. This point is discussed in detail in Appendix D.*

6. Experiments

Since computing the persistent homology of the Vietoris-Rips (VR) filtrations built on top of a large samples is infeasible, we resort to the subsampling strategy described in Section 3. More formally, let $X_N = \{x_1, \ldots, x_N\}$ be a large point cloud. We draw n subsamples, each of size $m \ll N$ points, from μ, the discrete uniform measure on X_N. First, we use a toy example to compare the time complexity of computing the persistent homology of the entire point cloud, with the complexity of the subsampling approach.

Example 12 (Toy)**.** *Let X_N be the sample of $N = 500$ points depicted in the left plot of Figure 3. The VR filtration built on top of the sample consists of 20,833,750 simplices and computing the persistence diagram, and hence the 1st persistence landscape of one-dimensional features in the middle plot, required 28.34 seconds on a Macbook Pro with 2.8 GHz processor and 16 GB RAM. The average landscape on the right plot is computed using $n = 10$ subsamples of size $m = 100$, each resulting in a VR filtration of 166,750 simplices, whose persistent homology was computed on average in 0.14 seconds. The 95% confidence band for the true average landscape is constructed using the multiplier bootstrap described in Chazal et al. (2014b). Both landscapes (the landscape of the full sample and the average landscape) show two peaks, corresponding to the two loops of the circles from which the data were sampled. However computing the the average landscape was 20 times faster.*

In each of the following two examples, we consider four point clouds and compare the corresponding average landscapes and closest subsample landscapes, induced by the persistent homology of the VR filtrations built on top of the subsamples.

Example 13 (3D Shapes)**.** *We use the publicly available database of triangulated shapes (Sumner & Popović, 2004). We select a single pose (#2) of four different classes: camel, elephant, flamingo, lion. The four shapes are represented in Figure 4. In practice, each shape consists of a 3D point cloud embedded in Euclidean space, with a number of vertices that ranges from $7K$ to $40K$. The data are normalized, so that the diameter of each shape is one. For $n = 100$ times we subsample $m = 300$ points from each shape, then we select the closest subsample to the corresponding original point cloud and compute $4 \times n$ persistence diagrams (dimension one), one for each subsample. See Figure 5: the plot on the left shows the landscapes corresponding to the closest subsamples of m points among the n different subsamples from each shape; the plot in the middle shows the empirical average landscapes within each class, computed as the pointwise average of n landscapes, with a 95% uniform confidence band for the true average landscape, constructed using the method described in Chazal et al. (2014b); the dissimilarity matrix on the right shows the pairwise ℓ_∞ distances between the average landscapes (scale from yellow to red), which, according to Theorem 5, represent a lower bound for the pairwise Wasserstein distances of the discrete uniform measures on the four different shapes.*

Example 14 (Magnetometer Data)**.** *For the second example, we consider the problem of distinguishing human activities performed while wearing inertial and magnetic sensor units. The dataset is publicly available at the UCI Machine Learning Repository*[1] *and is described in Barshan & Yüksek (2013), where it is used to classify 19 activities performed by eight people wearing sensor units on the chest, arms, and legs. For ease of illustration, we report here the results on four activities (walking, stepper, cross trainer, jumping) performed by a single person (#1). We use the data from the magnetometer of a single sensor (left leg), which measures the direction of the magnetic field in the space at a frequency of 25Hz. For each activity there are 7,500 consecutive measurements that we treat as a 3D point cloud in the Euclidean space. As an example, Figure 4 shows 500 points at random for two activities (walking and using a cross trainer). As in the previous example, for $n = 80$ times, we subsample $m = 200$ points from the point cloud of each activity, then construct the landscapes of the closest subsamples, the average landscapes (dimension one), and the dissimilarity matrix based on the ℓ_∞ distances of the average landscapes. See Figure 6. To the*

[1] http://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities

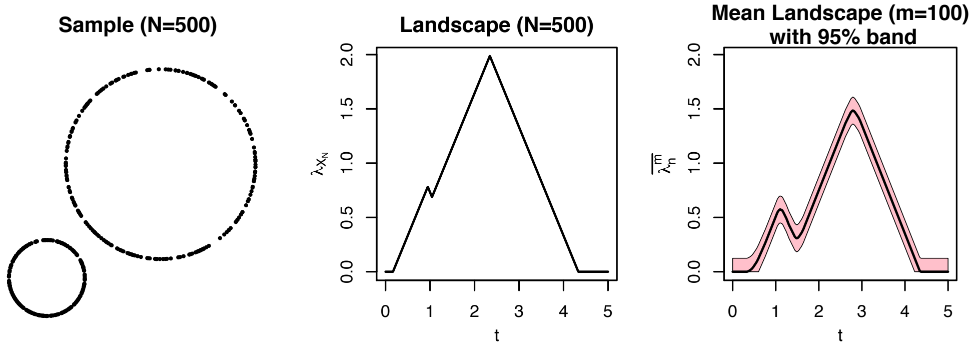

Figure 3. Left: 500 points from two circles of different radii. Middle: 1st landscape of one-dimensional features using the entire sample. Right: average landscape with 95% confidence band, constructed using $n = 10$ subsamples, each of size $m = 100$.

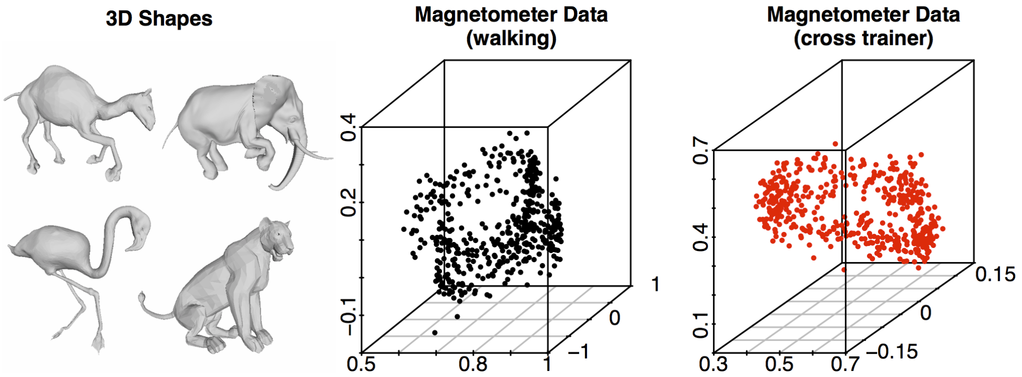

Figure 4. Left: Four 3D shapes. Middle and Left: 500 random points from the magnetometer data of the second experiment.

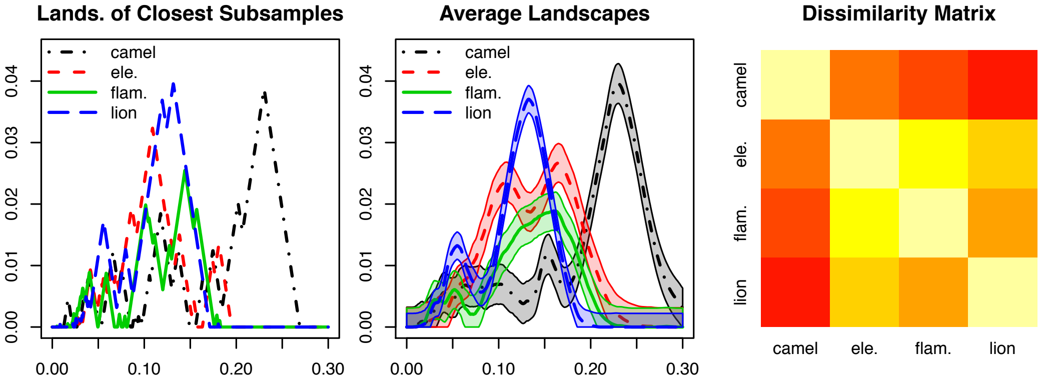

Figure 5. Subsampling methods applied to 3D shapes. For $n = 100$ subsamples of size $m = 300$, for each shape, we constructed the landscapes of the closest subsample (left), the average landscape with 95% confidence band (middle) and the dissimilarity matrix of the pairwise ℓ_∞ distance between average landscapes.

best of our knowledge, persistent homology has never been used to study data from accelerometers or magnetometers before. A remarkable advantage is that the methods of persistent homology are insensitive to the orientation of the input data, as opposed to other methods that require the exact calibration of the sensor units; see, for example, Altun et al. (2010) and Barshan & Yüksek (2013).

7. Conclusion

We presented a framework for approximating the persistent homology of a set using subsamples. The method is simple and computationally fast. Moreover, we provided stability results for the new summaries and bounds on the risk of the proposed estimators. In the future, we plan to investigate methods for further speeding up the computations.

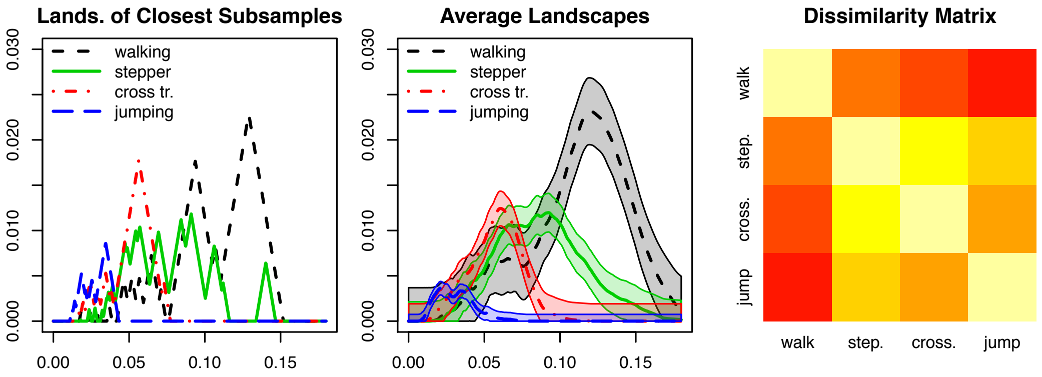

Figure 6. Subsampling methods applied to magnetometer data. For $n = 80$ subsamples of size $m = 200$, for each activity, we constructed the landscapes of the closest subsample (left), the average landscape with 95% confidence band (middle) and the dissimilarity matrix of the pairwise ℓ_∞ distance between average landscapes.

Acknowledgments

Research supported by ANR-13-BS01-0008, NSF CAREER Grant DMS 1149677, Air Force Grant FA95500910373 and NSF Grant DMS-0806009.

References

Alexander, Kenneth S. Rates of growth for weighted empirical processes. In *Proc. of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pp. 475–493, 1985.

Alexander, Kenneth S.˙ The central limit theorem for weighted empirical processes indexed by sets. *J. Multivar. Anal.*, 22(2):313–339, 1987a.

Alexander, Kenneth S.˙ Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probab. Theory Related Fields*, 75(3):379–423, 1987b.

Altun, Kerem, Barshan, Billur, and Tunçel, Orkun. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

Barshan, Billur and Yüksek, Murat Cihan. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal*, pp. bxt075, 2013.

Blumberg, Andrew J. Gal, Itamar, Mandell, Michael A. and Pancia, Matthew. Persistent homology for metric measure spaces, and robust statistics for hypothesis testing and confidence intervals. *Found. Comput. Math.*, pp. 1–45, May 2014.

Bubenik, Peter. Statistical topological data analysis using persistence landscapes. *arXiv preprint 1207.6437*, 2012.

Carlsson, Gunnar. Topology and data. *Bull. Amer. Math. Soc.*, 46(2):255–308, 2009.

Chazal, F., Fasy, B.T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Robust topological inference: Distance-to-a-measure and kernel distance. *arXiv preprint arXiv: 1412.7197*, 2014a.

Chazal, Frédéric, Cohen-Steiner, David, Glisse, Marc, Guibas, Leonidas J. and Oudot, Steve Y. Proximity of persistence modules and their diagrams. In *Proc. 25th Annu. Symp. Comp. Geom*, pp. 237–246. ACM, 2009.

Chazal, Frédéric, de Silva, Vin, Glisse, Marc, and Oudot, Steve. The structure and stability of persistence modules. *arXiv preprint 1207.3674*, 2012a.

Chazal, Frédéric, De Silva, Vin, and Oudot, Steve. Persistence stability for geometric complexes. *Geom. Dedicata*, pp. 1–22, 2012b.

Chazal, Frédéric, Fasy, Brittany Terese, Lecci, Fabrizio, Rinaldo, Alessandro, and Wasserman, Larry. Stochastic convergence of persistence landscapes and silhouettes. In *Proc. 30th Annu. Sympos. Comput. Geom*, 2014b.

Chazal, Frédéric, Glisse, Marc, Labruère, Catherine, and Michel, Bertrand. Convergence rates for persistence diagram estimation in topological data analysis. In *Proc. 31st Int. Conf. Mach. Learn.*, volume 32, pp. 10–18. JMLR W&CP, 2014c. arXiv preprint 1305.6239.

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

Cuevas, Antonio. Set estimation: another bridge between statistics and geometry. *Bol. Estad. Investig. Oper.*, 25 (2):71–85, 2009. ISSN 1889-3805.

Cuevas, Antonio and Rodríguez-Casal, Alberto. On boundary estimation. *Advances in Applied Probability*, pp. 340–354, 2004.

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

Edelsbrunner, Herbert and Harer, John. *Computational Topology: An Introduction*. AMS, 2010.

Fasy, Brittany Terese, Kim, Jisu, Lecci, Fabrizio, and Maria, Clement. Introduction to the R package TDA. *arXiv preprint arXiv: 1411.1830*, 2014a.

Fasy, Brittany Terese, Lecci, Fabrizio, Rinaldo, Alessandro, Wasserman, Larry, Balakrishnan, Sivaraman, and Singh, Aarti. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014b.

Giné, Evarist and Koltchinskii, Vladimir. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216, 2006. ISSN 0091-1798. doi: 10.1214/009117906000000070. URL http://dx.doi.org/10.1214/009117906000000070.

Giné, Evarist, Koltchinskii, Vladimir, and Wellner, Jon A. Ratio limit theorems for empirical processes. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pp. 249–278. Birkhäuser, Basel, 2003.

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

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

Sumner, Robert W and Popović, Jovan. Deformation transfer for triangle meshes. In *ACM Transactions on Graphics (TOG)*, volume 23, pp. 399–405. ACM, 2004.

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