# Persistent

Daniel Collin

February 10, 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, it can in fact be a useful tool in exploring data. It 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 these filtrations takes us to graded modules and the Structure Theorem for PIDs and the stability guarantees that we give requires us to dwell on a bit of Morse theory.

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 linearly 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 $\beta \in A$.*
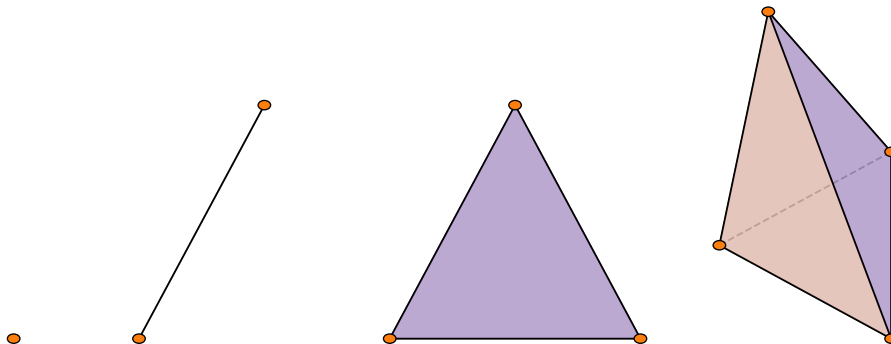
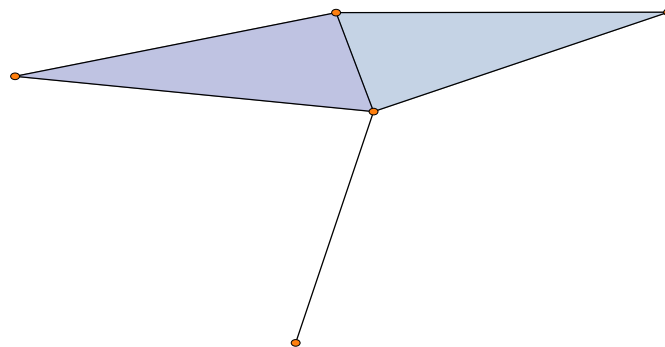**Figure 2.1:** *0-simplex (left), 1-simplex (middle left), 2-simplex (middle right) and 3-simplex (right).*



**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 $\alpha_i$ are coefficients in some ring $R$ and $\sigma_i$ 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 $k$th 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) □

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 $k$th 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.

# 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.
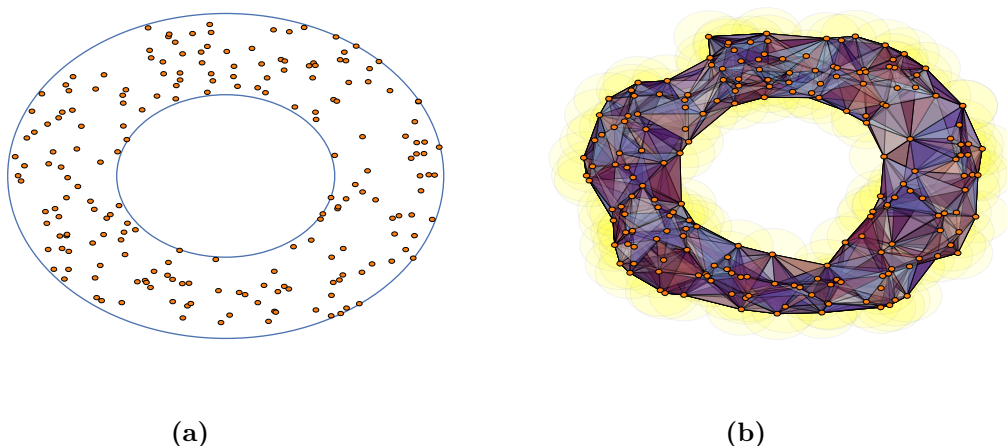


<div align="center">(a)</div>



<div align="center">(b)</div>

**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

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* $C_\epsilon$ *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 (really? cite, not sure about this).

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 the closed nbh around $k + 1$ vertices intersect 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* $R_\epsilon$ *is the abstract simplicial complex whose* $k$-*simplices are given by* $k + 1$ *points which are pairwise at most* $\epsilon$ *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:

$$R_1 \overset{x}{\hookrightarrow} R_2 \overset{x}{\hookrightarrow} \ldots \overset{x}{\hookrightarrow} R_{n-1} \overset{x}{\hookrightarrow} R_n$$

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 $\epsilon_i$ to resolution $\epsilon_j$.

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).
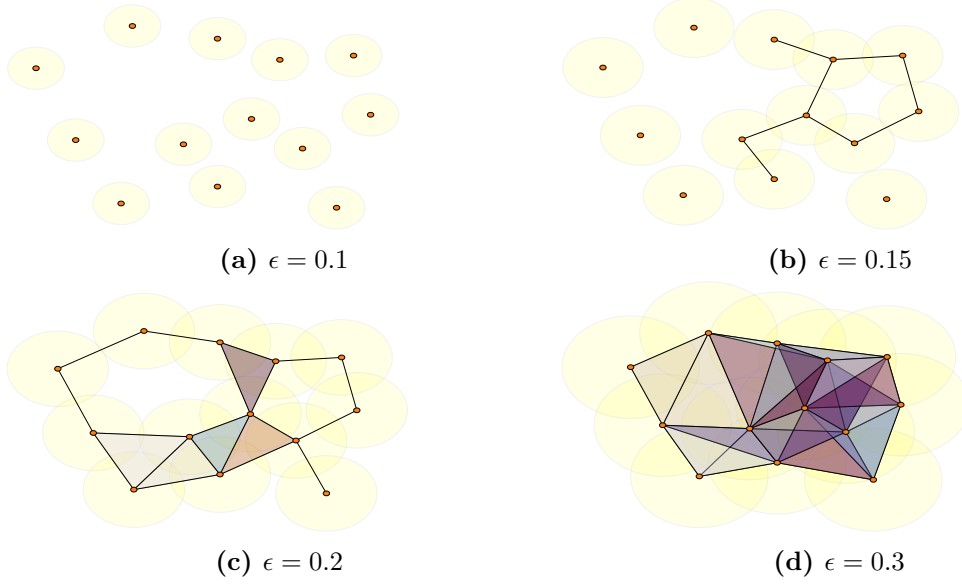
**(a)** $\epsilon = 0.1$

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

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

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

**Figure 3.2:** *The Vietoris-Rips complex at different $\epsilon$-values.*

**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 $\epsilon'$. 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 $R[x] = Rx^0 \oplus Rx^1 \oplus \ldots$

**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 $\epsilon$ 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 $\epsilon$

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 $\epsilon$.

## 3.3 Persistence Diagrams

Another way of illustrating is the persistence diagram which consists of..

## 3.4 Metrics

If we compute two barcodes, how do we compare them? There are two suitable metrics for this.. Wasserstein and Bottleneck.

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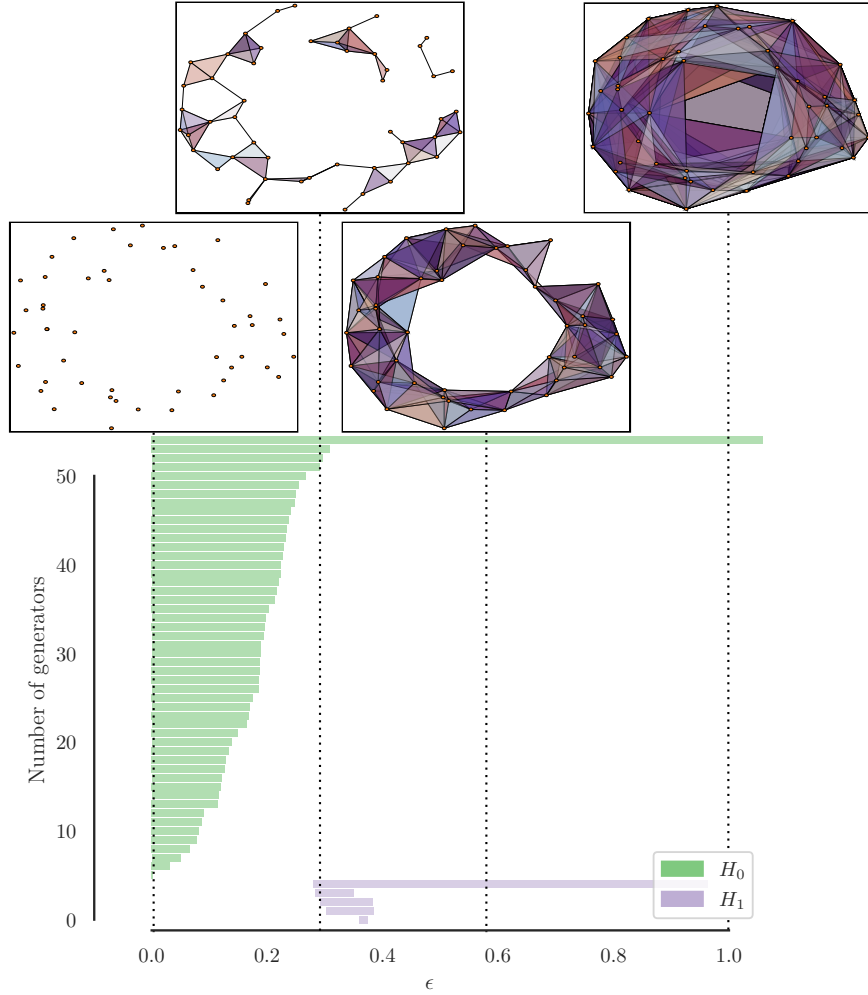
**Figure 3.3:** *Persistence barcode showing the birth and death of generators in the homology groups of a Vietoris-Rips complex approximated from points sampled from an annulus at different $\epsilon$.*

# 4 Data