STATISTICAL TOPOLOGICAL DATA ANALYSIS USING PERSISTENCE LANDSCAPES

PETER BUBENIK

Abstract. We define a new topological summary for data that we call the persistence landscape. In contrast to the standard topological summaries, the barcode and the persistence diagram, it is easy to combine with statistical analysis, and its associated computations are much faster. This summary obeys a Strong Law of Large Numbers and a Central Limit Theorem. Under certain finiteness conditions, this allows us to calculate approximate confidence intervals for the expected total squared persistence. With these results one can use t-tests for statistical inference in topological data analysis. We apply these methods to numerous examples including random geometric complexes, random clique complexes, and Gaussian random fields. We also show that this summary is stable and gives lower bounds for the bottleneck distance and the Wasserstein distance.

1. Introduction

Topological data analysis (TDA) consists of a growing set of methods that provide insight to the “shape” of data [29, 12]. These tools may be of particular use in understanding global features of high dimensional data that are not readily accessible using other techniques. The use of TDA has been limited by the difficulty of combining the main tool of the subject, the barcode or persistence diagram with statistics. Here we present an alternative approach, using a new summary that we call the persistence landscape. The main technical advantage of this descriptor is that it is a function and so we can utilize the vector space structure of its underlying function space. In fact, this function space is a separable Banach space and we apply the theory of random variables with values in such spaces. Furthermore, since the persistence landscapes are piecewise-linear functions, calculations with them are much faster than the corresponding calculations with barcodes or persistence diagrams, removing a second serious obstruction to the wider use of topological methods in data analysis.

Notable successes of TDA include the discovery of a subgroup of breast cancers [42], an understanding of the topology of the space of natural images [13] and the topology of orthodontic data [31], and the detection of genes with a periodic profile [25].

In the standard paradigm for TDA, one starts with data which one encodes as a finite set of points in \( \mathbb{R}^n \) or more generally in some metric space. Then one applies some geometric construction to which one applies tools from algebraic topology. The end result is a topological summary of the data. The standard topological descriptors are the barcode and the persistence diagram [28, 17], which give a multiscale representation of the homology [30].

Date: December 6, 2013.
of the geometric construction. Roughly, homology in degree 0 describes the connectedness of the data; homology in degree 1 detects holes or tunnels; homology in degree 2 captures voids; and so on. Of particular interest are the homological features that persist as the resolution changes. We give precise definitions and an illustrative example of this method, called persistent homology or topological persistence, in Section 2.

Now let us take a statistical view of this paradigm. We consider the data to be sampled from some underlying abstract probability space. Composing the constructions above, we consider our topological summary to be a random variable with values in some summary space $S$. In detail, the probability space $(\Omega, \mathcal{F}, P)$ consists of a sample space $\Omega$, a $\sigma$-algebra $\mathcal{F}$ of events, and a probability measure $P$. Composing our constructions gives a function $X : (\Omega, \mathcal{F}, P) \to (S, \mathcal{A}, P_*)$, where $S$ is the summary space which we assume has some metric, $\mathcal{A}$ is the corresponding Borel $\sigma$-algebra, and $P_*$ is the probability measure on $S$ obtained by pushing forward $P$ along $X$. We assume that $X$ is measurable and thus $X$ is a random variable with values in $S$.

Here is a list of what we would like to be able to do with our topological summary. Let $X_1, \ldots, X_n$ be a sample of independent, identically distributed random variables with the same distribution as $X$. We would like to have a good notion of the mean $\mu$ of $X$, and the empirical mean $\overline{X}_n$ of the sample. We would like be able to calculate $\overline{X}_n$ efficiently and prove that it converges to $\mu$. We would like to understand the difference $\overline{X}_n - \mu$, and be able to calculate confidence intervals related to $\mu$. Given two such samples for random variables $X$ and $Y$ with values in our summary space, we would like to be able to test the hypothesis that $\mu_X = \mu_Y$. In order to answer these questions we also need an efficient algorithm for calculating distances between elements of our sample space.

In this article, we construct a topological summary that we call the persistence landscape which satisfies these requirements.

Progress towards meeting these requirements has been made for the topological summaries the persistence diagram and the barcode [37, 45, 40, 17, 3]. A related statistical approach to TDA is given in [5]. The persistence landscape defined here has been used to study the maltose binding complex [35] and the bootstrap has been applied to persistence landscapes [16]. The persistence landscape is related to the previously defined well group [4]. Algorithms for persistence landscapes are given in [10].

2. Topological summaries

The two standard topological summaries of data are the barcode and the persistence diagram. We define a new closely-related topological summary, the persistence landscape.

2.1. Persistence modules. The main algebraic object of study in topological data analysis is the persistence module, which can be constructed from a finite set of points.

A persistence module $M$ consists of a vector space $M_a$ for all $a \in \mathbb{R}$ and linear maps $M(a \leq b) : M_a \to M_b$ for all $a \leq b$ such that $M(a \leq a)$ is the identity map and for all $a \leq b \leq c$,
\[ M(b \leq c) \circ M(a \leq b) = M(a \leq c). \] We will assume throughout that all of our vector spaces are finite dimensional.

Consider a set of points \( X = \{x_1, \ldots, x_n\} \) in the plane \( M = \mathbb{R}^2 \) as shown in the top left of Figure 1. To help understand this configuration, we “thicken” each point, by replacing each point, \( x_i \), with \( B_r(x_i) = \{ y \in M \mid d(x,y) \leq r \} \), a disk of fixed radius, \( r \), centered at \( x_i \). The resulting union, \( X_r = \bigcup_{i=1}^{n} B_r(x_i) \) for various values of \( r \) is shown in Figure 1. For each value of \( r \), we can calculate \( H(X_r) \), the homology of the resulting union of disks. To be precise, \( H(\cdot) \) denotes \( H_k(\cdot, \mathbb{F}) \), the singular homology functor in degree \( k \) with coefficients in a field \( \mathbb{F} \). So \( H(X_r) \) is a vector space that is the quotient of the \( k \)-cycles modulo those that are boundaries. As \( r \) increases, the union of disks grows, and the resulting inclusions induce maps between the corresponding homology groups. More precisely, if \( r \leq s \), the inclusion \( \iota_r^s : X_r \hookrightarrow X_s \) induces a map \( H(\iota_r^s) : H(X_r) \to H(X_s) \). The images of these maps are the persistent homology groups. The collection of vector spaces \( H(X_r) \) and linear maps \( H(\iota_r^s) \) is a persistence module.

Note that this construction works for any set of points in \( \mathbb{R}^n \) or more generally in a metric space.

The union of balls \( X_r \) has a nice combinatorial description. The \( \check{\text{C}}ech \) complex, \( \check{\mathcal{C}}_r(X) \), of the set of balls \( \{ B_x(r) \} \) is a simplicial complex whose vertices are the points \( \{ x_i \} \) and whose \( k \)-simplices correspond to \( k+1 \) balls with nonempty intersection (see Figure 1). This is also called the nerve. It is a basic result that if the ambient space is \( \mathbb{R}^n \), \( X_r \) is homotopy equivalent to its \( \check{\text{C}}ech \) complex \([8]\). So to obtain the singular homology of the union of balls, one can calculate the simplicial homology of the corresponding \( \check{\text{C}}ech \) complex. The \( \check{\text{C}}ech \) complexes \( \{ \check{\mathcal{C}}_r(X) \} \) together with the inclusions \( \check{\mathcal{C}}_r(X) \subseteq \check{\mathcal{C}}_s(X) \) for \( r \leq s \) form a filtered simplicial complex. Applying simplicial homology we obtain a persistence module. There exist efficient algorithms for calculating the persistent homology of filtered simplicial complexes \([28, 38, 18]\).

Remark. The \( \check{\text{C}}ech \) complex is often computationally expensive, so many variants have been used in computational topology. A larger, but simpler complex called the Vietoris-Rips complex is given by the flag complex of the 1-skeleton of the \( \check{\text{C}}ech \) complex. Other possibilities include witness complexes \([24]\), graph induced complexes \([26]\) and complexes built using kernel density estimators and triangulations of the ambient space \([9]\). Some of these are used in Section 4.

Given any real-valued function \( f : S \to \mathbb{R} \) on a topological space \( S \), we can define the associate a persistence module, \( M(f) \), where \( M(f)(a) = H(f^{-1}((\infty, a])) \) and \( M(f)(a \leq b) \) is induced by inclusion.

2.2. Persistence landscapes. In this section we define a number of versions of what we call the persistence landscape.
Let $M$ be a persistence module. For $a \leq b$, the corresponding \textit{Betti number} of $M$, is given by the dimension of the image of the corresponding linear map. That is,

$$\beta^{a,b} = \dim(\text{im}(M(a \leq b))).$$

The simplest version of the persistence landscape, which we call the \textit{rank function} is the function $\lambda : \mathbb{R}^2 \to \mathbb{R}$ given by

$$\lambda(b, d) = \begin{cases} 
\beta^{b,d} & \text{if } b \leq d \\
0 & \text{otherwise}.
\end{cases}$$

For an example see Figure 2.
Figure 2. The rank function (top left) and rescaled rank function (top right) for the homology in degree 1 of the example in Figure 1. The values of the functions on the corresponding region are given. The top left graph also contains the points of the corresponding persistence diagram. Below the top right graph is the corresponding barcode. We also have the corresponding persistence landscape (bottom left) and its 3d-version (bottom right). Notice that $\lambda_1$ gives a measure of the dominant homological feature at each point of the filtration.

Now let us change coordinates so that the resulting function is supported on the upper half plane. Let

$$m = \frac{b + d}{2}, \quad \text{and} \quad h = \frac{d - b}{2}. \tag{2.1}$$

The rescaled rank function is the function $\lambda : \mathbb{R}^2 \to \mathbb{R}$ given by

$$\lambda(m, h) = \begin{cases} 
\beta^{m-h, m+h} & \text{if } h \geq 0 \\
0 & \text{otherwise.}
\end{cases}$$

For an example see Figure 2.

Much of our theory will apply to these simple functions. However, the following version, which we will call the persistence landscape, will have some advantages.

First let us observe that for a fixed $t \in \mathbb{R}$, $\beta^{t-\bullet, t+\bullet}$ is a decreasing function. That is,

**Lemma 2.1.** For $0 \leq h_1 \leq h_2$,

$$\beta^{t-h_1, t+h_1} \geq \beta^{t-h_2, t+h_2}.$$ 


Proof. Since \( t - h_2 \leq t - h_1 \leq t + h_1 \leq t + h_2 \), \( M(t - h_2 \leq t + h_2) \leq M(t + h_1 \leq t + h_2) \cap M(t - h_1 \leq t + h_1) \cap M(t - h_2 \leq t - h_1) \). It follows that \( \beta^{t - h_2, t + h_2} \leq \beta^{t - h_1, t + h_1} \). □

**Definition 2.2.** The *persistence landscape* is a function \( \lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R} \), where \( \mathbb{R} \) denotes the extended real numbers. Alternatively, it may be thought of as a sequence of functions \( \lambda_k : \mathbb{R} \to \mathbb{R} \), where \( \lambda_k(t) = \lambda(k, t) \). Define
\[
\lambda_k(t) = \sup(m \geq 0 \mid \beta^{t - m, t + m} \geq k).
\]

For an example see Figure 2.

The persistence landscape has the following properties.

**Lemma 2.3.**
1. \( \lambda_k(t) \geq 0 \),
2. \( \lambda_k(t) \geq \lambda_{k+1}(t) \), and
3. \( \lambda_k \) is \( 1 \)-Lipschitz.

The first two properties follow directly from the definition. We prove the third in the appendix.

To help visualize the graph of \( \lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R} \), we can extend it to a function \( \bar{\lambda} : \mathbb{R}^2 \to \mathbb{R} \) by setting
\[
\bar{\lambda}(x, t) = \begin{cases} 
\lambda([x], t), & \text{if } x > 0, \\
0, & \text{if } x \leq 0.
\end{cases}
\]

For an example see Figure 2.

We remark that the non-persistent Betti numbers, \( \dim(M(t)) \), of a persistence module \( M \) can be read off from the diagonal of the rank function, the \( m \)-axis of the rescaled rank function, and from the support of the persistence landscape.

### 2.3. Barcodes and persistence diagrams

All of the information in a persistence module is completely contained in a multiset of intervals called a *barcode*. Mapping each interval to its endpoints we obtain the *persistence diagram*. There exists maps in both directions between these topological summaries and our persistence landscapes. For an example of corresponding persistence diagrams, barcodes and persistence landscapes, see Figure 2. Informally, the persistence diagram consists of the “upper-left corners” in our rank function. In the other direction, \( \lambda(b, d) \) counts the number of points in the persistence diagram in the upper left quadrant of \((b, d)\). Informally, the barcode consists of the “bases of the triangles” in the rescaled rank function, and the other direction is obtained by “stacking isosceles triangles” whose bases are the intervals in the barcode. We invite the reader to make the mappings precise. The fact that barcodes are a complete invariant of persistence modules is central to these equivalences.

**Remark.** Compared to the persistence diagram, the barcode has extra information on whether or not the endpoints of the intervals are included. This finer information is seen in the rank
function and rescaled rank function, but not in the persistence landscape. However when we pass the corresponding Lebesgue space in Section 2.4 this information disappears.

2.4. Norms for persistence landscapes. Recall that for a measure space \((S, \mathcal{A}, \mu)\), and a function \(f : S \to \mathbb{R}\) defined \(\mu\)-almost everywhere, for \(1 \leq p < \infty\), \(\|f\|_p = \left[ \int |f|^p d\mu \right]^{1/p}\), and \(\|f\|_\infty = \text{ess sup } f = \inf \{a \mid \mu \{s \in S \mid f(s) > a\} = 0\}\). For \(1 \leq p \leq \infty\), \(L^p(S) = \{f : S \to \mathbb{R} \mid \|f\|_p < \infty\}\) and define \(L^p(S) = L^p(S)/\sim\), where \(f \sim g\) if \(\|f - g\|_p = 0\).

On \(\mathbb{R}\) and \(\mathbb{R}^2\) we will use the Lebesgue measure. On \(\mathbb{N} \times \mathbb{R}\), we use the product of the counting measure on \(\mathbb{N}\) and the Lebesgue measure on \(\mathbb{R}\). For \(1 \leq p < \infty\) and \(\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}\),

\[
\|\lambda\|_p^p = \sum_{k=1}^{\infty} \|\lambda_k\|_p^p,
\]

where \(\lambda_k(t) = \lambda(k, t)\). By Lemma 2.3(2), \(\|\lambda\|_\infty = \|\lambda_1\|_\infty\). If we extend \(f\) to \(\overline{\lambda} : \mathbb{R}^2 \to \mathbb{R}\), as in (2.2), we have \(\|\lambda\|_p = \|\overline{\lambda}\|_p\), for \(1 \leq p \leq \infty\).

If \(\lambda\) is any of our persistence landscapes corresponding to a barcode that is a finite collection of finite intervals, then \(\lambda \in L^p(S)\) for \(1 \leq p \leq \infty\), where \(S\) equals \(\mathbb{N} \times \mathbb{R}\) or \(\mathbb{R}^2\).

Let \(\lambda_{bd}\) and \(\lambda_{mh}\) denote the rank function and the rescaled rank function corresponding to a persistence landscape \(\lambda\), and let \(D\) be the corresponding persistence diagram. Let \(\text{pers}_2(D)\) denote the sum of the squares of the lengths of the intervals in the corresponding barcode, and let \(\text{pers}_\infty(D)\) be the length of the longest interval.

**Proposition 2.4.**
(1) \(\|\lambda\|_1 = \|\lambda_{mh}\|_1 = \frac{1}{2}\|\lambda_{bd}\|_1 = \frac{1}{4}\text{pers}_2(D)\), and
(2) \(\|\lambda\|_\infty = \|\lambda_1\|_\infty = \frac{1}{2}\text{pers}_\infty(D)\).

**Proof.**
(1) To see that \(\|\lambda\|_1 = \|\lambda_{mh}\|\) we remark that both are the volume of the same solid. The change of coordinates (2.1) implies that \(\|\lambda_{mh}\|_1 = \frac{1}{2}\|\lambda_{bd}\|_1\). If \(D = \{(b_i, d_i)\}\), then each point \((b_i, d_i)\) contributes \(h_i^2\) to the volume \(\|\lambda_{mh}\|_1\), where \(h_i = \frac{d_i - b_i}{2}\). So \(\|\lambda_{mh}\|_1 = \sum_i h_i^2\). Finally, \(\text{pers}_2(D) = \sum_i (2h_i)^2 = 4 \sum_i h_i^2\).
(2) Lemma 2.3(2) implies that \(\|\lambda\|_\infty = \|\lambda_1\|_\infty\). If \(D = \{(b_i, d_i)\}\), then \(\|\lambda\|_\infty = \sup_i \frac{d_i - b_i}{2}\).

We remark that the quantities in (1) and (2) also equal \(W_2(D, \emptyset)^2\) and \(W_\infty(D, \emptyset)\) respectively (see Section 5 for definitions).

3. Statistics with landscapes

Now let us take a probabilistic viewpoint. First, we assume that our persistence landscapes lie in \(L^p(S)\) for some \(1 \leq p < \infty\), where \(S\) equals \(\mathbb{N} \times \mathbb{R}\) or \(\mathbb{R}^2\). In this case, \(L^p(S)\) is a separable Banach space. When \(p = 2\) we have a Hilbert space. However we will not use this Hilbert space structure. In some examples, the persistence landscapes will only be stable for some \(p > 2\) (see Theorem 5.5).
3.1. Landscapes as Banach space valued random variables. We consider our data $X$ as a random variable on some underlying probability space $(\Omega, \mathcal{F}, P)$, and so the corresponding persistence landscape $\lambda(X)$ is a Borel random variable with values in the separable Banach space $L^p(S)$.

Now let $X_1, \ldots, X_n$ be an independent and identically distributed sample, and let $\lambda(X_1), \ldots, \lambda(X_n)$ be the corresponding persistence landscapes. Using the vector space structure of $L^p(S)$, the mean landscape $\bar{\lambda}(X)_n$ is given by the pointwise mean.

$$\bar{\lambda}(X)_n(x,y) = \frac{1}{n} \sum_{i=1}^{n} \lambda(X_i)(x,y)$$

Let us give an interpretation of the mean landscape. If $B_1, \ldots, B_n$ are the barcodes corresponding to the persistence landscapes $\lambda(X_1), \ldots, \lambda(X_n)$ then for $k \in \mathbb{N}$ and $t \in \mathbb{R}$, $\bar{\lambda}(X)_n(k,t)$ is the average value of the largest radius interval centered at $t$ that is contained in $k$ intervals in the barcodes $B_1, \ldots, B_n$.

We would like to be able to say that the mean landscape converges to the expected persistence landscape. To say this precisely we need some notions from probability in Banach spaces.

3.2. Probability in Banach spaces. Here we present some results from probability in Banach spaces. For a more detailed exposition we refer the reader to [36].

Let $B$ be a real separable Banach space with norm $\|\cdot\|$. Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $V : (\Omega, \mathcal{F}, P) \to B$ be a Borel random variable with values in $B$. The composite $\|V\| : \Omega \xrightarrow{V} B \xrightarrow{\|\cdot\|} \mathbb{R}$ is a real-valued random variable.

Let $B^*$ denote the topological dual of continuous linear real-valued functions on $B$. For $f \in B^*$, the composite $f(V) : \Omega \xrightarrow{V} B \xrightarrow{f} \mathbb{R}$ is a real-valued random variable.

For a real-valued random variable $Y : (\Omega, \mathcal{F}, P) \to \mathbb{R}$, the mean or expected value, is given by $E(Y) = \int Y \, dP = \int_{\Omega} Y(\omega) \, dP(\omega)$.

We call an element $E(V) \in B$ the Pettis integral of $V$ if $E(f(V)) = f(E(V))$ for all $f \in B^*$.

**Proposition 3.1.** If $E\|V\| < \infty$, then $V$ has a Pettis integral and $\|E(V)\| \leq E\|V\|$.

Now let $(V_n)_{n \in \mathbb{N}}$ be a sequence of independent copies of $V$. For each $n \geq 1$, let $S_n = V_1 + \cdots + V_n$.

For a sequence $(Y_n)$ of $B$-valued random variables, we say that $(Y_n)$ converges almost surely to a $B$-valued random variable $Y$, if $P(\lim_{n \to \infty} Y_n = Y) = 1$.

**Theorem 3.2** (Strong Law of Large Numbers). $(\frac{1}{n} S_n) \to E(V)$ almost surely if and only if $E\|V\| < \infty$. 8
For a sequence \((Y_n)\) of \(B\)-valued random variables, we say that \((Y_n)\) converges weakly to a \(B\)-valued random variable \(Y\), if \(\lim_{n \to \infty} E(\varphi(Y_n)) = E(\varphi(Y))\) for all bounded continuous functions \(\varphi : B \to \mathbb{R}\).

A random variable \(G\) with values in \(B\) is said to be Gaussian if for each \(f \in B^*\), \(f(G)\) is a real valued Gaussian random variable with mean zero. The covariance structure of a \(B\)-valued random variable, \(V\), is given by the expectations \(E[(f(V) - E(f(V)))(g(V) - E(g(V)))]\), where \(f, g \in B^*\). A Gaussian random variable is determined by its covariance structure.

**Theorem 3.3** (Central Limit Theorem, [32]). Assume that \(B\) has type 2. (For example \(B = L^p(S), \text{ with } 2 \leq p < \infty\).) If \(E(V) = 0\) and \(E(\|V\|^2) < \infty\) then \(\frac{1}{\sqrt{n}}S_n\) converges weakly to a Gaussian random variable \(G(V)\) with the same covariance structure as \(V\).

**3.3. Convergence of persistence landscapes.** Now we will apply the results of the previous section to persistence landscapes.

Theorem 3.2 directly implies the following.

**Theorem 3.4** (Strong Law of Large Numbers for persistence landscapes). \(\overline{\lambda(X)}_n \to E(\lambda(X))\) almost surely if and only if \(E(\|\lambda(X)\|) < \infty\).

**Theorem 3.5** (Central Limit Theorem for persistence landscapes). Assume \(\lambda(X) \in L^p(S)\) with \(2 \leq p < \infty\). If \(E(\|\lambda(X)\|) < \infty\) and \(E(\|\lambda(X)\|^2) < \infty\) then \(\sqrt{n} \overline{\lambda(X)}_n - E(\lambda(X))\) converges weakly to a Gaussian random variable with the same covariance structure as \(\lambda(X)\).

**Proof.** Apply Theorem 3.3 to \(V = \lambda(X) - E(\lambda(X))\). \(\square\)

Now we apply a functional to the persistence landscapes to obtain a real-valued random variable that satisfies the usual Central Limit Theorem.

**Corollary 3.6.** Assume \(\lambda(X) \in L^p(S)\) where \(2 \leq p < \infty\) with \(E(\|\lambda(X)\|) < \infty\) and \(E(\|\lambda(X)\|^2) < \infty\). For any \(f \in L^q(S)\) with \(\frac{1}{p} + \frac{1}{q} = 1\), let

\[
Y = \int_S f\lambda(X).
\]

Then

\[
\sqrt{n}[\overline{Y}_n - E(Y)] \xrightarrow{d} N(0, \text{Var}(Y)),
\]

where \(d\) denotes convergence in distribution and \(N(\mu, \sigma^2)\) is the normal distribution with mean \(\mu\) and variance \(\sigma^2\).

**Proof.** Since \(V = \lambda(X) - E(\lambda(X))\) satisfies the Central Limit Theorem in \(L^p(S)\), for any \(g \in L^p(S)^*\), the real random variable \(g(V)\) satisfies the Central Limit Theorem in \(\mathbb{R}\) with limiting Gaussian law with mean 0 and variance \(E(g(V)^2)\). If we take \(g(h) = \int_S fh\), where \(f \in L^q(S)\), with \(\frac{1}{p} + \frac{1}{q} = 1\), then \(g(V) = Y - E(Y)\) and \(E(g(V)^2) = \text{Var}(Y)\). \(\square\)
3.4. **Confidence intervals.** The results of Section 3.3 allow us to obtain approximate confidence intervals for the expected values of functionals on persistence landscapes.

Assume that $\lambda(X)$ satisfies the conditions of Corollary 3.6 and that $Y$ is a corresponding real random variable as defined in (3.1). By Corollary 3.6 and Slutsky’s theorem we may use Student’s t-distribution to obtain the approximate $(1 - \alpha)$ confidence interval for $E(Y)$ using

$$Y_n \pm t_{(\frac{\alpha}{2}, n-1)} \frac{S_n}{\sqrt{n}},$$

where $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$, and $t_{(\frac{\alpha}{2}, n-1)}$ is the upper $\frac{\alpha}{2}$ critical value for the t distribution with $n - 1$ degrees of freedom.

3.5. **Statistical inference using landscapes I.** Here we apply the results of Section 3.3 to hypothesis testing using persistence landscapes.

Let $X_1, \ldots, X_n$ be an iid sample of the random variable $X$ and let $X_1', \ldots, X_{n'}'$ be an iid sample of the random variable $X'$. Assume that $\lambda(X), \lambda(X') \in L^p(S)$, where $2 \leq p < \infty$. Let $f \in L^q(S)$, where $\frac{1}{p} + \frac{1}{q} = 1$. Let $Y$ and $Y'$ be defined as in (3.1). Let $\mu = E(Y)$ and $\mu' = E(Y')$. We wish to test the null hypothesis that $\mu = \mu'$.

First we recall that the sample mean $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ is an unbiased estimator of $\mu$ and the sample variance $s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ is an unbiased estimator of Var($Y$) and similarly for $\bar{Y}'$ and $s_{Y'}^2$.

If we assume that $Y$ and $Y'$ have equal variance, then we can use the independent two-sample Student’s t-test. Let

$$t = \frac{\bar{Y} - \bar{Y}'}{s_{YY'} \sqrt{\frac{1}{n} + \frac{1}{n'}}},$$

where $s_{YY'}^2 = \frac{(n-1)s_Y^2 + (n'-1)s_{Y'}^2}{n + n' - 2}$

is an unbiased estimator of the common variance whether or not the two samples have a common mean. From this t statistic a p-value may be obtained from Student’s t-distribution with $n + n' - 2$ degrees of freedom.

One can test the equal variance assumption using Levene’s test. If this fails, we can repeat with the log transform of the data. If we do not assume equal variance then we can use Welch’s t-test.

3.6. **Choosing a functional.** To apply the above results, one needs to choose a functional, $f \in L^q(S)$. This choice will need to be made with an understanding of the data at hand. Here we present a couple of options.

There is a simple option if the persistence landscape $\lambda(X)$ has finite support. That is, the corresponding barcode is guaranteed to have a finite number of intervals, all of which are finite. Certain experimental data may have bounds on the number of intervals. For example, in the protein data considered using the ideas presented here in [35], the simplicial complexes have a fixed number of vertices. Infinite intervals can often be removed by considering
reduced homology or by applying extended persistence \[21, 11\] or by simply truncating
them at some fixed value. Under this assumption choose \( f \) to be the indicator function on
the support of \( \lambda(X) \). Then
\[
\int f\lambda(X) = \int \lambda(X) = \|\lambda(X)\|_1.
\]
A weaker assumption is that the parameter values for which the persistence landscape is
nonzero are bounded by \( \pm B \). In this case we have a nice choice of functional for the per-
sistence landscape, that is unavailable for the (rescaled) rank function. We can choose a
functional that is sensitive of the first \( K \) dominant homological features. That is, take
\[
f(k, t) = \begin{cases} 1 & \text{if } t \in [-B, B] \text{ and } k \leq K \\ 0 & \text{otherwise.} \end{cases}
\]
Then
\[
\int f\lambda(X) = \sum_{k=1}^{K} \int \lambda_k(X) = \sum_{k=1}^{K} \|\lambda_k(X)\|_1.
\]
Under this weaker assumption we can also take \( f_k(t) = \frac{1}{k^r} \chi_{[-B, B]} \), where \( r > 1 \). Then
\[
\int f\lambda(X) = \sum_{k=1}^{\infty} \frac{1}{k^r} \|\lambda_k(t)\|_1.
\]
One way to enforce that the parameter values for which the persistence landscape is nonzero
are bounded is to threshold the underlying persistence module \( M \) as follows. For some
\( B > 0 \), let \( M_B \) be the persistence module that is equal to \( M \) on \( [-B, B] \) but equal to zero
outside this interval. Let \( \lambda_B(M) = \lambda(M_B) \) and let \( D_B(M) = D(M_B) \), where \( D(M) \) denotes
the persistence diagram of \( M \).

3.7. Statistical inference using landscapes II. The functionals suggested in Section \[3.6\]
in the hypothesis test given in Section \[3.5\] may not have enough power to discriminate
between two groups with different persistence in some examples.

To increase the power, one can apply a vector of functionals and then apply Hotelling’s \( T^2 \)
test. For example, consider
\[
Y = (\int (\lambda_1 - \lambda'_1), \ldots, \int (\lambda_K - \lambda'_K)), \text{ where } k \ll n_1 + n_2 - 2.
\]
This alternative will not be sufficient if the persistence landscapes are translates of each
other, as in the homology in degree 0 in Figure \[8\]. An additional approach is to compute
the distance between the mean landscapes of the two groups and obtain a p-value using a
permutation test. This is done in the Section \[4.5\]. For another example, see \[10\]. This test
has been applied to persistence diagrams and barcodes \[19, 43\].

4. Examples

The persistent homologies in this section were calculated using javaPlex \[44\] and Perseus \[41\].
Another publicly available alternative is Dionysus \[39\]. In Section \[4.4\] we use Matlab code
courtesy of Eliran Subag that implements an algorithm from \[46\].
Figure 3. 200 points were sampled from a pair of linked annuli. Here we show the points and a corresponding union of balls and 1-skeleton of the Čech complex. This was repeated 100 times. Next we show two of the degree one persistence landscapes and the mean degree one persistence landscape.

4.1. Linked annuli. We start with a simple example to illustrate the techniques.

Following [40], we sample 200 points from the uniform distribution on the union of two annuli. We then calculate the corresponding persistence landscape in degree one using the Vietoris-Rips complex. We repeat this 100 times and calculate the mean persistence landscape. See Figure 3.

Note that in the degree one barcode of this example, it is very likely that there will be one large interval, one smaller interval born at around the same time, and all other intervals are smaller and die around the time the larger two intervals are born.

4.2. Random geometric complexes. The (non-persistent) homology of random geometric complexes has been studied in [33, 34, 6].
We sample 100 points from the uniform distribution on the unit cube $[0, 1]^3$, and calculate the persistence landscapes in degrees 0, 1 and 2 of the corresponding Vietoris-Rips complex. In degree 0, we use reduced homology. We repeat this 1000 times and calculate the corresponding mean persistence landscapes. See Figure 4.

Since the number of simplices is bounded and the filtration is bounded, these persistence landscapes have finite support. As discussed in Section 3.6, we can choose the functional given by the indicator function on this support. We obtain the real random variable $Y = \|\lambda(X)\|_1 = \frac{1}{4} \text{pers}_X(D(X))$, where $D(X)$ is the persistence diagram corresponding to $\lambda(X)$. Following Section 3.4, we calculate the approximate 95% confidence intervals of $E(Y)$ in degrees 0, 1 and 2 to be [0.1534, 0.1545], [0.0064, 0.0066] and [0.0002, 0.0003].

Remark. The graphs in Figure 4 may be thought of as a persistent homology version of the graph in Figure 2 of [34].

4.3. Erdős-Rényi random clique complexes. The (non-persistent) homology of the random complexes in this section has been studied [34].

Let $G(n)$ be the following random filtered graph. There are $n$ vertices with filtration value 0. Each of the possible $\binom{n}{2}$ edges has a filtration value which is chosen independently from a uniform distribution on $[0, 1]$. Let $X(n)$ be the clique complex of $G(n)$. In Figure 5 we show the mean persistence landscapes of a sample of 10 independent copies of $G(100)$ in degrees.
Figure 5. The mean persistence landscapes in degrees 0–3 from 10 copies of the random clique complex \( G(n) \). Note that we have rescaled the filtration by a factor of 100.

0, 1, 2, and 3, where in degree 0 we use reduced homology. For computational reasons, we only considered the subcomplex of \( G(100) \) with filtration values at most 0.55.

The graphs in Figure 5 are a persistent homology version of Figure 1 in [34]. In fact the latter graphs are given by the support of the graphs in Figure 5.

As in the example in Section 4.2, we let \( Y = \|\lambda(X)\|_1 = \frac{1}{2} \text{pers}_2(D(X)) \). The approximate 95% confidence intervals of \( E(Y) \) in degrees 0, 1, 2 and 3 are estimated to be [0.0034, 0.0039], [0.751, 0.777], [1.971, 2.041] and [2.591, 2.618].

4.4. Gaussian random fields. The topology of Gaussian random fields is of interest in statistics. The Euler characteristic of superlevelsets of a Gaussian random field may be calculated using the Gaussian Kinematic Formula [2]. The persistent homology of Gaussian random fields has been considered [11] and its expected Euler characteristic has been obtained [7].

Here we consider a stationary Gaussian random field on \([0, 1]^2\) with autocovariance function \( \gamma(x, y) = e^{-400(x^2+y^2)} \). See Figure 6. We sample this field on a 100 by 100 grid, and calculate the persistence landscape of the sublevel set. For homology in degree 0, we truncate the infinite interval at the maximum value of the field. We calculate the mean persistence landscapes in degrees 0 and 1 from 100 samples (see Figure 6 where we have rescaled the filtration by a factor of 100).
Figure 6. The graph of a Gaussian random field on $[0, 1]^2$ (top left) and its corresponding mean persistence landscapes (middle row) in degree 0 and 1. The 0-isosurface of a Gaussian random field on $[0, 1]^3$ (top right) and the corresponding mean persistence landscapes in degrees 0, 1 and 2.

Since we have sampled the field on a finite grid, there is a bound on the values of $k$ for which $\lambda_k(t)$ is nonzero. However since the field is unbounded, so is the support of $\lambda$. So the situation is more delicate than that in Sections 4.2 and 4.3.

One way to resolve this difficulty is to consider $\lambda_B$ and $D_B$ for some $B \gg 0$ as defined in Section 3.6. We let $Y = \|\lambda_B(X)\|_1 = \frac{1}{4} \text{pers}_2(D_B(X))$. The approximate 95% confidence intervals of $E(Y)$ in degrees 0 and 1 are estimated to be $[33.87, 35.37]$ and $[14.54, 15.39]$.

In the Gaussian random field literature, it is more common to consider superlevel sets. However, by symmetry, the expected persistence landscape in this case is same except for a change in the sign of the filtration. Furthermore $E(Y)$ is the same.

We repeat this calculation for the similar Gaussian random field on $[0, 1]^3$, this time using reduced homology. See Figure 6. This time we sample on a $25 \times 25 \times 25$ grid. For $Y = \|\lambda_B(X)\|_1 = \frac{1}{3} \text{pers}_2(D_B(X))$, the approximate 95% confidence intervals of $E(Y)$ in degrees 0, 1 and 2 are estimated to be $[110.72, 112.99]$, $[115.97, 117.65]$ and $[48.98, 50.27]$. 
4.5. **Torus and sphere.** We will use statistical inference on persistence landscapes to discriminate between iid samples of 1000 points from a torus and a sphere in $\mathbb{R}^3$ with the same surface area, using the uniform surface area measure [27] (see Figure 7).

For these points, we construct a filtered simplicial complex as follows. First we triangulate the underlying space using the Coxeter–Freudenthal-Kuhn triangulation. Next we smooth our data using a triangular kernel. We evaluate this kernel density estimator at the vertices of our simplicial complex. Finally, we filter our simplicial complex by taking the flag complex on upper excursion sets of the vertices. That is, for filtration level $r$, we include a simplex in our triangulation if and only if the kernel density estimator has values less than or equal
to $r$ at all of its vertices (see Figure 7). We then calculate the persistent homology of this filtered simplicial complex (see Figure 7).

Since the support of the persistence landscapes is bounded, we can use the integral of the landscapes to obtain a real valued random variable that satisfies (3.2).

First we apply Levene’s test to check the equal variance assumption. Then we use Student’s t-test to test the null hypothesis that these random variables have equal mean. For the landscapes in dimensions 0 and 2 we cannot reject the null hypothesis. In dimension 1 we do reject the null hypothesis with a p-value of $4.6 \times 10^{-6}$.

We can also choose a functional that only integrates the persistence landscape $\lambda(k, t)$ for certain ranges of $k$. In dimension 1, with $k = 1$ or $k = 2$ there is a statistically significant difference (p-values of $10^{-8}$ and $7 \times 10^{-7}$), but not for $k > 2$. In dimension 2, there is not a significant difference for $k = 1$, but there is a significant difference for $k > 1$ (p-value $< 10^{-4}$). For the last case, we use Welch’s t-test.

Now we increase the difficulty by adding a fair amount of Gaussian noise to the point samples (see Figure 8). This time we calculate the $L^2$ distances between the mean landscapes. We use the permutation test with 10,000 repetitions to determine if this distance is statistically significant. There is a significant difference in dimension 0, with a p value of 0.0111. This is surprising, since the mean landscapes look very similar. However, on closer inspection, they are shifted slightly (see Figure 8). Note that we are detecting a geometric difference, not a topological one. Less surprisingly, there is also a significant difference in dimensions 1 and 2, with p values of 0.0000 and 0.0000, respectively.
Figure 8. On the left we have 1000 points sampled from a torus, from the perspective that makes it easiest to see the hole in the middle. In the middle we have points sampled from the sphere. We calculate persistent homology as above. In columns 1, 2 and 3, we have the mean persistence landscape in dimension 0, 1 and 2, respectively, with the torus in row 2 and the sphere in row 3. The top left is a graph of the difference between the mean landscapes in dimension 0.
5. Landscape distance and stability

In this section we define the landscape distance and use it to show that the persistence landscape is a stable summary statistic. We also show that the landscape distance gives lower bounds for the bottleneck and Wasserstein distances. We defer the proofs of Theorems 5.1, 5.2 and 5.3 to the appendix.

Let $M$ and $M'$ be persistence modules as defined in Section 2.1 and let $\lambda$ and $\lambda'$ be their corresponding persistence landscapes as defined in Section 2.2. For $1 \leq p \leq \infty$, define the $p$-landscape distance between $M$ and $M'$ by

$$\Lambda_p(M, M') = \|\lambda - \lambda'\|_p.$$  

Similarly, if $\lambda$ and $\lambda'$ are the persistence landscapes corresponding to persistence diagrams $D$ and $D'$, then we define

$$\Lambda_p(D, D') = \|\lambda - \lambda\|_p.$$  

Given a real valued function $f : X \to \mathbb{R}$ on a topological space $X$, let $M(f)$ denote the corresponding persistence module defined at the end of Section 2.1.

**Theorem 5.1** ($\infty$-Landscape Stability Theorem). Let $f, g : X \to \mathbb{R}$. Then

$$\Lambda_\infty(M(f), M(g)) \leq \|f - g\|_\infty.$$  

We remark that there are no assumptions on $f$ and $g$, not even the q-tame condition of [15].

Let us consider a persistence diagram to be an equivalence class of multisets of pairs $(b, d)$ with $b \leq d$, where $D \sim D \sqcup \{(t, t)\}$ for any $t \in \mathbb{R}$. Each persistence diagram has a unique representative $\hat{D}$ without any points on the diagonal. By allowing ourselves to add as many points on the diagonal as necessary, there exists bijections between any two persistence diagrams.

Any bijection $\varphi : D \xrightarrow{\cong} D'$ can be represented by $\varphi : x_j \mapsto x'_j$, where $j \in J$ with $|J| = |\hat{D}| + |\hat{D}'|$. Let $x_j = (b_j, d_j)$, $x'_j = (b'_j, d'_j)$ and $\epsilon_j = \|x_j - x'_j\|_\infty = \max(|b_j - b'_j|, |d_j - d'_j|)$. Let $\ell_j = d_j - b_j$ be the persistence of $x_j$.

The bottleneck distance [20] between persistence diagrams $D$ and $D'$ is given by

$$W_\infty(D, D') = \inf_{\varphi : D \xrightarrow{\cong} D'} \sup_j \epsilon_j.$$  

It follows that $W_\infty(D, \emptyset) = \frac{1}{2} \sup_j \ell_j$, where $\emptyset$ is the empty persistence diagram.

The $\infty$-landscape distance is bounded by the bottleneck distance.

**Theorem 5.2.** For persistence diagrams $D$ and $D'$,

$$\Lambda_\infty(D, D') \leq W_\infty(D, D').$$  

19
For $p \geq 1$, the $p$-Wasserstein distance \cite{22} between $D$ and $D'$ is given by

$$W_p(D, D') = \inf_{\varphi : D \xrightarrow{\sim} D'} \left[ \sum_{j=1}^{n} \varepsilon_j^p \right]^{\frac{1}{p}},$$

where the infimum is taken over all bijections from $D$ to $D'$.

We remark that if $D = \{x_j\}$ and $D' = \{x_j + \varepsilon_j\}$, then the Wasserstein distance gives equal weighting to the $\varepsilon_j$ while the landscape distance gives a stronger weighting to $\varepsilon_j$ if $x_j$ has larger persistence.

The landscape distance is most closely related to a weighted version of the Wasserstein distance that we now define. The persistence weighted $p$-Wasserstein distance between $D$ and $D'$ is given by

$$\overline{W}_p(D, D') = \inf_{\varphi : D \xrightarrow{\sim} D'} \left[ \sum_{j=1}^{n} \ell_j \varepsilon_j^p \right]^{\frac{1}{p}}.$$

Note that it is asymmetric.

**Theorem 5.3.** If $n = |D| + |D'|$ then

$$\Lambda_p(D, D')^p \leq \min_{\varphi : D \xrightarrow{\sim} D'} \left[ \sum_{j=1}^{n} \ell_j \varepsilon_j^p + \frac{2}{p+1} \sum_{j=1}^{n} \varepsilon_j^{p+1} \right].$$

If $W_p(D, D') \leq 1$ then for the minimizer for $W_p(D, D')$, $\varepsilon_j \leq 1$ for all $j$. So we get the following corollary.

**Corollary 5.4.** If $W_p(D, D') \leq 1$ then $\Lambda_p(D, D')^p \leq 2(W_\infty(D, \emptyset) + \frac{1}{p+1})W_p(D, D')^p$. Thus $W_p(D, D')^p \geq \min(1, \frac{1}{2}[W_\infty(D, \emptyset) + \frac{1}{p+1}]^{-1})\Lambda_p(D, D')^p$.

From Theorem 5.3 using the terminology from \cite{22} and following the proof of the Wasserstein Stability Theorem therein, we have the following.

**Theorem 5.5** ($p$-Landscape stability theorem). Let $X$ be a triangulable, compact metric space that implies bounded degree-$k$ total persistence for some real number $k \geq 1$, and let $f$ and $g$ be two tame Lipschitz functions. Then

$$\Lambda_p(D(f), D(g))^p \leq C\|f - g\|_\infty^{-k},$$

for all $p \geq k$, where $C = C_X \max\{\text{Lip}(f)^k, \text{Lip}(g)^k, \text{Lip}(f)^{k+1}, \text{Lip}(g)^{k+1}\}(W_\infty(D, \emptyset) + \frac{1}{p+1})$.

So the $p$-landscape condition is stable if $p > k$, where $X$ has bounded degree-$k$ total persistence. This is the same condition as for the stability of the $p$-Wasserstein distance in \cite{22}.

**Appendix A. Proofs**

**Proof of Lemma 2.3.** We will prove that $\lambda_k$ is 1-Lipschitz. That is, $|\lambda_k(t) - \lambda_k(s)| \leq |t - s|$, for all $s, t \in \mathbb{R}$. 

20
Let \( s, t \in \mathbb{R} \). Without loss of generality, assume that \( \lambda_k(t) \geq \lambda_k(s) \geq 0 \). If \( \lambda_k(t) \leq |t-s| \), then \( \lambda_k(t) - \lambda_k(s) \leq \lambda_k(t) \leq |t-s| \) and we are done. So assume that \( \lambda_k(t) > |t-s| \).

Let \( 0 < \varepsilon < \lambda_k(t) - |t-s| \). By Lemma 2.1, our assumption, and Definition 2.2 (A.1)

\[
\beta^{t-(\lambda_k(t)-\varepsilon),t+(\lambda_k(t)-\varepsilon)} \geq \beta^{t-|t-s|,t+|t-s|} \geq k.
\]

Since \(-|t-s| \leq t-s \leq |t-s|\), \( s - |t-s| \leq t \leq s + |t-s| \). Also, \(-\lambda_k(t) + |t-s| + \varepsilon < 0 < \lambda_k(t) - |t-s| - \varepsilon \). Using these inequalities, we have,

\[
t - \lambda_k(t) + \varepsilon \leq s - \lambda_k(t) + |t-s| + \varepsilon < s < s + \lambda_k(t) - |t-s| - \varepsilon \leq t + \lambda_k(t) - \varepsilon.
\]

Together with Lemma 2.1 and (A.1), we see that

\[
\beta^{s-(\lambda_k(t)-|t-s|-\varepsilon),s+(\lambda_k(t)-|t-s|-\varepsilon)} \geq \beta^{t-(\lambda_k(t)-\varepsilon),t+(\lambda_k(t)-\varepsilon)} \geq k.
\]

Therefore \( \lambda_k(s) \geq \lambda_k(t) - |t-s| \). Thus \( \lambda_k(t) - \lambda_k(s) \leq |t-s| \).

Theorems 5.1 and 5.2 follow from the next result which is of independent interest. For persistence modules \( M \) and \( M' \) let \( d_I(M,M') \) denote the interleaving distance between \( M \) and \( M' \) [14].

**Theorem A.1.** \( \Lambda_\infty(M, M') \leq d_I(M, M') \).

*Proof.* Assume that \( M \) and \( M' \) are \( \varepsilon \)-interleaved. Then for \( t \in \mathbb{R} \) and \( m \geq \varepsilon \), the map \( M(t-m \leq t+m) \) factors through the map \( M'(t-m+\varepsilon \leq t+m-\varepsilon) \). So \( \beta^{t-m+\varepsilon,t+m-\varepsilon}(M') \geq \beta^{t-m,t+m}(M) \), and thus \( \lambda'(k,t) \geq \lambda(k,t) - \varepsilon \) for all \( k \geq 1 \). It follows that \( \| \lambda - \lambda' \|_\infty \leq \varepsilon \). \( \square \)

*Proof of Theorem 5.1.* Combining Theorem A.1 with the Stability Theorem of [11], we have \( \Lambda_\infty(M(f), M(g)) \leq d_I(M(f), M(g)) \leq \|f - g\|_\infty \). \( \square \)

*Proof of Theorem 5.2.* For a persistence diagram \( D \), consider the persistence module given by the corresponding sum of interval modules [15], \( M(D) = \oplus_{(a,b) \in D} \mathbb{I}(a,b) \). Combining Theorem A.1 with [15] Theorem 4.9] we have \( \Lambda_\infty(M(D), M(D')) \leq d_I(M(D), M(D')) \leq W_\infty(D, D') \). \( \square \)

*Proof of Theorem 5.3.* Let \( \varphi : D \rightarrow D' \) with \( \varphi(x_j) = x'_j \). Let \( \lambda = \lambda(D) \) and \( \lambda' = \lambda(D') \). So \( \Lambda_p(D, D')^p = \| \lambda - \lambda' \|_p^p \).

\[
\| \lambda - \lambda' \|_p^p = \int |\lambda(k,t) - \lambda'(k,t)|^p \, dt = \sum_{k=1}^n \int |\lambda_k(t) - \lambda'_k(t)|^p \, dt = \int \sum_{k=1}^n |\lambda_k(t) - \lambda'_k(t)|^p \, dt.
\]

21
Fix $t$. Let $u_j(t) = \lambda(\{x_j\})(1,t)$ and $v_j(t) = \lambda(\{x'_j\})(1,t)$. For each $t$, let $u_{(1)}(t) \leq \cdots \leq u_{(n)}(t)$ denote an ordering of $u_1(t), \ldots, u_n(t)$ and define $v_{(k)}(t) = \lambda_k(t)$ for $1 \leq k \leq n$ similarly. Then $u_{(k)}(t) = \lambda_k(t)$ and $v_{(k)}(t) = \lambda'_k(t)$ (see Figure 2). We obtain the result from the following where the two inequalities are proven in Lemmata A.2 and A.3.

\[
\|\lambda - \lambda'\|_p = \int \sum_{k=1}^n |u_{(k)}(t) - v_{(k)}(t)|^p dt
\leq \int \sum_{k=1}^n |u_k(t) - v_k(t)|^p dt
= \sum_{j=1}^n \int |u_j(t) - v_j(t)|^p dt
\leq \sum_{j=1}^n \ell_j \varepsilon_j^p + \frac{2}{p+1} \sum_{j=1}^n \varepsilon_j^{p+1}. \quad \square
\]

**Lemma A.2.** Let $u_1, \ldots, u_n \in \mathbb{R}$ and $v_1, \ldots, v_n \in \mathbb{R}$. Reorder them $u_{(1)} \leq \cdots \leq u_{(n)}$ and $v_{(1)} \leq \cdots \leq v_{(n)}$. Then

\[
\sum_{j=1}^n |u_{(j)} - v_{(j)}|^p \leq \sum_{j=1}^n |u_j - v_j|^p.
\]

**Proof.** Assume $u_1 < \cdots < u_n$, $v_1 < \cdots < v_n$, and $p \geq 1$. Let $u$ and $v$ denote $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$. Let $\Sigma_n$ denote the symmetric group on $n$ letters and let $f_n : \Sigma_n \to \mathbb{R}$ be given by $f_n(\sigma) = \sum_{j=1}^n |u_j - v_{\sigma(j)}|^p$. We will prove by induction that if $f_n(\sigma)$ is minimal then $\sigma$ is the identity, which we denote by 1.

For $n = 1$ this is trivial. For $n = 2$ assume without loss of generality that $u_1 = 0$, $u_2 = 1$ and $0 \leq v_1 < v_2$. Let $1$ and $\tau$ denote the elements of $\Sigma_2$. Then $f(1) = v_1^p + |1 - v_2|^p$ and $f(\tau) = v_2^p + |1 - v_1|^p$. Notice that $f(1) < f(\tau)$ iff $v_1^p < |1 - v_1|^p$. The result follows from checking that $g(x) = x^p - |1 - x|^p$ is an increasing function for $x \geq 0$.

Now assume that the statement is true for some $n \geq 2$. Assume that $f_{n+1}(\sigma^*)$ is minimal. Fix $1 \leq i \leq n+1$. Let $u' = (u_1, \ldots, \hat{u}_i, \ldots, u_{n+1})$ and $v' = (v_1, \ldots, \hat{v}_{\sigma^*(i)}, \ldots, v_{n+1})$, where $\hat{\cdot}$ denotes omission. Since $f_{n+1}(\sigma^*)$ is minimal for $u$ and $v$, it follows that $\sum_{j=1, j \neq i}^n |u_j - v_{\sigma^*(j)}|$ is minimal for $u'$ and $v'$. By the induction hypothesis, for $1 \leq j < k \leq n+1$ and $j, k \neq i$, $\sigma^*(j) < \sigma^*(k)$. Therefore $\sigma^* = 1$. Thus, by induction, the statement is true for all $n$.

Hence $\sum_{j=1}^n |u_{(j)} - v_{(j)}|^p \leq \sum_{j=1}^n |u_j - v_j|^p$ if $u_{(1)} < \cdots < u_{(n)}$ and $v_{(1)} < \cdots < v_{(n)}$. The statement in the lemma follows by continuity. \quad \square

**Lemma A.3.** Let $x = (b, d)$ and $x' = (b', d')$ where $b \leq d$ and $b' \leq d'$. Let $\ell = d - b$ and $\varepsilon = \|x - x'\|_\infty$. Then $\|\lambda(\{x\}) - \lambda(\{x'\})\|_p \leq \ell \varepsilon^p + \frac{2}{p+1} \varepsilon^{p+1}$.

**Proof.** Let $\lambda = \lambda(\{x\})$ and $\lambda' = \lambda(\{x'\})$. First $\lambda_k = \lambda'_k = 0$ for $k > 1$; so $\|\lambda - \lambda'\|_p = \|\lambda_1 - \lambda'_1\|_p$. Second $\lambda_1(t) = (h - |t - m|)_+$, where $h = \frac{d-b}{2}$, $m = \frac{b+d}{2}$, and $y_+ = \max(y, 0)$, and similarly for $\lambda'_1$ (see Figure 2).
Fix $x$ and $\varepsilon$. As $x'$ moves along the square $\|x - x'\|_\infty = \varepsilon$, $\|\lambda_1 - \lambda_1'\|_p$ has a maximum if $x' = (a - \varepsilon, b + \varepsilon)$. In this case $\|\lambda_1 - \lambda_1'\|_p = 2 \int_0^h \varepsilon^p \, dt + 2 \int_0^\varepsilon t^p \, dt = \ell \varepsilon^p + \frac{2}{p+1} \varepsilon^{p+1}$. □

Acknowledgments

The author would like to thank Robert Adler, Frederic Chazal, Herbert Edelsbrunner, Giseon Heo, Sayan Mukherjee and Stephen Rush for helpful discussions. Thanks to Junyong Park for suggesting Hotelling’s $T^2$ test. The author also gratefully acknowledges the support of AFOSR grant FA9550-13-1-0115.

References

[1] Robert J. Adler, Omer Bobrowski, Matthew S. Borman, Eliran Subag, and Shmuel Weinberger. Persistent homology for random fields and complexes. In Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown, volume 6 of Inst. Math. Stat. Collect., pages 124–143. Inst. Math. Statist., Beachwood, OH, 2010.

[2] Robert J. Adler and Jonathan E. Taylor. Random fields and geometry. Springer Monographs in Mathematics. Springer, New York, 2007.

[3] Sivaraman Balakrishnan, Brittany Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Statistical inference for persistent homology. 03 2013.

[4] Paul Bendich, Herbert Edelsbrunner, Dmitriy Morozov, and Amit Patel. Homology and robustness of level and interlevel sets. Homology Homotopy Appl., 15(1):51–72, 2013.

[5] Andrew J. Blumberg, Itamar Gal, Michael A. Mandell, and Matthew Pancia. Persistent homology for metric measure spaces, and robust statistics for hypothesis testing and confidence intervals. 06 2012.

[6] Omer Bobrowski and Robert J. Adler. Distance functions, critical points, and topology for some random complexes. arXiv:1107.4775 [math.AT], 2011.

[7] Omer Bobrowski and Matthew Strom Borman. Euler integration of Gaussian random fields and persistent homology. J. Topol. Anal., 4(1):49–70, 2012.

[8] Karol Borsuk. On the imbedding of systems of compacta in simplicial complexes. Fund. Math., 35:217–234, 1948.

[9] Peter Bubenik, Gunnar Carlsson, Peter T. Kim, and Zhi-Ming Luo. Statistical topology via Morse theory persistence and nonparametric estimation. In Algebraic methods in statistics and probability II, volume 516 of Contemp. Math., pages 75–92. Amer. Math. Soc., Providence, RI, 2010.

[10] Peter Bubenik and Pawel Dlotko. Interleavings of categories and generalized factor persistence. 2013.

[11] Peter Bubenik and Jonathan A. Scott. Categorification of persistent homology. arXiv:1205.3669 [math.AT], 2012.

[12] Gunnar Carlsson. Topology and data. Bull. Amer. Math. Soc. (N.S.), 46(2):255–308, 2009.

[13] Gunnar Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. Int. J. Comput. Vision, 76(1):1–12, 2008.

[14] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In Proceedings of the 25th annual symposium on Computational geometry, SCG ’09, pages 237–246, New York, NY, USA, 2009. ACM.

[15] Frederic Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. arXiv:1207.3674 [math.AT], 2012.

[16] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. On the bootstrap for persistence diagrams and landscapes. 11 2013.

[17] Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Optimal rates of convergence for persistence diagrams in topological data analysis. 05 2013.

[18] Chao Chen and Michael Kerber. An output-sensitive algorithm for persistent homology. Comput. Geom., 46(4):435–447, 2013.
[19] Moo K. Chung, Peter Bubenik, and Peter T. Kim. Persistence diagrams in cortical surface data. In Information Processing in Medical Imaging (IPMI) 2009, volume 5636 of Lecture Notes in Computer Science, pages 386–397, 2009.

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

[21] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Extending persistence using Poincaré and Lefschetz duality. Found. Comput. Math., 9(1):79–103, 2009.

[22] David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have $L_p$-stable persistence. Found. Comput. Math., 10(2):127–139, 2010.

[23] William Crawley-Boevey. Decomposition of pointwise finite-dimensional persistence modules. arXiv:1210.0819 [math.RT], 2012.

[24] Vin de Silva and Gunnar Carlsson. Topological estimation using witness complexes. Eurographics symposium on point-based graphics, 2004.

[25] Mary-Lee Dequéant, Sebastian Ahnert, Herbert Edelsbrunner, Thomas M. A. Fink, Earl F. Glynn, Gaye Hattem, Andrzej Kudlicki, Yuriy Mileyko, Jason Morton, Arcady R. Mushegian, Lior Pachter, Maga Rowicka, Anne Shiu, Bernd Sturmfels, and Olivier Pourquié. Comparison of pattern detection methods in microarray time series of the segmentation clock. PLoS ONE, 3(8):e2856, 08 2008.

[26] Tamal Krishna Dey, Fengtao Fan, and Yusu Wang. Graph induced complex on point data. In Proceedings of the Twenty-ninth Annual Symposium on Computational Geometry, SoCG ’13, pages 107–116, New York, NY, USA, 2013. ACM.

[27] Persi Diaconis, Susan Holmes, and Mehrdad Shajeshahani. Sampling from a manifold. arXiv:1206.6913 [math.ST], 2012.

[28] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. Discrete Comput. Geom., 28(4):511–533, 2002. Discrete and computational geometry and graph drawing (Columbia, SC, 2001).

[29] Robert Ghrist. Barcodes: the persistent topology of data. Bull. Amer. Math. Soc. (N.S.), 45(1):61–75, 2008.

[31] Allen Hatcher. Algebraic topology. Cambridge University Press, Cambridge, 2002.

[32] Matthew Kahle. Random geometric complexes. Discrete Comput. Geom., 45:553–573, 2011.

[34] Matthew Kahle and Elizabeth Meckes. Limit theorems for betti numbers of random simplicial complexes. arXiv:1009.4130 [math.AT], 2010.

[35] Violeta Kovacev-Nikolic, Giseon Heo, Dragan Nikolić, and Peter Bubenik. Topological statistical analysis of the maltose binding protein. 2013.

[36] Michel Ledoux and Michel Talagrand. Probability in Banach spaces. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[37] J. Hoffmann-Jørgensen and G. Pisier. The law of large numbers and the central limit theorem in Banach spaces. Ann. Probability, 4(4):587–599, 1976.

[38] Nikola Milosavljević, Dmitriy Morozov, and Primož Škraba. Zigzag persistent homology in matrix multiplication time. In Computational geometry (SCG’11), pages 216–225. ACM, New York, 2011.

[39] Dmitriy Morozov. Dionysus: a C++ library with various algorithms for computing persistent homology. Software available at http://www.mrzv.org/software/dionysus/ 2012.

[40] Elizabeth Munch, Paul Bendich, Katharine Turner, Sayan Mukherjee, Jonathan Mattingly, and John Harer. Probabilistic fréchet means and statistics on vineyards. 07 2013.

[41] Vidit Nanda. Perseus: the persistent homology software. Software available at http://www.math.rutgers.edu/~vidit/perseus/index.html 2013.

[42] Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proc. Nat. Acad. Sci., 108(17):7265–7270, 2011.

[43] Andrew Robinson and Katharine Turner. Hypothesis testing for topological data analysis. 10 2013.
[44] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. Javaplex: A research software package for persistent (co)homology. Software available at [http://code.google.com/javaplex](http://code.google.com/javaplex) 2011.

[45] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. 06 2012.

[46] Andrew T. A. Wood and Grace Chan. Simulation of stationary Gaussian processes in $[0, 1]^d$. *J. Comput. Graph. Statist.*, 3(4):409–432, 1994.

[47] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005.