Análisis topológico de datos

Daniel Garcés Paniagua

Trabajo de fin de grado de Matemáticas

Universidad de Zaragoza

de

Resumen

En el campo del Análisis de Datos el uso de herramientas matemáticas tanto para el desarrollo de algoritmos como para evaluar la precisión de los procesos es fundamental. Tradicionalmente se emplean métodos y conecptos basados en las áreas de Optimización, Álgebra Lineal y Estadística. En este Trabajo de Fin de Grado exploraremos el Análisis Topológico de Datos a través de una de las herramientas más utilizadas en esta disciplina: la homología persistente.

En el primer capítulo dedicaremos una seccion inicial a introducir el concepto de complejo simplicial y hablaremos brevemente de alguna de las aplicaciones y propiedades de estos objetos. Continuaremos construyendo los grupos de cadenas y consecuentemente los emplearemoss junto a los operadores borde para definir los complejos de cadenas y grupos de homología.

Comenzaremos el segundo capítulo dando unos cuantos ejemplos de filtraciones y cómo relacionarlas a fin de dar sentido a la definición de homología persistente, la cual presentaremos junto a un teorema que determina su estructura. En una última sección daremos una manera sencilla de computar dicha homología persistente a partir de la matriz del operador borde utilizando el concepto de barcoding, muy útil a la hora de presentar la información obtenida mediante este proceso.

Finalmente, como cierre aplicaremos algunos de los conceptos tratados en las secciones anteriores sobre un ejemplo real: Utilizaremos algunas de las herramientas existentes en Python para calcular la distancia "bottleneck" entre imágenes de tejido de cáncer de próstata. De este modo pretendemos mostrar una de las posibles utilidades de la homología persistente como herramienta para el análisis de datos.

En el esfuerzo de documentación para este trabajo se han seguido los libros [1], [2] según el curso *Geometry and Topology in Data Analysis* impartido por Ulrich Bauer en la Universidad Técnica de Múnich. Además se han complementado algunos conceptos algebraicos con [3].

Contents

Chapter 1

Simplicial Homology

In this part we will give a brief explanation of the context in which we will apply our tools, and then provide basic definitions coupled with some illustrative examples. In a final section we will construct intuitively the idea of homology, from which the main tool in Topological Data Analysis emanates.

1.1 Introduction

Suppose we have a dataset to work with. As per usual in Data Analysis, one interprets the data sets as a finite set of points in an euclidean space $\mathbb{K}$.

We can construct from such a dataset structures following certain criteria (e.g: distance in a metric space or any algorithm of our choice) associated to a topological space, and then study some topological invariants in order to further understand the properties of our data.

This approach is somewhat different from traditional methods in Data Analysis. In this field, the majority of tools employed come from the field of Statistics and Probability, with some Linear Algebra and Optimization added into the mix for the creation of algorithms. In this context, the introduction of topological concepts into the discipline opens up a different perspective for interpretation of our data, for which traditional methods are not suited.

1.2 Simplicial complexes

Let us, as one usually does, introduce first the main objects of our study and a few concepts related to them in order to have a basic understanding of the matters we will be dealing with.

Definition 1.1. An *abstract simplicial complex* K is a collection of finite sets closed under inclusion. Every set $S \in K$ is called a *simplex*, and every element $v \in S$ is called a *vertex*.

Further definitions inmediately arise from this one:

1. A is a *face* of S if $A \subseteq S$. In this context S is said to be a *coface* of A.

2. The *dimension* of a simplex is defined as:

$$dim(S) := \#S - 1$$

3. Naturally, simplices of dimension n are called n-simplices.

4. One defines subsequently the dimension of a simplicial complex K as the highest dimension of its simplices.

5. If A is a *face* of S with $dim(S) = dim(A) + 1$, then A is a *facet* of S.

Example. The set

$$A = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\}\}$$

is an abstract simplicial complex. Observe that for each element $a \in A$ every subset is itself contained in A, thus A is closed under inclusion.

Abstract simplicial complexes are only one approach to define our concept. An alternative, more visual and intuitive way to define a simplicial complex is geometrically, from points inside a certain space X.

Definition 1.2. A *geometric simplex* σ of dimension n is the convex hull generated by a set of $n+1$ affinely independent points V in a real vector space $\mathbb{K}$.

Dimension	Shape
$n=0$	Point
$n=1$	Segment
$n=2$	Triangle
$n=3$	Tetrahedron

Table 1.1: Geometrical shapes of the simplices of dimensions $n = 0,1,2,3$.

1. Such generating points $v \in V$ are called the *vertices* of σ.
2. Given a simplex σ, τ is a *face* of σ if it is the convex hull of a subset $U \subseteq V$.
3. A collection G of geometric simplices that only intersect over whole faces and is closed under the face relation is called a *geometric simplicial complex*.

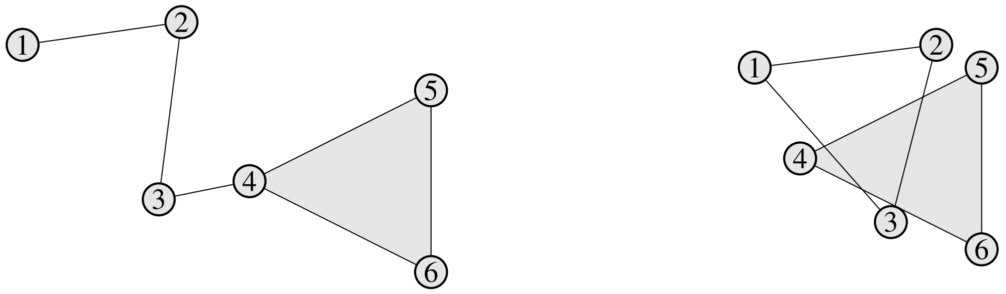


Figure 1.1: To the left, an example of a geometrical simplicial complex and to the right of something that is *not* a geometric simplicial complex.

Geometric and abstract simplicial complexes, as one can imagine, are related to each other. To clarify this relation, we need another concept:

Definition 1.3. Let K be an abstract simplicial complex. Its *vertex set* is defined as the union of all its simplices

$$VertK := \cup_{S\in K} S$$

Remark. We will use from now on the same notation to refer to the set of vertices of geometric simplicial complexes, whose definition is the natural one.

Example. For our previous example of an abstract simplicial complex we have:

$$A = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\}\} \ , \ VertA = \{1,2,3\}$$

Definition 1.4. Given G a geometrical simplicial complex, the set of vertex sets of the simplices in G is an abstract simplicial complex, called the *vertex scheme* of G.

$$Scheme(G) := \{Vert(G) \mid S \text{ is a simplex of } G\}$$

A geometrical simplicial complex G whose vertex scheme is isomorphic to an abstract simplicial complex A is called a *geometric realization of* A.

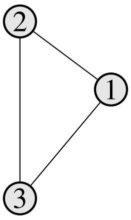


Figure 1.2: A geometric realization of our example of an abstract simplicial complex A

The following proposition and proof gives us further insight on how to construct a geometric realization of an abstract simplicial complex.

Proposition 1.1. *Given a d-dimensional abstract simplicial complex A , it is possible to construct a geometric realization in $\mathbb{R}^{2d+1}$*

Proof. For this proof we will use the following auxiliary lemma:

Lemma. Consider the *moment curve* $M = \{m(t) = (t, t^2, ..., t^d) \mid t \in \mathbb{R}\}$. The points in M are in general position, i.e, any subset of at most $d+1$ points is affinely independent.

Proof. (of the Lemma) Every hyperplane intersects the moment curve in a finite set of at most d points. If a hyperplane intersects the curve in exactly d points, then the curve crosses the hyperplane at each intersection point. Thus, every finite point set on the moment curve is in affine general position. (Check in [4]). □

Back to the proof of Proposition 1.1, we can construct an embedding of the vertices of A in general linear position by mapping each d-dimensional complex into $d+1$ distinct points of the $2d+1$-dimensional moment curve $\phi : VertA \longrightarrow \mathbb{R}^{2d+1}$. Since by our Lemma we know its points are in general position, any two collections of $d+1$ points in general position span disjoint affine subspaces.

The images of the vertices of A span geometric simplices in $\mathbb{R}^{2d+1}$. We get that

$$G = \{conv(\phi(S)) \mid S \in A\}$$

is a geometrical simplicial complex realizing A □

Due to this relation, in many cases we will simply talk generically about simplicial complexes meaning any of the definitions depending on the context.

The concept of simplicial complex is very broad and many shapes satisfy our definition. We will be interested however in constructing simplicial complexes from a fixed set of points in a systematic way so that the resulting complex is uniquely determined in order to study its properties.

Example. One such example which will be very useful later on is the Vietoris-Rips complex. It is defined as follows:

Definition 1.5. Given (X, d) a finite metric space, one defines the *Vietoris-Rips complex* at scale t as:

$$Rips_t(X) := \{Q \subseteq X \mid Q \neq \emptyset \, , \, diamQ \leq t\}$$

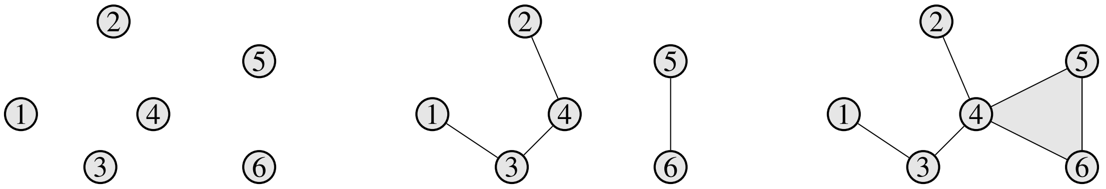


Figure 1.3: $Rips_t(X)$ for respectively $t = 0, 2, \sqrt{5}$

Remark. Something to note about the Vietoris-Rips complex is that it does not admit empty skeletons: if n points are connected pairwise, then the $n-1$-dimensional simplex spanned by them is included in the Vietoris-Rips complex. This makes it a *clique complex*: the maximal complex given its 1-simplices, and thus it is only dependant on them. This does not happen for every such construction of complexes as we will see in the next chapter.

Having defined and exemplified simplicial complexes, we will now delve into their properties when we endow them with a topology.

Definition 1.6. Let $K \subseteq \mathbb{R}^d$ be a finite geometric simplicial complex. The *polyhedron* $|K|$ is the subspace comprised of the union of all simplices in K endowed with the subspace topology.

Definition 1.7. A *triangulation* of a space $X \subseteq \mathbb{R}^d$ is a pair (K, f) consisting of a geometric simplicial complex K together with an homeomorphism $|K| \xrightarrow{f} X$.

If such a thing exists, then X is said to be *triangulable*.

The definition above implies that any triangulable topological space X can be studied via its triangulation. Hence, one expects that finding certain properties of the triangulation will give us information about different aspects of X.

This is already very useful as it is. However, if we stick to our definition we have that a topological space admits many triangulations. We have to see which properties of a topological space are essential in terms of its triangulations in order to simplify our study.

For this goal, let us first define relations between simplicial complexes.

Definition 1.8. For K, L two geometric simplicial complexes, a continuous map $f : |K| \to |L|$ that maps each simplex of K affinely onto some simplex in L is called a *simplicial map*.

Remark. Let (K, f), (L, g) be two triangulations of the same topological space X. Then, by definition, we get $|K| \cong |L|$ by the homeomorphic simplicial map $f \circ g^{-1}$.

The next remark gives us a natural result in regards to the relations between definitions of simplicial complexes.

Remark. For K an abstract simplicial complex, its geometric realization is unique up to simplicial homeomorphism.

We will now address a very important kind of relation between simplicial complexes.

Definition 1.9. A *subdivision* of K is a simplicial complex L such that $|K| = |L|$ and every simplex of L is contained in some simplex of K.

Example. The most straightforward example of subdivision of an abstract simplicial complex is the barycentric subdivision.

Definition 1.10. The *flag complex* $Flag(P)$ for a poset $(P, \leq)$ is the abstract simplicial complex where:

1. The vertex set is P.

2. Its simplices are the flags (or totally ordered subsets) of P.

For any simplicial complex K, the *face poset* is defined as $Pos(K) = (K, \subseteq)$. Then, for an abstract simplicial complex A, its *barycentric subdivision* is defined as:

$$Sd(A) = Flag(Pos(A))$$

where the vertex set is A and the simplices are the flags of A.

$$A = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\}\}$$

$$\begin{aligned} Sd(A) = \Big\{ &\{1\},\{2\},\{3\},\{\{1,2\}\},\{\{1,3\}\}, \\ &\{\{2,3\}\},\{1,\{1,2\}\},\{1,\{1,3\}\}, \\ &\{2,\{1,2\}\},\{2,\{2,3\}\},\{3,\{2,3\}\}, \\ &\{3,\{2,3\}\}\Big\} \end{aligned}$$

Figure 1.4: An example of abstract barycentric subdivision

The barycentric subdivision has a definition as well for geometrical simplicial complexes. For a geometric simplicial complex G, we add as vertices the barycenters of every simplex in G, and as new simplices we get those spanned by the barycenters of each flag. Formally:

$$Vert(Sd(G)) = \big\{ z(\sigma) \mid z \text{ is the barycenter for } \sigma \in G \big\}$$

$$Scheme(Sd(G)) = \big\{ \{z(\sigma_1), z(\sigma_2), ..., z(\sigma_k)\} \mid \{\sigma_1, \sigma_2, ..., \sigma_k\} \text{ is totally ordered} \big\}$$

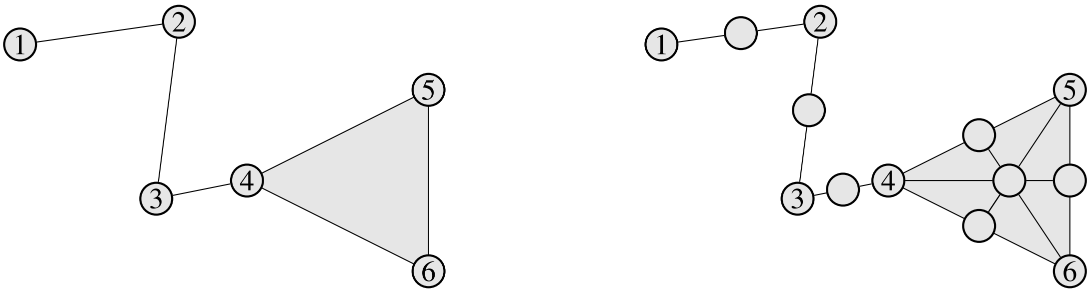


Figure 1.5: An example of geometric barycentric subdivision

One has that the geometric barycentric subdivision of a geometric realization is a geometric realization of the abstract barycentric subdivision of its vertex scheme.

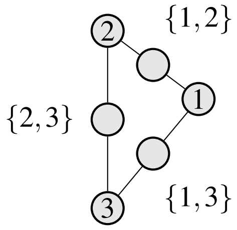


Figure 1.6: Barycentric subdivision of the geometric realization of A

Even though this reduces the list of simplicial complexes quite a lot, it is still incredibly vast. We would want to further classify them, and one way is through the usage of homotopy equivalence.

Definition 1.11. Two maps $f, g : X \longrightarrow Y$ are called *homotopic*,$f \simeq g$, if there is a continuous deformation from one to the other, i.e, there exists a continuous map $F : X \times [0,1] \longrightarrow Y$ with:

$$f(x) = F(x,0)$$

$$g(x) = F(x,1)$$

Definition 1.12. Two spaces X,Y are *homotopy equivalent*, $X \simeq Y$ if there are maps $f : X \longrightarrow Y$ and $g : Y \longrightarrow X$ such that $g \circ f \simeq id_X$, $f \circ g \simeq id_Y$.

In particular, two simplicial complexes K, L are homotopy equivalent if there are simplicial maps $f : |K| \longrightarrow |L|$, $g : |L| \longrightarrow |K|$ such that $g \circ f \simeq id_{|K|}$, $f \circ g \simeq id_{|L|}$

Remark. Homotopy equivalence $\simeq$ is, as one may suspect, an equivalence relation.

Definition 1.13. Let $i : A \hookrightarrow X$ be an inclusion and $r : X \longrightarrow A$ a retraction such that $r \circ i = id_A$. Then A is a *retract* of X.

If also $i \circ r \simeq id_X$, the homotopy is a *deformation retraction* and A a *deformation retract* of X. Moreover, if A is fixed throughout the deformation retraction then we say it is a *strong deformation retraction* and A is a *strong deformation retract* of X.

Homotopy equivalency is rich enough as a topic on its own. However, we will focus our attention on a different aspect of our objects. This aspect will be realised in the concept of simplicial homology and will nontheless be an invariant under homotopy.

1.3 Homology groups

Consider a simplicial complex K. As we stated at the end of the previous section, our goal is to find an invariant under homotopy that gives us information about the inherent structure of our complex. This will be achieved in the concept of homology group of a simplicial complex K.

Going further, two different but somewhat equivalent procedures can be followed. One can either provide an orientation for the vertices of the d-simplices up to even permutation and define the d-chain group as the free abelian group generated by the oriented d-chains, which is equivalent to define the d-chain group as a module over $\mathbb{Z}$. Alternatively, one can define a d-chain space as the vector space over $\mathbb{F}_2$, that is, a $\mathbb{Z}_2$-module.

Both approaches are means to the same end, providing their own advantages and disadvantages. We will focus in this chapter on the algebraic approach of chain groups, but we will be recurring to the chain spaces later on for easier computations. For the equivalent definitions with vector spaces over $\mathbb{F}_2$ see [1].

As a first step, we will need to dotate our simplicial complex K with a group structure in order to be able to operate within it.

Definition 1.14. Let K be a simplicial complex. The *d-th chain group* of K is the free abelian group generated by the d-simplices of K. It is denoted by $C_d(K)$.

Remark. What we denote by σ and $-\sigma$ can be interpreted as the orientation of the simplex σ the following way:

Suppose that we have a specific order for the vertex set of a d-simplex σ. For any given set of $d+1$ vertices $\{v_i\}_{i=0}^{d}$ there are $d!$ possible orderings, meaning that imposing an order right away would get us many different elements.

However, since all possible orderings are equivalent up to even permutations to either $(v_0, v_1, ..., v_d)$ or $(v_1, v_0, ..., v_d)$ we can consider that a d-simplex σ is *positively oriented* if the order of its vertices is equivalent to the order $(v_0, v_1, ..., v_d)$ and *negatively oriented* if its order is equivalent to $(v_1, v_0, ..., v_d)$.

Having already defined group structures for our simplicial complex K, one defines a *d-chain* of K as any subset of d-simplices of K.

Now that we have as many chain groups as dimensions of simplices in a complex, by relating simplices with their faces, we have a map that connects the d-th chain groups to their immediate superior and inferior.

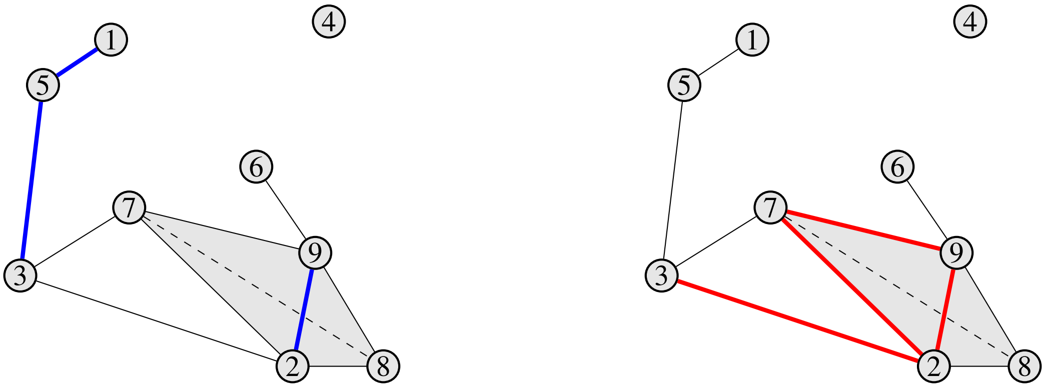


Figure 1.7: This simplicial complex has marked as blue and red two different 1-chains

Definition 1.15. Let K be a simplicial complex, $C_d(K)$, $C_{d-1}(K)$ its respective d-chain and $d-1$-chain groups. We define the *d-th boundary homomorphism* of K as the map:

$$\partial_d : C_d(K) \longrightarrow C_{d-1}(K) \quad , \quad (v_0, v_1, \ldots, v_d) \longmapsto \sum_{i=0}^{d} (-1)^i (v_0, v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_d)$$

For the case $d = 0$ one takes as $C_{-1}(K)$ the zero group.

Proposition 1.2. *∂_d is indeed a group homomorphism.*

Now we have in $(C_d(K), \partial_d)_d$ a set of abelian groups and a set of homomorphisms that join them consecutively.

$$\ldots \xrightarrow{\partial_{d+2}} C_{d+1}(K) \xrightarrow{\partial_{d+1}} C_d(K) \xrightarrow{\partial_d} C_{d-1}(K) \xrightarrow{\partial_{d-1}} \ldots$$

As with any group homomorphism, from ∂_d some subgroups of $C_d(K)$ can be derived. We will be interested in the next two in particular.

Definition 1.16. The *d-th cycle group of* K, $Z_d(K)$ is defined as

$$Z_d(K) := ker\partial_d \subset C_d(K)$$

Elements of $Z_d(K)$ are called d-cycles of K. $Z_d(K)$ is the group whose elements are all the cycles of d-chains of K.

The *d-th boundary group of* K, $B_d(K)$ is defined as

$$B_d(K) = Im\partial_{d+1}$$

Elements of $B_d(K)$ are called d-boundaries of K. It is the group of d-chains of K for which $\exists\phi \in C_{d+1}(K)$ such that $\tau = \partial_{d+1}(\phi)$, that is, those d-chains which bound a larger $d+1$-chain in the complex.

The following statement is easily deduced:

Remark. $B_d(K) \subseteq Z_d(K)$, i.e, $\partial_d \circ \partial_{d+1} = 0$.

This property is essential to our set $(C_d(K), \partial_d)_d$, so much so that from now on we will refer to any sequence of modules together with homomorphisms that follow this defining property as a *complex*. In particular, $(C_d(K), \partial_d)_d$ is the chain complex of K.

Having defined both the cycle group and the boundary subgroup, it is natural to study the quotient group and try to give an interpretation to it. One has that every two d-cycles of K describe the same hole whenever they differ by a d-boundary.

We have then that the equivalence relation set by the boundary group give us the number of d-dimensional holes of a simplicial complex K.

This gives us finally the definition of homology.

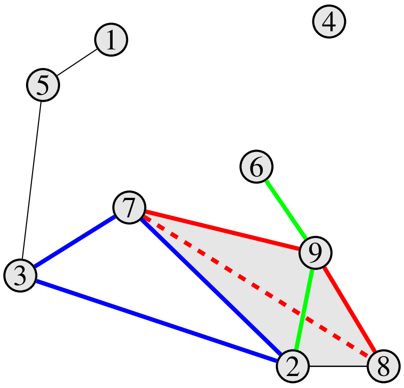


Figure 1.8: In red a 1-boundary, in blue a 1-cycle and in green a 1-chain which is neither a boudnary nor a cycle

Definition 1.17. The *d-th homology of K* is defined as

$$H_d(K) := Z_d(K)/B_d(K)$$

This quotient groups give us insight into the shape of our simplicial complex by way of its rank. Since $H_d(K)$ is a finitely generated abelian group, we know that it is isomorphic to $\mathbb{Z}_{{p_1}^{a_1}} \oplus \ldots \oplus \mathbb{Z}_{{p_k}^{a_k}} \oplus \mathbb{Z}^n$ for some prime numbers $p_1, \ldots, p_k$, $a_1, \ldots, a_k \in \mathbb{Z}$. The rank of the free abelian group above $\mathbb{Z}^n$ is called the *d-th Betti number of K*, denoted $\beta_d(k)$

This remark will clarify the meaning of what we have devised throughout the section.

Remark. For $d = 0, 1, 2$ the Betti number counts connected components, holes and voids of K respectively.

As we said at the beginning of the section, simplicial homology is an invariant under homotopy and is as such compatible with all our preliminary classifications of simplicial complexes.

As an end to this chapter, we will prove the next theorem, which will show how to go from a simplicial map to a map between chain complexes, and from it subsequently to a homomorphism between homology groups.

Theorem 1.1. *Let $f : K \longrightarrow L$ be a simplicial map. Then, we can induce a homomorphism between homology groups $H_*(f) : H_*(K) \longrightarrow H_*(L)$.*

Proof. We will begin by inducing a map $f_\sharp^d : C_d(K) \longrightarrow C_d(L)$ from f. Consider the map:

$$f_\sharp^d(\sigma) = \begin{cases} f(\sigma) & \text{if } f(\sigma) \text{ is a } d\text{-simplex} \\ 0 & \text{otherwise} \end{cases}$$

as this is well-defined for all d over the basis formed by the d-simplices and thus can be extended to the whole d-chain group. We will abuse the notation and simply write $f_\sharp$.

We have to see that indeed $f_\sharp$ is compatible with the structure of chain complexes, that is, that the following diagram commutes:

$$\begin{tikzcd}
\ldots \arrow[r] & C_{d+1}(K) \arrow[r, "\partial_{d+1}^K"] \arrow[d, "f_\sharp^{d+1}"'] & C_d(K) \arrow[r, "\partial_d^K"] \arrow[d, "f_\sharp^d"'] & C_{d-1}(K) \arrow[r, "\partial_{d-1}^K"] \arrow[d, "f_\sharp^{d-1}"'] & \ldots \\
\ldots \arrow[r] & C_{d+1}(L) \arrow[r, "\partial_{d+1}^L"'] & C_d(L) \arrow[r, "\partial_d^L"'] & C_{d-1}(L) \arrow[r, "\partial_{d-1}^L"'] & \ldots
\end{tikzcd}$$

Let's see that $\partial^L \circ f_\sharp(\sigma) = f_\sharp \circ \partial^K(\sigma)$.
Let $\sigma = \{v_0, v_1, ..., v_k\}$ and $\tau_i = \{v_0, v_1, ..., \hat{v}_i, ..., v_d\}$ be its i-th facet. We have

$$\partial^L \circ f_\sharp(\sigma) = \sum_i (-1)^i f_\sharp(\tau_i)$$

Let $\gamma = f(\sigma)$ and consider each case:

- $\dim\gamma = \dim\sigma$:

  Then we have by definition $\gamma = f_\sharp(\sigma)$ and each facet of σ is bijectively mapped to one of γ, so

  $$\partial^L \circ f_\sharp(\sigma) = \partial^L(\gamma) = \sum_i (-1)^i f_\sharp(\tau_i) = f_\sharp(\sum_i (-1)^i \tau_i) = f_\sharp \circ \partial^K(\sigma)$$

- $\dim\gamma = \dim\sigma - 1$

  Then $f_\sharp(\sigma) = 0$ and $f(v_i) = f(v_j)$ for two unique i, j with $i \neq j$, meaning $f(\tau_i) = \pm f(\tau_j)$. All other facets go to a $k-2$-simplex that dies out, meaning:

  $$f_\sharp(\partial^K(\sigma)) = f_\sharp(\sum_l (-1)^l \tau_l) = \sum_l (-1)^l f_\sharp(\tau_l) = \sum_{l \neq j} (-1)^l f_\sharp(\tau_l) \pm (-1)^j f_\sharp(\tau_l) =$$
  $$(-1)^i f_\sharp(\tau_i) \pm (-1)^j f(\tau_j)$$

  Now, for each case we have:

  (a) $(-1)^i = (-1)^j$:
  In this case both i, j are of the same parity, in any case making v_i and v_j separated by an odd number of permutations. Therefore $f(\tau_i) = -f(\tau_j)$ and

  $$(-1)^i f_\sharp(\tau_i) \pm (-1)^j f(\tau_j) = f(\tau_i) - f(\tau_j) = 0$$

  (b) $(-1)^i \neq (-1)^j$: (w.l.o.g. $(-1)^j = 1$)
  In this case i, j are of distinct parity, and thus v_i and v_j are separated by an even number of permutations, inducing the same orientation. Hence $f(\tau_i) = f(\tau_j)$ and

  $$(-1)^i f_\sharp(\tau_i) \pm (-1)^j f(\tau_j) = -f(\tau_i) f(\tau_j) = 0$$

  making $f_\sharp \circ \partial^K(\sigma) = 0$.

- $\dim\gamma \leq \dim\sigma - 2$

  Again we have $f_\sharp(\sigma) = 0$ but now each facet of σ is mapped to a lower-dimensional simplex, and so the other end amounts to:

  $$f_\sharp(\partial^K(\sigma)) = f_\sharp(\sum_i (-1)^i \tau_i) = \sum_i (-1)^i f_\sharp(\tau_i) = 0$$

Now that we have checked that $f_\sharp$ is a chain map, we have to see that it maps cycles to cycles and boundaries to boundaries.

Note that for $\theta \in Z_d(K)$, by definition $\partial_d^K(\theta) = 0$ and thus $\partial^L \circ f_\sharp(\theta) = f_\sharp \circ \partial^K(\theta)$, meaning that we have $f_\sharp(\theta) \in Z_d(L)$.

For a boundary $\tau = \partial_d^K(\sigma)$ we have $f_\sharp(\tau) = f_\sharp \circ \partial^K(\sigma) = \partial^L \circ f_\sharp(\sigma) = \partial^L(f_\sharp(\theta))$, making $f_\sharp(\tau)$ again a boundary in $C_{d-1}(L)$.

Therefore, the induced map

$$f_\sharp^* : H_*(K) \longrightarrow H_*(L)(K) \quad , \quad \theta + B_d(K) \longmapsto f_\sharp(\theta) + B_d(L)$$

is a homomorphism. □

If instead of working over $\mathbb{Z}$-modules we do it over $\mathbb{Z}_2$, then $f_\sharp^*$ is a linear map. If f were as well a homotopy equivalence then it can be proven that the induced homomorphism $f_\sharp^*$ is an isomorphism.

Chapter 2

Persistent homology

In the previous chapter we defined the concept of homology of a simplicial complex. Now, we must see how to apply it to our context in the field of Data Analysis.

As we have hinted at before, our approach to study the point clouds resulting from our data will be to construct a series of simplicial complexes following a determined parametric criterion and study the way their homology develops as our parameter changes, this is what we will call persistent homology.

Related to persistent homology, we will introduce our main tool, which consists of a combinatorial encoding from which we will be able to recover the homology groups and Betti numbers for every different simplicial realization of our data.

In this chapter we will put in mathematical terms such concepts and give some interesting properties of them.

2.1 Filtrations and Interleavings

First in this section, let us define the systematic way of constructing simplicial complexes we alluded to in the previous chapter and will be using from now on.

Definition 2.1. A *filtration of simplicial complexes* (indexed over $\mathbb{R}$) is a family $K_\bullet = \{K_t\}_{t\in\mathbb{R}}$ of simplicial complexes such that $s \leq t$ implies $K_s \subseteq K_t$.

Example. The Vietoris-Rips complex from Example 1.2 parametrized over $\mathbb{R}$: $Rips_\bullet = \{Rips_t\}_{t\in\mathbb{R}}$ is a filtration.

Let's now see another example of such a construction:

Definition 2.2. Let $\mathscr{F} = (F_i)_{i\in I}$ be a collection of sets. The *nerve* of $\mathscr{F}$ is the collection:

$$Nrv\mathscr{F} := \Big\{J \subseteq I : \bigcap_{j\in J} F_j \neq \emptyset,\ J \neq \emptyset\, finite\Big\}$$

Proposition 2.1. *The nerve of a set $\mathscr{F}$ is an abstract simplicial complex.*

Proof. Let us prove that the nerve is closed under inclusion.

Let $J \in Nrv\mathscr{F}$ and take $K \subseteq J$. If $J \in Nrv\mathscr{F}$, then $\bigcap_{j\in J} F_j \neq \emptyset$ and J is finite.

Now, $K \subseteq J$, so as a consequence we have that K is again finite and $\bigcap_{k\in K} F_k \supseteq \bigcap_{j\in J} F_j \neq \emptyset$, hence $\bigcap_{k\in K} F_k \neq \emptyset$ and $K \in Nrv\mathscr{F}$. □

Definition 2.3. Let (X,d) be a metric space with $X \subset \mathbb{R}^d$ a finite point set. The *Čech complex of* X for radius r is the nerve complex of the balls centered at the points in X:

$$\check{C}ech_r(X) := \{Q \subset X : \bigcap_{x\in Q} D_r(x) \neq \emptyset\}$$

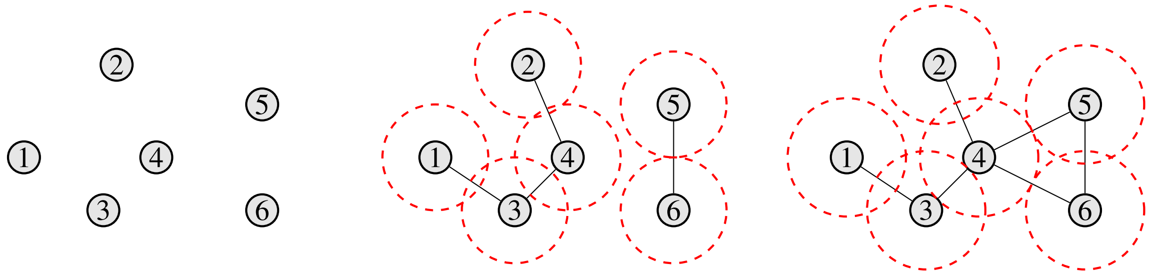


Figure 2.1: $\check{C}ech_r(X)$ for respectively $r = 0, 1, \frac{\sqrt{5}}{2}$

Notice that for both the Čech and Vietoris-Rips complexes as we scale up the parameters t and r respectively, new simplices keep being added and our new resulting complexes are susceptible to change their structure, all while being computed from the same datacloud. That core idea behind a filtration will be our focus.

The Čech and Vietoris-Rips filtrations are just two very useful examples. Another one, which will be the one we use in Chapter 3 is the so-called level set filtrations.

Definition 2.4. Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ be a function. The *superlevel set* of f for a parameter $t \in \mathbb{R}$ is the subset of $\mathbb{R}^d$

$$L(f)_t = \Big\{ x \in \mathbb{R}^d : f(x) \geq t \Big\}$$

Note that $s \leq t$ implies $L_t(f) \subseteq L_s(f)$.

In an analogous way one can define the *sublevel set* of f.

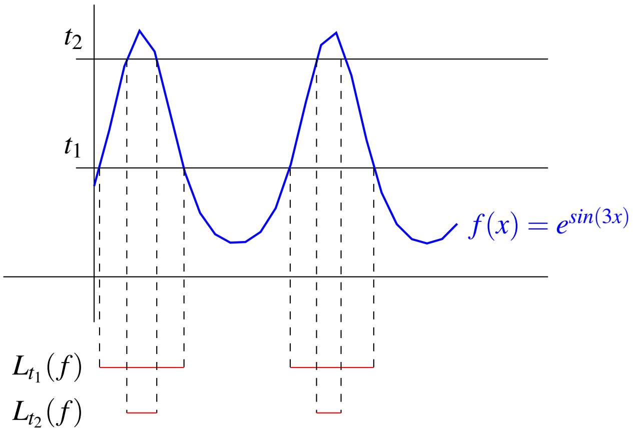


Figure 2.2: An example of superlevel set for $f : \mathbb{R} \longrightarrow \mathbb{R}$ for two different parameters t_1, t_2

If f is restricted to a finite domain $K \subseteq \mathbb{R}^d$ (a datacloud for example) one can define the *superlevel set filtration* $L(f)_\bullet$ as:

$$L(f)_t = \Big\{ conv(X) \,\Big|\, f(x) \geq t, \forall x \in X \Big\}$$

Going back to the Čech and Vietoris-Rips filtrations, consider them both computed from the same data cloud X (That is the case represented in Figures 1.3 and 2.1)

Even though they are different, they are close to each other in a sense that we will precise now.

Definition 2.5. Two filtrations $K_\bullet, L_\bullet$ are δ-*interleaved* for $\delta \geq 0$ if $K_t \subseteq L_{t+\delta}$ and $L_t \subseteq K_{t+\delta}$ $\forall t \in \mathbb{R}$. The *interleaving distance* between $K_\bullet$ and $L_\bullet$ is defined as:

$$d_I(K_\bullet, L_\bullet) = inf\{\delta \geq 0 : K_\bullet,\, L_\bullet \text{ } are \text{ } \delta - interleaved\}$$

Back to our example, it is quick to see that $\check{C}ech_t(X) \subseteq Rips_{2t}(X)$. For the other inclusion we have theorem formulated by Jung in 1901, but first we will give a preliminary result that is used in the proof:

Proposition 2.2. *(Special Karush-Kuhn-Tucker conditions) Let $P, Q, E \subset \mathbb{R}$ be disjoint finite sets of points. Let S be a sphere that encloses Q, contains P, and excludes E. Then S is the smallest such sphere iff its center is an affine combination of the points $x \in Q \cup P \cup E$,*

$$z = \sum_x \lambda_x x\,,\ 1 = \sum_x \lambda_x$$

such that

1. $\lambda_x = 0$ *whenever x does not lie on S.*
2. $\lambda_x \leq 0$ *whenever $x \in E$.*
3. $\lambda_x \geq 0$ *whenever $x \in Q$.*

Theorem 2.1. *Let $Q \subset \mathbb{R}^d$ be a set with diameter t. Then Q is contained in a closed ball with radius $r \leq \vartheta t$ for $\vartheta = \sqrt{\frac{d}{2(d+1)}}$*

Proof. Let S be the smallest enclosing sphere of our set Q with center z and radius r. We will prove that S is contained in the ball of radius ϑt

The special Karush-Kuhn-Tucker conditions imply that z can be expressed as an affine combination $\sum_{i=1}^n \lambda_i q_i$ for $q_1, ..., q_n$ affinely independent points lying in Q, therefore $n \leq d+1$.

Now, let $x_i = q_i - z$, we have $||x_i|| = r$ for all $i = 0, ..., n$ and $\sum_{i=1}^n \lambda_i x_i = 0$. Thus, we have

$$diam(Q)^2 \geq ||x_i - x_k||^2 = ||x_i||^2 + ||x_k||^2 - 2 < x_i, x_k >= 2(r^2 - < x_i, x_k >)$$

For a fixed k we have

$$2r^2 = \sum_{i=1}^n \lambda_i 2r^2 = \sum_{i=1}^n \lambda_i(||x_i - x_k||^2 - 2 < x_i, x_k >) =$$

$$= \lambda_k ||x_k - x_k||^2 + \sum_{i=1, i\neq k}^n \lambda_i ||x_i - x_k||^2 - 2\sum_{i=1}^n \lambda_i < x_i, x_k >=$$

$$= \sum_{i=1, i\neq k}^n \lambda_i ||x_i - x_k||^2 - 2 < \sum_{i=1}^n \lambda_i x_i, x_k >= \sum_{i=1, i\neq k}^n \lambda_i ||x_i - x_k||^2$$

Thus, summing for $1 \leq k \leq n$ we get

$$n(2r^2) = \sum_{i,k=1, i\neq k}^n \lambda_i ||x_i - x_k||^2 \leq \sum_{i,k=1, i\neq k}^n \lambda_i diam(Q)^2 = \sum_{k=1,}^n (1-\lambda_k) diam(Q)^2 = (n-1) diam(Q)^2$$

Since $n \leq d+1$ we have

$$r \leq diam(Q)\sqrt{\frac{d}{2(d+1)}} = t\sqrt{\frac{d}{2(d+1)}}$$

□

Corollary. We have that

$$\check{C}ech_t(X) \subseteq Rips_{2t}(X) \subseteq \check{C}ech_{\vartheta 2t}$$

which does not give us an interleaving distance for our filtrations. However, if we reparametrize them by $t = e^\lambda$, since we have that $2e^{t+log2\vartheta} = 2\vartheta t \geq 2t$, the *logarithmic Čech and Vietoris-Rips filtrations* $\check{C}ech_{e^\bullet}$, $Rips_{2e^\bullet}$ verify

$$\check{C}ech_{e^\lambda}(X) \subseteq Rips_{2e^\lambda}(X) \subseteq Rips_{2e^{\lambda+log(2\vartheta)}}(X)\ ,\ Rips_{2e^\lambda}(X) \subseteq \check{C}ech_{2e^{\lambda+log2\vartheta}}$$

and so they are $log2\vartheta$-interleaved.

As we have now defined what a filtration is, now we want to compute the homology groups of each simplicial complex involved in the process. This will be what we call persistent homology, and will allow us to observe which of them last longer throughout the filtration and thus is more inherent to the point structure.

2.2 Persistent homology

In order to give a more easy to follow explanation of persistent homology, this next two sections are written approaching homology as built from the chain space over $\mathbb{Z}_2$. In Section 1.3 we have constructed homology working over $\mathbb{Z}$, the constructions for $\mathbb{Z}_2$ are similar but in case of doubt, see [1] and [2] for more details.

Definition 2.6. A *persistence module* is a map from a totally ordered set T that assigns:

1. $t \mapsto V_t$ where all V_t are modules over the same ring R
2. $s \leq t \mapsto V_s \longrightarrow V_t$: for every ordered pair a morphism between R-modules

The next remark gives us interpretations of concepts in the previous section under our new lens.

Remark. A simplicial filtration $K_\bullet$ indexed over $\mathbb{R}$ induces, together with inclusion maps between steps, a persistence module for chain spaces $\mathbb{R} \longrightarrow C_d(K_\bullet)$ sending elements and ordered pairs:

$$r \mapsto C_d(K_r)$$

$$r \leq s \mapsto C_d(K_r) \hookrightarrow C_d(K_s)$$

Some clarification is needed in order to understand what we have in our hands right now. As we can construct a chain complex over each element of our filtration, we as well have a relation between chain complexes throughout the filtration.

$$\begin{tikzcd}
\dots \arrow[r, "\partial_{d+2}"] & C_{d+1}(K_{j-1}) \arrow[r, "\partial_{d+1}"] & C_d(K_{j-1}) \arrow[r, "\partial_d"] \arrow[d, "\iota_{j-1}"] & C_{d-1}(K_{j-1}) \arrow[r, "\partial_{d-1}"] & \dots \\
\dots \arrow[r, "\partial_{d+2}"] & C_{d+1}(K_j) \arrow[r, "\partial_{d+1}"] & C_d(K_j) \arrow[r, "\partial_d"] \arrow[d, "\iota_j"] & C_{d-1}(K_j) \arrow[r, "\partial_{d-1}"] & \dots \\
\dots \arrow[r, "\partial_{d+2}"] & C_{d+1}(K_{j+1}) \arrow[r, "\partial_{d+1}"] & C_d(K_{j+1}) \arrow[r, "\partial_d"] & C_{d-1}(K_{j+1}) \arrow[r, "\partial_{d-1}"] & \dots
\end{tikzcd}$$

Similarly, via the inclusions present in the filtration we have induced maps between homologies $\forall d \in \mathbb{N} \cup \{0\}$

$$\begin{tikzcd}
H_d(K_{j-1}) \arrow[d, "\iota_{j-1}^*"] \\
H_d(K_j) \arrow[d, "\iota_j^*"] \\
H_d(K_{j+1})
\end{tikzcd}$$

Now we can present the definition of persistent homology.

Definition 2.7. The *persistent homology* of a simplicial filtration $K_\bullet$ is the persistence module

$$\mathbb{R} \longrightarrow H_*(K_\bullet)$$

assigning to each real number r the homology of its corresponding step of the filtration and to ordered pairs the induced linear maps from the inclusions .

$$r \mapsto H_*(K_r)$$
$$r \leq s \mapsto H_*(K_r) \hookrightarrow H_*(K_s)$$

Note that the map between homologies of the filtration isn't an inclusion but the induced chain map from the inclusion, which may not, and many times will not, be injective.

By following our linear maps, in persistence homology we find a tool to study the development of topological holes in a filtration. As the Betti numbers β_* of the complexes grow or shrink with every step holes are being created and removed. This, paired with a given basis of the homology, allows us to identify the holes that are characteristic of the data cloud's structure, those that *persist* for long periods, from those that arise occasionally and can be considered as noise.

Finally, in the next section we will show a way to encode the information about the lifespan of these holes throughout the filtration. All will be based in the next fundamental theorem, which we will prove for the case of discrete persistence modules, i.e., those indexed by a discrete totally ordered set.

Definition 2.8. Let $M_1, M_2 : N \longrightarrow Vect_{\mathbb{Z}_2}$ be persistence modules. The *direct sum* $M_1 \oplus M_2 : N \longrightarrow Vect_{\mathbb{Z}_2}$ is the persistence module with

$$(M_1 \oplus M_2)(i) = (M_1)(i) \oplus (M_2)(i)$$

and

$$(M_1 \oplus M_2)(i \leq j) = (M_1)(i \leq j) \oplus (M_2)(i \leq j)$$

Theorem 2.2. *(Structure theorem for persistence modules) Let (M, α) be a persistence module over a field $\mathbb{F}$ such that $\forall j \geq 0$ the module M_i is finite-dimensional and $\forall j \gg 0$ all α_j are invertible. Then there exists a unique pair consisting of:*

1. *A finite set $Bar(M, \alpha)$ of intervals $[i, j)$ with $i \in \mathbb{Z}_{\geq 0}$, $j \in \mathbb{Z}_{\geq 0} \cup \{+\infty\}$*
2. *A function $\mu : Bar(M, \alpha) \longrightarrow \mathbb{Z}_{>0}$*

such that $(M, \iota) \cong \bigoplus_{[i,j] \in Bar(M,\alpha)} (I^{i,j}, c^{i,j})^{\mu[i,j]}$, where $(I^{i,j}, c^{i,j})$ is the persistence module of the form:

$$0 \to 0 \to \dots \underset{i}{\to} \mathbb{F} \xrightarrow[i+1]{id} \mathbb{F} \xrightarrow[i+2]{id} \dots \xrightarrow[j]{id} \mathbb{F} \xrightarrow[j+1]{} 0 \to \dots$$

Proof. Consider a persistence module (M, ι) satisfying the above conditions. Let n be the largest integer such that i_j is invertible $\forall j \geq n$. Take the truncation of the persistence module

$$M_0 \xrightarrow{\iota_1} M_1 \xrightarrow{\iota_2} M_2 \xrightarrow{\iota_1} M_3 \xrightarrow{\iota_4} \dots \xrightarrow{\iota_n} M_n$$

From this persistence module we construct the graded ring $M = \bigoplus_{i=0}^{n} M_i$ and define the shift map $t : M \longrightarrow M$ as:

$$(x_0, x_1, \dots, x_n) \mapsto t(x_0, x_1, \dots, x_n) = (0, \alpha_0(x_0), \alpha_1(x_1), \dots, \alpha_{n-1}(x_{n-1}))$$

Now, we have that M is a graded module over the polynomial ring $\mathbb{F}[t]$. Since $\mathbb{F}$ is a field, we have that polynomial rings over a field are principal ideal domains. Therefore, we have that M is a finitely generated module over a principal ideal domain and so its structure is well known. It is of the form:

$$M = \bigoplus_{i} (t^i\mathbb{F}[t])^{\mu[i,+\infty)} \bigoplus_{[i,j)} (t^i\mathbb{F}[t]/ < t^j >)^{\mu[i,j)}$$

Hence, by defining the set of coefficients present in this sum we can observe that at a given step k our persistence module will have a module which will correspond to the k-th degree elements, being of the form

$$M_k = \bigoplus_{k\in[i,j)} \mathbb{F}^{\mu[i,j)}$$

and the maps between steps of the persistence module are those given by applying t, i.e. the corresponding α.

□

Although the filtrations we have seen so far are indexed over $\mathbb{R}$, one can see that for most of the values the resulting simplicial complexes are the same. Only for a finite number of values we get an actual change in the structure. For this reason, our filtration can be discretised by reindexing to only keep the critical steps and so they are proved to have such a structure.

Remark. For this proof to work the field condition for our persistence module is essential. This means that $\mathbb{Z}$-modules are not suitable for this proof

The structure theorem allows us to state this final definition, as we have now a way to relate :

Definition 2.9. Let $i, j \in I$ with $i \leq j$. The (i, j)*-persistent homology* is defined as

$$H_*^{i,j} = \frac{Z_*^i}{B_*^j \cap Z_*^i} = \frac{ker\partial_*^i}{im\partial_*^j \cap ker\partial_*^i}$$

That is, cycles in K_i quotiented by those who become boundaries in K_j. Equivalence classes are those non-bounding cycles that "persist" throughout the marked steps of the filtration.

2.3 Computation of persistence homology: Persistence barcodes

While the usefulness of persistent homology is clear, we have not yet given a way to easily compute it so that we can actually work with it. With the aid of the structure theorem for persistence modules we can recover the persistent homology from a certain finite set of intervals and a multiplicity function.

In the following section we will give an algorithm to compute this structural multiset: the persistence barcode, an elaborate on the details of this construction.

First, one needs a certain kind of filtration in order to apply our algorithm.

Definition 2.10. Let K be a simplicial complex. A *simplexwise filtration* $K_\bullet$ of K is a filtration $\{K_i\}_{1\leq i\leq n}$ such that $K_i = K_{i-1} \cup \{\sigma_i\}$ for $1 \leq i \leq n$, where $K = K_n$ and $\sigma_i \in K$ is a simplex.

The following proposition is of importance for our purpose, as it tells us how to go from a regular filtration to a simplexwise one.

Proposition 2.3. *One can turn a regular filtration into a simplexwise filtration by decomposing each step into adding a single simplex following any order that verifies:*

$$\sigma_a \textit{ goes before } \sigma_b \Longrightarrow \sigma_a \in K_i, \sigma_b \notin K_i \vee dim\sigma_a \leq dim\sigma_b$$

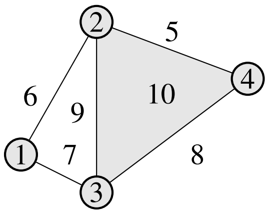
From a simplexwise filtration we then construct the following object:

Definition 2.11. The *d-th boundary matrix* D_d of a simplexwise filtration $K_\bullet$ is the matrix of the boundary map $\partial_d(K)$ with respect to the ordered basis $\{\sigma_i\}_{1\leq i\leq n}$. It is filled as follows:

$$(D_d)_{i,j} = 1 \Leftrightarrow \sigma_i \subseteq \sigma_j$$

Remark. Note that in the previous definition not all indexes are present. Only those corresponding to those of dimension d and $d-1$ appear in the matrix and this may lead to confusion.

To fix this, we will abuse the notation a bit and will write $\partial : C_* \longrightarrow C_*$ referring to the collection of all boundary maps of the chain complex of K, and respectively D to the $n \times n$ matrix containing all facet relations, i.e, $D_{i,j} = 1 \Leftrightarrow \sigma_i$ is a **facet** of σ_j.



$$\begin{bmatrix} 0&0&0&0&0&\boxed{1}&\boxed{1}&0&0&0\\ 0&0&0&0&\boxed{1}&\boxed{1}&0&0&\boxed{1}&0\\ 0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0\\ 0&0&0&0&\boxed{1}&0&0&\boxed{1}&0&0\\ 0&0&0&0&0&0&0&0&0&\boxed{1}\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&\boxed{1}\\ 0&0&0&0&0&0&0&0&0&\boxed{1}\\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

Figure 2.3: To the left a simplexwise filtration of a simplicial complex and to the right its boundary matrix D

The choice of ordered basis in our context will be given by the filtration chosen (Rips, Čech,...etc.) as well as any order satisfying the criteria for a simplexwise decomposition. This ordering of the basis is nonimportant in terms of the homology itself, since a different basis should still span the same homology, but is nontheless neccesary in order to implement our method for computing the barcode.

The following algorithm takes the matrix D and returns its reduced form R, from which we will construct the persistence barcode of the simplexwise filtration. This algorithm is similar to that used for obtaining the Smith normal form of a matrix for coefficients in $\mathbb{Z}_2$, but in our case it respects the placement of pivots in order for us to be able to extract as well a basis for the homology.

Remark. In the algorithm the notation used corresponds to:

1. For M an $I \times I$ matrix and $i \in I$, M_i is the ith column of M.

2. $$pivotM_i = min\{j \in I : m_ik = 0 \forall k > j\}$$

3. $$pivotsM = \{pivotM_i : i \in I\} \setminus \{0\}$$

Algorithm 1. (Matrix reduction algorithm):

```
INPUT: $D : n \times n$ matrix
OUTPUT: $R$ reduced
$R \leftarrow D$,
while $\exists i < j : pivotR_i = pivotR_j$ do
    $R_j \leftarrow R_i + R_j$
end while
return $R$
```

Some explanation is needed in order to interpret this result.

Every zero column R_i of the reduced matrix R represents a cycle in the step K_i, as we, via the operations being made to D in the algorithm, have been able to nullify its boundaries by adding elements already present in the filtration. As such, finding a nonzero entry in a certain row i signifies that σ_i is a boundary in the j-th simplicial complex of the filtration K_j.

The meaning behind finding a pivot in row i at column R_j is that the addition of σ_i creates a homology class that disappears once we add σ_j.

Thus, intervals $[i,j)$ represent homology classes that appear in K_i and disappear from K_j on, while intervals $[j,\infty)$ represent persisting homology classes born in K_j

From the output of this algorithm we get a reduced matrix R whose pivots are unique. This matrix is very rich in information and allows us to define the persistent barcode as we present next.

Definition 2.12. Let $K_\bullet$ be a simplexwise filtration. The *persistence barcode* of $K_\bullet$ is the collection of intervals

$$Barc(H_*(K_\bullet)) = \{[i,j) : R_j \neq 0, i = pivotR_j\} \cup \{[j,\infty) : R_j = 0, j \notin pivotsR\}$$

where R is the reduced matrix produced by the matrix reduction algorithm.

The *homological dimension d* of a bar $[a,b)$ in $Barc(H_*(K_\bullet))$ is the dimension of the simplex σ_a

Example. The barcoding of the boundary matrix D of the simplexwise filtration presented in Figure 2.3 will help clarify this procedure:

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & \boxed{1} & \boxed{1} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \boxed{1} & \boxed{1} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \boxed{1} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \boxed{1} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\qquad
\begin{aligned}
&Barc(H_0(K)) = \Big\{[1,+\infty),[2,6),[3,7),[4,5)\Big\}\\
&Barc(H_1(K)) = \Big\{[8,+\infty),[9,10)\Big\}\\
&Barc(H_2(K)) = \emptyset
\end{aligned}$$

Figure 2.4: The reduced matrix from the previous example together with its barcode

The algorithm has performed the following operations to reduce R:

$$R_8 \leftarrow R_5 + R_8$$

$$R_8 \leftarrow R_7 + R_8$$

$$R_8 \leftarrow R_6 + R_8$$

$$R_9 \leftarrow R_7 + R_9$$

$$R_9 \leftarrow R_6 + R_9$$

Check how the operations work out in the end to get an intuition of the workings behind our algorithm.

The running time for our algorithm is at most cubic to the number of simplices in the filtration. Note that the matrix D is sparse but this does not imply that R is, although very often that is the case.

As we announced, barcoding is not only useful for representation. The following proposition tells us how to interpret the persistence barcode in order to extract information from it:

Proposition 2.4. *Let $K_\bullet$ be a simplexwise filtration.*

1. *For $i \in \mathbb{N}$, $\beta_d(K_i)$ is equal to the number of bars $I \in Barc(H_*(K_\bullet))$ of dimension d such that $i \in I$*
2. *For $i \leq j \in \mathbb{N}$, the rank of the linear map $H_d(K_i) \longrightarrow H_d(K_i)$ is equal to the number of bars $I \in Barc(H_*(K_\bullet))$ of dimension d such that $i, j \in I$*

Notice we haven't computed homology. The next proposition tells us how to obtain the persistent homology of $K_\bullet$ from the persistence barcode.

Proposition 2.5. *Let $K_\bullet$ be a simplexwise filtration, $I = [i, j)$ or $[i, \infty)$ and $\mathbb{F}(I)_\bullet : N \longrightarrow Vect_{\mathbb{Z}_2}$ the persistence module*

$$\mathbb{F}(I)_t = \begin{cases} \mathbb{F} & if\ t \in I \\ 0 & otherwise \end{cases}, \quad \mathbb{F}(I)_{s \leq t} = \begin{cases} Id_{\mathbb{F}} & if\ s,t \in I \\ 0 & otherwise \end{cases}$$

Then

$$H_*(K_\bullet) \cong \bigoplus_{I \in Barc(H_*(K_\bullet))} \mathbb{F}(I)_\bullet$$

Proof. It is a special case of the structure theorem for persistent modules. If we consider the finite set in the statement of the theorem as a multiset where repetition of intervals is given by the function μ, then we have $Barc(H_*(K_\bullet)) = Bar(H_*(K_\bullet), \iota^*)$ □

One other very important property of persistence barcodes is that they are *stable invariants*. Small perturbations of the data generating them won't lead to drastic changes on the resulting barcode.

As when talking about interleavings of filtrations, we need a notion of distance for barcodes to make such a claim as stability.

Definition 2.13. Let B, B' be barcodes. A δ*-matching* between barcodes B and B', noted $\gamma : B \nrightarrow B'$ is a bijection $\gamma' : U \longrightarrow U'$ between subsets $U \subseteq B$, $U' \subseteq B'$ such that the following holds:

1. If a bar $[b, d)$ of any barcode is unmatched, then $d - b < 2\delta$
2. If $[b, d) \in B$ is matched to $[b', d') \in B'$, then

$$b \in [b' - \delta, b' + \delta] \wedge d \in [d' - \delta, d' + \delta]$$

Definition 2.14. The *bottleneck distance* between barcodes B and B' is defined as

$$d_{bot}(B, B') := inf\{\delta \in [0, \infty) : \exists\ B \longrightarrow B', \delta - matching\}$$

The bottleneck distance is symmetric. For barcodes B_1, B_2 and B_3, a δ_1-matching $\gamma_1 : B_1 \longrightarrow B_2$ and a δ_2-matching $\gamma_2 : B_2 \longrightarrow B_3$, the composition $\gamma_2 \circ \gamma_1 : B_1 \longrightarrow B_3$ is a $(\delta_1 + \delta_2)$-matching. It follows that bottleneck distance satisfies the triangle inequality.

Persistence modules are stable with respect to the bottleneck distance. Proofs for certain cases of persistence modules can be found in [2], although we will not go delve further.

To see how the bottleneck distance of a persistence barcode relates to the interleaving distance, we give without proof the following theorem:

Theorem 2.3. *Let $H_*(K_\bullet)$, $H_*(L_\bullet)$ be persistent homologies of filtrations $K_\bullet$ and $L_\bullet$ respectively. Then,*

$$d_bot(BarcH_*(K_\bullet), H_*(L_\bullet)) \leq d_I(K_\bullet, L_\bullet)$$

Proof. See [6]. □

Chapter 3

An example of barcoding and bottleneck distance

As to give closure to our work, we will dedicate this final part to present an example of the usage of the main libraries and fundamental tools of Topological Data Analysis.

To make it more interesting, we have based our example on the study used in the paper by Paul Lawson et al that can be found here [7]. The abstract of this paper reads:

[...] The Gleason score is currently the most powerful prognostic predictor of patient outcomes; however, it suffers from problems in reproducibility and consistency due to the high intra-observer and inter-observer variability amongst pathologists. In addition, the Gleason system lacks the granularity to address potentially prognostic architectural features beyond Gleason patterns. We evaluate prostate cancer for architectural subtypes using techniques from topological data analysis applied to prostate cancer glandular architecture. In this work we demonstrate the use of persistent homology to capture architectural features independently of Gleason patterns.[...] Our results indicate the ability of persistent homology to cluster prostate cancer histopathology images into unique groups with dominant architectural patterns consistent with the continuum of Gleason patterns. [...]

In this study, they compute sublevel set filtrations out of samples of a dataset with a total of 5.182 images of stained prostate cancer tissue. They combine this with other methods of classification in order find subtypes in their architecture. This paper is a good example of a real-life application of Topological Data Analysis that mixes both TDA and traditional techniques.

Back to our example, we have taken a 512×512 image from the mentioned dataset and wish to construct a sublevel set filtration from it. To do so, we will sue the Lower Star Image Filtration function found in the Ripster.py package [8], itself included in the library GUDHI for Python. To apply it we have to first convert our image to greyscale as we will use the intensity function f which ranges from 0 to 255, where the lower the number the darker the pixel. In addition to this, in order to reduce the number of connected components, we add some blur to the image.

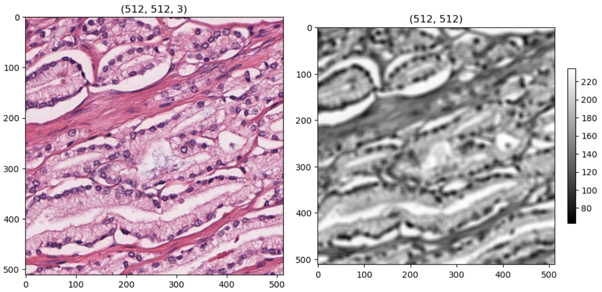


Figure 3.1: On the left the original image and on the right its conversion to greyscale after blurring

From the resulting image, a simplicial complex is constructed by assigning to each pixel a vertex and connecting it to each adjacent pixel. The corresponding pixel gives a value of f for each vertex and edges are assigned the value corresponding to the maximum of its endpoints, this way we can construct adjacency simplicial complexes with respect to a sublevel set filtration of f. As our example will be only centered in connected components we will only add 0 and 1-simplices.

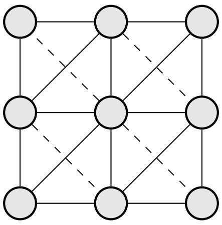


Figure 3.2: An example of our simplicial complex constructed for a 3×3 image

We compute the superlevel set filtration of this function f and get its persistence barcode in the shape of a persistence diagram, where points represent intervals with beginning and endpoints those corresponding to its coordinates.

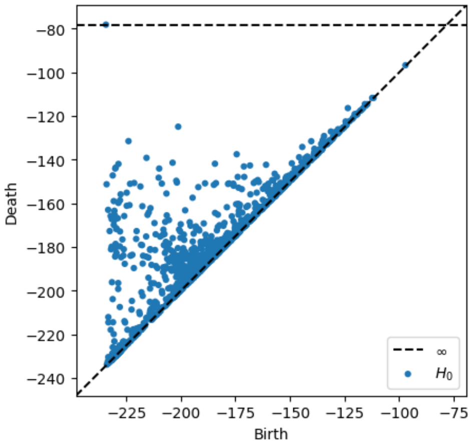


Figure 3.3: The persistence diagram of our image of tissue

The blue dot in the upper left corner corresponds to the final connected component that absorbs all others when our simplicial complex becomes connected.

We will now repeat the process with two other images, one that resembles the original and one that is clearly different.

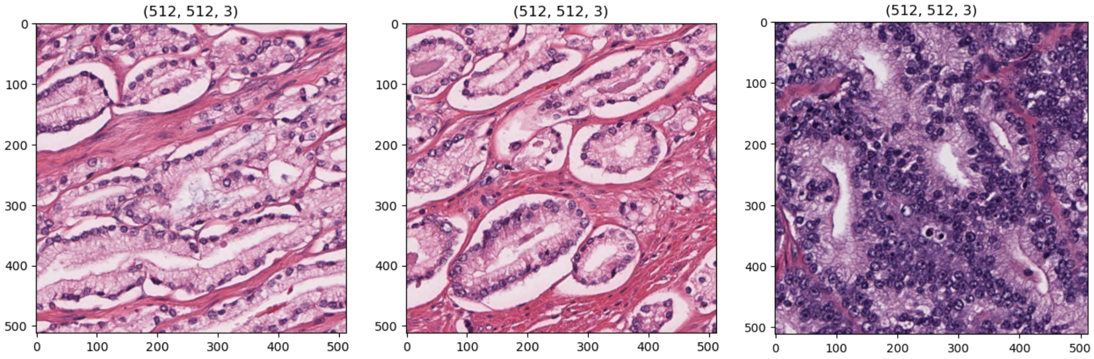


Figure 3.4: Images 1, 2 and 3

Using the bottleneck function we get the overlapped persistence diagrams. This show the distances between matched intervals as a segment in red, the one that determines the bottleneck distance, and some other green segment, corresponding to those matched bars with distance lower than the bottleneck. Note that in this process the points corresponding to infinite death time are ignored.

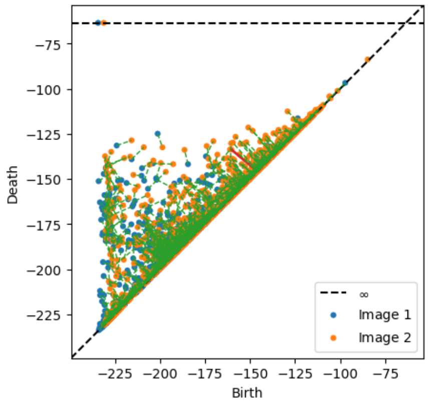


Figure 3.5: Matching of the persistence diagrams for images 1 and 2. Their bottleneck distance is 12.512161254882812

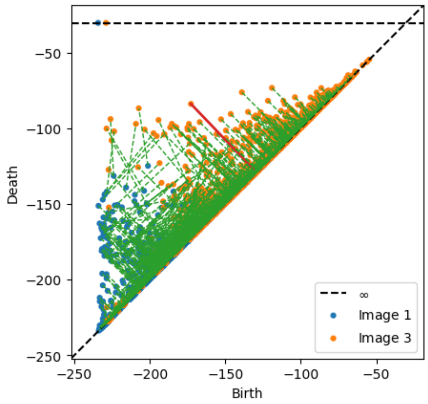


Figure 3.6: Matching of the persistence diagrams for images 1 and 3. Their bottleneck distance is 43.21422576904297

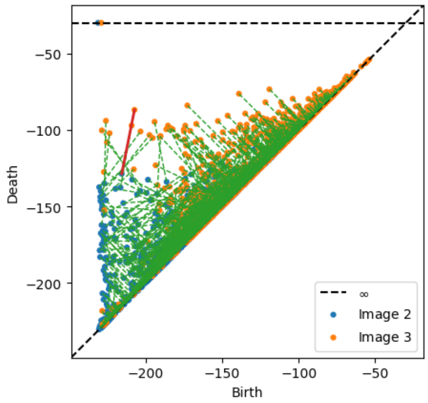


Figure 3.7: Matching of the persistence diagrams for images 2 and 3. Their bottleneck distance is 41.813507080078125

As we can see, the bottleneck distance is able to differentiate images. This and many more applications make persistent homology a very relevant tool and explain the amount of research effort being poured into TDA as of today.

Bibliography

[1] JEAN-DANIEL BOISSONNAT, FRÉDÉRIC CHAZAL, MARIETTE YVINEC, *Geometric and Topological Inference.*, Cambridge University Press, 2018. ffhal-01615863v2f https://inria.hal.science/hal-01615863v2.

[2] HERBERT EDELSBRUNNER, *A Short Course in Computational Geometry and Topology*, Springer Cham, 2014 ISBN: 978-3-319-05956-3 https://link.springer.com/book/10.1007/978-3-319-05957-0.

[3] EMILE JACQUARD, VIDIT NANDA, ULRIKE TILLMANN, *The space of barcode bases for persistence modules*, Journal of Applied and Computational Topology, 2022, you can find it at https://link.springer.com/article/10.1007/s41468-022-00094-6.

[4] HERBERT EDELSBRUNNER, *Algorithms in combinatorial geometry*, Berlin, [etc.] : Springer, cop. 1987. https://link.springer.com/book/10.1007/978-3-642-61568-9.

[5] WILLIAM CRAWLEY-BOEVEY, *Decomposition of pointwise finite-dimensional persistence modules*, https://arxiv.org/abs/1210.0819

[6] ULRICH BAUER, MICHAEL LESNICK, *Induced matchings and the algebraic stability of persistence barcodes*, Journal of Computational Geometry, https://doi.org/10.20382/jocg.v6i2a9

[7] LAWSON P, SHOLL AB, BROWN JQ, FASY BT, WENK C., *Persistent Homology for the Quantitative Evaluation of Architectural Features in Prostate Cancer Histology.*, Sci Rep. 2019 Feb 4;9(1):1139. doi: 10.1038/s41598-018-36798-y. PMID: 30718811; PMCID: PMC6361896. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6361896/.

[8] TRALIE ET AL., *Ripser.py: A Lean Persistent Homology Library for Python*, Journal of Open Source Software (2018), 3(29), 925, https://doi.org/10.21105/joss.00925