Persistent

Daniel Collin

March 15, 2021

Contents

1 Introduction

Ordinary statistical analysis and machine learning are often used tools to understand and explore the increasing amounts of data that are present in the modern digital landscape. While these approaches continue to see great success, there is perhaps some value in exploring other avenues in mathematics that could prove useful in understanding data.

Persistent homology, being a tool of topological data analysis, provides a way of quantifying and measuring the global shape of the data, rather than local geometries or patterns. While homology initially might be seen as something esoteric relegated to the realms of abstract mathematics, attempts have been made to use it as a tool for understanding data. Persistent homology is coarse enough to withstand noise that is often present in data (cite), while at the same time sophisticated enough to capture features which are particular to that dataset (cite).

The basic principle is actually quite intuitive. We impose a simplicial complex on the dataset, that in some suitable sense should approximate a reasonable underlying topology in which the dataset lives, and then we compute the homology of this space. However, since there are many ways of approximating a simplicial complex on a set of points we consider not only one simplicial complex but rather a filtration of simplicial complexes parametrized by a given distance.

While the high-level idea is not very complicated, the devil is in the details when it comes to persistent homology. The homology of the filtrations takes us to graded modules and the Structure Theorem for Principal Ideal Domains.

This thesis will serve as both an introduction the workings of persistent homology as well as an example of persistent homology applied to a real dataset (what dataset? TBA. Perhaps something with the insects).

2 Homology

Before go into what *persistent* homology it is well worth our time to clearly state what we mean by homology. (Why? Can this be skipped by experienced readers or are our definitions non-standard? Do we mostly follow hatcher?). In a general sense, homology is a particular of invariant of topological spaces. This has categorical reasons and others. Importantly we need to define simplicial complexes. There are other ways of defining this, notably singular homology, but for the computational aspect of persistent homology we do not have to dwell on this. For completion, we refer the reader to Hatcher for a more traditional treatment of homology.

2.1 Simplices

First we start with the simplex.

Definition 2.1.1. *An n-simplex is the smallest possible convex set in $\mathbb{R}^m$ containing $n+1$ points $v_0, \ldots, v_n$ such that the vectors $v_1 - v_0, \ldots, v_n - v_0$ are linearily independent. The points $v_0, \ldots, v_n$ are known as the vertices of the simplex.*

Definition 2.1.2. *The standard n-simplex is the n-simplex with vertices being the unit vectors along coordinate axes*

$$\Delta^n := \{(t_0, \ldots, t_n) \in R^{n+1} \mid \sum_i t_i = 1, t_i \geq 0 \quad \forall i\}$$

Definition 2.1.3. *A face of a simplex is the convex hull of a subset of its vertices.*

2.2 Simplicial complex

Definition 2.2.1. *A simplicial complex K is a finite collection of simplices such that*

1. *$\sigma \in K$ and $\tau \subset \sigma$ implies that $\tau \in K$*
2. *$\sigma_1, \sigma_2 \in K$ implies that $\sigma_1 \cap \sigma_2$ is either empty or a face of both.*

This is the geometric definition of a simplicial complex. However, since we are working with topological spaces it is advantageous to think of an abstract simplicial complex without concerning ourselves with the geometric connotations:

Definition 2.2.2 (book)**.** *An abstract simplicial complex A is a finite collection of sets such that $\alpha \in A$ and $\beta \subseteq \alpha$ implies that $\beta \in A$.*

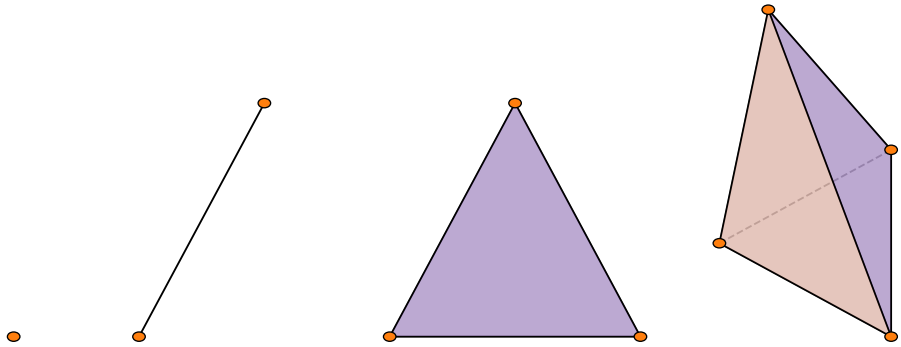


Figure 2.1: *0-simplex (left), 1-simplex (middle left), 2-simplex (middle right) and 3-simplex (right).*

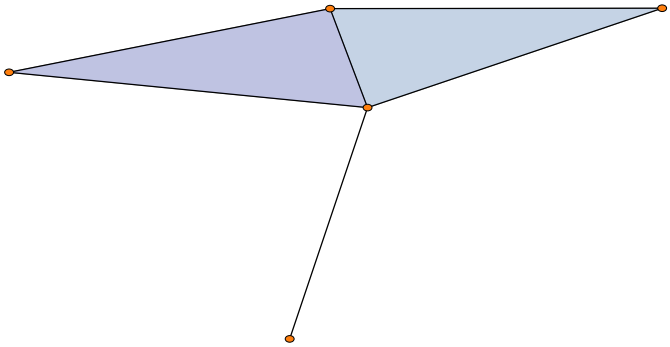


Figure 2.2: *Example of a simplicial complex consisting of two 2-simplices glued together with an attached 1-simplex.*

This abstract definition coincides with the geometric definition by calling the elements of A its simplices. The simplices of A are no longer geometric objects in Euclidean space, but simply combinatorial objects consisting of vertex sets.

It is easy to see how one can go from a geometric simplicial complex to an abstract simplicial complex simply by forgetting everything but the vertices themselves. However, most of the time our interest lies in the opposite direction: how do we go from an abstract simplicial complex to a geometric one? This is done by the geoemtric realization of A.

Theorem 1. *Every abstract simplicial complex of dimension d has a geometric realization in $\mathbb{R}^{2d+1}$.*

Proof. See ?. □

From here on we will simply refer to abstract simplicial complexes as a simplicial complex unless stated otherwise.

2.3 Simplicial homology

For a simplicial complex K of dimension n we define a free abelian group C_k on the oriented $k-$simplices of K. The elements of C_k are called k-chains and are formal sums of the type $\sum \alpha_i \sigma_i$ where α_i are coefficients in some ring R and σ_i are k-dimensional simplices. Furthermore, we have a collection of homomorphisms, known as boundary maps, which together with the groups form a chain complex. The kth boundary map

$$\partial_k : C_k \to C_{k-1}$$

takes a k-simplex to its boundary

$$\partial_k \sigma = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k]$$

where $\hat{v}_i$ signifies that this vertex has been omitted. This is a homomorphism so

$$\partial_k \sum \alpha_i \sigma_i = \sum \alpha_i \partial_k \sigma_i$$

Now a simplicial chain complex is a collection of chain groups together with their corresponding boundary maps as a sequence:

$$\ldots \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} C_{k-2} \xrightarrow{\partial_{k-2}} \ldots$$

Note that the boundary maps compose to become the zero map.

Theorem 2. *The composed boundary map $\partial_{k+1} \circ \partial_k$ is the zero map.*

Proof. (Proof is in Hatcher. Maybe write it down, it's short and shows how boundary maps work.) □

From this definition we know that from every simplicial complex K we can associate a simplicial chain complex (this is a functor). We then define the kth homology group of K as the quotient group

$$H_k(K) = Ker(\partial_k)/Im(\partial_{k+1})$$

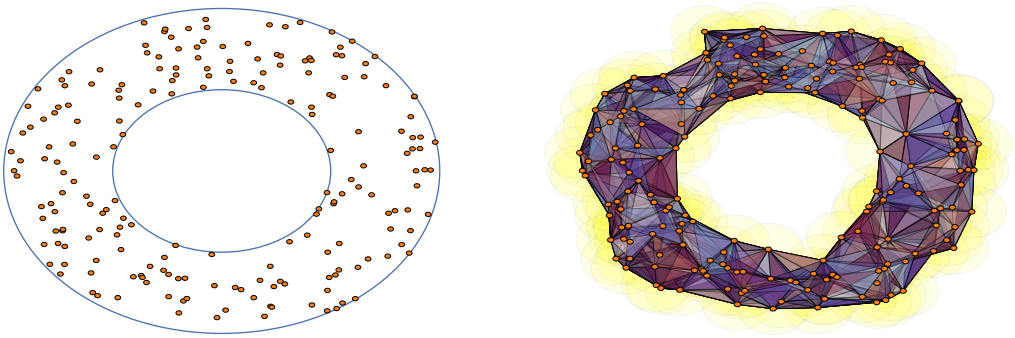
This is also a functor, so any simplicial map $K_1 \to K_2$ induces a map on homology $H_*(K_1) \to H_*(K_2)$.

TODO: Something about the coefficients. Expand on the sections above with examples of what cycles are etc. Describe cubical complexes and how the theory above applies in the same way for these.

3 Persistent

In the world of data we rarely have a topological description of the space our dataset lives in. We could endow our the space which our data lives in with a topology, but just giving it the discrete topology the homology of that space would not be very informative.

What if there is an underlying topological space with a non trivial topology? Consider for example the points sampled from an annulus in Figure 3.1a.



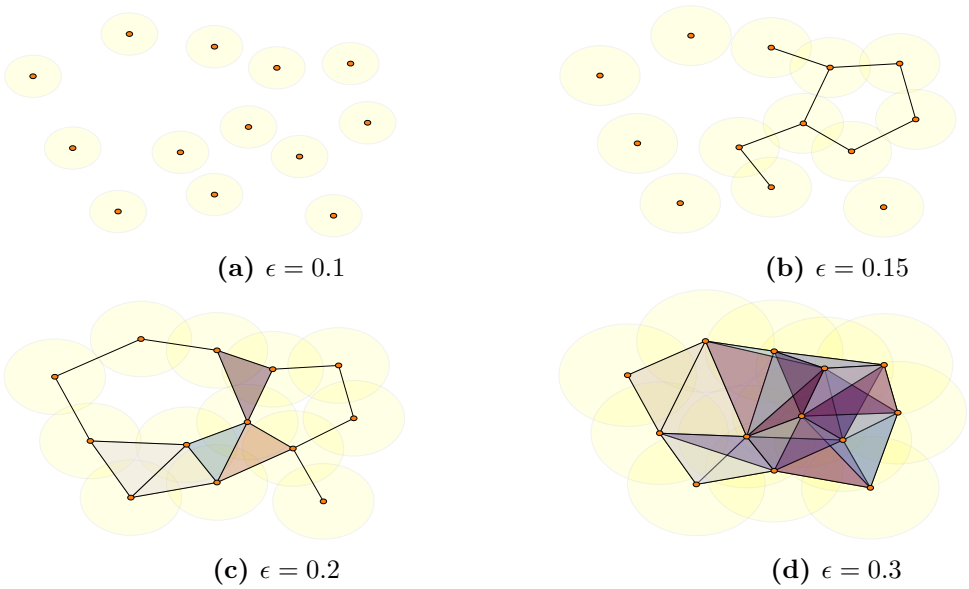
(a) **(b)**

Figure 3.1: *Imposing a simplicial complex* ***(b)*** *on data sampled from an annulus* ***(a)****.*

If we know our space is an annulus we know what the homology is of this space, it contains a single cycle, but with raw data (figure without annulus) this can be harder to tell. This is where persistent homology comes in, a way of gaining information about the homological structure of the data space.

The basic idea is quite simple. Using the theory of simplicial homology we can impose an abstract simplicial complex on our dataset as in Figure 3.1b. A natural way of doing this is defining some form of metric on our space, not necessarily metric in the sense of a metric space, such that when points are sufficiently close to each other we say they belong to the same simplex.

However, there is a problem with the idea in its naive form. How large is "sufficiently close"? If we use too large of a distance we end up with all points in a single simplex and retrieve no valuable homological information. On the other hand, if the distance is too small we end up with a simplicial complex with very few connections between vertices and this too could prove uninformative. Persistent homology addresses this by simply



(a) $\epsilon = 0.1$ **(b)** $\epsilon = 0.15$

(c) $\epsilon = 0.2$ **(d)** $\epsilon = 0.3$

Figure 3.2: *The Vietoris-Rips complex at different ϵ-values.*

considering *all* of them and encoding the lifetime of homological features occuring in something called a *barcode diagram*.

3.1 Views

The perhaps most natural way to impose an abstract simplicial complex on a set of points is the Cech complex

Definition 3.1.1 (Cech complex)**.** *For a given selection of points $\{x_\alpha\}$ in some Euclidean space $\mathbb{R}^n$ the Cech complex C_ϵ is given by the abstract simplicial complex whose k-simplices are given by $k+1$ points in the collection of points whose closed balls of radius $\epsilon/2$ have a point in common.*

The Cech complex is a special case of something called the nerve of a topological space. Through the Nerve theorem (cite) this guarantees that the Cech complex has the same homotopy type as the underlying space given some assumptions (what are they?). A well known result in algebraic topology is that if two spaces have the same homotopy type, they in particular have the same homology groups (cite).

However, the Cech complex is for practical purposes not feasible to compute (cite). The reason being that we need to keep the entire simplicial complex in memory and this can be quite expensive (elaborate this).

A sort of compromise is the Vietoris-Rips complex as seen in Figure 3.2. This complex is a simplification where we do not look for points in common, but rather say that if $k+1$ vertices intersect pairwise they form a k-simplex.

Definition 3.1.2 (Vietoris-Rips complex)**.** *For a given selection of points $\{x_\alpha\}$ in some Euclidean space $\mathbb{R}^n$ the Vietoris-Rips complex R_ϵ is the abstract simplicial complex whose k-simplices are given by $k+1$ points which are pairwise at most ϵ apart.*

The Vietoris-Rips complex does not come with the same guarantee if fidelity to the underlying space as the Cech complex does. However, it is entirely defined by the vertices and the edges of the simplicial complex, allowing it to be stored as a simple graph (elaborate why the edges and vertices are enough).

Given a monotonically increasing sequence of resolutions $(\epsilon_i)_{i\in I}$ we can associate to a finite set of points X the Vietoris-Rips complexes $(R_i)_{i\in I}$. Then there are natural inclusions:

$$R_1 \xhookrightarrow{x} R_2 \xhookrightarrow{x} \dots \xhookrightarrow{x} R_{n-1} \xhookrightarrow{x} R_n$$

We then look at the image of the induced inclusions $x : H_*(R_i) \to H_*R_j$ where $i < j$. These inclusions tell us what homological features persist going from resolution ϵ_i to resolution ϵ_j.

This lends some credibility to the Vietoris-Rips construction as an approximation of the underlying space since it establishes a relationship between it and the Cech complex through a result due to de Silva (cite).

Lemma 3. *Given $\epsilon > 0$ there is a chain of inclusions*

$$R_\epsilon \hookrightarrow C_{\epsilon\sqrt{2}} \hookrightarrow R_{\epsilon\sqrt{2}}$$

This tells us that any feature preserved in the inclusion $R_\epsilon \to R_{\epsilon\sqrt{2}}$ is also present in the Cech complex at resolution $\epsilon\sqrt{2}$ and so in the underlying topological space by theorem ?. In fact, any feature that is preserved up to resolution $\epsilon' \geq \epsilon\sqrt{2}$ is present in the Cech complex at resolution ϵ'. We are now ready to state what persistent homology formally is.

Definition 3.1.3. *Given a persistent complex, a sequence of chain complexes with chain maps from $x : C_*^i \to C_*^{i+1}$ we define the (i,j)-persistent homology $H_*^{i\to j}(C)$, where $i < j$, to be the image of the induced homomorphism on homology $x_* : H_*(C_*^i) \to H_*(C_*^j)$.*

In the case of the obvious filtrations created for Rips or Cech complexes x will be the homomorphism induced by inclusion, but the definition is general enough that this is not necessarily the case.

Definition 3.1.4. *The persistent Betti numbers are given by the ranks of the abelian groups $H_*^{i,j}$, in other words the number of generators of $H_k^{i,j}$ for all k and for all $i < j$.*

Definition 3.1.5. *Let R be a ring. We say R is a **graded ring** if it can be decomposed as*

$$R = \bigoplus_i R_i$$

Note that given a ring R the polynomial ring $R[x]$ is always a graded ring, since it can be decomposed into $R[x] = Rx^0 \oplus Rx^1 \oplus \dots$

Definition 3.1.6. *Let $R = \bigoplus_i R_i$ be a graded ring and M a left $R-$module. We say that M is a **graded R-module** if*

$$M = \bigoplus_i M_i$$

where M_i are abelian subgroups of M, such that $R_i M_j \subseteq M_{i+j}$.

We can now see that a persistent chain complex C_* with coefficients in a ring R can be given a graded module structure by considering the graded ring $R[x]$ where x are the chain maps associated with the persistent chain complex. The monomial $x^k \in R[x]$ sends C_*^i to C_*^{i+k} by k repeated applications of $x : C_*^i \to C_*^{i+1}$ and so we get $Rx^k C_*^i \subseteq C_*^{i+k}$.

Now taking the homology $H_*(C)$ we retain this graded module structure, but it is not necessarily free. However, if we take our coefficient ring R to be a field $\mathbb{F}$ then the Structure Theorem for PIDs (cite) gives us the following result:

Theorem 4. *For a finite persistence module C with coefficients in a field $\mathbb{F}$,*

$$H_*(C;\mathbb{F}) \cong \bigoplus_i x^{t_i} \cdot \mathbb{F}[x] \oplus (\bigoplus_j x^{r_j} \cdot (\mathbb{F}[x]/(x^{s_j} \cdot \mathbb{F}[x])))$$

This theorem in the case of persistence modules has quite an intuitive intepretation. The free part consists of features which appear at resolution indexed by t_i and continue to exist for all future resolutions. The torsion part consists of the features which appear at resolution indexed by r_j and disappear at resolution $r_j + s_j$.

While the restriction to a field $\mathbb{F}$ somewhat limits the usefulness of persistence homology, often in practice we prefer working in $\mathbb{Z}_2$ due to computational aspects and hence it poses no real problem.

3.2 Barcodes

With our algebraic description (in ref theorem above) of persistence we are now able to state the first invariant of persistent homology. This invariant is known as a **barcode**.

This is a visual depiction of $H_*(C;\mathbb{F})$ where each bar depicts the birth and death of a particular generator in one of the homology groups.

Theorem 5. *The rank of the persistent homology group $H_k^{i\to j}(C;\mathbb{F})$ is equal to the number of intervals in the barcode of $H_k(C;\mathbb{F})$ in the interval of parameters $[i,j]$.*

Proof. (TODO: Show this not very difficult proof) □

(TODO: Add a remark here explaining why this is interesting.)

Example. In Figure 3.3 we see a barcode generated from points sampled from an annulus. Note that for small values of ϵ there are many generators of H_0, this is because

the vertices have not been connected into a single component yet. We see that there some short intervals appearing for H_1 at around $\epsilon = 0.3$ and we can see that these are not the hole that would represent the annulus, but rather noise that appears before ϵ has become large enough. But we see at around $\epsilon = 0.6$ that the simplicial complex now captures the shape of the annulus and indeed the barcode reports that we have one generator of H_0, the only connected component, and one generator of H_1 which is the hole in the middle of the annulus. We see that this hole in the middle of the annulus is gone when $\epsilon = 1$ which highlights that it is difficult to find an optimal ϵ.

3.3 Persistence Diagrams

Another way of illustrating persistent homology is the persistence diagram as seen in Figure 3.4. This is an alternative to the barcode in Figure 3.3 where we instead plot the ϵ-value on both axises and the further a point is from the diagonal line the longer the generator in the homology group survived. When we have a lot of birth-death pairs this is a preferable way of visualizing the persistent homology, since unlike the barcode it does not grow vertically with the number of generators.

Just like in the barcode in Figure 3.3 we can see that the only two generators that live for a considerable amount of time is a single connected component in H_0 and a single hole in H_1. This is consistent with the topology that we expect from an annulus.

At around $\epsilon = 0$ we see a lot of H_0 generators being born and dying after another. Since the number of generators of H_0 tells us the number of connected components in the topology this clearly illustrates how the sampled points go from being isolated islands to being incorporated in a larger simplex.

3.4 Metrics

Given that we compute the persistent homology between two spaces, how can we compare them? There are two suitable metrics that are often used for doing this, namely the **q-Wasserstein distance** and the **Bottleneck distance**.

Definition 3.4.1. *The Bottleneck distance between two persistence diagrams X, Y is*

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

Definition 3.4.2. *The q-Wasserstein distance between two persistence diagrams X, Y is*

$$W_q(X, Y) = (\inf_{\beta: X \to Y} \sum_{x \in X} ||x - \beta(x)||^q)^{\frac{1}{q}}$$

The Bottleneck distance is of particular interest since it gives a stability guarantee through the following theorem

Theorem 6. *Given two filtering functions f, g and a simplicial complex K we have that*

$$W_\infty(f, g) \leq ||f - g||_\infty$$

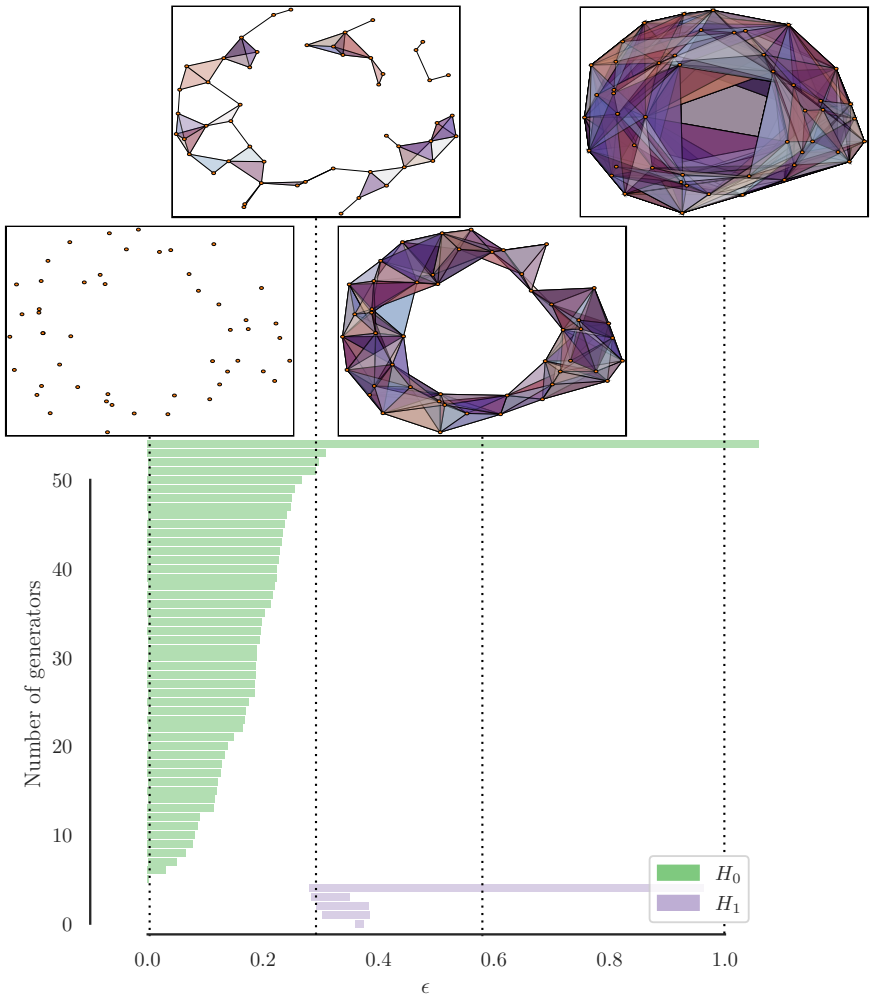


Figure 3.3: *Persistence barcode showing the birth and death of generators in the homology groups of a Vietoris-Rips complex approximated from points sampled from an annulus at different ϵ.*

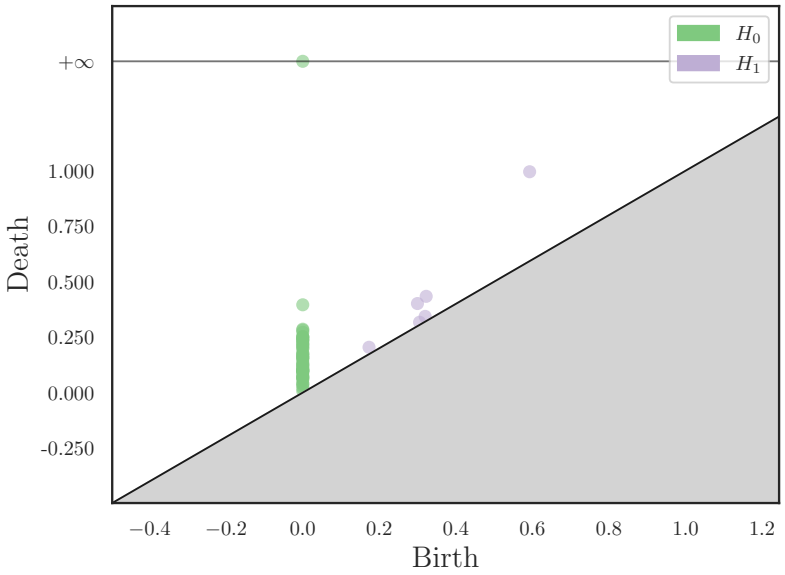


Figure 3.4: *A persistence diagram over the birth and death of generators in the homology groups of a Vietoris-Rips complex approximated from points sampled from an annulus. The closer a point is to the diagonal line the shorter it lived. The diagram is truncated towards infinity, so generators that lived for a long enough time are considered to be at infinity.*

In other words, any small perturbations of the filtering functions will at most be as large as the difference between the functions themselves.

3.5 Computation of

Some aspects of the computational part of this. How is it done in practice? Mention an example but do not dwell too much on this. Perhaps go into smith normal form and how it all translates to linear algebra?

4 Analysis

It is hypothesized that the size of a Bombus Terrestris has an impact on its behavior (cite Emily et al). In particular, it is likely that this difference in behavior translates to a difference in the shape of the eyes of Bombus Terrestris. We investigate this hypothesis as a means of showcasing the use of persistent homology in the wild in order to see whether the topological summaries given by persistent homology support this hypothesis.

4.0.1 Data

The data consists of binary 3D volumes of the corneas acquired by microCT scans of different species. The main focus of the analysis will be Bombus Terrestris, but in total there are 20 samples consisting of ? different species. We use the additional samples to act as a control upon our topological findings. See figure ? for a full breakdown of the samples.

4.0.2 Methodology

Since our samples consist of binary 3D volumes they can be considered as cubes in a 3-dimensional grid, where a value of 1 indicates the presence of a cube and a value of 0 indicates the absence of one. We can exploit this inherent structure in the data and instead of considering simplicial complexes as described in Section ? we can instead impose the structure of a *cubical complex* on the data samples. The benefits are computational.

Cubical complexes

Describe cubical complexes. Describe how they work. Describe how nothing changes other than the complex construction: from there on the results and definitions regarding persistent homology are the same.

Lore Lore Lorem ipsum.

We also need to impose a metric structure on the space given by each sample. If we only consider the binary situation then threshold in which persistent homology examines the sample will just lead to all cubes appearing at a threshold of 1. Instead, we consider the Euclidean Distance Transform (EDT).

Definition 4.0.1. *The Euclidean Distance Transform is a function from* $Z \times Z \to \mathbb{R} \times \mathbb{R}$ *in the following way: Give each n-dimensional cube a value depending on how far away it is from the background (where values are 0).*

2-dimensional example here.

In order for not all topological features to appear at the same time we slightly perturb each value with some random uniform noise. In Figure ? we see what the sample ??? looks like at different thresholds.

Our filtration then, as the threshold increases, will describe the structure starting at the densest parts of the cornea and expanding until the entire cornea is considered. The resulting barcode then describes to us the local geometries and the innermost levels of the cornea as well as the global topology when the entire cornea is considered. Our reason for doing this is because the cornea considered as a whole has a trivial topology: it has no holes and consists of a single connected component. But this allows us to look at it at different scales.

The resulting topological summaries we find are barcodes. While these are in themselves interesting, in order to answer whether the size of a Bombus Terrestris has an impact on the topology of its cornea we compare the samples in a distance matrix where the metric is the 1-Wasserstein distance. We choose the Wasserstein distance because it is sensitive to small changes in the persistence diagrams whereas the bottleneck distance only considers the largest differences.

We then analyze this distance matrics using standard tools for data analysis. We are interested in two things:

1. is there a correlation between the size of the bumblebees and their persistent homology?

2. is persistent homology able to identify subgroups of bumblebees, and if so is it related to their size?

Answering these questions will allow us to evaluate whether persistent homology provides information which relates to the hypothesis. In order to check the correlation we additionally compute the distance matrix between the samples' ITW (whatever this abbreviated, the size between wings in insect morphology) and use the Mantel test to see the correlation between the wasserstein distance matrix (describe mantel test). In order to identify groupings in the samples we use hierarchical clustering (describe this as well).

4.0.3 Results

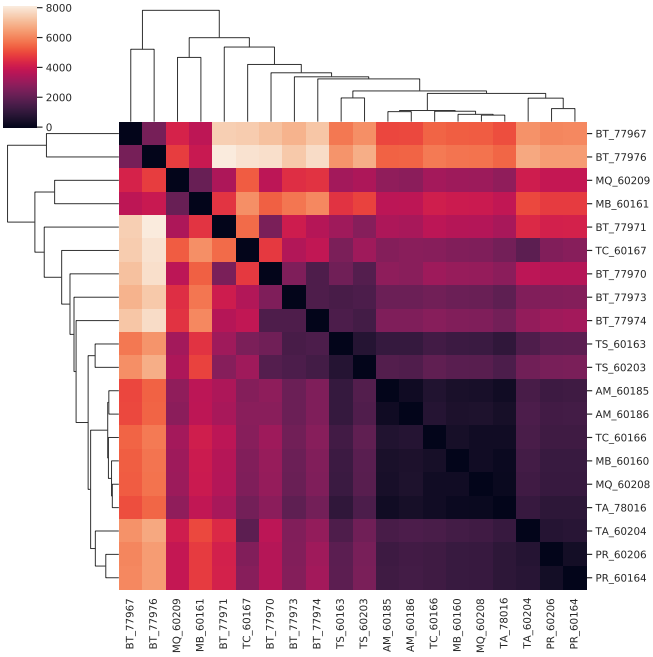


Figure 4.1

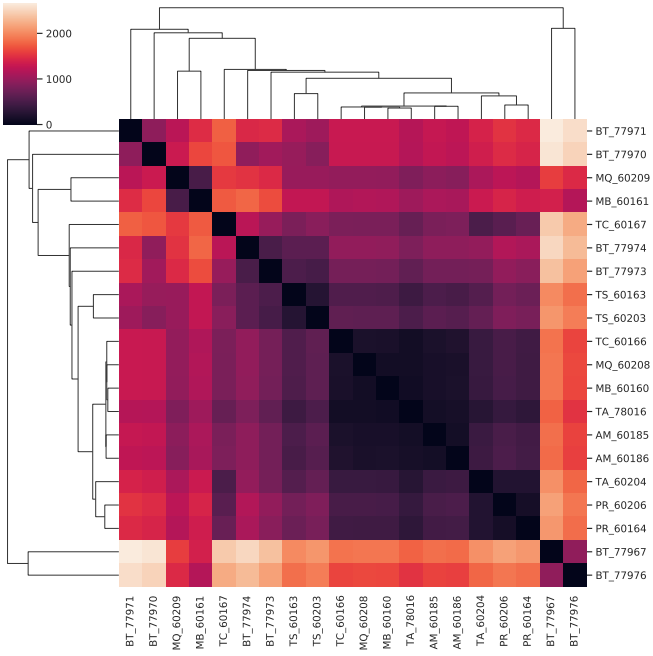


Figure 4.2

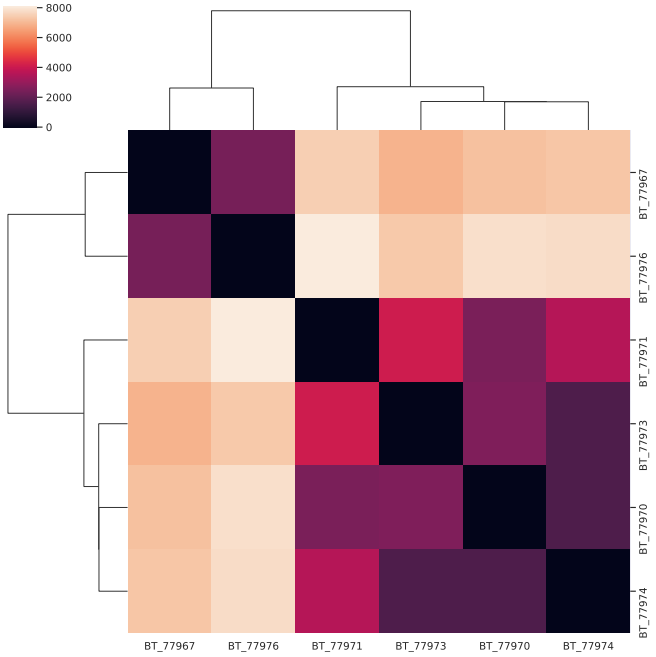


Figure 4.3

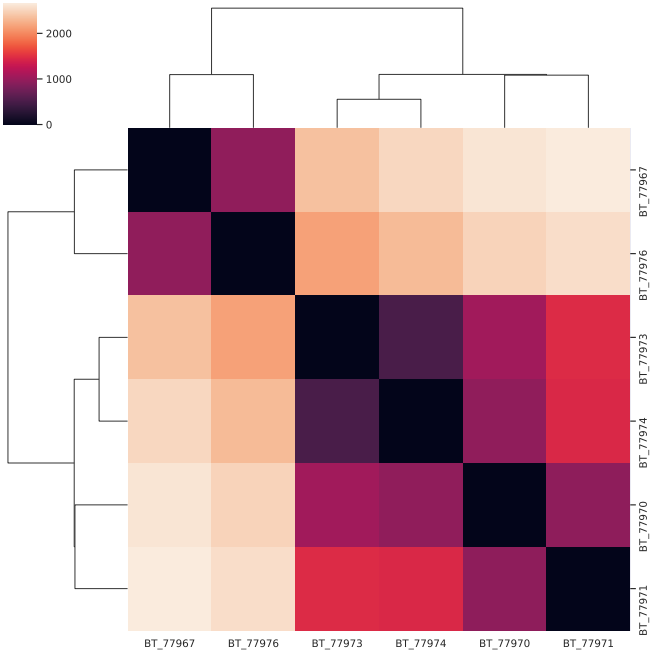


Figure 4.4

4.1 Brain network (work in progress)

(WIP) In this case analysis our main object of study will be two synthetic networks generated based on the striatum. One network consists of 50 001 vertices and the other network of 999 vertices. It has to be said that the smaller network suffers from being too small to give even the most central neurons all the neighbours it should have.

Definition 4.1.1. *A directed clique is a directed graph $G = (V, E)$ such that every vertex has at least an outgoing or incoming edge to every other vertex in the graph.*

Definition 4.1.2. *Let G=(V,E) be a directed graph. The directed flag complex dFl(G) is defined to be the ordered simplicial complex whose k-simplicies are all directed cliques with vertices $v_0, \ldots, v_k$ such that $\forall i : v_i \in V$ and $\forall i, j : i < j \implies (v_i, v_j) \in E$. The vertices v_0, v_k are called the source and the sink of a k-simplex.*

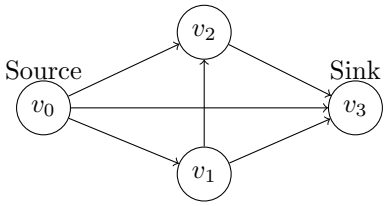


Figure 4.5: *A 3-simplex in a directed flag complex.*

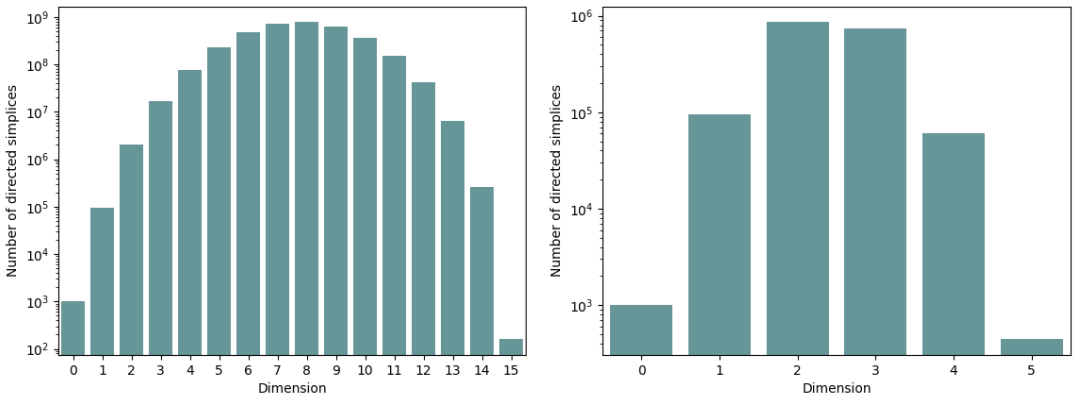


Figure 4.6: *The number of simplices in each dimension for the directed flag complex generated by (a) synthetic network from Snudda, (b) random network generated with the same edge probability creation as the first network, both with 999 vertices.*

In Figures 4.6 and 4.7 we see that the synthetic brain networks have much more higher order structure in terms of high dimensional simplices than a network generated solely

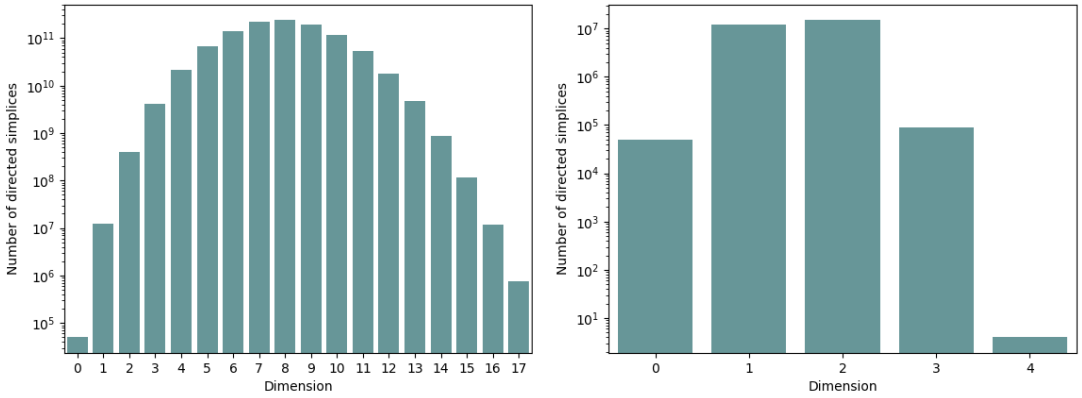


Figure 4.7: *The number of simplices in each dimension for the directed flag complex generated by (a) synthetic network from Snudda, (b) random network generated with the same edge probability creation as the first network, both with 50 001 vertices.*

based on edge connectivity. For instance, we see in 4.7 the presence of 17-dimensional cells in the synthetic network, which means directed cliques consisting of 18 participating neurons, whereas in the random network we see at most 4-dimensional cell.

In other to further investigate these higher order cells in the synthetic networks we can look at their persistent homology. However, a priori the directed brain network does not have any weights, and so it is not obvious what a filtration $f : V \to \mathbb{R}_+$ would look like. So we impose a metric space structure on the directed graph by giving the value of a directed edge between two vertices the Euclidean distance between the two neurons in the simulated model. This means that at low threshold values the filtration will only look at connections made by neurons very close to each other, but as the threshold increases we look at a larger and larger part of the network.

So what is a generator of a homology group in a brain network? It would have to be a k-simplex which is not the boundary of a $k+1$-simplex, which translated to the brain network means a clique of neurons that are in themselves an isolated source-sink network and not part of any other network.

Due to computational aspects it is not feasible to compute the persistent homology of the synthetic network with 50 001 vertices, so we restrict ourselves to a subnetwork of the full network consisting only of dSPN neurons as seen in Figure 4.8. We also look at a full synthetic network generated with only 999 vertices in Figure 4.9.

We see that the formation of higher order (> 5) homology generators mostly happens over small distances, which reaffirms the notion of the brain having a small-world structure.

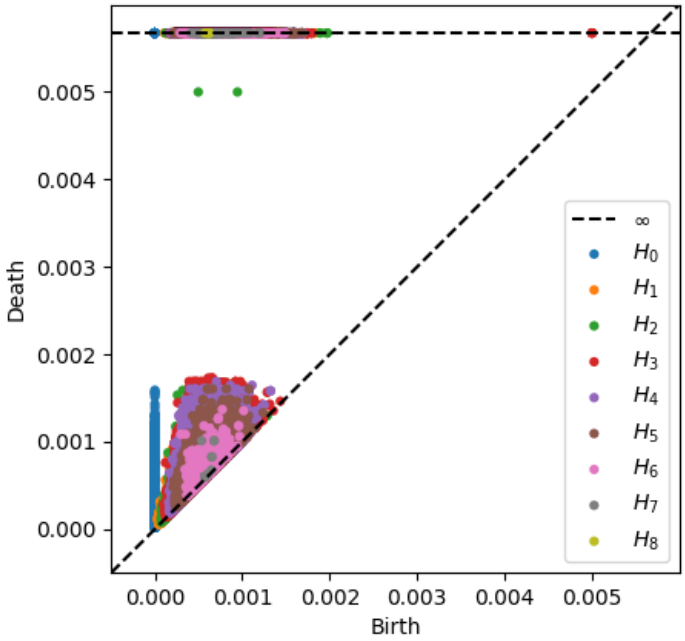


Figure 4.8: *Persistence diagram of the subnetwork of dSPNs extracted from a synthetic network of 50 001 vertices.*

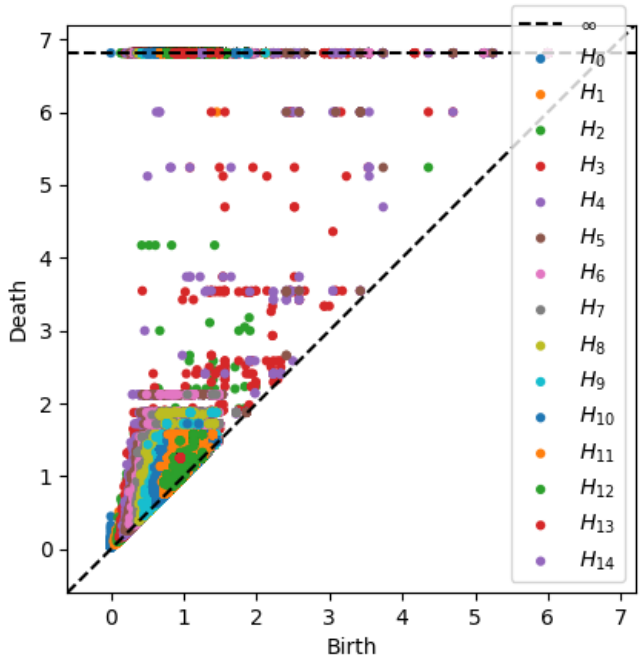


Figure 4.9: *Persistence diagram of the entire synthetic network consisting of 999 vertices. (this is scaled 1000 larger than in actual data, generate new diagram)*