TOPOLOGICAL DATA ANALYSIS: THEORY AND EXAMPLES

BRODY WAGERSON, ALLISON RAMASAMI

Contents

1. Introduction

In this paper we aim to develop some of the theory behind topological data analysis (TDA) and its main tool, persistent homology. We start from the usual scenario with a point cloud in n-dimensional Euclidean space and follow the basic work flow in TDA: that is, to create a nested sequence of simplicial complexes using either the Čech or Rips complex. Then, we compute the persistent homology of this sequence and quantify the birth and death of homological features in a persistence diagram. We give an example using the `TDA` package in R. One of the pillars that makes TDA possible is the stability theorem: we explain this and work out a simple example verifying the theorem. Finally, we discuss how one could go about applying statistical methods to quantify the uncertainty associated with persistence diagrams, via a method given by Fasy, et al. We continue our example by computing confidence bands for persistence diagrams.

2. Simplicial Complexes

We would like to first note that nearly all of the discussion over the next four sections come from Edelsbrunner and Harer's *Computational Topology* [EH10].

There are many ways to represent a topological space but in practice and for our purpose we use simplicial complexes as the data structure. Simplicial complexes are just sets of what are called simplices, that follow some intuitive rules. A simplex (or plural simplices) is a generalization of the notion of triangle or tetrahedron to higher dimensions. Concretely this is defined as follows.

Date: April 24, 2024.

What we would like to do is define a k-simplex, in order to do so we need a few more definitions to get us there. Let $u_0, ..., u_k$ be points in $\mathbb{R}^d$. A point $x = \sum_{i=0}^{k} \lambda_i u_i$ with $\lambda_i \in \mathbb{R}$, is an affine combination of the u_i if λ_i sum to 1. The affine hull is the set of affine combinations. It is a k-plane if the $k+1$ points are affinely independent, by which we mean that any two affine combinations, $x = \sum_{i=0}^{k} \lambda_i u_i$ and $y = \sum_{i=0}^{k} \mu_i u_i$ are the same iff $\lambda_i = \mu_i$ for all i. The $k+1$ points are affinely independent iff the k vectors $u_i - u_0$ for $1 \leq i \leq k$ are linearly independent. In $\mathbb{R}^d$ we can have at most d linearly independent vectors and therefore at most $d+1$ affinely independent points.

An affine combination, $x = \sum_{i=0}^{k} \lambda_i u_i$ is a convex combination if all λ_i are non-negative. The convex hull is the set of convex combinations. A k-simplex is the convex hull of $k+1$ affinely independent points, defined as $\tau = \text{conv}\{u_0, .., u_k\}$. We can say u_i span τ. Then a vertex is a 0-simplex, edge is 1-simplex, triangle is a 2-simplex and tetrahedron is a 3-simplex.

A face of a k-simplex is the convex hull of a non-empty subset of the u_i and it is proper if the subset is not the entire set. Since a set of size $k+1$ has 2^{k+1} subsets including the empty set, a k-simplex has $2^{k+1} - 1$ faces, all of which are proper except for τ itself.

The interior of a k-simplex σ is σ minus all of its proper faces. This is equivalent to the statement that x is in the interior of σ if and only if all its coefficients λ_i are positive. It follows from this that every point $x \in \sigma$ is in the interior of exactly one face of σ.

We sometimes write $\tau \leq \sigma$ if τ is a face and $\tau < \sigma$ if it is a proper face of σ.

Now that we have a definition of k-simplex and face of a simplex, we can define a geometric simplicial complex.

Definition 2.1 (Geometric Simplicial Complex)**.** A *geometric simplicial complex* is a finite collection of simplices K such that

(1) $\sigma \in K$ and $\tau \leq \sigma$ implies $\tau \in K$.
(2) $\sigma, \tau \in K$ implies $\sigma \cap \tau$ is either empty or a face of both.

The dimension of K is the maximum dimension of any of its simplices. In addition, we define the underlying space of a geometric simplicial complex $|K|$ as simply the union of all its simplices in the space it lives in, $\mathbb{R}^n$. The vertex set of K, Vert K is simply the collection of all 0-simplices of K.

There is a way to generalize the idea of a simplicial complex without talking about $\mathbb{R}^n$. The idea is that the simplices in a simplicial complex are completely determined by their vertices. So instead of defining a simplex as the convex hull of affinely independent points, we can instead define a simplex as simply a set of points. We give the definition of an abstract simplicial complex below:

Definition 2.2. An *abstract simplicial complex* is a collection of finite sets A such that $\alpha \in A$, $\beta \subseteq \alpha$ implies $\beta \in A$.

We note that condition (2) of the definition for a geometric simplicial complex is no longer necessary here, because it is satisfied by construction. We define a geometric realization of an abstract simplicial complex as simply a geometric simplicial complex K that has all the same simplices as A and also satisfies condition (2) of a geometric simplicial complex.

Now that these definitions are out of the way, we can talk about simplicial maps. First, we define barycentric coordinates. Recall from earlier that each point x in a simplex σ belongs to the interior of one face in σ. In particular, if K is a simplicial complex with vertices $u_0, \cdots, u_n$ and $\sigma = \operatorname{conv}\{u_0, \cdots, u_k\}$, then $x = \sum_{i=0}^{k} \lambda_i u_i$. Then the barycentric coordinates of x are

$$b_i(x) = \begin{cases} \lambda_i, & 0 \leq i \leq k \\ 0, & k < i \leq n \end{cases}$$

So this gives us a way to describe the position of a point in $|K|$ given the vertices of the simplicial complex K. Now, we can define a vertex map.

Definition 2.3 (Vertex Map)**.** Let K, L be simplicial complexes. A map $\varphi : \operatorname{Vert} K \to \operatorname{Vert} L$ is a *vertex map* if it has the property that vertices of a simplex in K map to vertices of a simplex in L.

Note that φ need not be injective: the vertices of a 3-simplex could be mapped to a single 0-simplex without violating the vertex map condition. What a vertex map between K and L does is induce a map between the simplices of these complexes. This is called a simplicial map.

Definition 2.4 (Simplicial Map)**.** Let K, L be simplicial complexes and φ be a vertex map. Then we define a *simplicial map* $f : |K| \to |L|$ as

$$f(x) = \sum_{i=0}^{n} b_i(x)\varphi(x)$$

We generally drop the fact that we are talking about the underlying space and simply say $f : K \to L$. The idea with the simplicial map is that it maps simplices to other simplices of the same or lower dimension.

3. Homology

In order to talk about persistent homology, we first have to talk about homology. The aim of topological data analysis and in particular persistent homology is to approximate the underlying manifold that our data points have been sampled from. Persistent homology gives us a way to summarize the topological features of our data, these topological features are then encoded in what is called a persistence diagram. We will cover simplicial homology with $\mathbb{Z}_2$ coefficients, as this is what we will need.

Suppose we have a simplicial complex K. Then we define p-chains:

Definition 3.1 (Group of p-chains)**.** If K is a simplicial complex and $p \in \mathbb{Z}$, a *p-chain* is a formal sum of p-simplices of K, with coefficients in $\mathbb{Z}_2$. That is,

$$\sigma = \sum_{i=1}^{k} a_i \sigma_i$$

where each $a_i \in \mathbb{Z}_2$ and σ_i is a p-simplex. We denote the collection of all p-chains of a simplicial complex C_p, which turns out to be an abelian group under addition.

Since each coefficient is either 0 or 1, we can think of a p-chain as a subset of p-simplices taken from the simplicial complex. While this idea doesn't generalize to $\mathbb{Z}$ coefficients, it is convenient to visualize homology like this. So for any simplicial complex K, we have a collection of chain groups $C_0, C_1, \ldots, C_n$. We now want to define the boundary operator, ∂_p:

Definition 3.2 (Boundary Operator)**.** If K is a simplicial complex, we define the boundary operator $\partial_p : C_p \to C_{p-1}$ on a p-simplex $\sigma = [u_0, \ldots, u_p]$ as

$$\partial_p \sigma = \sum_{k=1}^{p} [u_0, \ldots, \hat{u}_k, \ldots, u_p]$$

where $\hat{u}_k$ means that this particular element has been removed from the simplex. The boundary operator on a p-chain extends linearly: $\partial_p(\sigma + \tau) = \partial_p \sigma + \partial_p \tau$.

We note that the condition that the boundary operator be linear means that ∂_p is a group homomorphism from C_p to C_{p-1}. This forms something we call a chain complex: it is a sequence of chain groups connected by boundary homomorphisms.

$$\begin{tikzcd}
0 & C_0 \arrow[l, "\partial_0"'] & C_1 \arrow[l, "\partial_1"'] & C_2 \arrow[l, "\partial_2"'] & \cdots \arrow[l, "\partial_3"']
\end{tikzcd}$$

Here, by 0 we mean the trivial group with a single element, 0.

Now, we cover the p-cycle and p-boundary groups, which are subgroups of the p-chain group.

Definition 3.3 (Groups of p-cycles and p-boundaries)**.** A *p-cycle* is a p-chain with empty boundary, $\partial_p \sigma = 0$. The collection of all p-cycles is denoted Z_p, and is a subgroup of C_p.

A *p-boundary* is a p-chain that is the boundary of a $p+1$-chain, $\sigma = \partial_{p+1}\tau$, $\tau \in C_{p+1}$. The collection of all p-boundaries is denoted B_p, and is again a subgroup of C_p.

Before we proceed, we have an extremely important lemma. This lemma is what makes homology possible!

Lemma 3.4 (Fundamental Lemma of Homology)**.** $\partial_p \partial_{p+1} \sigma = 0$ *for every integer p and every $p+1$-chain σ.*

We will not go over the proof of this lemma here, but what this means is that every p-boundary is a p-cycle: i.e. B_p is a subgroup of Z_p. We can see that this is true by observing that if we have some p-boundary, then by the Fundamental Lemma of Homology, its boundary will be 0. A diagram of this relationship is shown below.

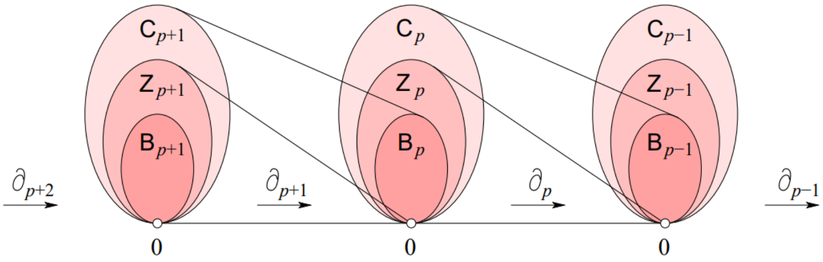

FIGURE 1. A diagram of the subgroup relations of C_p, Z_p, and B_p along with the boundary homomorphisms that connect them.

With these definitions, we can now state what the homology groups are.

Definition 3.5 (Homology Groups)**.** The *pth homology group* is the pth cycle group modulo the pth boundary group, $H_p = Z_p/B_p$. The pth Betti number is the rank of this group, $\beta_p = \operatorname{rank} H_p$.

The rank of a group is simply the smallest number of generators one needs in order to generate the entire group. We note that B_p is a normal subgroup of Z_p since Z_p is abelian, so the quotient group is well defined.

Before we move on to persistent homology, we want to note an interesting consequence of the homology groups. Suppose we have a simplicial map $f : K \to L$. It takes simplices of K to simplices in L. We can extend this to a map $f_\#$ between p-chains of K and p-chains of L. Specifically, if $\sigma = \sum a_i\sigma_i \in C_p$, then $f_\#(\sigma) = \sum a_i\tau_i$, where $\tau_i = f(\sigma_i)$ if $\dim f(\sigma_i) = p$ and $\tau_i = 0$ if $\dim f(\sigma_i) < p$. It turns out that $f_\#$ actually commutes with the boundary operators of K and L, respectively: $f_\# \circ \partial_K = \partial_L \circ f_\#$. This fact is somewhat nontrivial to prove since simplices in K can be sent to simplices of a lower dimension in L.

The point of noting all of this is that because $f_\#$ commutes with the boundary, we can conclude that it takes cycles to cycles and boundaries to boundaries. Therefore, it induces a map between the homology groups of K and L:

Definition 3.6 (Induced Map)**.** Suppose $f : K \to L$ is a simplicial map. Then it induces a map $f_* : H_p(K) \to H_p(L)$ between the pth homology groups by

$$f_*(\sigma + B_p(K)) = f_\#(\sigma) + B_p(L).$$

Note that this map is well defined precisely because $f_\#$ takes cycles to cycles and boundaries to boundaries. This is actually very surprising: if we have any simplicial map f between simplicial complexes, we automatically get a map between their homologies! With this, we can now move on to persistent homology.

4. Persistent Homology and Persistence Diagrams

Normally, we talk about the homology of a single, static object. If we instead want to analyze how the homology of an object changes over time, what tools do we need to do this? It turns out there is a way to do this if you have something called a *filtration* of simplicial complexes. Once you have such an object, you can find its persistent homology groups, which tell you about how the homological features evolve and die over time. This information can be expressed in a persistence diagram, which encodes how long a particular homological feature lasted as a point on a 2D grid. To start off, we have some definitions.

Definition 4.1 (Filtration)**.** A *filtration* is an indexed set S_i of sub objects of a given algebraic structure S, with the index i running over some totally ordered index set I, subject to the condition that if $i, j \in I$, then

$$i \leq j \implies S_i \subseteq S_j.$$

Conceptually, we can imagine a filtration as taking an object and adding pieces onto it to get a bigger object, labeling these snapshots S_i. We can also think of a filtration as a sequence of spaces S_i with inclusion maps $f^{i,j} : S_i \to S_j$ where $i < j$ between the spaces. Now, we want to concentrate on a filtration of simplicial complexes

$$\varnothing = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$$

and the simplicial inclusion maps $f^{i,j} : K_i \to K_j$. Recall that a simplicial map $f : K \to L$ induces a corresponding map between the pth homology groups: $f_* : H_p(K) \to H_p(L)$. Therefore, we can use the inclusion maps to induce a sequence of homology groups

$$0 = H_p(K_0) \to H_p(K_1) \to \cdots \to H_p(K_n) = H_p(K)$$

where each homology group is connected to the next by the induced group homomorphisms $f_p^{i,j} : H_p(K_i) \to H_p(K_j)$. Now we can define the persistent homology groups.

Definition 4.2 (Persistent Homology Groups)**.** The *pth persistent homology groups* are the images of the homomorphisms induced by inclusion, $H_p^{i,j} = \operatorname{im} f_p^{i,j}$, for $0 \leq i \leq j \leq n$. The corresponding pth persistent Betti numbers are the ranks of these groups, $\beta_p^{i,j} = \operatorname{rank} H_p^{i,j}$.

Essentially, what $H_p^{i,j}$ tells you is which homological features in K_i persisted until K_j: hence the name persistent homology. Formally, if $\gamma \in H_p(K_i)$, we say γ is born at K_i if $\gamma \notin H_p^{i-1,i}$, or in other words γ wasn't in $H_p(K_{i-1})$, but appeared in $H_p(K_i)$. We say that if γ was born at K_i, that it dies entering K_j if $f^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$ but $f^{i,j}(\gamma) \in H_p^{i-1,j}$, or in other words, γ was still a significant feature at $H_p(K_{j-1})$, but was absorbed back into the original feature at $H_p(K_j)$. Note that the rules for when a feature is born and dies obeys something called the elder rule:

Definition 4.3 (Elder Rule)**.** At a juncture, the older of the two merging paths continues and the younger path ends.

We can see that if one feature is born earlier than another feature, then when they become part of the same homology group, it is the older feature that we say dies first.

Now that it makes sense to talk about homological features being born, persisting, and dying, we can start to talk about persistence diagrams. What a persistence diagram allows us to do is summarize the details of the persistent Betti numbers of a filtration.

Definition 4.4 (Persistence Diagrams)**.** Define $\mu_p^{i,j}$ to be the number of homology classes that are born at K_i and die entering K_j. Then the *pth persistence diagram* of the filtration is the multiset of points (i, j) with multiplicity $\mu_p^{i,j}$ along with all the points (i, i) on the diagonal with infinite multiplicity.

One can recover the persistent Betti numbers from the persistence diagram due to the following lemma:

Lemma 4.5 (Fundamental Lemma of Persistent Homology)**.** *Let* $\varnothing = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$ *be a filtration. For every pair of indices* $0 \leq k \leq l \leq n$ *and every dimension p, the pth persistent Betti number is* $\beta_p^{k,l} = \sum_{i \leq k} \sum_{j > l} \mu_p^{i,j}$.

So we can get the actual Betti number by looking at the upper left quadrant of the persistence diagram, starting from (k, l), and counting the number of points in that region to get the Betti number.

We can now generalize the definitions we made previously to a simplicial complex K and a real valued map $f : K \to \mathbb{R}$ that satisfies a special condition called monotonicity.

Definition 4.6. A function $f : K \to \mathbb{R}$ is monotonic if when $\sigma, \tau \in K$,

$$\sigma \leq \tau \implies f(\sigma) \leq f(\tau).$$

A function being monotonic is equivalent to the sublevel set $f^{-1}(\infty, a]$ always being a subcomplex of K. What this means in practice is that the sublevel sets of f form a filtration! In other words, we have a series of simplicial complexes

$$\varnothing = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$$

where each K_i corresponds to a specific sublevel set of f. Therefore, it makes sense to talk about the persistent homology and persistence diagrams of f by taking the persistent homology of this filtration. We denote the persistence diagram of a monotonic function like this $\mathrm{Dgm}_p(f)$, as thinking about persistent homology this way will be very useful for the upcoming stability theorems.

5. Stability Theorems

Before we can talk about the stability theorems, we just need to review what the L^∞ distance is.

Definition 5.1 (L^∞ Distance)**.** Suppose $x = (x_1, x_2)$ and $y = (y_1, y_2)$. Then the L^∞ *distance* between the two points is

$$\|x - y\|_\infty = \max\{|x_1 - y_1|, |x_2 - y_2|\}.$$

Suppose $f, g : X \to \mathbb{R}$ are functions. Then the L^∞ *distance* between these functions is

$$\|f - g\|_\infty = \sup_{x \in X}\{|f(x) - g(x)|\}.$$

What the stability theorems are is a way to quantify the idea that small changes made to a function f lead to small changes in the corresponding persistence diagram $\mathrm{Dgm}_p(f)$. But what is a small change in the persistence diagram? In order to quantify this, we will define something called the bottleneck distance between two persistence diagrams.

Definition 5.2 (Bottleneck Distance)**.** Let X and Y be two persistence diagrams. To define the distance between them, consider bijections $\eta : X \to Y$ and record the supremum of the distances between corresponding points for each. Then the *bottleneck distance* between the diagrams is defined as

$$W_\infty(X,Y) = \inf_{\eta: X\to Y} \sup_{x \in X} \|x - \eta(x)\|_\infty$$

We can think of the bottleneck distance between two diagrams as drawing a box with side length $2W_\infty(X,Y)$ around each point in X such that it includes at least one point of Y. Note that the points of X and Y include the points on the diagonal, so these can be used in the pairing between the two diagrams. This idea is illustrated in the diagram below, where the white points are points of X and the black points are points of Y.

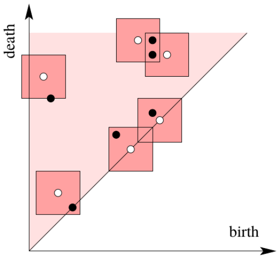

Figure 2. A visualization of the bottleneck distance between the diagrams X (in white) and Y (in black). Note how the points on the diagonal are used.

The reason this distance between persistence diagrams is defined this way is that it allows the stability theorems to become possible. Now we can state the first stability theorem:

Theorem 5.3 (Stability Theorem for Filtrations)**.** *Suppose K is a simplicial complex and $f, g : K \to \mathbb{R}$ are monotonic functions. Then for each dimension p, the bottleneck distance between the diagrams $X = \mathrm{Dgm}_p(f)$ and $Y = \mathrm{Dgm}_p(g)$ is bounded from above by the L^∞ distance between the two functions, $W_\infty(X,Y) \leq \|f - g\|_\infty$.*

This is quite a powerful result, and gives us the sense in which small changes to the function cause small changes in the persistence diagram. To give an idea of the proof, we utilize the straight line homotopy between f and g, $f_t = (1-t)f + tg$, $0 \leq t \leq 1$. In particular, f_t is monotonic for any t, so we can consider the space of persistence diagrams of f_t for $0 \leq t \leq 1$. If we consider the paths that points in the persistence diagrams trace out in this space, it is a polygonal path with one endpoint at the persistence diagram of f and reaching up to a point in the persistence diagram of g or merging with the diagonal of the diagram earlier than that. We can use this to finally derive that the length that this path traces out is bounded by the L^∞ distance between the two functions.

It turns out that this idea can be generalized to not just simplicial complexes and monotonic functions, but triangulable spaces and something we will call a "tame" function. First, what is a triangulable space?

Definition 5.4 (Triangulation)**.** A *triangulation* of a topological space $\mathbb{X}$ is a simplicial complex K that is homeomorphic to $\mathbb{X}$, that is there is a homeomorphism $h : K \to \mathbb{X}$. If a space $\mathbb{X}$ has a triangulation, we say $\mathbb{X}$ is triangulable.

We can redefine many of the ideas for persistent homology in terms of a triangulable space instead of a simplicial complex. Suppose $f : \mathbb{X} \to \mathbb{R}$ is a function. Then in a similar vein to monotonic functions, we can create a filtration using the sublevel sets of f, $\mathbb{X}_a = f^{-1}(-\infty, a]$. Once again, we have inclusion maps $f^{a,b} : \mathbb{X}_a \to \mathbb{X}_b$ for $a \leq b$, which induce group homomorphisms $f_p^{a,b} : H_p(\mathbb{X}_a) \to H_p(\mathbb{X}_b)$ between the homology groups of the sublevel sets. We once again define $H_p^{a,b}$ as the image of the map $f_p^{a,b}$, and $\beta_p^{a,b}$ to be the rank of $H_p^{a,b}$.

In order to define what a tame function is, we first need to define what a homological critical point is.

Definition 5.5 (Homological Critical Point)**.** A point $a \in \mathbb{R}$ is a *homological critical point* if there is no $\epsilon > 0$ for which $f_p^{a-\epsilon, a+\epsilon}$ is an isomorphism for each dimension p.

For $f_p^{a,b}$ to be an isomorphism means that the homology of the sublevel sets did not change from a to b, so a homological critical point is simply a point where the homology changes. Now, we can define a tame function.

Definition 5.6 (Tame Function)**.** A function f is *tame* if it has only finitely many homological critical values and all homology groups of all sublevel sets have finite rank.

The requirements here are very weak: it is only in theoretical cases that the tameness condition will fail. We can now state the more general analogue of the stability theorem for triangulable spaces:

Theorem 5.7 (Stability Theorem for Tame Functions)**.** *Suppose $\mathbb{X}$ is triangulable and and $f, g : \mathbb{X} \to \mathbb{R}$ are tame functions. Then for each dimension p, the bottleneck distance between the diagrams $X = \mathrm{Dgm}_p(f)$ and $Y = \mathrm{Dgm}_p(g)$ is bounded from above by the L^∞ distance between the two functions, $W_\infty(X, Y) \leq \|f - g\|_\infty$.*

6. Topological Data Analysis

The setup for topological data analysis is this: Suppose we have a compact manifold M, and we sample finitely many points uniformly at random off of M: call the collection of these points S. Our question is if we are given S, can we estimate M? What we do in topological data analysis is transform S in such a way that we can find its homology, and use the homological features of S to say something about the homological features of M. This is a better way to look at data because homology is invariant under deformation, so skewing the data in some way will not change its homology significantly.

The first question that immediately comes up is how do we find the homology of a discrete set of points? Its homology is trivial if we do nothing to the data, so we need to transform it somehow. The way we transform this data is we put a closed ϵ-ball around each point of S, and take the union of all these ϵ-balls: we call this $S(\epsilon)$. If we have enough points and the manifold is "nice" enough (positive reach), this ends up being a good approximation to the original manifold M. But we've only talked about simplicial homology: how are we supposed to find the homology of this object? Surprisingly, there is a way to do this with just simplicial homology.

We consider two simplicial complexes that can be formed from our set S: the Čech complex and the Rips complex.

Definition 6.1 (Čech Complex)**.** Let $X = \{x_1, x_2, .., x_n\}$ be a finite set of points sampled from $\mathbb{R}^d$ and $\epsilon > 0$. The *Čech complex* $\check{C}_\epsilon(X)$ is the simplicial complex with vertex set X and n-simplices the subsets $x_0, ..., x_n \subseteq X$ such that

$$\overline{B_\epsilon(x_1)} \cap \overline{B_\epsilon(x_2)} \cap \cdots \cap \overline{B_\epsilon(x_n)} \neq \emptyset$$

where each epsilon ball is defined as $B_\epsilon(x_i) = \{y \in \mathbb{R}^d \mid \|y - x\| \leq \epsilon\}$.

The reason we define this is that we actually have a very surprising result with the Čech complex:

Theorem 6.2 (Nerve Theorem)**.** *The homotopy types of $X(\epsilon)$ and $C_\epsilon(X)$ are the same.*

It turns out that when two topological spaces have the same homotopy type, they have the same homology groups. So if we want to find the homology of $S(\epsilon)$, all we have to do is calculate the Čech complex and find its simplicial homology! Unfortunately, the Čech complex is tough to compute. To tell whether there are any 10-simplices you have to inspect all subsets of size 10. In general, computing the entire Čech complex requires exponential run time in the size of X, which is extremely slow. The next complex is not as precise, but is easier to compute.

Definition 6.3 (Rips Complex)**.** Let $X = \{x_1, \ldots, x_n\}$ be a finite set of points sampled from a compact subset of $\mathbb{R}^d$. The Vietoris-Rips Complex or Rips complex, $V_\epsilon(X)$, is the simplicial complex with vertex set X and n-simplices the subsets $\{x_1, \ldots, x_n\} \subseteq X$ such that $d(x_i, y_i) \leq \epsilon$, $0 \leq i, j \leq n$.

Unfortunately, we don't get the Nerve Theorem for the Rips Complex, as the Rips complex and Čech complex are significantly different for similar ϵ values. What we do get is a good approximation, because of the following containment:

$$C_{\epsilon/2}(X) \subseteq V_\epsilon(X) \subseteq C_\epsilon(X).$$

So what we have so far is that we can estimate the homology of M through S by finding the homology of the Rips complex with S as the vertex set. But we have a problem here: how do we pick the right ϵ that will be the closest to the true manifold M? The solution is that instead of fixing an ϵ, we let ϵ vary and compute the persistent homology of the resulting filtration. Specifically, if we define $d_S : \mathbb{R}^d \to \mathbb{R}$ by

$$d_S(x) = \inf_{s \in S} \|x - s\|,$$

then $d_S^{-1}(-\infty, \epsilon] = S(\epsilon)$: that is, the sublevel sets of the d_S function are the sets $S(\epsilon)$. So we can form a filtration consisting of the sublevel sets of d_S and compute its persistent homology. Furthermore, we are guaranteed by the Stability Theorem for Tame Functions that if $S(\epsilon)$ approximates M well, then the persistence diagram for d_S will approximate the diagram for d_M well.

7. Confidence Sets for Persistence Diagrams

In this section, we describe the method we used for quantifying the uncertainty associated with the persistence diagram we get from our sample. The main tool we use to compute these confidence bands is the `hausdInterval` function from the `TDA` package in R. This function uses the method of sub sampling to compute a confidence interval for the Hausdorff distance between a point cloud and the underlying manifold from which X was sampled. By the stability theorem, this confidence band for the Hausdorff interval is also a valid confidence interval for the persistence diagram generated from the point cloud. The following theorem from [FLR$^+$14] proves that validity of this method.

Let $b = b_n$ be such that $b = o(\frac{n}{\log(n)})$ and $b_n \to \infty$. We draw N subsamples $S^1_{b,n}, \ldots, S^N_{b,n}$ each of size b, from the data where $N = \binom{n}{b}$. Let $T_j = H(S^j_{b,n}, S_n)$,$j = 1, \ldots, N$. Define

$$L_b(t) = \frac{1}{N}\sum_{j=1}^{N} I(T_j > t)$$

and let $c_b = 2L^-1_b(\alpha)$. Recalling the definition of ρ from (7) in [FLR$^+$14],

Theorem 7.1. *Assume $\rho > 0$. Then, for all large n,*

$$P(W_\infty(\hat{\mathcal{P}}(X), \mathcal{P}(X)) > c_b) \leq P(H(S_n, M) > c_b) \leq \alpha + O\left(\frac{b}{n}\right)^{\frac{1}{4}}$$

So the function called by `hausdInterval` will subsample m points S_i from our data S (without replacement) and compute the Hausdorff distance between the original S and the sub sample. The result is a sequence of values $B_i = H(S, S_i)$. Let q be the $1 - \alpha$ quantile of these B values and let $c = 2 * q$. The interval $[0, c]$ is a valid $(1 - \alpha)$ confidence interval for the Hausdorff distance between S and the underlying manifold, as shown by Theorem 7.1. The function `hausdInterval` returns the value c as just described and the confidence interval is $[0, c]$. What makes the stability theorem so nice is it tells us the bottleneck distance between diagrams is less than or equal to the Hausdorff distance between manifolds that generate those diagrams. So once we have a confidence interval for a manifold we have a confidence interval for the persistence diagram generated from that manifold.

8. Application to the Circle

In this section we use a simple synthetic example of a manifold in $\mathbb{R}^2$: namely, the unit circle. Our ability to compute confidence bands for persistence diagrams rests on the stability theorem, so we first verify the stability theorem and then compute confidence bands for our example, all the while visualizing what we are doing. First, we sample 500 points uniformly at random on the unit circle.

To represent noisy data that may be sampled from this unit circle, we will have 2 more data sets. For these data sets, we take points along the unit circle that are perturbed by a value taken from a normal distribution with mean 0 and standard deviation 0.1 and 0.3, respectively.

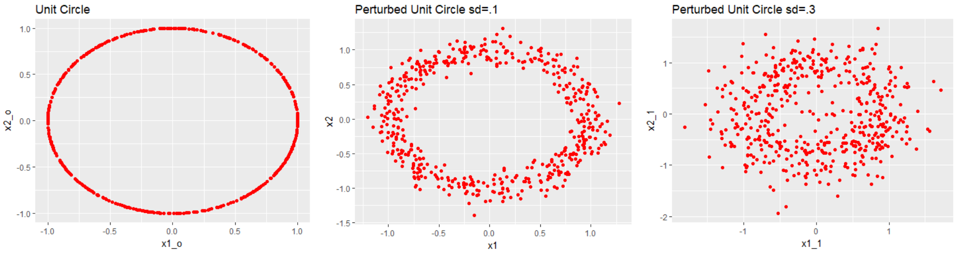

Figure 3. Our samples of 500 points on the unit circle for no perturbation, perturbation with sd=0.1, and perturbation with sd=0.3.

Next we can use the `TDA` package provided in R to compute the homology of these three manifolds and output the persistence diagrams.

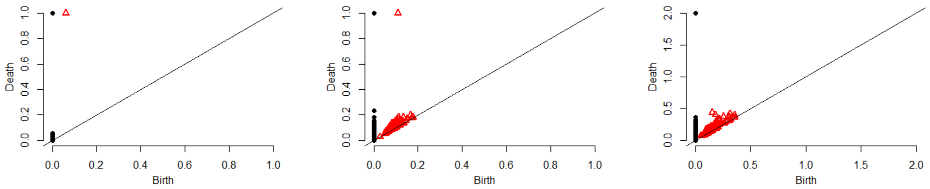

Figure 4. The persistence diagrams for 500 points on the unit circle for no perturbation, perturbation with sd=0.1, and perturbation with sd=0.3.

In the diagrams, black circles represent changes in the 0th homology (connected components) and red triangles represent changes in the 1st homology (1-dimensional holes, like a circle). We say a topological feature is significant if it is far away from the line $y = x$: i.e. it persists for a very long time. For the first two cases, it looks like the diagrams show one connected component and one 1-dimensional hole that are highly significant, since they are very far away from the line $y = x$. In the last case, we only see one connected component and no 1-dimensional hole: the noise is just too overwhelming and we lose that feature. We also see the noisy cases have many features that are born and die quickly and fall along the line $y = x$: because these features do not persist for very long, they are not usually significant topological features. In order to quantify which features are significant and which are not, we will need to compute confidence bands for our diagrams. First, we verify the stability theorem. For this we use only the first and second circles. As noted earlier, the stability theorem when working in a Euclidean metric space simplifies to the bottleneck distance between persistence diagrams must be less than or equal to the Hausdorff distance between the point sets that generated those diagrams. So we first calculate the bottleneck distance between diagrams, then calculate the Hausdorff distance between manifolds, and see if in fact the bottleneck distance is less than the Hausdorff distance.

First we compute the Hausdorff distance between the two data sets. The `TDA` package gives us a function called `distFct` that will compute the one way Hausdorff distance between the data sets. In order to get the two way Hausdorff distance, we use the `distFct` twice: we switch the parameters on the second one and take the maximum of the two. The value we got was 0.40459, which makes sense since we perturbed the first data set by a random value taken from a normal distribution with mean 0 and sd = 0.1. The Hausdorff distance is meant to measure how far two manifolds are from each other, and 0.40459 is about 4 standard deviations and a good guess for the largest of those random values chosen from a normal distribution.

Next using the built in function `bottleneckdist` from the `TDA` package, we find the bottleneck distance between the two diagrams to be 0.04871. Clearly $0.04871 \leq 0.40459$, so the stability theorem is satisfied. Now, we are all set to compute confidence bands.

The method of computing confidence intervals we used relies on the stability theorem, since it relies on the fact that the distance between any two persistence diagrams will be less than or equal to the Hausdorff distance between the sets that generate those diagrams. We bound our confidence bands between diagrams by the Hausdorff distance of the manifolds that generate them. Using the subsampling method from [FLR+14], "we bound $H(S, M)$ to obtain a bound on $W_\infty(P(S), P(M))$. In particular we obtain a confidence set for $W_\infty(P(S), P(M))$ by deriving a confidence set for $H(S, M)$". The `TDA` package gives us a built in function to compute these Hausdorff intervals called `hausdinterval`, which outputs a value c. These intervals can be visualized by placing a box of side length $2c$ around each point on the diagram or by placing a band of width $\sqrt{2}c$ around the line $y = x$. If a point lies inside the region between the band and $y = x$, it is considered noise and not statistically significant. For each diagram the respective c values are as follows: 0.1311, 0.469, and 0.948. The diagrams below show the previously computed persistence diagrams with the confidence band placed on top.

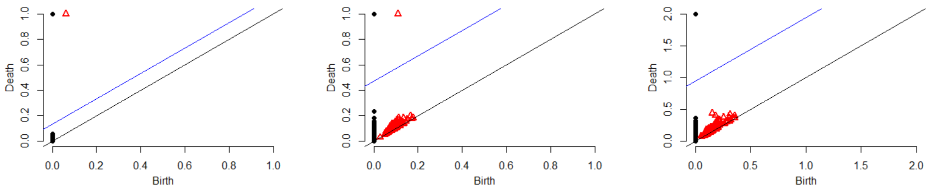

FIGURE 5. The persistence diagrams and confidence bands for 500 points on the unit circle for no perturbation, perturbation with sd=0.1, and perturbation with sd=0.3.

Just like we observed previously, what we find is there is one significant connected component and one significant 2-dimensional hole for the first two diagrams, and only one significant connected component for the third diagram. What can be noticed though is that as the variance in the noisy cases increases so does the size of the confidence band.

References

[EH10] Herbert Edelsbrunner and John L. Harer. *Computational topology: an introduction.* American Mathematical Society, 2010.

[FLR+14] Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *Ann. Statist.*, 42(6):2301–2339, 12 2014.