Persistent

Daniel Collin

March 29, 2021

Abstract

PLACEHOLDER

Acknowledgements

PLACEHOLDER

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. Broadly speaking, homology is an invariant of topological spaces which is concerned with cycles in the space which are not boundaries. Or more abstractly, homology captures the notion n-dimensional holes in the space. In *persistent* homology we generally work without predefined topological spaces and start with a basic dataset which at most contains some metric structure. Therefore the classical definitions involving singular homology are not something we will dwell on, but rather we refer the reader to Hatcher's excellent exposition in [4].

The main concept we will be working with is simplicial homology since computationally we can approximate a simplicial complex on our topological spaces. This way we can compute the homological invariants without any a priori information about what topological space we are working in.

2.1 Simplices

We start with what will constitute our atoms in simplicial homology, namely the simplices.

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

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

As seen in Figure 2.1 the $0, 1, 2$ and 3-dimensional simplices are familiar shapes consisting of vertices, edges, triangles and tetrahedrons.

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

Since higher dimensional simplices are made up of simplices of lower dimensions we can always decompose a simplex into its faces, in other words the lower dimensional simplices that make up the simplex. For example, a 2-simplex can be decomposed into the three edges that make up the triangle.

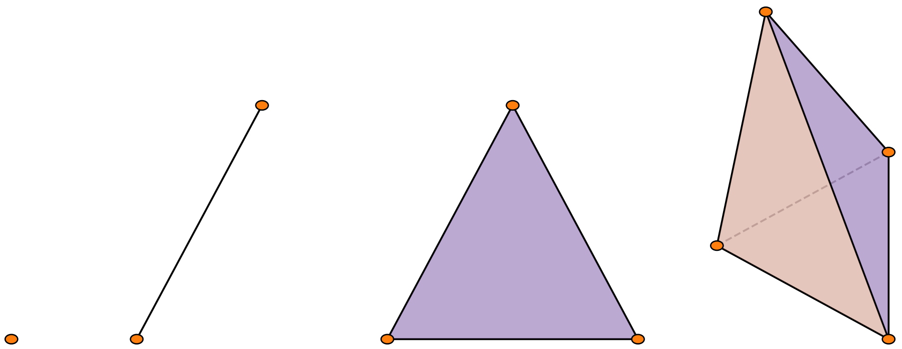


Figure 2.1: *0-simplex (left), 1-simplex (middle left), 2-simplex (middle right) and 3-simplex (right).*

2.2 Simplicial complex

By gluing together simplices at their faces as seen in Figure **??** we can construct higher-order objects which we call simplicial complexes.

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 a 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. It is possible to define a simplicial complex by only considering the ordering of the vertices and what higher-dimensional simplices they make up:

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 $\beta \in A$.

Definition 2.2.3 (book)**.** An (finite) abstract simplicial complex K is a collection of (finite) sets such that $\alpha \in K$ and $\beta \subseteq \alpha$ implies that $\beta \in A$.

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

Furthermore, we need an ordering on the simplicial complex.

Definition 2.2.4. An **ordered abstract simplicial complex** is an abstract simplicial complex together with a partial order $(\geq)$ which restricts to a total order on each simplex.

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

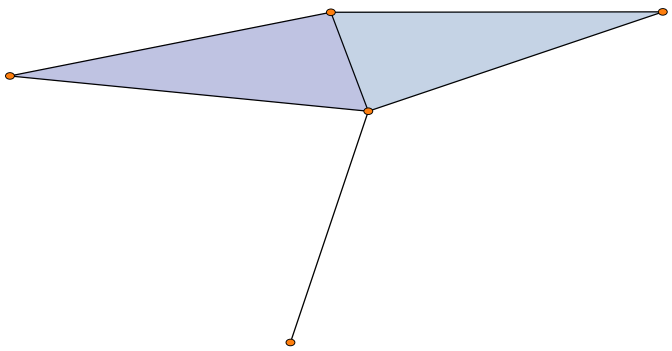


Figure 2.2: *Example of a simplicial complex consisting of two 2-simplices glued together with an attached 1-simplex.*

2.3 Simplicial homology

Homology can generally thought of as being the characterization of cycles which are not boundaries. In the context of simplicial complexes we first need some machinery in order to define exactly what cycles and boundaries are.

Definition 2.3.1 ([11])**.** The kth **chain module** $C_k(K)$ on a simplicial complex K is the free module with basis given by the k-dimensional simplices in K with coefficients in some ring $\mathcal{R}$ with additive unit 0 and multiplicative unit 1. In other words, the elements of $C_k(K)$ are formal sums

$$\sum_{i=1} r_i \sigma_i$$

where $r_i \in \mathcal{R}$ and σ_i is a k-dimensional simplex in K.

Definition 2.3.2 ([10, p. 2])**.** A **chain complex** over a ring R is a family of R-modules $\{C_k\}_{k\in\mathbb{Z}}$ with R-module maps $\partial_k : C_i \to C_{k-1}$ such that the composition $\partial_{k-1} \circ \partial_k$ is the zero map. We call the maps ∂_k the **differentials** of the chain complex.

Theorem 1. *Given a simplicial complex K and a ring $\mathcal{R}$ the sequence of chain modules*

$$\dots \xrightarrow{\partial_{k+1}} C_k(K) \xrightarrow{\partial_k} C_{k-1}(K) \xrightarrow{\partial_{k-1}} C_{k-2}(K) \xrightarrow{\partial_{k-2}} \dots \xrightarrow{\partial_1} C_0(K)$$

with differentials defined as

$$\partial_k : C_k(K) \to C_{k-1}(K)$$
$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_k]$$

is a chain complex.

Proof. It suffices to show that $\partial_{k-1} \circ \partial_k$ is the zero map. (Proof missing, TODO) □

Example 2.3.1. Given a simplicial complex K consisting of a triangle without interior as in Figure 2.3, a chain in $C_1(K)$ would be a linear combination of edges. For example, an element of $C_1(K)$ is $[v_0, v_1] + [v_1, v_2]$ which is highlighted in green.

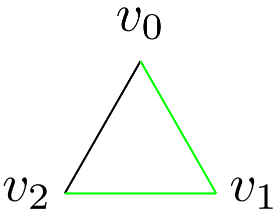


Figure 2.3: *A simplicial complex in which the 1-chain* $[v_0, v_1] + [v_1, v_2]$ *is highlighted in green.*

Example 2.3.2. Given a 2-simplex as in Figure 2.4 we get the differential of the interior of the simplex as

$$\partial_2([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_1, v_2]$$

which is the boundary of the simplex. For this reason we refer to the differential of a chain complex of a simplicial complex as the **boundary map** or **boundary operator**.

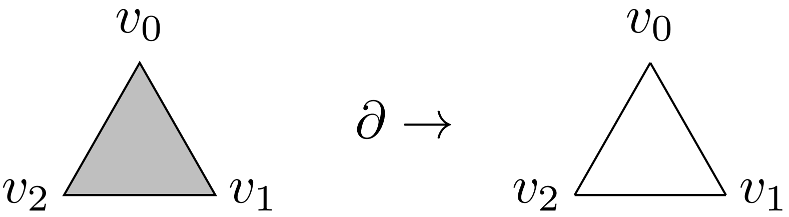


Figure 2.4

We are now ready to state the definitions of cycles and boundaries.

Definition 2.3.3 ([10, p. 4])**.** Given a chain complex C_* the k**-cycles** Z_k and the k**-boundaries** B_k of K are the $\mathcal{R}$-modules

$$Z_k := \ker \partial_k$$

$$B_k := \text{im } \partial_{k+1}$$

As mentioned our purpose of constructing this simplicial chain complex is to identify cycles which are not boundaries of higher dimensional simplices. Hence, a vital result is this corollary of (cite thm above).

Corollary 1.1. *The k-boundaries are a submodule of the k-cycles.*

Proof. Let $\sigma \in B_k = \operatorname{im} \partial_{k+1}$ then for some $\tau \in C_{k+1}$ we have that $\partial_{k+1}(\tau) = \sigma$. Hence,

$$\partial_k(\sigma) = \partial_k \partial_{k+1}(\tau) = (\partial_k \circ \partial_{k+1})\tau = 0$$

and so $\sigma \in \ker \partial_k = Z_k$. □

This tells us that there are cycles which are not boundaries and cycles which are. This motivates the following definition of homology.

Definition 2.3.4. Given a simplicial chain complex C_* the homology group H_k is defined as

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

Hence, the kth homology group captures precisely those cycles which are not in the image of the higher dimensional differential. In other words, the non-trivial elements are the cycles which are not boundaries.

Example 2.3.3. Figure 2.5 of square + triangle, with triangle filled in. Let us compute

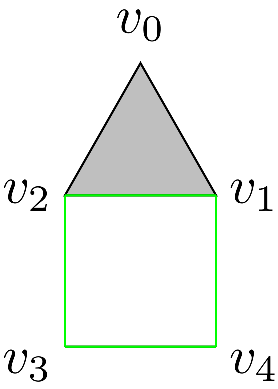


Figure 2.5

the homology of this simplicial complex consisting of a 2-simplex glued together with a square made of 1 simplices.

There are two 1-cycles, the boundary of the filled triangle and the square. Since the boundary of the filled triangle is in $Im(\partial_2)$ we get that this element belongs to the trivial class in H_1. But there is one non-trivial element given by the square without interior. So H_1 contains one non-trivial element generated by the cycle in green.

Definition 2.3.5. The kth Betti number is the rank of the module H_k.

3 Persistence

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.

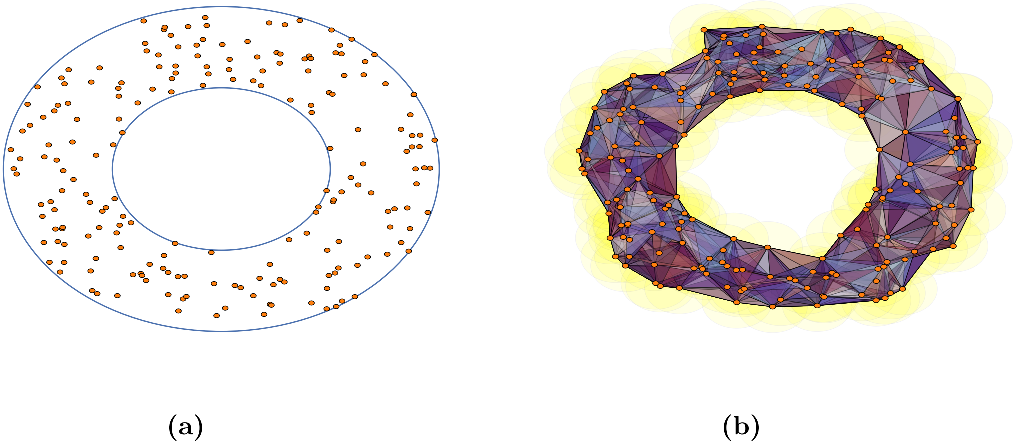
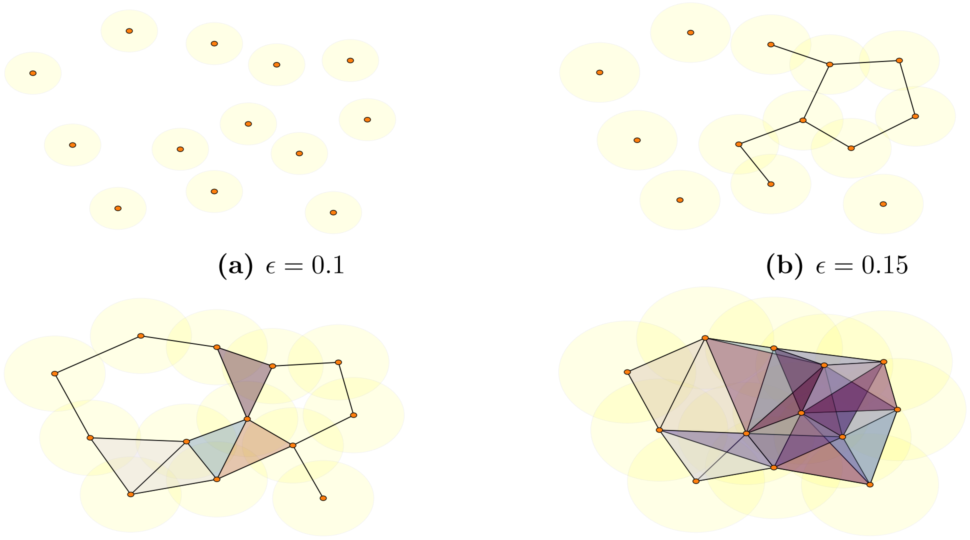


Figure 3.1: *Imposing a simplicial complex* ***(b)*** *on data sampled from an annulus* ***(a)****.*

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 Filtrations

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 2. *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 formally what persistent homology is. While the Vietoris-Rips complex is a common used way of approximating a simplicial complex on a data-set we can be more general than that.

Definition 3.1.3. A **filtration** of a simplicial complex K is a totally ordered set of subcomplexes $K^i \subseteq K$ for $i \in \mathbb{N}$ such that if $i \leq j$ then $K^i \subseteq K^j$.

Note that the Cech and Vietoris-Rips complexes are two instances of filtrations, but with this definition we are not restricted to them alone. This means we can compute persistent homology for *any* filtration.

Definition 3.1.4. For $p > 0$ the **p-persistent kth homology module** of K^i is given as

$$H_k^{i,p} = Z_k^i / (B_k^{i+p} \cap Z_k^i)$$

This module is well-defined since the inclusion $K^i \hookrightarrow K^{i+p}$ induces inclusions $K_k^i \hookrightarrow K_k^{i+p}$ hence we have inclusions $Z_k^i \hookrightarrow K_k^i \hookrightarrow K_k^{i+p}$ and so Z_k^i is a submodule of K_k^{i+p}.

Since we can associate each simplicial subcomplex K^i in a filtration with a simplicial chain complex K^{i_*} we can associate with each filtration a complex of simplicial chain complexes.

3.2 Persistence Module

While we have arrived at a definition (refer) that serves as a sufficient framework for persistent homology, it is still particular in the sense that we are talking about simplicial complexes and filtrations on them. It is possible to make the notion of persistent homology even more general which allows us to understand its algebraic structure even better. This does not mean that we should entirely discard our anchoring of persistent homology in the realm of simplicial complexes, as it relates closely to how we will do persistent homology in practice, but rather we should let this more abstract approach serve as the theoretical underpinning which opens up the possibility of other types of approximation of data than simplicial complexes.

In order to transition from simplicial complexes to a more general framework we need some definitions that has us end up in a less particular construction of persistent homology.

Definition 3.2.1 ([10, p. 2])**.** Let C_*, D_* be chain complexes over some ring R. A **chain map** $u : C_* \to D_*$ is a family of R-module homomorphisms $u_k : C_k \to D_k$ such that the following diagram commutes

$$\begin{tikzcd}
\cdots \arrow[r, "\partial_{k+2}"] & C_{k+1} \arrow[r, "\partial_{k+1}"] \arrow[d, "u_{k+1}"] & C_k \arrow[r, "\partial_k"] \arrow[d, "u_k"] & C_{k-1} \arrow[r, "\partial_{k-1}"] \arrow[d, "u_{k-1}"] & \cdots \\
\cdots \arrow[r, "\partial_{k+2}"] & D_{k+1} \arrow[r, "\partial_{k+1}"] & D_k \arrow[r, "\partial_k"] & D_{k-1} \arrow[r, "\partial_{k-1}"] & \cdots
\end{tikzcd}$$

Definition 3.2.2. A **persistence complex** is a family of chain complexes C_*^i together with chain maps $\iota^i : C^i \to C^{i+1}$ that go between them in the following way

$$\begin{tikzcd}
 & \vdots \arrow[d, "\partial_{k+2}"] & \vdots \arrow[d, "\partial_{k+2}"] & \\
\cdots \arrow[r, hook, "\iota^{i-1}"] & C^i_{k+1} \arrow[r, hook, "\iota^i"] \arrow[d, "\partial_{k+1}"] & C^{i+1}_{k+1} \arrow[r, hook, "\iota^{i+1}"] \arrow[d, "\partial_{k+1}"] & \cdots \\
\cdots \arrow[r, hook, "\iota^{i-1}"] & C^i_k \arrow[r, hook, "\iota^i"] \arrow[d, "\partial_k"] & C^{i+1}_k \arrow[r, hook, "\iota^{i+1}"] \arrow[d, "\partial_k"] & \cdots \\
 & \vdots & \vdots &
\end{tikzcd}$$

Definition 3.2.3 ([11])**.** A **persistence module** M is a family of R-modules M^k together with module homomorphisms $\phi : M^k \to M^{k+1}$.

With the definition of the persistence module we arrive at an alternate definition of persistent homology, the persistent homology of a persistence complex.

Definition 3.2.4. (Alternate Definition) For $p > 0$ the p-persistent homology of a persistence complex (C_*, ι) is denoted H_*^p and is defined to be the images of the induced homomorphisms $\iota_*^{p-1} \circ \iota_*^{p-2} \circ \cdots \circ \iota_*^i : H_*(C_*^i) \to H_*(C_*^p)$.

In the light of this definition, we see that the p-persistent homology of a persistence complex is a persistence module where the module homomorphisms are the maps induced by the chain maps $\iota : C_*^i \to C^{i+1}$. The constructions given by this definition (refer) and the previous definition (refer) are in fact isomorphic

Lemma 3. *Let $\iota_k^{i,p} : H_k^i \to H_k^p$ be the module homomorphism that takes a class in H^i to the class which contains that class in H^p. Then $Im(\iota_k^{i,p}) \simeq H_k^p$.*

Proof. Note that the kernel of $\iota^{i,p}$ are exactly those classes of cycles which become boundaries at some index $i, i+1, \ldots, p$, hence $\ker(\iota^{i,p}) = (B^{i+p} \cap Z^i)$. So by the first isomorphism theorem for modules we get that

$$Im(\iota^{i,p}) \simeq H_k^i / \ker(\iota^{i,p}) \simeq H_k^i/(B^{i+p} \cap Z_k^i) \simeq (Z_k^i/B_k^i)/(B_k^{i+p} \cap Z_k^i) \simeq H_k^p$$

where last isomorphism follows from the fact that $B_k^i \subseteq B_k^{i+p} \cap Z_k^i$. □

Definition 3.2.5 ([11])**.** We say a persistence complex (C_*^k, f^k) (persistence module (M^k, ϕ^k)) is of **finite type** if each component C_n^k (M_n^k) is a finitely generated R-module and the maps f^k (ϕ_k) are isomorphisms for $k \geq N$ for some integer N.

When we start with a finite simplicial complex K we get that $C_*(K)$ is consists of finitely generated R-modules, since the number of simplices in each dimension is finite, and so the resuling persistence complex and persistence modules are of finite type.

The most important theoretic result is just around the corner, but before that we need to recall some definitions regarding graded rings and modules.

Definition 3.2.6. 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 $R[x] = Rx^0 \oplus Rx^1 \oplus \ldots$

Definition 3.2.7. 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 submodules of M, such that $R_i M_j \subseteq M_{i+j}$.

We can now see that if we have a persistence module M over some ring R and we give R a graded structure by considering $R[t]$ then a graded module structure on M is given by

$$\alpha(M) = \bigoplus_{k=0}^{\infty} M^k$$

The monomial t^p sends $M^k \to M^{k+p}$ by p repeated applications of t, in other words t shifts the elements up in the graduation by its power

$$t \cdot (m^0, m^1, m^2, \dots) = (0, \phi^0(m^0), \phi^1(m^1), \phi^2(m^2), \dots)$$

and so we get that $R[t]_p M^k = Rt^p M^k \subseteq M^{k+p}$ which satisfies the condition we gave in our definition of a graded module.

Note that by taking the homology of a persistence complex we get a persistence module, and so the graded structure α on p-persistent homology is well-defined. However, it is not necessarily the case that the graded persistence module retains the free structure of the persistence complex. In fact, classifying the p-persistent homology modules over an arbitrary ring R turns out to be equivalent to classifying finitely generated non-negatively graded $R[t]$-modules, which is known to be a hard problem. By restricting our ring R to be a field, however, we get the following result:

Theorem 4 ([11]). *For a persistence complex C_* of finite type over a field $\mathbb{F}$,*

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

Proof. See (cite) for now. Maybe we will bring in this proof, but it requires the structure theorem for PIDs. □

This theorem has an intuitive explanation in terms of filtrations: the free part consists of generators which appear at filtration t_i and continue to exist for all future filtrations. The torsion part consists of the generators which appear at filtration indexed by r_j and disappear at filtration indexed by $r_j + s_j$. Note how the decomposition provides the p-persistent homology for all p and so satisifies our initial problem of how we can decide at which granularity we wish to perform a filtration by considering all of the filtrations at the same time.

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

3.3 Visualizing Persistence

3.3.1 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.

3.3.2 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

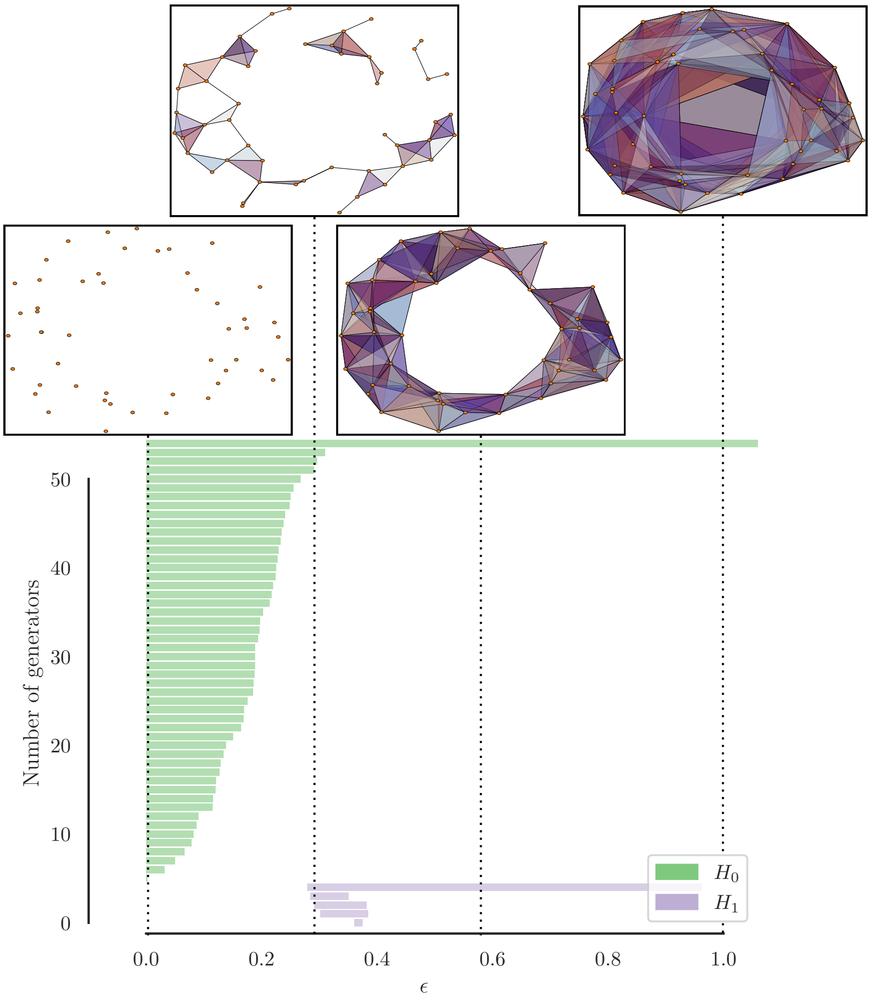


Figure 3.3: *Persistence barcode showing the birth and death of generators in the homology groups of a Vietoris-Rips complex approximated from points sampled from an annulus at different* ϵ*.*

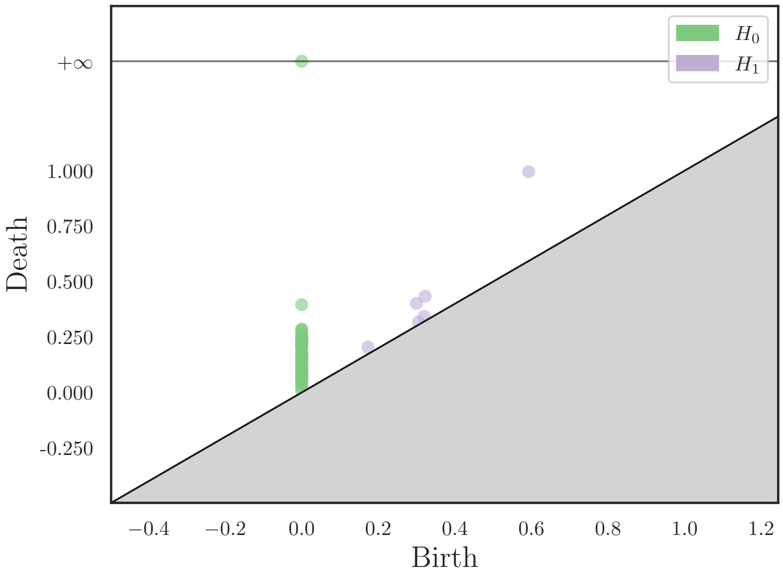


Figure 3.4: *A persistence diagram over the birth and death of generators in the homology groups of a Vietoris-Rips complex approximated from points sampled from an annulus. The closer a point is to the diagonal line the shorter it lived. The diagram is truncated towards infinity, so generators that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

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

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

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 Two Applications of Persistent Homology

Since our purpose with thesis is not only to give an introduction to persistent homology in terms of theory, but also display how it can be used with actual real-world data, we have investigated two different situations where the theory we have expanded upon so far serve as our main tool for data analysis.

In the first case we quantify differences in morphology between different-sized individuals of the bumblebee *Bombus terrestris* by computing the persistent homology of 3D volumes of their corneas. To our knowledge this is the first use of persistent homology in data pertaining to insects, although in [7, 1] materials and in [3, 2] reconstructions of 3D volumes are investigated with approaches that are similar in spirit.

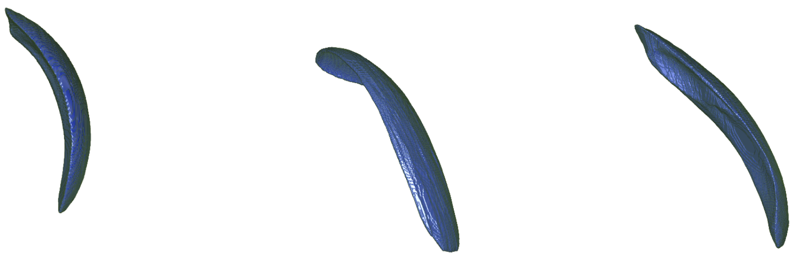
In the second case we try to understand the network structure of the striatum, a part of the basal ganglia in the brain. Due to the sheer computational power needed to compute persistent homology for this data our analysis is more of a holistic summary of the resulting simplicial structure rather than focusing solely on persistent homology. Our approach is largely inspired by [8], in which a similiar analysis is done but for a different part of the brain.

4.1 Corneas of Bombus terrestris

It has been found that the size of individuals of the species *Bombus terrestris* affects aspects of their visual capabilities [9]. By applying persistent homology we can investigate whether this difference in size also translates to a difference in persistent homology, and so by proxy a difference in topology. If so, this could serve to strengthen the hypothesis that larger individuals have superior, or at the very least different, visual capabilities than smaller individuals. Persistent homology is a good candidate for this purpose as metrics on persistence diagrams are indifferent to differences in scale but rather measures differences in shape.

4.1.1 Data

The data consists of binary 3-dimensional volumes (see Figure 4.1 for renderings of some of the samples) of the corneas acquired by micro-CT scans of the samples described in Table 4.1. The main focus of the analysis will be on samples from the bumblebee *Bombus terrestris*, but in total there are 20 samples belonging to 8 different species of insects. The additional samples from other species will be used to verify our topological findings.



(a) *TA_60204* **(b)** *MQ_60209* **(c)** *BT_77970*

Figure 4.1: *Example renderings of cornea volumes.*

4.1.2 Methodology

Since our samples consist of binary 3-dimensional volumes each voxel[1] can be considered as cube 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 Chapter 2, we can instead impose the structure of a *cubical complex* on the data samples.

Cubical complexes

We follow [5] in defining the cubical complex.

Definition 4.1.1. An elementary interval is a unit interval $[k, k+1]$ or a degenerate interval $[k, k]$ for $k \in \mathbb{N}$.

Definition 4.1.2. In an n-dimensional space a cube is the cartesian product of n elementary intervals. The dimension of the cube is exactly the number of intervals in the product which are not degenerate.

Hence, analoguous to simplices a 0-cube is a vertex, a 1-cube is an edge, a 2-cube is a square and a 3-cube is an actual cube in the geometric sense.

From here on the same construction as in simplicial homology holds [2]. We can in the entirely same way as for simplicial homology define a cubical complex and homology on cubical chain complexes.

Example 4.1.1. Just like in a simplicial complex the hole highlighted in green in Figure 4.2 is indeed a non-trivial generator of H_1.

We also need to impose a metric structure on the space given by each sample. Computing the persistent homology of a binary volume will just lead to all cubes in the complex appearing at a threshold of 1. Instead, we consider the Euclidean Distance Transform (EDT) as a way of imposing a metric structure on the binary volume.

[1] A voxel is the 3-dimensional equivalent of a pixel.

[2] For computational reasons we can also consider the dual cubical complex where we let voxels be the vertices. See [?] for a thorough explanation of this.

ID	ITW	Species
AM_60185	2.90	Apis mellifera
AM_60186	2.95	Apis mellifera
BT_77967	5.42	Bombus terrestris
BT_77970	4.00	Bombus terrestris
BT_77971	4.02	Bombus terrestris
BT_77973	1.97	Bombus terrestris
BT_77974	2.97	Bombus terrestris
BT_77976	5.47	Bombus terrestris
MB_60160	3.25	Melipona bicolor
MB_60161	3.25	Melipona bicolor
MQ_60208	3.64	Melipona quadrifasciata
MQ_60209	3.64	Melipona quadrifasciate
PR_60164	1.49	Plebia remota
PR_60206	1.49	Plebia remota
TA_60204	1.17	Tetragonista angustula
TA_78016	1.17	Tetragonista angustula
TC_60166	1.94	Tetragona clavipes
TC_60167	1.94	Tetragona clavipes
TS_60163	2.10	Trigona spinipes
TS_60203	2.10	Trigona spinipes

Table 4.1: *Table over the data samples used in the analysis. The ID column gives a unique ID to each sample and the ITW column gives the intertegular width of each sample.*

Definition 4.1.3. Given a subset $Y \subset \mathbb{R}^n$ we define

$$EDT(x) = \inf_{y \in \partial Y} ||x - y||_2$$

where ∂Y is the boundary of Y.

So for a given binary volume we embed it in $\mathbb{R}^n$ and apply the Euclidean distance transform to get a distance metric between voxels.

Example 4.1.2. To calculate the EDT of the binary image in Figre 4.3 we simply calculate the difference vector from a cell of value 1 to the closest cell with value 0. For example, to get $\sqrt{5}$ in the top left corner we need to walk one step in to the left and two steps upwards which translates to the vector $(-1, 2)$ which has Euclidean norm $\sqrt{1+4} = \sqrt{5}$.

Our filtration then will describe the structure starting at the boundary of the cornea, the hollow shell surrounding the volume, and then as the threshold increases the cubical complex will include more and more of the denser parts within the volume. An illustration of the thresholding at different values is seen in Figure 4.4.

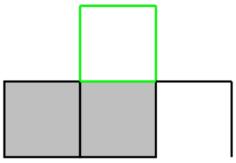
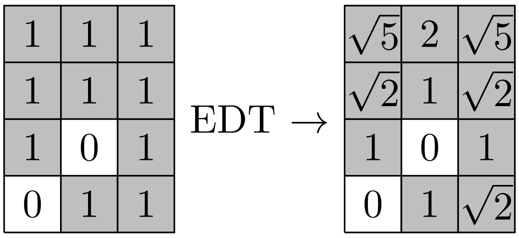


Figure 4.2: *A cubical complex consisting of two 2-cubes, a 2-cube without the interor and two 1-cubes.*



1	1	1
1	1	1
1	0	1
0	1	1

EDT →

$\sqrt{5}$	2	$\sqrt{5}$
$\sqrt{2}$	1	$\sqrt{2}$
1	0	1
0	1	$\sqrt{2}$

Figure 4.3

The resulting topological summaries we find are barcodes. While these are in themselves interesting, in order to answer whether the size of an individual of the specis *Bombus terrestris* has an impact on the topology of its cornea we compare the samples in a distance matrix, where the metric between pairs 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:

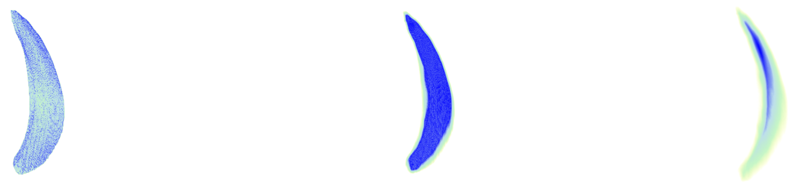
1. Is there a correlation between the size of the bumblebees and their persistent homology?

2. Can we with persistent homology identify subgroups of bumblebees, and if so are these subgroups related to their size?

Answering these questions will allow us to evaluate whether persistent homology provides information which relates to the hypothesis.

(Work in progress, proper definitions for clustering and Mantel) In order to identify groupings in the samples we use *hierarchical single-linkage clustering.* Each sample starts out in its own cluster. We then cluster that sample with the sample to which, in terms of persistent homology, it has the lowest distance. We then proceed inductively, and consider the distance from a cluster to another cluster to be the smallest distance among the distances of samples within the two clusters.

To investigate the relationship between between size and topology we additionally compute the distance matrix between the samples' ITWs (intertegular widths) and use the Mantel test to see the correlation between the Wasserstein distance matrix. The Mantel test is a non-parametric test of the correlation between two distance matrices



(a) $\epsilon < 2$ (b) $\epsilon < 10$ (c) $\epsilon < 100$

Figure 4.4: *EDT thresholding of BT_77976. Cooler colors indicate denser parts of the volume relative to the rest of the volume.*

given. The idea is quite simple: we take the two matrices and permute the columns of one of the matrices to compute the probabilities that the values fall within a certain range. If the two matrices are correlated we expect the value to change drastically.

4.1.3 Results

The clusterings on the entire data-set in Figure 4.5 and Figure 4.6 reveals that there are two samples of *Bombus terrestris* that are notably different from the other samples of the same species both in their first and second persistent homologies. Coincidentally, these two samples are the samples with the largest ITW in the entire data-set (5.42, 5.47).

We further see that species are mostly grouped together indicating that distances given by 1-Wasserstein between persistent homologies is a capable discriminant when it comes to species. However, there are some oddities such as *Melipona bicolor* and *Melipona quadrifascisata* not being clustered by species.

If we look more closely at *Bombus terrestris* in Figure 4.7 and Figure 4.8 we see that persistent homology in both first and second dimension leads to clusterings where the two large samples are considered in a cluster for themselves. However, H_2 from H_1 leads to a slightly different clustering where BT_77971 is considered different from the remaining smaller bumblebees.

Finally, we compute the Mantel test and see a non-trivial correlation coefficient for the samples from the species *Bombus terrestris* in Table 4.2 for the 1-Wasserstein distance in all dimensions. For the significant results we see that we have roughly the same correlation coefficient (0.57) in dimensions 1 and 2 and a weaker one (0.26) in dimension 0.

Notably, none of the bottleneck distances produce any significant results. Perhaps this is not too surprising as the bottleneck distance only provides information about the largest distance in the matching between persistence diagrams. This works well for simple topological shapes, but perhaps too much geometric information is lost about the intermediate stages of the persistent homology filtrations.

While the purpose of this analysis is mostly a showcase of persistent homology in

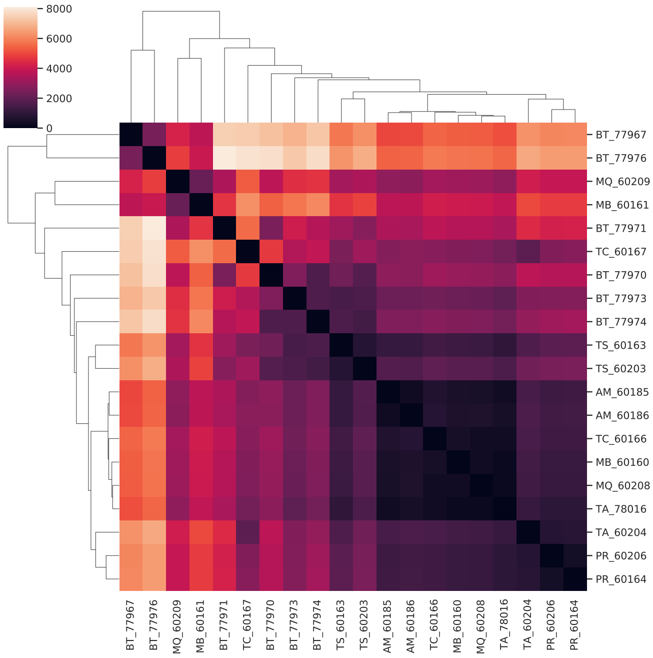


Figure 4.5: *Hierarchical single-link clustering of the 1-Wasserstein distance matrix derived from the persistent homologies of the entire data-set in* H_1.

the wild, these results do support to the idea that different sized *Bombus terrestris* do not only have larger eyes, but also that they are topologically different. Our clustering results group the larger individuals together and we find a moderate correlation between the higher order persistent homologies of the samples and their intertegular widths.

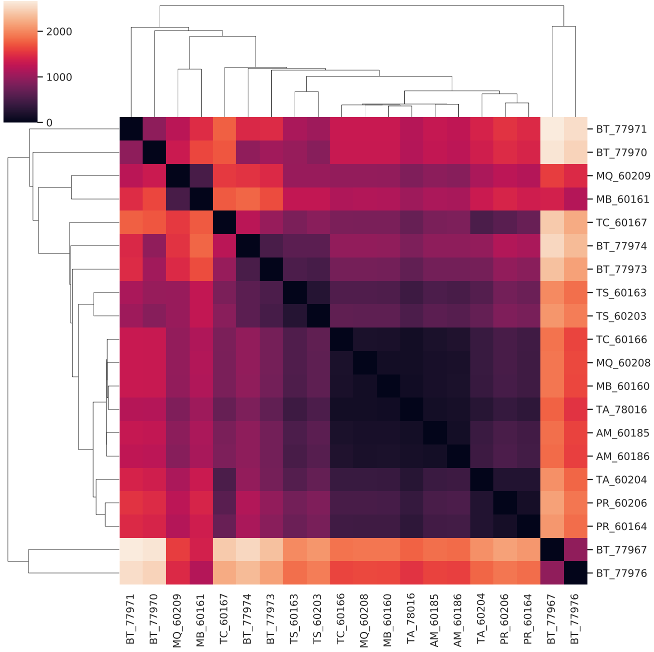


Figure 4.6: *Hierarchical single-link clustering of the 1-Wasserstein distance matrix derived from the persistent homologies of the entire data-set in* H_2.

Dimension	Metric	Correlation	P-value
0	Bottleneck	0.014	0.9096
0	1-Wasserstein	0.26	**0.013**
1	Bottleneck	-0.11	0.4304
1	1-Wasserstein	0.57	**0.0004**
2	Bottleneck	-0.12	0.3415
2	1-Wasserstein	0.57	**0.0003**

Table 4.2: *Table displaying the statistics computed in the Mantel test of the pairwise distances in different dimensions of persistent homology and the ITW for the species Bombus terrestris.*

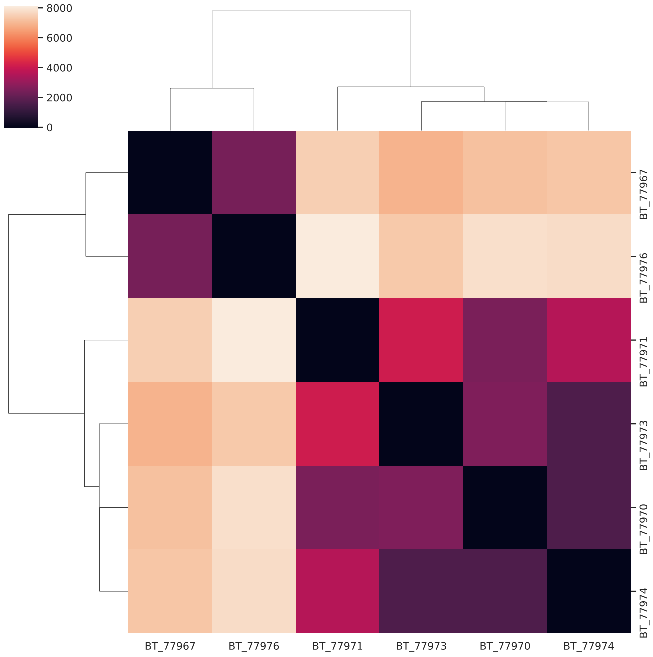


Figure 4.7: *Hierarchical single-link clustering of the 1-Wasserstein distance matrix derived from the persistent homologies of samples from the species Bombus terrestris in H_1.*

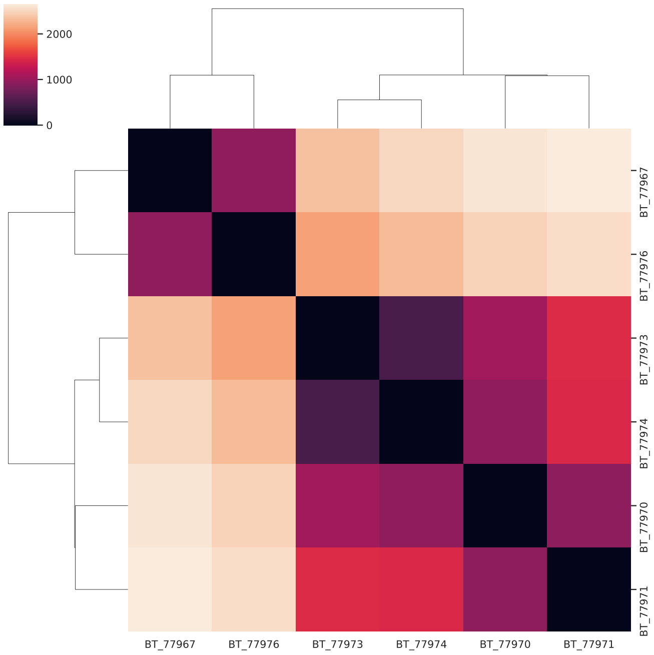


Figure 4.8: *Hierarchical single-link clustering of the 1-Wasserstein distance matrix derived from the persistent homologies of samples from the species Bombus terrestris in* H_2.

4.2 Brain network (work in progress)

Intro. Lorem ipsum.

4.2.1 Data

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.

We follow the results of [8] in showing the non-random structure in brain networks.

4.2.2 Methodology

The networks we work with are directed graphs, meaning that a connection from $A \to B$ does not imply a connection from $B \to A$.

Definition 4.2.1. A directed graph $G = V, E$ consists of an ordered vertex set $V = (v_0, \dots)$ and an ordered set of edges E whose elements are of the form (v_i, v_j) where the edge in the opposite direction $(v_j, v_i)$is not necessarily in the set.

It is not entirely clear how we should define a simplicial complex on a directed graph, but it is of importance that we try to do so. The assymetry in brain networks are important to capture the structure of how the neurons bind together, and should we disregard the fact that the network is directed important qualities of the network might be lost. While the Vietoris-Rips complex is directly definable on an undirected graph there are a number of different ways to go about defining a similar complex on a directed graph (cite digraph clique article). In our case we follow Reimann et al. (cite) and we extend the notion of the Vietoris-Rips complex, which is a special case of a flag complex, to something called a *directed flag complex* which we construct by imposing a certain partial order on the vertices of the directed graph.

Definition 4.2.2 ([6])**.** 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.

Recall that an ordered (abstract) simplicial complex arises from an ordinary simplicial complex where we give a

Definition 4.2.3 ([6])**.** Let G=(V,E) be a directed graph. The directed flag complex dFl(G) is defined to be the ordered simplicial complex whose k-simplices are all directed cliques with vertices $v_0, \dots, v_k$ 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.

It is important to try to understand this construction because it will have meaning when we try to interpret the homology of the resulting networks. In Figure 4.9 we see what a 3-simplex looks like in a *directed flag complex.* We then need to think about what a chain of these directed simplices look like. For example, a 2-chain of directed simplices can be see in Figure **??**. Then what is a cycle? It is a cavity built in by these directed simplices. In the 1-dimensional case it is not very difficult, it will look something like this.

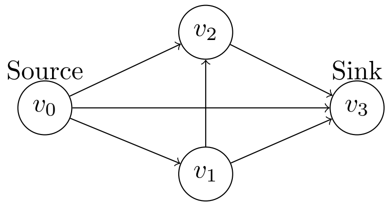


Figure 4.9: *A 3-simplex in a directed flag complex.*

4.2.3 Results

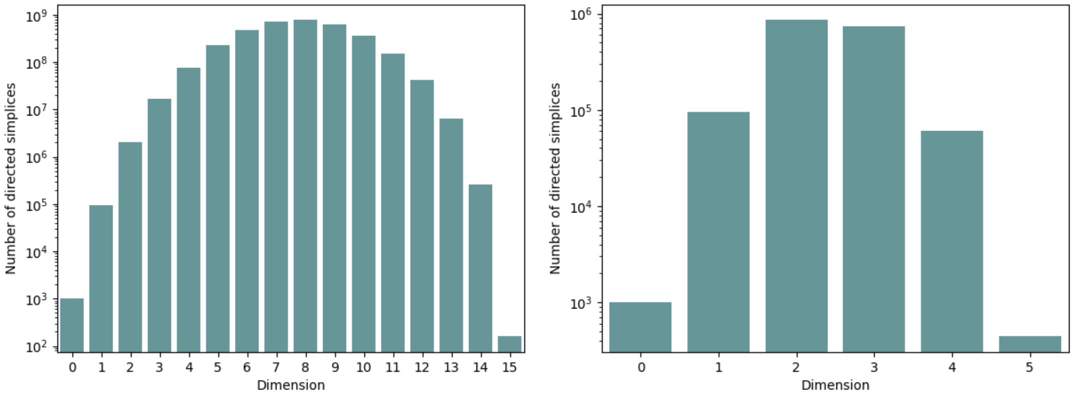


Figure 4.10: *The number of simplices in each dimension for the directed flag complex generated by (a) synthetic network from Snudda, (b) random network generated with the same edge probability creation as the first network, both with 999 vertices.*

In Figures 4.10 and 4.11 we see that the synthetic brain networks have much more higher order structure in terms of high dimensional simplices than a network generated solely based on edge connectivity. For instance, we see in 4.11 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

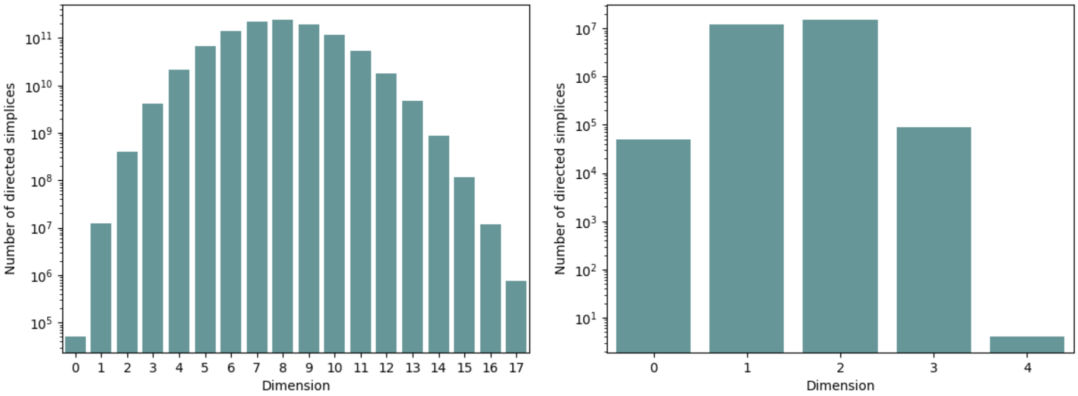


Figure 4.11: *The number of simplices in each dimension for the directed flag complex generated by (a) synthetic network from Snudda, (b) random network generated with the same edge probability creation as the first network, both with 50 001 vertices.*

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 4.12. We also look at a full synthetic network generated with only 999 vertices in Figure 4.13.

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.

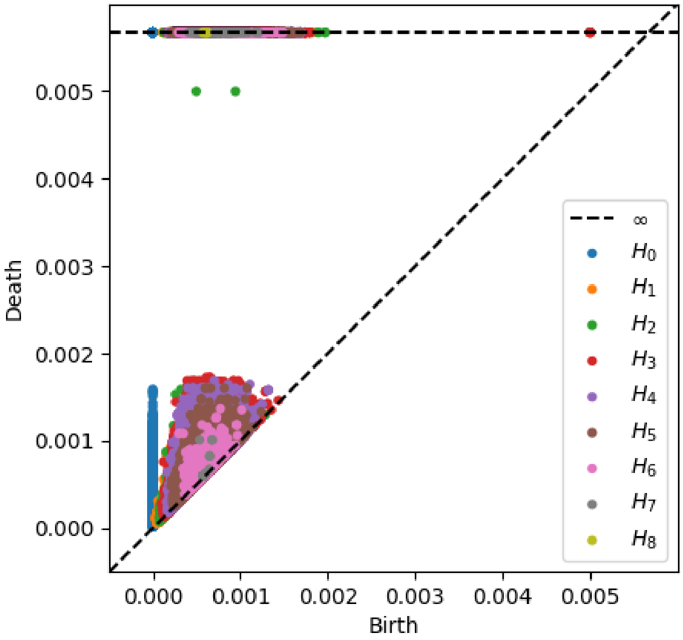


Figure 4.12: *Persistence diagram of the subnetwork of dSPNs extracted from a synthetic network of 50 001 vertices.*

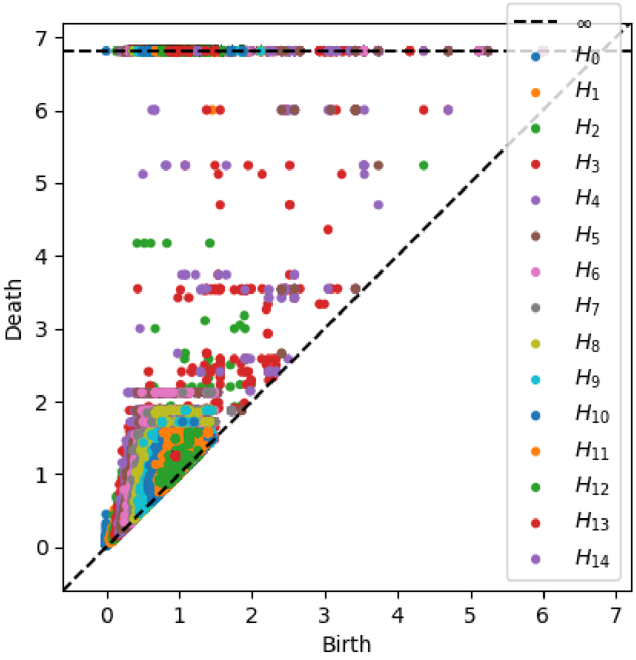


Figure 4.13: *Persistence diagram of the entire synthetic network consisting of 999 vertices. (this is scaled 1000 larger than in actual data, generate new diagram)*

5 Conclusion

Our aim in this thesis was to provide both an introduction to persistent homology as well as providing two examples of persistent homology being used in the wild in order to facilitate analysis of data.

We start by giving an introduction to homology, but focusing solely on simplicial homology while ignoring the more classical treatments which involve singular homology. Since simplicial homology can be considered conceptually more simple than other ways of defining homology this also leads to intuitive, immediate expressions of homology in terms of holes in simplicial complexes.

We then go on by defining persistent homology which gives us the tool with which we will perform our analysis. We explain

In the case of cornea 3D volumes gathered from microCT scans of the eyes of the species *Bombus terrestris* we showed how persistent homology can be applied to volumetric data and how it can be used to perform a clustering and similarity analysis. In the case of synthetic networks from the striatrum we showed how persistent homology can be performed on graphs, in particular on directed graphs which contain assymetrical information about the network.

Using persistent homology we were able to relate the already observed fact that individuals of the species *Bombus terrestris* have differ in morphology when it comes to larger and smaller individuals. The fact this is done with topological methods mean that the correlation we found is likely non-trivial.

By using these non-traditional ways of exploring data we show that perhaps there is some merit to considering persistent homology as a way of enhancing a traditional data analysis.

Bibliography

[1] Olaf Delgado-Friedrichs, Vanessa Robins, and Adrian Sheppard. Morse theory and persistent homology for topological analysis of 3d images of complex materials. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4872–4876, 2014.

[2] Antonio Gutierrez, Maria Jose Jimenez, David Monaghan, and Noel E. O'Connor. Topological evaluation of volume reconstructions by voxel carving. *Computer Vision and Image Understanding*, 121:27–35, 2014.

[3] Antonio Gutierrez, David Monaghan, María José Jiménez, and Noel E. O'Connor. Persistent homology for 3d reconstruction evaluation. In Massimo Ferri, Patrizio Frosini, Claudia Landi, Andrea Cerri, and Barbara Di Fabio, editors, *Computational Topology in Image Context*, pages 139–147, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[4] Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge New York, 2002.

[5] Tomasz Kaczynski, Konstantin Mischaikow, and Marian Mrozek. *Computational homology*, volume 157. Springer, New York, 2004.

[6] Daniel Luetgehetmann, Dejan Govc, Jason Smith, and Ran Levi. Computing persistent homology of directed flag complexes. *Algorithms*, 13(1), January 2020.

[7] Chul Moon, Scott A. Mitchell, Jason E. Heath, and Matthew Andrew. Statistical inference over persistent homology predicts fluid flow in porous media. *Water Resources Research*, 55(11):9592–9603, 2019.

[8] Michael W. Reimann, Max Nolte, Martina Scolamiero, Katharine Turner, Rodrigo Perin, Giuseppe Chindemi, Paweł Dłotko, Ran Levi, Kathryn Hess, and Henry Markram. Cliques of neurons bound into cavities provide a missing link between structure and function. *Frontiers in Computational Neuroscience*, 11:48, 2017.

[9] Gavin J. Taylor, Pierre Tichit, Marie D. Schmidt, Andrew J. Bodey, Christoph Rau, and Emily Baird. Bumblebee visual allometry results in locally improved resolution and globally improved sensitivity. *eLife*, 8, February 2019.

[10] Charles Weibel. *An introduction to homological algebra*. Cambridge University Press, Cambridge England New York, 1994.

[11] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, February 2005.