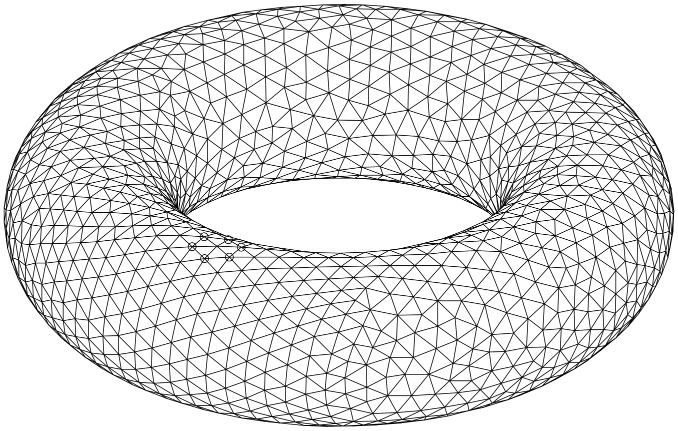
Topological Data Analysis



Written by Quirijn Meijer *under the supervision of*
prof. dr. Ieke Moerdijk *and* prof. dr. Hans Bodlaender.

Utrecht University

July 4th, 2018

Contents

1 Introduction

In the field of mathematics, topology is the study of spaces through their invariants, the properties that are preserved under continuous deformations of the spaces. From this point of view spaces are the same after being bent or stretched because even if the space is warped, the identifying properties remain unchanged. For example, as long as the deformation is continuous, no holes can be created or closed. This specific invariant will be of importance and is captured in the theory of homology. Topology is hence a qualitative approach to geometry rather than a quantitative one. Due to its abstractness it is considered arcane, inspiring the following passage from the mathematically trained novelist Alexander Solzhenitsyn:

> *„Topology! The stratosphere of human thought! In the twenty-fourth century it might possibly be of use to someone."*
>
> The First Circle (1968).

Until the end of the last millennium topology was indeed thought to be securely positioned within the confines of pure mathematics. Since then however, the emergence of applied topology was seen, with topological data analysis in its vanguard. Akin to this topological data analysis is the centerpiece of this research project. In brief, topological data analysis uses tools from topology to study the geometrical properties of a dataset. In particular the theory of homology was reinterpreted to accomodate a measure of persistence, so that persistent geometrical holes can be distinguished from noise. The goal of this work is to peer into this area, acquainting the reader with the theory and its applications. Assuming knowledge from first courses in point-set topology and abstract algebra, it is structured as follows.

Chapter 2 collects the theory needed to bridge the gap between the presumed knowledge of the reader and the theoretical basis of the subsequent chapters of this work. This preliminary chapter contains all essential definitions and theorems, and those only briefly required are given in their respective sections.

Chapter 3 introduces the reader to point clouds, filtrations of simplicial complexes, and the salient persistent homology theory defined thereon. Persistent homology is the tool from topological data analysis that is most thoroughly studied and used in this work.

Chapter 4 is based on a small script in the R language that makes it possible to easily compute summaries of the homological information detected in a given point cloud dataset.

Chapter 5 illustrates how the theory of persistent homology has been applied thus far. Examples are taken from engineering, statistics and neuroscience in such a way that a variety of manners in which it can be utilised is illuminated.

Chapter 6 contains a short discussion of other important results in topological data analysis and glances on the research currently being done. In addition a section is dedicated to a possibly novel data clustering algorithm, the idea of which was perceived during this research project.

In essence this work encompasses all theory and examples required to understand this new method of data analysis, starting from the preliminaries.

2 Preliminaries

Algebraic topology, and topological data analysis specifically, requires knowledge from a slew of areas in mathematics. In this chapter the relevant notions from these areas are presented in rapid succession. Starting with category theory and ending with algebraic topology the intermezzo of abstract algebra is a natural occurance. As the name of the area of algebraic topology does betray, algebra fills a vital role in its studies.

2.1 Category Theory

Category theory is a framework meant to formalise mathematical structure by defining objects as belonging to categories and bundling them with their morphisms. The abstraction this formalisation provides is of tremendous power as it allows statements to be made about entire categories of objects rather than the individual objects themselves. In order to start exploring this framework the components that make up a category $\mathcal{C}$ are discussed.

First a class of objects of any nature is required and will be denoted by $\mathrm{Ob}(\mathcal{C})$. Furthermore morphisms between the objects in this class are collected in a separate class $\mathrm{Mor}(\mathcal{C})$. In addition the definition of a binary composition rule for morphisms is required. For A, B and C objects with $f : A \to B$, $g : B \to C$ morphisms between them the result of this composition should be a morphism $g \circ f : A \to C$. With these constituents a category can be defined.

Definition 2.1. (Category). A category $\mathcal{C}$ is the combination of a class of objects, a class of morphisms between these objects and a composition rule so that the following axioms hold.

1. The operation that takes the composition of morphisms is associative. If $f : A \to B$, $g : B \to C$ and $h : C \to D$ are morphisms then $h \circ (g \circ f) = (h \circ g) \circ f$;
2. For every object A there exists an identity morphism $1_A : A \to A$ so that for any two morphisms $f : A \to B$ and $g : B \to A$ the equalities $f \circ 1_A = f$ and $1_A \circ g = g$ hold.

Among categories a distinction is made. If $\mathrm{Ob}(\mathcal{C})$ and all collections $\mathrm{Mor}(X, Y)$ of morphisms between objects $X, Y \in \mathrm{Ob}(C)$ are sets rather than proper classes, $\mathcal{C}$ is called a *small category*. Otherwise it is called a *large category*. All following statements apply regardless of the involved categories being small or large.

Definition 2.2. (Isomorphisms in Categories). Two objects A and B in $\mathcal{C}$ are called isomorphic if there exist morphisms $f : A \to B$ and $g : B \to A$ such that $g \circ f = 1_A$ and $f \circ g = 1_B$. The morphisms f, g are called isomorphisms.

Generalising from morphisms in categories one arrives at the idea of mappings between categories. Two such are treated in the next two definitions.

Definition 2.3. (Covariant Functor). For $\mathcal{C}_1$ and $\mathcal{C}_2$ two categories, let F be a map that assigns to each object X in $\mathcal{C}_1$ an object $F(X)$ in $\mathcal{C}_2$ and to each morphism $f : A \to B$ in $\mathcal{C}_1$ a morphism $F(f) : F(A) \to F(B)$ in $\mathcal{C}_2$. Then F is called a covariant functor from the category $\mathcal{C}_1$ to the category $\mathcal{C}_2$ if the following two conditions hold.

1. If f is an identity morphism, then $F(f)$ is an identity morphism;
2. If a composition $f \circ g$ is well-defined, then $F(f \circ g) = F(f) \circ F(g)$.

An alternative functor is defined as follows.

Definition 2.4. (Contravariant Functor). Let G be as the map in the definition of the covariant functor with the difference that each morphism $f : A \to B$ in $\mathcal{C}_1$ is mapped to a morphism $G(f) : G(B) \to G(A)$ in $\mathcal{C}_2$. Then G is said to be a contravariant functor from the category $\mathcal{C}_1$ to the category $\mathcal{C}_2$ if the following two conditions hold.

1. If f is an identity morphism, then $G(f)$ is an identity morphism;
2. If a composition $f \circ g$ is well-defined, then $G(f \circ g) = G(g) \circ G(f)$.

Before concluding with an important theorem, a transformation between functors is defined.

Definition 2.5. (Natural Transformation). Given two covariant functors F and G from $\mathcal{C}_1$ to $\mathcal{C}_2$, a natural transformation $\tau : F \to G$ is a function that assigns to each object A in $\mathcal{C}_1$ a morphism $\tau(A) : F(A) \to G(A)$ in $\mathcal{C}_2$ so that for every morphism f in $\mathcal{C}_1$ the below diagram commutes.

$$\begin{tikzcd}
F(A) \arrow[r, "F(f)"] \arrow[d, "\tau(A)"'] & F(B) \arrow[d, "\tau(B)"] \\
G(A) \arrow[r, "G(f)"'] & G(B)
\end{tikzcd}$$

If both F and G are contravariant functors a natural transformation between them is defined analogously.

Natural transformations are often referred to as morphisms of functors. If for every X in $\mathcal{C}_1$ the *component* $\tau(X)$ is an isomorphism in $\mathcal{C}_2$, τ is said to be a *natural isomorphism.* Natural isomorphisms can be used to define the parallel of isomorphisms between categories. An equivalence of the categories $\mathcal{C}_1$ and $\mathcal{C}_2$ consists of two functors $F : \mathcal{C}_1 \to \mathcal{C}_2$, $G : \mathcal{C}_2 \to \mathcal{C}_1$ and two natural isomorphisms $G \circ F \to 1_{\mathcal{C}_1}$, $F \circ G \to 1_{\mathcal{C}_2}$ for $1_{\mathcal{C}_i}$ the identity functors of the categories. Equivalent categories share results as they are translated through the equivalence. Remaining on the topic of isomorphisms, the promised theorem is given.

Theorem 2.6. *Let $F : \mathcal{C}_1 \to \mathcal{C}_2$ be a functor. If two objects in $\mathcal{C}_1$ are isomorphic, then so are their images under F in $\mathcal{C}_2$. If the images of the objects are not isomorphic, then neither are the two objects.*

Proof. Assume F is a covariant functor. If A and B are two isomorphic objects in $\mathcal{C}_1$ there must exist two morphisms $f : A \to B$ and $g : B \to A$ in $\mathcal{C}_1$ for which $g \circ f = 1_A$ and $f \circ g = 1_B$. It follows from the definition of the covariant functor that $F(g) \circ F(f) = F(g \circ f) = F(1_A) = 1_{F(A)}$ and $F(f) \circ F(g) = 1_{F(B)}$, demonstrating that $F(A)$ and $F(B)$ are isomorphic. The proof of the contrapositive statement is then easily derived using proof by contradiction, and the theorem is proven for a contravariant functor in an analogous way. □

For a more thorough exposition to the elements of category theory the reader is referred to *Categories for the Working Mathematician* [1] by Saunders Mac Lane.

2.2 Abstract Algebra

The most important result for topological data analysis from the area of abstract algebra is *the structure theorem.* Its formulation involves rings and modules, and in order to properly state it the relevant definitions are collected in this section. Throughout, R is assumed to be a commutative ring with unit, and the first step will be the definition of a polynomial ring.

Definition 2.7. (Polynomial Ring). The set of all polynomials $f(t) = \sum_{i=0}^{\infty} a_i t^i$ with a finite number of non-zero coefficients a_i in R forms a commutative ring with unit. It is called the polynomial ring over R in one variable and denoted $R[t]$.

With this line masquerading as a connection between the two, the next definition is given.

Definition 2.8. (Zero Divisor). An element r of R is called a left zero divisor if there exists a non-zero element x in R so that $rx = 0$. If there exists an element y in R so that $yr = 0$, r is called a right zero divisor. Unless the distinction needs to be made both are referred to as zero divisors.

If an element in R is not a zero divisor it is called a *regular* element. The concept of zero divisors is needed in order to define principal ideal domains. For giving the definition the notion of an ideal is required, as well as what it means for an ideal to be a principal ideal.

Definition 2.9. (Ideal). Let $(R, +)$ be the additive group of the ring $(R, +, \cdot)$, and let I be a subset of R. Then I is called an ideal if the following holds.

1. $(I, +)$ is a subgroup of $(R, +)$;
2. For all $x \in I$ and $r \in R$, $x \cdot r \in I$.

As R is commutative both one-sided products are in I if one of them is. In general if commutativity is not assumed an ideal can be a left or right ideal depending on which one-sided product is assumed to be in I.

Definition 2.10. (Finitely Generated, Principal Ideal). Let $a_1, a_2, ..., a_n$ be a finite sequence of elements in R. The ideal I generated by these elements is:

$$(a_1, a_2, ..., a_n) = \{ \textstyle\sum_{i=1}^{n} r_i a_i \ \mid \ r_i \in R \}.$$

If I is generated by only one such element a, I is called a principal ideal.

Note that a finitely generated module is defined in much the same way as a finitely generated ring. The last three definitions are now brought together.

Definition 2.11. (Principal Ideal Domain). A principal ideal domain is a ring R in which all ideals are principal, and which admits no zero divisors.

It is a commonly known result that any two elements in a principal ideal domain have a greatest common divisor, which is exactly why the use of such a ring is required later. Before expanding on this another structure that a ring may possess is described.

Definition 2.12. (Graded Ring). A graded ring is a ring R that is isomorphic to a direct sum of abelian groups R_i indexed over $\mathbb{Z}$, i.e.

$$R \cong \textstyle\bigoplus_{i \in \mathbb{Z}} R_i,$$

so that multiplication in the latter is defined by bilinear pairings $R_m \otimes R_n \to R_{m+n}$.

Elements that occur in a factor R_n of the decomposition are called *homogeneous*, and the *degree* of all elements in R_n is defined as being n. If the ring under examination is a polynomial ring $R[t]$ the familiar notion of the degree of a single variable monomial corresponds to such a grading. It is called the *standard grading* and is precisely defined by $R_n = Rt^n$ for all $n \in \mathbb{N}$.

If a ring is equipped with a grading this leads to the possibility of its ideals being graded.

Definition 2.13. (Graded Ideal). Let R be a graded ring with components indexed over $\mathbb{Z}$. A graded ideal I of R is an ideal of R so that:

$$I \cong \textstyle\bigoplus_{i \in \mathbb{Z}} (I \cap R_i).$$

Like rings and their ideals, modules too can be equipped with a grading. More specifically this is possible if the module is taken over a graded ring and its grading can be related to the module.

Definition 2.14. (Graded Module). A graded module M over a graded ring R is a module that is isomorphic to a direct sum of submodules M_i of M indexed over $\mathbb{Z}$,

$$M \cong \bigoplus_{i \in \mathbb{Z}} M_i,$$

in such a way that the action of R on M is defined by bilinear pairings $R_m \otimes M_n \to M_{m+n}$.

The grading on both a graded ring and a graded module may be shifted. This shifting is defined in generality for graded objects, so as to capture both meanings.

Definition 2.15. (Grading Shift). Let X be a graded object. Then the shifting operator Σ^j acts on each component X_i of X by $\Sigma^j X_i = X_{i+j}$.

A shift in attention towards the main result of this section may now indeed take place.

Theorem 2.16. *(Structure Theorem). If R is a principal ideal domain then every finitely generated R-module M is isomorphic to a direct sum of finitely many cyclic R-modules. In other words, the module M can be uniquely decomposed into the direct sum:*

$$\left(\bigoplus_{i=1}^{m} R\right) \oplus \left(\bigoplus_{i=1}^{n} R/(r_i)\right)$$

in which $m, n \in \mathbb{N}$ and the r_i are non-unital and non-zero elements of R so that for every $i \in \mathbb{N}$ the divisibility relation $r_i \mid r_{i+1}$ is satisfied.

The proof can be found in the twelfth chapter of [2]. Intuitively it grants a decomposition of the module into two parts, one part being a free vector space, the other containing all torsional elements.

The following corollary can be derived from the structure theorem and will later be used to extract a summary from the topological properties the data was found to be in possession of.

Corollary 2.17. *If R is a graded principal ideal domain then every finitely generated and graded R-module can be uniquely decomposed into a direct sum of the form:*

$$\left(\bigoplus_{i=1}^{m} \Sigma^{x_i} R\right) \oplus \left(\bigoplus_{i=1}^{n} \Sigma^{y_i} R/(r_i)\right)$$

for $m, n \in \mathbb{N}$, all x_i, y_i taken from $\mathbb{N}$ and the r_i homogeneous non-unital and non-zero elements so that for every $i \in \mathbb{N}$ the divisibility relation $r_i \mid r_{i+1}$ is satisfied.

It is proven analogously to the structure theorem with additional steps that keep track of the degrees of the generating elements that effectuate the shifting.

2.3 Algebraic Topology

One of the central concepts in topology is that of a homeomorphism. If a homeomorphism between two topological spaces exists the spaces are considered to be the same under the lens of topology. It is through the homeomorphism that the source can be continuously deformed into the target space, preserving all invariants. If the restriction of this preservation of invariants is made to be less strict however, more spaces may be equated. This leads to the definition of homotopy types.

Definition 2.18. (Homotopy). A homotopy is a family of maps $f_t : X \to Y$ between two topological spaces, indexed over the unit interval I, so that the associated map $F : I \times X \to Y$ defined by $F(x, t) = f_t(x)$ is continuous. Two continuous maps $f_0, f_1 : X \to Y$ are called homotopic, denoted $f_0 \simeq f_1$, if there exists a homotopy f_t connecting them.

This definition speaks of homotopic maps but does not unveil what it means for spaces to be of the same homotopy type. The definition of the latter requires a homotopy between maps.

Definition 2.19. (Homotopy Equivalence). A continuous map $f : X \to Y$ between topological spaces is called a homotopy equivalence if there exists a continuous map $g : Y \to X$ such that $g \circ f \simeq 1_X$ and $f \circ g \simeq 1_Y$. If such maps exist the spaces X and Y are to be said homotopy equivalent, or of the same homotopy type. This is denoted $X \simeq Y$.

There is one specifically named homotopy type. A space that has the homotopy type of a point is called *contractible*. A space is contractible precisely when its identity map is *null-homotopic*, meaning it is homotopic to a constant map.

Besides homotopy types, the definition of a homotopy leads to other tools to classify topological spaces. One such tool is the functor that assigns an algebraic group called the fundamental group to a space. More precisely put this functor takes in a space and a basepoint from the space and maps this pair to a group that encodes the loops of the space based at this point up to homotopy. The complete construction requires another set of definitions.

Definition 2.20. (Path). If X is a topological space, a path in X is a continuous map $f : I \to X$.

Such a map excavates a path from the starting point $f(0)$ to the end point $f(1)$. If these points are equal the path is called a *loop* and the point is called its *basepoint*. Furthermore, two paths f_0, f_1 with shared starting and end points are *path homotopic* if there exists a homotopy between them such that for all t, $f_t(0) = f_0(0) = f_1(0)$ and $f_t(1) = f_0(1) = f_1(1)$. It can be proven that path homotopies define an equivalence relation on all paths in a space. An equivalence class under this relation with representative f is denoted $[f]$ and is called the *homotopy class* of f.

The set $\pi_1(X, x)$ of all homotopy classes of loops based in $x \in X$ form the set in the definition of the fundamental group at basepoint x of a space X. The accompanying group operation follows from the composition of paths, defined as follows for two paths f_0, f_1 so that $f_0(1) = f_1(0)$:

$$f_0 \cdot f_1(t) = \begin{cases} f_0(2t) & 0 \le t \le 1/2 \\ f_1(2t-1) & 1/2 \le t \le 1 \end{cases}$$

Proposition 2.21. *(Fundamental Group). The set $\pi_1(X, x)$ forms a group when combined with the binary operation $[f][g] = [f \cdot g]$ that maps the product of two homotopy classes to the homotopy class of their composition.*

A proof of this construction is most likely given in any book on algebraic topology. One such specific proof can be found in the first chapter of Allen Hatcher's *Algebraic Topology* [3], where it is also proven that the fundamental group is a topological invariant preserved up to isomorphism under homotopy equivalence.

The fundamental group records what loops are path homotopic. For loops that cannot be deformed into each other there must be something in the space that obstructs the process. For all intents and purposes this must be a hole, meaning the fundamental group holds information about the holes of a space. This same observation leads to the discussion of the importance of the chosen basepoint. As it turns out the choice of basepoint makes no difference if the space is connected. The fundamental group at any basepoint of such a space is unique up to isomorphism. If the space is disconnected no such claims can be made for apparent reasons.

The subscript in π_1 suggests the fundamental group is one of multiple. This is indeed the case, as it is the first of the homotopy groups π_n. The discussion of these groups however is not within scope. Instead a different method, one that is more easily computed, for encoding the holes of a space in algebraic objects is treated. Its introduction is preluded by those of simplices and chains.

Definition 2.22. (Standard n-Simplex). Let $\{e_0, ..., e_n\}$ be the standard basis of $\mathbb{R}^{n+1}$. The standard n-simplex is defined as being the polytope:

$$\Delta_n = \{ \textstyle\sum_{i=0}^n \lambda_i e_i \mid \sum_{i=0}^n \lambda_i = 1, \lambda_i \in [0,1] \}$$

in $\mathbb{R}^{n+1}$, in which the λ_i are called the barycentric coordinates.

Through the standard simplices a different type of simplex may be defined.

Definition 2.23. (Affine Singular n-Simplex). For $v_0, ..., v_n$ not necessarily independent elements of $\mathbb{R}^m$, $[v_0, ..., v_n] : \Delta_n \to \mathbb{R}^m$ is the map that takes $\Sigma_{i=0}^n \lambda_i e_i$ to $\Sigma_{i=0}^n \lambda_i v_i$. The image, or equivalently the map, is called the affine singular n-simplex.

As polytopes, simplices may have several faces. Seeing as affine singular n-simplices are defined by a convex span of elements each face, specifically itself an affine singular $(n-1)$-simplex, can be obtained by omitting one of the spanning elements. This leads to the following definition.

Definition 2.24. (Face Map). The i-th face F_i^n of an affine singular n-simplex is defined by the map $[e_0, ..., \hat{e}_i, ..., e_n] : \Delta_{n-1} \to \Delta_n$ that sets the corresponding values λ_i in Δ_n to 0.

Using these notions the simplices can be embedded in arbitrary topological spaces. It is often in this more general setting where they find their full utility.

Definition 2.25. (Singular n-Simplex). Let X be a topological space. A singular n-simplex of X is a map $\sigma_n : \Delta_n \to X$. The i-th face $\sigma \circ F_i^n$ of a singular n-simplex is denoted $\sigma^{(i)}$.

At this point the question of the role of the simplices in the process of detecting holes of a topological space looms. Before going on to define n-chains their definition and the subsequent specification of singular homology should be motivated. The most concise explanation of this is that the simplices enclose parts of the space when the latter is approximated by a construction consisting of such simplices. Depending on what the simplices are found to be enclosing, conclusions about the shape of the space can be drawn. Before being able to consider what the simplices circumscribe a definition of a chain of simplices is required.

Definition 2.26. (Singular n-Chain Group, n-Chain). The singular n-chain group $\Delta_n(X)$ of a topological space X is the free abelian group generated by the singular n-simplices $\sigma_n : \Delta_n \to X$ of X. An element of this group is called an n-chain.

Each chain is therefore a formal finite sum of simplices preceded by integer coefficients. Considering the faces of a simplex bound the area within the simplex and are in fact themselves simplices, the existence of a map that maps a simplex to its boundary sounds plausible. It indeed exists. It takes a simplex and returns a formal sum of its oriented faces.

Definition 2.27. (Boundary, Boundary Homomorphism). For σ_n a singular n-simplex of X, the boundary of σ_n is $\partial_n \sigma_n = \Sigma_{i=0}^n (-1)^i \sigma_n^{(i)}$, an $(n-1)$-chain. As the singular n-simplices are the basis for $\Delta_n(X)$ this can be extended to a homomorphism called the boundary homomorphism by:

$$\partial_n : \Delta_n(X) \to \Delta_{n-1}(X),\ \partial_n \left(\textstyle\sum_\sigma c_\sigma \sigma \right) = \sum_\sigma c_\sigma \partial_n \sigma.$$

It is readily shown that for any n, $\partial^2 = \partial_n \partial_{n+1} = 0$. This property is what makes the sequence of singular n-chain groups along with their boundary homomorphisms a *singular chain complex*:

$$\cdots \xrightarrow{\partial_{n+1}} \Delta_n(X) \xrightarrow{\partial_n} \Delta_{n-1}(X) \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_1} \Delta_0(X) \xrightarrow{\partial_0} 0$$

with the convention that $\partial_0 = 0$. The singular homology groups are derived from this complex.

Definition 2.28. (Singular Homology Groups). Let X be a topological space, $Z_n(X) = \ker \partial_n$ and $B_n(X) = \text{im } \partial_{n+1}(X)$. The n-th singular homology group of X is given by:

$$H_n(X) = Z_n(X)/B_n(X) = (\ker \partial_n) \,/\, (\text{im } \partial_{n+1}).$$

The elements of $\ker \partial_n$ and im ∂_{n+1} are commonly referred to as n-cycles and n-boundaries respectively. Two n-cycles are said to be *homologous* if their difference is an n-boundary, and the equivalence classes that arise from this relation are denoted $[\![c]\!]$ for c a representative n-cycle.

From the definition alone it may not be apparent how the singular homology groups encode the holes of a space. The intuition is that the n-cycles are all *closed*, but only the boundaries enclose part of the space. Taking the cycles modulo the boundaries then amounts to dividing out the boundaries, retaining only those cycles that do not bound anything. In other words, only the n-dimensional holes remain in the n-th singular homology group. The number of holes for each dimension can informally be measured on account of the following definition.

Definition 2.29. (Betti Numbers). If X is a topological space and $H_i(X)$ is finitely generated its rank β_i is called the i-th Betti number of X.

The homotopy groups, and therefore the fundamental group, are connected to the homology groups through the Hurewicz theorem. This theorem is named for context but will not be discussed. Instead a light will be shone on the functorial properties of the process of taking the homology groups, before restricting the scope to that of the less general simplicial homology theory.

Definition 2.30. (Induced Homomorphism). Let $f : X \to Y$ be a map between topological spaces. Then through any singular n-simplex $\sigma_n : \Delta_n \to X$ of X the composition $f \circ \sigma_n : \Delta_n \to Y$ is uniquely extended to a homomorphism $f_\Delta : \Delta_n(X) \to \Delta_n(Y)$ by:

$$f_\Delta\left(\textstyle\sum_\sigma c_\sigma \sigma\right) = \textstyle\sum_\sigma c_\sigma (f \circ \sigma)$$

called the induced homomorphism of f.

The induced homomorphism is in fact a special type of map between the singular chain complexes of X and Y called a *chain map* for which $f_\Delta \circ \partial = \partial \circ f_\Delta$. If two aligned chain complexes are pictured and an arrow for f_Δ between each pair of n-chain groups is added, every cell in the resulting diagram commutes. This fact is captured in the following proposition.

Proposition 2.31. *The induced homomorphism $f_\Delta : \Delta_n(X) \to \Delta_n(Y)$ is a chain map.*

The proof is straightforward but left out. It, and the proof of the following corollary, can be found in the fourth chapter of *Topology and Geometry* by Glen Bredon [4].

Corollary 2.32. *A map $f : X \to Y$ between topological spaces induces a sequence of homomorphisms $f_n : H_n(X) \to H_n(Y)$ between the singular homology groups that make H_n into a functor.*

From the preceding corollary and the theory of categories it follows that the singular homology groups of a space form a topological invariant.

Corollary 2.33. *If $f : X \to Y$ is a homeomorphism then $f_n : H_n(X) \to H_n(Y)$ is an isomorphism.*

As a final addition to the fundamentals of homology, *homology with coefficients* is introduced. In the above treatment the coefficients in the n-chains were taken from $\mathbb{Z}$, but all results hold in more generality if they are taken from an abelian group G instead. If the latter is done this is indicated by the notation $H_*(X; G)$.

With this remark the scope of the treatment is adjusted to a nicer set of spaces called simplicial complexes, built up entirely out of conjoined affine (non-singular) simplices. The homology theory for these spaces is called simplicial homology accordingly, and it can be shown that the simplicial homology groups coincide with their singular counterparts. This is explicitly done in [3,4].

Definition 2.34. (Affine Simplex). Let $v_0, ..., v_n$ be a set of affinely independent elements of $\mathbb{R}^{n+1}$. The affine n-simplex σ_n spanned by the v_i is given by:

$$\sigma_n = (v_0, ..., v_n) = \{\ \textstyle\sum_{i=0}^{n} \lambda_i v_i \ \mid\ \sum_{i=0}^{n} \lambda_i = 1, \lambda_i \in [0,1]\ \}.$$

Although the notation does not differ from that of an ideal in the previous section it should be clear from the context which is meant. For one difference, unlike ideals, affine simplices have faces.

Definition 2.35. (k-Face). For $k \leq n$, the k-face of an affine n-simplex $(v_0, ..., v_n)$ is the affine k-simplex spanned by a subset of $\{v_0, ..., v_n\}$ with cardinality k.

The complex is then formed as prescribed by the next definition.

Definition 2.36. (Simplicial Complex). A (geometric) simplicial complex K is a collection of affine simplices such that:

1. If σ is a simplex in K, every face of σ is a simplex in K;
2. For two simplices σ, η in K, their intersection $\sigma \cap \eta$ is either empty or a face of both σ and η.

This definition excludes the possibility of two simplices in the complex being joined at anything other than a completely shared face. Naturally there is an underlying space of this construction. It is called the *polyhedron* of K and is given by $|K| = \bigcup_{\sigma \in K}\{\sigma\}$. If a space is homeomorphic to this polyhedron the simplicial complex K is called the *triangulation* of the space. It is through this notion that the approximation of a space through affine simplices is given meaning.

The main advantage simplicial homology holds over singular homology is that it's easier to compute directly. In order to compute the homology groups of a simplicial complex all simplices are given an orientation. This orientation manifests itself as an ordering of the vertices that define the simplex, respecting the rule that two orderings define the same orientation if and only if they differ by an even permutation. The chain complex and simplicial homology groups are then calculated analogously to the singular chain complex and homology groups. Because the space already consists of simplices, no embedding of any form takes place.

The final part of this chapter is dedicated to the theory specific to simplicial complexes.

Definition 2.37. (Simplicial Map). Let $f : K \to L$ be a map between simplicial complexes and let $f' : |K| \to |L|$ be the associated map between their polyhedra that is obtained in the canonical way. If f is an affine map and f' takes the vertices of any simplex of K to the vertices of a simplex of L, f is called a simplicial map.

The last requirement is identical to the condition that for every chain c of n-simplices in the simplicial complex K, $f(c) = f\left(\Sigma_{i=0}^{n} \lambda_i v_i\right) = \Sigma_{i=0}^{n} \lambda_i f(v_i)$. This guarantees that the image of any affine simplex in K is again an affine simplex in L.

Definition 2.38. (Carrier). For x an element of the polyhedron of a simplicial complex K, the carrier $\mathrm{carr}(x)$ of x is defined as being the smallest simplex of K that contains x.

Using the carrier the idea of a simplicial approximation to a map between polyhedra can take shape.

Definition 2.39. (Simplicial Approximation). Let $f : |K| \to |L|$ be a continuous map between polyhedra of simplicial complexes. A simplicial map $g : K \to L$ is called a simplicial approximation to f if for every $x \in |K|$, $g(x) \in \mathrm{carr} f(x)$.

Proposition 2.40. *If g is a simplicial approximation to f, $f \simeq g$.*

Proof. Both $f(x)$ and $g(x)$ lie in $\mathrm{carr} f(x)$ and therefore so must a line segment connecting them. The map $F(x,t) = tf(x) + (1-t)g(x)$ is a linear homotopy. □

This concludes the exposition of preliminary concepts.

3 Persistent Homology

In general the area of topological data analysis seeks to employ topological machinery in order to study the intrinsic shape of data. As such the techniques are primarily fit for tasks such as clustering, dimensionality reduction and extracting geometric information from the dataset. Because of their topological nature the tools are decidedly insensitive to the choice of metric and are able to accurately distinguish between noise and persistent features of the data. The main theory that illustrates these qualities is an adaption of homology theory called persistent homology, which is introduced in this chapter. Before rigorously building this theory the first section delves into one of the domains of persistent homology and will sketch the process of its application.

3.1 Point Cloud Data

Loosely speaking point cloud data is a numerical dataset that allows itself to be visualized in a Euclidean space. It is therefore intuitively possible to think about its shape in lower dimensions, or about its higher dimensional properties through a projection onto a lower dimensional subspace. One possible approach to dimensionality reduction is the statistical technique of principal component analysis that allows for the condensation of higher dimensional properties into principal components that retain the most variable information. Persistent homology on the other hand revolves around extracting purely geometric information, and the dimensionality reduction achieved takes shape as easy to interpret summaries. More precisely these summaries show what homological properties persist as a space fitted to the datapoints is grown. In addition they register at what points in the growing process holes appear and disappear meaning the analyst can distinguish what properties are noise or indicative of actual features of the space, and by extension of the dataset itself. The next section addresses the question of how a space can be fitted and grown, but before that can be done a short contextual introduction to the main object of study is required.

Definition 3.1. (Point Cloud Data). If $\mathbb{X}$ is a subspace of $\mathbb{R}^n$, a set of points $X \subseteq \mathbb{X}$ is called a point cloud. A dataset that can be represented as a point cloud is referred to as point cloud data.

Below is an example of a point cloud in $\mathbb{R}^2$, for $\mathbb{X}$ an annulus.

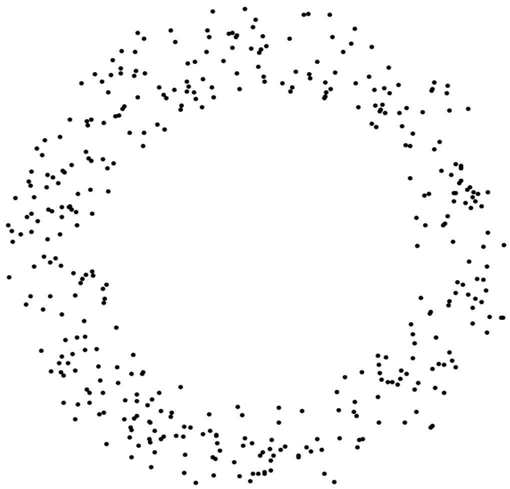


Figure 1: A point cloud X.

Point cloud data occurs naturally in many different fields. Examples include configuration spaces, spatial data or even sets as simple as the quantitative results of a customer survey. It follows from the definition that even if the dataset is not originally in the form of a point cloud it may be transformed into one if a suitable map can be constructed. In the case of matrices for example, a map as strong as a diffeomorphism exists between $\mathrm{M}(n, \mathbb{R})$ and $\mathbb{R}^{n^2}$, meaning even matrices can be represented as point clouds.

Alternatively the point cloud can be obtained as a sample taken from an underlying space that is to be approximated. The topological features of the point cloud data are computed by proxy and in order to start doing this the underlying space $\mathbb{X}$ will indeed need to be approximated regardless of the origin of the data. Inspired by the discussion of simplicial homology in the first chapter the point cloud is completed to a simplicial complex, and more specifically a filtration of such complexes is considered.

3.2 Constructing Complexes

Given any point cloud X and the want to construct an object resembling a space, the first course of action is to exaggerate the presence of each point in the cloud. This fattening of the points is done by enclosing them in closed balls of a fixed positive radius with respect to for instance the Euclidean metric and defined as follows:

$$X_\epsilon = \bigcup_{x \in X} B(x, \tfrac{\epsilon}{2}) \subset \mathbb{R}^n.$$

The below figure shows a small example of such a set of balls.

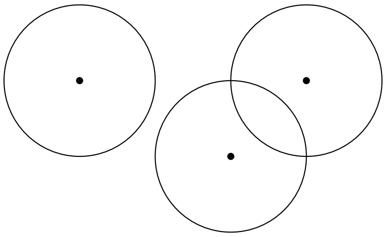


Figure 2: A set X_ϵ.

For suitably chosen values of ϵ the balls in X_ϵ intersect and based on these intersections points in the cloud can be connected. After having made this connection the balls are discarded and the connections that remain are a frame for the space fitted to the point cloud. Precisely put the points in X are treated as the vertices of a simplicial complex and based on the overlap found in X_ϵ connecting k-simplices may be added to the construction. The complexes that result from such processes are named differently depending on what rules prescribe when the k-simplices are added. Two important formulae are treated in this section, starting with the Čech complex.

Definition 3.2. (Čech Complex). Given a point cloud X, the Čech complex $C_\epsilon(X)$ on X is the simplicial complex whose k-simplices are determined by unordered finite sequences of points $\{x_i\}_{i=0}^k$ in X for which the $B(x_i, \frac{\epsilon}{2})$ in X_ϵ have a point of common intersection.

In other words a k-simplex in the Čech complex can only occur if the balls surrounding all $k+1$ vertices have a non-empty intersection. This is a strong condition, but it is justified as the importance of the Čech complex is underscored by the following theorem.

Theorem 3.3. *(Nerve Theorem). The Čech complex $C_\epsilon(X)$ is of the same homotopy type as X_ϵ.*

The nerve theorem derives its name from the more general concept of the nerve of a covering of a topological space, of which the Čech complex is a specific instance. Because homology can be shown to be invariant under homotopy the theorem is assurance that the Čech complex is a topologically faithful model of X_ϵ with regard to homology. The proof of the theorem itself is not at all trivial and can be found in [5].

Although faithful the use of the Čech complex has several drawbacks. The computation of the complex takes exponential time in the size of the point cloud, and the matters of storage and use bring about problems of their own. For these reasons many applications resort to a different complex, the Vietoris-Rips complex, that approximates the qualities of the Čech complex.

Definition 3.4. (Vietoris-Rips Complex). Given a point cloud X, the Vietoris-Rips complex $VR_\epsilon(X)$ on X is the simplicial complex whose k-simplices are determined by unordered finite sequences of points $\{x_i\}_{i=0}^{k}$ in X for which the $B(x_i, \frac{\epsilon}{2})$ in X_ϵ have pairwise non-empty intersections.

The next figure serves to illustrate the difference between the complexes. Whereas the Vietoris-Rips complex on the left shows a 2-simplex for this value of ϵ, the Čech complex on the right has no common point of intersection for the three balls. The highest dimensional simplices in the Čech complex are therefore the 1-simplices connecting the vertices.

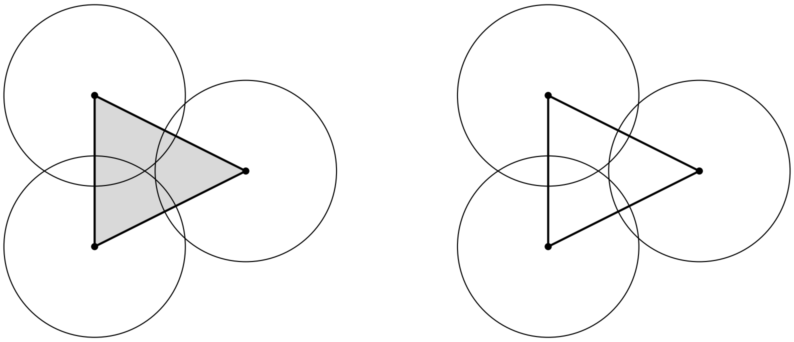


Figure 3: A comparison of the complexes in $\mathbb{R}^2$.

The main difference with the Čech complex is that the Vietoris-Rips complex extends a k-simplex to a $(k+1)$-simplex when a new point is within a distance of ϵ from the vertices of the k-simplex. It is for this reason that the Vietoris-Rips complex can be seen as a complex built entirely by adding 1-simplices to the vertices in the point cloud. This aspect is given more attention shortly.

Proposition 3.5. *For a point cloud X in $\mathbb{R}^n$, $\epsilon > 0$ and $\epsilon' \geq \epsilon\sqrt{2n(n+1)^{-1}}$,*

$$VR_\epsilon(X) \subset C_{\epsilon'}(X) \subset VR_{\epsilon'}(X).$$

The second inclusion follows directly from the definition of the complexes, and the proof of the first inclusion is found in [6]. Although the behaviour of the complexes may differ this proposition is a guarantee that the Vietoris-Rips complex at least approximates the Čech complex. Regardless, besides a passing remark about 1-simplices, nothing about what makes the Vietoris-Rips complex more feasible to work with has been explained. The explanation requires the concept of a graph from discrete mathematics and is ushered in by this definition.

Definition 3.6. (Graph). A graph $G = (N, E)$ consists of a set of nodes N and a set $E \subseteq N \times N$ called the edges of the graph. If the pairs are ordered the graph is called directed, if not it is called undirected. An undirected graph without parallel edges between two nodes, or loops connecting a node to itself, is called a simple graph.

A *subgraph* G' of G is a graph of which the nodes and edges are subsets of those of G.

Graphs are informally related to simplicial complexes as the former can be treated as 1-dimensional simplicial complexes. This connection won't be used past this section in which a specific type of graph is related to a subcomplex in order to define a class of simplicial complexes.

Definition 3.7. (Complete Graph). A complete graph is a simple undirected graph in which every node is uniquely connected to every other node.

The next definition is a small addition to the theory of simplicial complexes.

Definition 3.8. (n-Skeleton). The n-skeleton of a simplicial complex is the union of all simplices in the complex of dimension n or lower.

Using it the new class of simplicial complexes can be defined.

Definition 3.9. (Flag Complex). A flag complex is a simplicial complex in which every simplex corresponds to a complete subgraph in the 1-skeleton of the simplicial complex.

The Vietoris-Rips complex is a flag complex, meaning it can be stored in graph datastructures which are extensively studied in computer science. Instead of having to store the entirety of the complex it can be constructed when needed and analysed in the same way a graph would be. Even if the representation of the Čech complex is more truthful, access to the efficient toolbox of graph theory is a large advantage the Vietoris-Rips complex has over the Čech complex.

The key insight for persistent homology then is that the radius ϵ can be varied. Let K_ϵ denote a simplicial complex of choice constructed with parameter ϵ. By letting the value of ϵ increase the balls around the points in the dataset expand and new connections are made, growing K_ϵ. It is apparent from the above definitions that if ϵ_2 is larger than ϵ_1 and both are non-zero, $K_{\epsilon_1} \subset K_{\epsilon_2}$. Moreover the identity map $\iota_{1,2} : K_{\epsilon_1} \hookrightarrow K_{\epsilon_2}$ doubles as a natural inclusion map. This leads to the following definition.

Definition 3.10. (Filtration on Point Cloud). Let X be a point cloud and $m \in \mathbb{N}$. For $\{\epsilon_i\}_{i=1}^m$ an increasing sequence of parameter values the ordered sequence $\mathcal{F} = \{K_{\epsilon_i}(X)\}_{i=1}^m$ of simplicial complexes is called a filtration of simplicial complexes on X.

Rather than looking at the simplicial homology groups of individual complexes in the filtration, maps from the homology groups of K_{ϵ_i} into the homology groups of K_{ϵ_j} for $i < j$ are considered. This is the subject of the next section.

3.3 Persistence Modules

Persistent homology is defined on the very filtrations that were discussed at the end of the last section and is no more than a formalisation of the explanations given previously. Let Δ_*^i denote the chain complex of the i-th simplicial complex in the filtration and denote the associated boundary operators by ∂_*^i. The persistent homology groups are then defined as follows.

Definition 3.11. (Persistent Homology). For $p \in \mathbb{N}$ the k-th p-persistent homology group of the i-th simplicial complex in the filtration is defined as:

$$\mathcal{H}_k^{i,p} = Z_k^i \;/\; (B_k^{i+p} \cap Z_k^i) = \left(\ker \partial_k^i\right) / (\operatorname{im} \partial_{k+1}^{i+p} \cap \ker \partial_k^i).$$

The motivations for this definition were foreshadowed throughout the preceding work. For a fixed simplicial complex in the filtration the k-th p-persistent homology group records all k-dimensional holes that persist between the simplicial complex and its p-th supercomplex in the filtration. These holes are again informally counted by a parametrised variant of the Betti numbers.

Definition 3.12. (Persistent Betti Numbers). If $\mathcal{H}_k^{i,p}$ is finitely generated the k-th p-persistent Betti number of the i-th simplicial complex in the filtration is the rank $\beta_k^{i,p}$ of $\mathcal{H}_k^{i,p}$.

There is an important observation that leads to the possibility of summarising the information given by the persistent Betti numbers in a planar region. The chain complexes of the filtration can be combined into one object leading to the introduction of new algebraic structures through which not only the persistent homology groups can be related to simplicial homology more directly, but the structure theorem becomes applicable as well. The summary can then be extracted by utilising the structure theorem. The first queued definition is that of the persistence complex.

Definition 3.13. (Persistence Complex). A family of chain complexes $\{C_*^i\}_{i\in\mathbb{N}}$ together with a sequence of chain maps $\varphi^i : C_*^i \to C_*^{i+1}$ is called a persistence complex, denoted $\mathcal{C} = \{C_*^i, \varphi^i\}_{i\in\mathbb{N}}$.

The chain complexes belonging to the filtration of simplicial complexes naturally form a persistence complex when the φ^i are taken to be the homomorphisms induced by the inclusion maps $\iota_{i,i+1}$. The result is a multi-dimensional complex, part of which is shown below.

$$\begin{tikzcd}
 & \vdots & \vdots & \vdots & \vdots \\
\cdots \arrow[r, "\partial_3^2"] & \Delta_2^2 \arrow[u, "\varphi^3"] \arrow[r, "\partial_2^2"] & \Delta_1^2 \arrow[u, "\varphi^3"] \arrow[r, "\partial_1^2"] & \Delta_0^2 \arrow[u, "\varphi^3"] \arrow[r, "\partial_0^2"] & 0 \arrow[u, "\varphi^3"] \\
\cdots \arrow[r, "\partial_3^1"] & \Delta_2^1 \arrow[u, "\varphi^2"] \arrow[r, "\partial_2^1"] & \Delta_1^1 \arrow[u, "\varphi^2"] \arrow[r, "\partial_1^1"] & \Delta_0^1 \arrow[u, "\varphi^2"] \arrow[r, "\partial_0^1"] & 0 \arrow[u, "\varphi^2"] \\
\cdots \arrow[r, "\partial_3^0"] & \Delta_2^0 \arrow[u, "\varphi^1"] \arrow[r, "\partial_2^0"] & \Delta_1^0 \arrow[u, "\varphi^1"] \arrow[r, "\partial_1^0"] & \Delta_0^0 \arrow[u, "\varphi^1"] \arrow[r, "\partial_0^0"] & 0 \arrow[u, "\varphi^1"]
\end{tikzcd}$$

When computing the simplicial homology of the persistence complex a new object that comprises all k-th persistent homology arises. The relevant definitions are given in generality first.

Definition 3.14. (Persistence Module). A persistence module $\mathcal{M} = \{M^i, \phi^i\}_{i\in\mathbb{N}}$ is a family of R-modules M^i together with a sequence of homomorphisms $\phi^i : M^i \to M^{i+1}$.

Both the persistence complex and the newly introduced persistence module are said to be of *finite type* if each chain group respectively module is finitely generated and the sequences of associated maps stabilize beyond a certain index i. The stabilisation criterion is equivalent to the requirement that for all $i < j$ the maps φ^j or ϕ^j respectively are isomorphisms. Rid of generality this means that when filtrations of simplicial complexes on a finite point cloud are considered the complexes remain unchanged after a certain value for the radius of the balls is reached.

The definition of persistent homology in the language of persistence modules then follows from the following alternative but equivalent definition of persistent homology, wherein H_k^i denotes the k-th simplicial homology group of the i-th simplicial complex in the filtration.

Definition 3.15. (Persistent Homology, Alternative). For $p \in \mathbb{N}$ the k-th p-persistent homology group of the i-th simplicial complex in the filtration is defined as the image of H_k^i under the injection $\eta_k^{i,p} : H_k^i \to H_k^{i+p}$ that maps homology classes into the homology classes containing them. More concretely, the persistent homology group is isomorphic to this image:

$$\mathcal{H}_k^{i,p} \cong \text{im } \eta_k^{i,p}.$$

The next definition then ties persistent homology to persistence modules and is justified precisely because every abelian group is in particular a $\mathbb{Z}$-module.

Definition 3.16. (k-th Persistence Module). The k-th persistence module $\mathcal{H}_k$ of a filtration of simplicial complexes is the family of k-th simplicial homology groups H_k^i together with the sequence of maps $\eta_k^{i,i+1} : H_k^i \to H_k^{i+1}$

Notably every persistent homology group can be constructed from its corresponding persistence module through composition of the injective maps:

$$\mathcal{H}_*^{i,p} \cong \operatorname{im}\left[\eta_*^{p-1,p} \circ \eta_*^{p-2,p-1} \circ \cdots \circ \eta_*^{i,i+1}\right].$$

The construction of persistent homology with coefficients is possible and is a direct extension of simplicial homology with coefficients. Furthermore it bears repeating that if the point cloud X is finite the k-th persistence modules are guaranteed to be of finite type since the number of simplices on a finite point cloud is easily bounded. For real world datasets this assumption is no obstruction.

3.4 Barcodes

Let $\mathcal{M} = \{M^i, \phi^i\}_{i\in\mathbb{N}}$ be a persistence module over R. After equipping the polynomial ring $R[t]$ with the standard grading the following functor between the category of persistence modules over R of finite type and the category of finitely generated non-negatively graded modules over $R[t]$ can be defined:

$$\mathcal{M} \mapsto \bigoplus_{i=0}^{\infty} M^i,$$

where t acts by $t \cdot (m^0, m^1, ...) = (0, \phi^0(m^0), \phi^1(m^1), ...)$. It can be shown that this functor can be inverted and in fact induces an equivalence of the categories, as is done in [7]. As the equivalence translates results the doors to the structure theorem are now opened.

Let the persistent homology modules be taken over a field $\mathbb{F}$. Because the polynomial ring $\mathbb{F}[t]$ is a principal ideal domain all graded ideals are homogeneous and generated by t^n for some n. Through the above equivalence of categories it follows that the corollary to the structure theorem decomposes each persistent homology module into:

$$\mathcal{H}_*(X;\mathbb{F}) \cong \left(\bigoplus_{i=1}^{m} \Sigma^{x_i}\mathbb{F}[t]\right) \oplus \left(\bigoplus_{i=1}^{n} \Sigma^{y_i}\mathbb{F}[t]/(t^{n_i})\right).$$

The use of a field effectuates this decomposition, and moreover the indeterminates of the polynomial ring now encode the appearance and disappearance of the homology generators. This information can be extracted to create a graphical representation of the evolution of the generators.

Definition 3.17. (Bar). A bar is an ordered pair (i, j) with $i \in \mathbb{Z}$, $j \in \mathbb{Z} \cup \{\infty\}$ and $i < j$.

The lifespan of every homology generator can be represented as a bar through the map:

$$\text{Bar} : \mathcal{H}_*(X;\mathbb{F}) \to \mathbb{Z} \times \mathbb{Z} \cup \{\infty\}$$

that maps each free component in the decomposition to the respective pair (x_i, ∞), and every torsional component to $(y_i, y_i + n_i)$. These bars define the barcodes.

Definition 3.18. (Barcode). The k-th barcode of a filtration $\mathcal{F}$ on X is defined as the direct sum of the bars in the image B_k of $\mathcal{H}_k(X;\mathbb{F})$ under the Bar map:

$$\text{Barcode}_k(\mathcal{F}) = \bigoplus_{\text{bar}\in B_k} \text{bar}.$$

The full barcode of $\mathcal{F}$ is defined as the direct sum of the k-th barcodes:

$$\text{Barcode}(\mathcal{F}) = \bigoplus_{k=0}^{q} \text{Barcode}_k(\mathcal{F}),$$

where q is the index at which the sequence of maps in the persistence module stabilizes.

It is worthy of note that the barcode function is a bijection between the isomorphism classes of $\mathbb{F}[t]$-modules and the finite sets of bars. This is discussed in [8].

Barcodes can be plotted and by doing so the graphical representation of the persistence is achieved. By inspecting the barcode plot one can not only discern in which dimensions the data admits holes, but can also directly read off how long these holes persist as the simplicial complexes in the filtration are grown. In other words the barcodes inspire a classification for which holes in the data are artifacts of noise and which are truly geometric properties of the underlying space. The figure at the end of this section contains an illustration of a barcode plot.

There are several alternatives to this summary. One is the persistence landscape treated in the next section, and another is the persistence diagram. Because the persistence diagram is equivalent to the barcode it is defined in this section as an addendum.

Definition 3.19. (Persistence Diagram). A persistence diagram is a plot in two dimensions of the identity function, and each bar of $\mathcal{F}$ as a point in the plane.

The reason for the identity function to be included is to separate the regions where points can and cannot be plotted. As the disappearance of a generator must always be at a higher radius than the appearance, no points can be plotted under the identity line. It is customary to style the grouped points in the persistence diagram as different symbols or with different colours so that the dimensions can be told apart. The below figure showcases what a persistance diagram and barcode for $\mathbb{X} \subset \mathbb{R}^2$ a circle may look like.

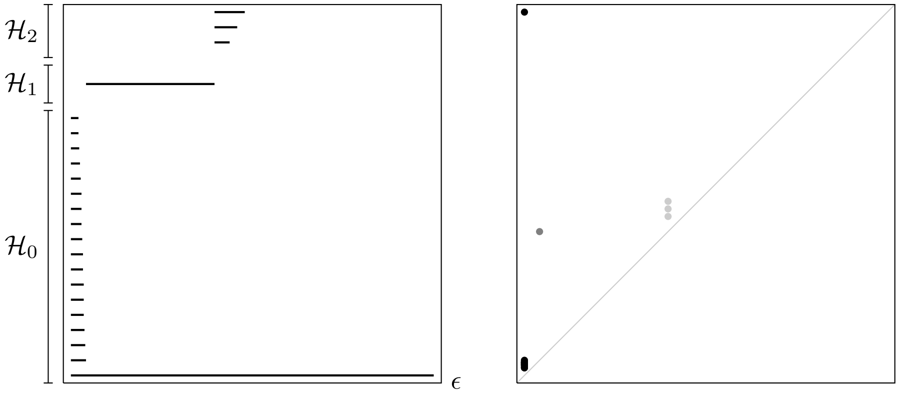


Figure 4: A barcode (left) and persistence diagram (right).

The two summaries are easily transformed into one another. Every bar in the barcode is a line between two points (x_1, y) and (x_2, y). This corresponds to an appearance at x_1 and a disappearance at x_2, meaning this bar corresponds to the point (x_1, x_2) in the persistence diagram. Moving back to the barcode from the persistence diagram raises the question of what value to choose for y. The answer is that it does not matter, as long as it is distinct from the points that were processed before.

3.5 Persistence Landscapes

Although the barcode has strong qualities as a summary, it has a drawback. There is no possibility of computing for instance an average of two barcodes, because barcodes that result from repeated sampling reside in isolated environments. The persistences landscapes are different summaries that are elements of a normed vector space. Because of this structure, statistical methods are no longer barred from the topological summaries. The definitions in this section require some familiarity with measure theory, as for example [9] provides.

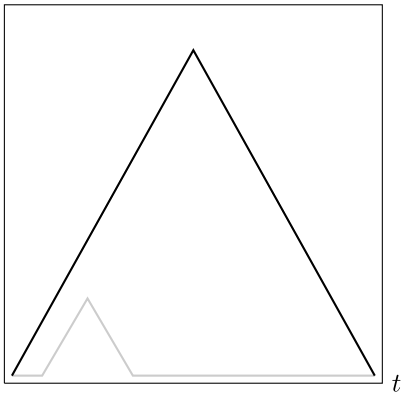
Rather than explicitly using one point cloud, the setting is generalised to that of a probability space $(\mathbb{X}, \mathcal{S}, \mathbb{P})$. The σ-algebra $\mathcal{S}$ contains all possible point clouds in $\mathbb{X}$, and in this context $\mathbb{P}$ is the measure that prescribes the probability of obtaining each point cloud as a sample from $\mathbb{X}$. The point cloud used in previous sections can be interpreted as a single element of $\mathcal{S}$. If the point cloud is taken from this probability space, the persistence landscape Y that is computed from it is a random variable and an element of a different probability space $(\mathbb{Y}, \mathcal{B}, \mathbb{P}_*)$ in which $\mathcal{B}$ is the Borel σ-algebra on the space of persistence landscapes $\mathbb{Y}$, and $\mathbb{P}_*$ is the pushforward of $\mathbb{P}$ along Y. After defining the persistence landscapes, a norm on $\mathbb{Y}$ will be defined.

Definition 3.20. (Persistence Landscape). The persistence landscape of the k-th persistence module is a function $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R} \cup \{-\infty, \infty\}$ defined by:

$$\lambda(n,t) = \sup\{x \geq 0 \mid \beta_k^{t-x,t+x} \geq n\},$$

where the supremum of the empty set is taken to be 0. Equivalently it can be thought of as a sequence of piecewise linear functions $\lambda_n(t) = \lambda(n,t)$.

The intuition that underlies the persistence landscape is that $\lambda(n,t)$ records the largest positive number x so that at least n holes persist from radius $t-x$ to $t+x$. If n is kept fixed, $\lambda_n(t)$ is piecewise linear because $\lambda_n(t)$ can be seen as the distance from t to the nearest point on the t-axis where the value of the parametrised Betti number $\beta_k^{t,t}$ passes n. The below figure shows what the first partial persistence landscapes $\lambda(1,t)$ of $\mathcal{H}_0$ and $\mathcal{H}_1$ may look like if $\mathbb{X}$ is again taken to be a circle. The lighter of the two landscapes corresponds to $\mathcal{H}_0$.



If the full plots of the $\lambda(n,t)$ need to be made, each persistence landscape λ can be extended to a function $\bar{\lambda} : \mathbb{R}^2 \to \mathbb{R} \cup \{-\infty, \infty\}$ by defining:

$$\bar{\lambda}(x,t) = \begin{cases} \lambda(\lceil x \rceil, t) & x > 0 \\ 0 & x \leq 0 \end{cases}$$

on the real plane. The usage of this extension makes the plot look like an actual landscape which is significantly easier to read.

The next lemma collects some properties of the persistence landscapes. It is proven in [10].

Lemma 3.21. *For all $n \in \mathbb{N}$ and $t \in \mathbb{R}$, the following is true:*

1. $\lambda_n(t) \geq 0$;
2. $\lambda_n(t) \geq \lambda_{n+1}(t)$ *and;*
3. $\lambda_n(t)$ *is* 1*-Lipschitz continuous.*

There is an interesting connection between the persistence diagram and persistence landscapes. The landscapes can be obtained from the persistence diagram by first adding the two line segments parallel to the axes that connect each point to the identity plot and then clockwise rotating the result 45°. The connections between the summaries are further explored in [10].

The main advantage over the previous summaries is that the persistence landscapes are Lebesgue integrable, which is precisely what allows the machinery of calculus, probability and statistics to be applied to this type of summary. If $\mathbb{R}^n$ is equipped with the Lebesgue measure, and $\mathbb{N} \times \mathbb{R}$ with the product of the counting measure and the Lebesgue measure, the p-norm for $0 \leq p < \infty$ is given by:

$$\|\lambda\|_p = \sum_{i=1}^{\infty} \|\lambda(i,t)\|_p = \sum_{i=1}^{\infty} \left[\int |\lambda(i,t)|^p \, \mathrm{d}\mu(t) \right]^{1/p}.$$

The value of $\|\bar{\lambda}\|_p$ coincides with that of $\|\lambda\|_p$.

Many more results on landscapes were derived in the paper mentioned twice above, among which a central limit theorem, hypothesis tests and other methods for statistical inference. These results are too many to recount, as this chapter is coming to an end.

4 Summaries in R

The computation of homology is well-trodden ground [3,4] and much about the computation of persistent homology and its summaries has been written as well [8,10]. The subject of this chapter is a small script available on:

```
https://github.com/QuirijnMeijer/TDA
```

that offers functions to generate the topological summaries over point cloud data discussed in the previous chapter. It is based on the methods of the TDA package [11] that offers an R interface for several C++ libraries for topological data analysis.

4.1 Barcodes

After executing the script, the barcode of the point cloud X is plotted by the following command.

```
Barcode(X, MaxDim, MaxScale)
```

The MaxDim and MaxScale parameters are used to bound the maximum dimension in which features are calculated and the maximum number of parameters used for the filtration respectively. Their standard values are set to MaxDim = 2, MaxScale = 5. Using the standard values, the barcode of 100 points sampled uniformly from a torus with major radius 2 and minor radius 1 is plot by:

```
T <- torusUnif(100, 1, 2)
Barcode(T)
```

and is shown below. Note that the TDA package labels the radius parameter as time.

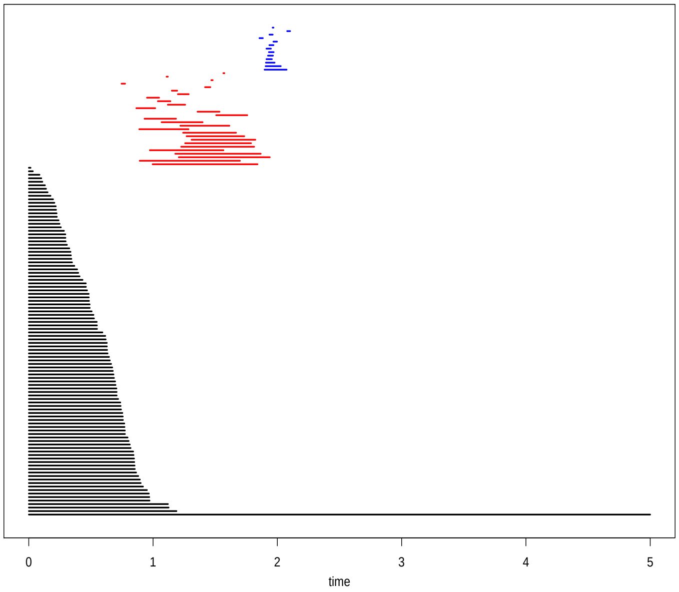


Figure 5: A barcode generated in R.

The bars are colour-coded by dimension and given in ascending order, as was done in section 3.4.

4.2 Persistence Diagrams

The persistence diagram is plotted using the exact same parameters with the same standard values as the barcode. It is noteworthy that because of the correspondence between these summaries discussed earlier, they are internally handled in the same way.

```
PersistenceDiagram(X, MaxDim, MaxScale)
```

If the point cloud sample `T` is reused, the command and plot are given below.

```
PersistenceDiagram(T)
```

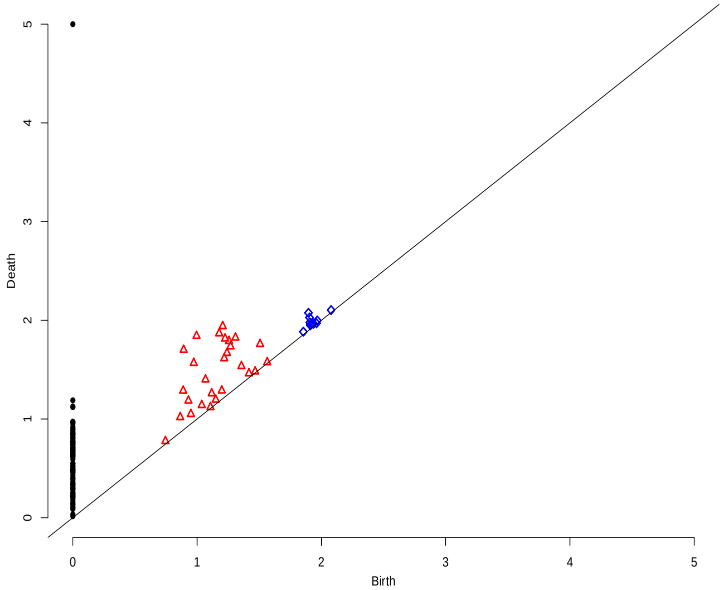


Figure 6: A persistence diagram generated in R.

Inspired by the birth-death process from the area of stochastic processes, the appearance and disappearance of homology generators is referred to as a birth or death respectively. The points in the plot are styled depending on what feature group each point belongs to. The black points represent 0-dimensional generators or connected components, the red triangles the 1-dimensional generators or loops, and the blue diamonds the 2-dimensional generators or voids. Because the 2-torus admits no higher dimensional holes, no higher dimensional generators appear.

4.3 Persistence Landscapes

Lastly, a persistence landscape is produced by the next command.

```
Landscape(X, i, j, MaxDim, MaxScale)
```

More specifically this command produces the i-th persistence landscape $\lambda(i, t)$ of the j-th persistence module. The `MaxDim` and `MaxScale` parameters are the same as the parameters for the barcode and persistence diagram, with the same standard values.

Again reusing the point cloud sampled from the torus, and first setting the graphical parameters in R so that both plots are shown next to each other, the first landscape plots $\lambda(1, t)$ of $\mathcal{H}_1$ and $\mathcal{H}_2$ are obtained as follows:

```
par(mfrow = c(1, 2))
Landscape(T, 1, 1)
Landscape(T, 1, 2)
```

and shown in the figure below.

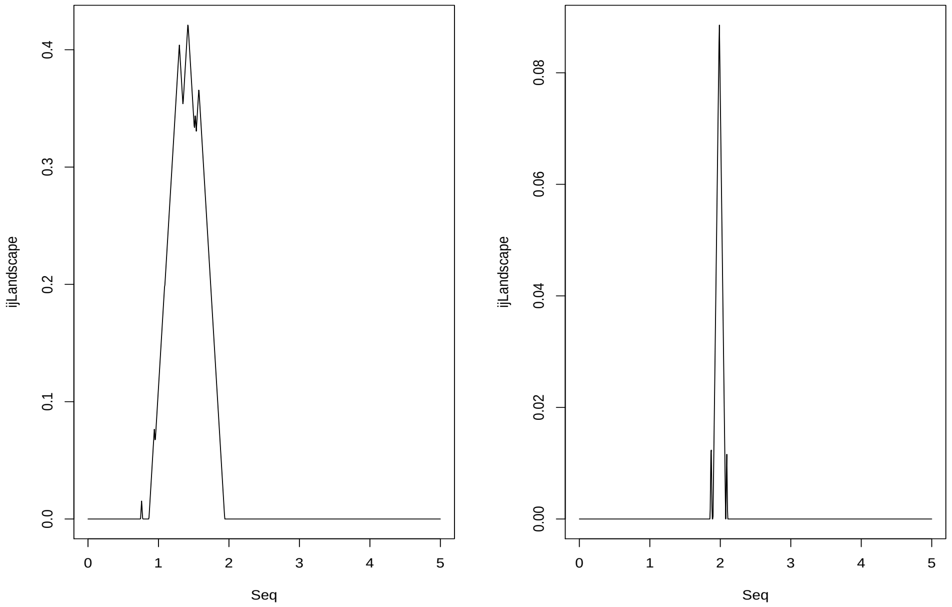


Figure 7: The first landscapes of $\mathcal{H}_1$ (left) and $\mathcal{H}_2$ (right) generated in R.

By default the landscapes are plotted separately, but by omitting the par-command and instead using the command:

```
par(new = TRUE)
```

between every two calls to the Landscape function, the landscapes can be overlaid. This does however require more attention to be paid to the scaling of the graphs.

The Seq-label on the x-axis is a result of the sequence used to determine at what points the landscape function should be evaluated and can generally be disregarded. If the need arises however, it can be manually adjusted by modifying the script.

5 Survey of Applications

Since its emergence topological data analysis has been applied many times over [6,12,13,14]. This chapter surveys a diverse selection of such applications. In order to get a cursory overview of how persistent homology can be applied to engineering problems, statistical methods and finally directly in the analysis of datasets, the following three papers are reviewed.

- Topological Trajectory Classification with Filtrations of Simplicial Complexes and Persistent Homology (Pokorny et al.) [15].
- Persistent Homology for Learning Densities with Bounded Support (Pokorny et al.) [16].
- Topological Analysis of Population Activity in Visual Cortex (Singh et al.) [17].

During the treatment of these papers the opportunity to introduce several new concepts relating to topological data analysis is taken. This chapter is therefore not only an exposition to the applications of persistent homology but also explains some related notions as they are used in practice.

5.1 Motion Planning for Robotics

Autonomous robots require the capability of reasoning about possible trajectories through their surrounding areas, avoiding obstacles and preferring efficient paths. The relevance of geometry to the study of robotics becomes apparent through this requirement, and if the luxury of being able to choose from a collection of homotopically equivalent paths can be afforded this relevance can be extended to include algebraic topology. Whereas the space of possible motion is often modelled by a graph, this paper studies the possibility of representing this space by a filtration of simplicial complexes. Persistent homology then allows for algorithmic and continuous reasoning by the robot.

Consider a multi-joint robotic arm. Such a system has various configurations that are collected in a *configuration space*. Every element of this space represents a possible state of the arm and each dimension corresponds to a joint. A path γ in the configuration space is a continuous variation of the joints that starts in the configuration $\gamma(0)$ and ends in $\gamma(1)$. The paths of most importance correspond to those trajectories on which the robot's movement is not inhibited by any obstacles, and every one of these paths lie in the *collision-free* subspace of the configuration space. For real world applications it is a natural choice to identify the configuration space C with $\mathbb{R}^n$ and after doing so the collision-free subspace is denoted by C_f.

If no analytic model of C_f is available or it is infeasible to compute, it can be approximated through random samples. One way to do this is by using spatial motion capture data from the robot, leading to the familiar point cloud datasets. In this context the subspace $\mathbb{X}$ from which the points X are assumed to be sampled is C_f. Homological information is then extracted from this sample with the intention of using it to distinguish non-homotopic paths, modelling obstructions as holes. Rather than reason about each path individually the robot can then reason about its trajectories through the determined homotopy class representatives, reducing the cost of planning.

The paper uses a different type of simplicial complex that was not introduced in the previous chapters of this work. Although the authors recognise the use of the Vietoris-Rips complex when the curse of dimensionality sets in for higher dimensions, they advocate the use of the Delaunay-Čech complex for lower dimensions because it is of the same homotopy type as X_ϵ [18]. It is defined as an abstract simplicial complex, a more general construction than the one presented in the chapter on preliminaries. Both the definitions of the Vietoris-Rips complex and the Čech complex can be reformulated in terms of the abstract simplices defined below.

Definition 5.1. (Abstract Simplex, Abstract Simplicial Complex). An abstract simplicial complex is a set V whose elements are called vertices and a collection K of finite subsets of V called the abstract simplices of the complex such that:

1. If σ is an abstract simplex in K, every non-empty subset of σ is an abstract simplex in K;
2. For every $v \in V$, $\{v\}$ is an abstract simplex in K.

It is convention to denote an abstract simplicial complex by K, the set of abstract simplices.

A *face* of an abstract simplex is simply a subset of the simplex as a set. It is sometimes useful to add the dimension of a simplex to its notation. The dimension of an abstract simplex is its cardinality, and if the dimension of σ is n this is denoted $\sigma^{(n)}$.

The Delaunay-Čech complex derives its name from the Delaunay triangulation on which it is based, and this triangulation in turn is defined through Voronoi cells. These concepts must be introduced before the new complex can be defined. For the definition of the Voronoi cells the convention of measuring the distance between a point z and a subset A of a metric space (Z, d) by:

$$d(z, A) = \inf\{d(z, a) \mid a \in A\}$$

is adopted. The Voronoi cells are then defined as follows.

Definition 5.2. (Voronoi Cell, Voronoi Diagram). Let (Z, d) be a metric space and let $\{P_i\}_{i \in N}$ be a finite sequence of non-empty subsets of Z indexed over N. Every P_i is called a site. The Voronoi cell associated with P_k for $k \in N$ fixed is given by:

$$V_k = \{z \in Z \mid d(z, P_k) \leq d(z, P_j) \; \forall j \in N - \{k\}\}.$$

The sequence $\{V_i\}_{i \in N}$ of Voronoi cells is called a Voronoi diagram.

A Delaunay triangulation is an abstract simplicial complex conditioned on the Voronoi cells.

Definition 5.3. (Delaunay Triangulation). The Delaunay triangulation of a metric space (Z, d) for a given Voronoi diagram is the abstract simplicial complex:

$$D(Z) = \{\sigma \subseteq Z \mid \textstyle\bigcap_{z \in \sigma} V(z) \neq \varnothing\}$$

where $V(z)$ denotes the Voronoi cell containing z.

Note that because the Voronoi cells and diagrams are not unique, there is in general no unique Delaunay triangulation for a given space. In order to guarantee uniqueness of the complexes the sequence of sites that defines the Voronoi diagram needs to be kept fixed.

Definition 5.4. (Delaunay-Čech Complex). For a fixed Voronoi diagram and $\epsilon > 0$, the Delaunay-Čech complex $DC_\epsilon(Z)$ on a metric space (Z, d) is defined as the set of all abstract simplices that $D(Z)$ and $C_\epsilon(Z)$ viewed as an abstract simplicial complex share. From the definition of the Čech complex:

$$DC_\epsilon(Z) = \{\sigma \in D(Z) \mid \textstyle\bigcap_{z \in \sigma} B(z, \frac{\epsilon}{2}) \neq \varnothing\}.$$

The main advantage the Delaunay-Čech complex has over other complexes of the same homotopy type is that its 2-skeleton can be directly determined using the Delaunay triangulation, without having to compute the entire complex. This leads to a large gain in performance.

Rather than using a predetermined set of values for the scale parameter ϵ the authors allow the value to vary continuously over $\mathbb{R}^+$ and define the filtration using a discrete Morse function, a concept closely related to differential topology. A full understanding of the definitions requires some knowledge of smooth manifolds, as can be acquired from [19].

Definition 5.5. (Morse function). A real-valued function on a smooth manifold is called a Morse function if for all its critical points x the Hessian matrix of the function at x is non-degenerate.

Discrete Morse functions are these functions their counterparts on abstract simplicial complexes. If τ is a face of a simplex σ, a discrete Morse function g generally obeys the rule that $g(\tau) < g(\sigma)$, allowing precisely one exception locally. This is made concrete in the next definition.

Definition 5.6. (Discrete Morse function). Let K be an abstract simplicial complex. A function $g : K \to \mathbb{R}$ is called a discrete Morse function if for all simplices $\sigma^{(n)} \in K$:

1. $g(\alpha^{(n)}) \geq g(\beta^{(n+1)})$ for at most one $\beta^{(n+1)}$ of which $\alpha^{(n)}$ is a face and;
2. $g(\alpha^{(n)}) \leq g(\beta^{(n-1)})$ for at most one face $\beta^{(n-1)}$ of $\alpha^{(n)}$.

Define a discrete Morse function $f : D(X) \to \mathbb{R}$ on the Delaunay complex by:

$$f(\varnothing) = 0,\ f(\sigma) = \min\{\epsilon \mid \textstyle\bigcap_{x\in\sigma} B(x, \frac{\epsilon}{2}) \neq \varnothing\}.$$

Then every complex $DC_\epsilon(X)$ can be defined as the *sublevel set* $f^{-1}((-\infty, \epsilon])$ of f, and if X is a finite point cloud there are only finitely many values for ϵ where this pre-image changes. By letting the sequence $\{\epsilon_i\}_{i=1}^m$ of these points be the index set of the filtration, a filtration that adheres to the definition given in the third chapter of this work is obtained. Although the result is the same as before, the use of a discrete Morse function opens the study of the dataset to new tools.

Next, assume γ is a path in X_ϵ. In order to reason about it using the filtration it needs to be represented as a piecewise linear curve in the ϵ-complex that is moreover homotopy equivalent to γ in X_ϵ. This process is called *trajectory discretization.* The easiest way to discretize a path heuristically is by first choosing a natural number m and defining $v_i = \gamma(i/m)$ for $i \in \mathbb{N}_m$. Let:

$$\gamma' : X_\epsilon \to DC_\epsilon(X)$$

then be the map that maps each v_i to the 0-simplex of $DC_\epsilon(X)$ that is closest to v_i in X_ϵ, and each path segment $\gamma((\frac{i}{m}, \frac{i+1}{m}))$ to the shortest path of 1-simplices between $\gamma'(v_i)$ and $\gamma'(v_{i+1})$. The image $\gamma'([0, 1])$ is then called the discretization of γ. From now on it is assumed that all paths are discretized.

Assuming C_f is path-connected, the fundamental group $\pi_1(C_f)$ of C_f is independent of the chosen basepoint. If γ_1, γ_2 are two distinct paths in C_f with $\gamma_1(0) = \gamma_2(0)$ and $\gamma_1(1) = \gamma_2(1)$, they are homotopic in C_f if for $\gamma(t) = \gamma_1(t)\cdot\gamma_2(1-t)$ the homotopy class $[\gamma]$ of $\pi_1(C_f)$ is trivial. This observation shows that the fundamental group contains all information on homotopic paths that is required for motion planning, but as mentioned before it is hard to compute due to its possibly complicated structure. Because the homology groups are a coarser invariant it is possible to use them to reason about the homotopy classes without necessarily having to compute the fundamental group. If the 1-cycle γ' is the discretized path of γ in a simplicial complex, then if $[\![\gamma']\!]$ is not trivial in the first simplicial homology group H_1 of the complex, γ_1 is not homotopic to γ_2. The first homology group allows the robot to distinguish between homotopy classes even if the exact partitioning is not known.

Let $\{\gamma_i\}_{0\leq i\leq n}$ be a set of paths in C_f with the same starting and end points, and let ϵ' be the smallest scale parameter so that these paths can all be discretized in $DC_{\epsilon'}(X)$. As the previous paragraph suggests only the persistence of the 1-dimensional homology generators is studied. In addition the coefficients are taken from $\mathbb{Z}_2$ to make computation more efficient. Define:

$$\gamma_{i,j}(t) = \gamma'_i(t)\cdot\gamma'_j(1-t)$$

to be the composition of the discretized paths γ'_i and γ'_j into a loop. Then if two loops $\gamma_{0,i}$ and $\gamma_{0,j}$ with $i \neq j$ do not represent the same class in $\mathcal{H}_1^{l,l+p}$ the homology class represented by $\gamma_{i,j}$ must be non-trivial, and therefore $\gamma'_i \not\simeq \gamma'_j$ in all $DC_\epsilon(X)$ with $\epsilon' \leq \epsilon_l \leq \epsilon \leq \epsilon_{l+p}$. An intuitive way to think about this is that if at least one of the loops $\gamma_{0,i}$ or $\gamma_{0,j}$ enclose a hole, then surely $\gamma_{i,j}$ must enclose that same hole. This hole is then in the way when an attempt to detract $\gamma_{i,j}$ to a point is made, and $[\gamma_{i,j}]$ is non-trivial in $\pi_1(C_f)$. This leads to a set of homology classes $\{[\![\gamma_{0,i}]\!]\}_{0 \leq i \leq n}$. Using this set the robot can reason about homotopically inequivalent paths with a certain robustness introduced by the persistence of the generators.

The paper further explores the introduction of cost functions that assign a cost to each path, or the idea of letting the starting and end points vary in connected components of C_f. While of importance to robotics it adds little to the exposure of applied persistent homology and is therefore not treated.

5.2 Bandwidth Selection in Kernel Density Estimation

Another example of where persistent homology finds application is non-parametric statistics, in probability density estimation. If an unobservable density is estimated based on a sample of datapoints its support can be conditioned so that it meets certain qualitative geometric requirements. One such condition can be a restriction on the Betti numbers. By for instance formulating a bound on the zeroth Betti number, the number of connected components in the support of the density estimator can be restricted. The paper reviewed in this section takes this approach based on spherical kernels with bounded support. Before applying topological methods, kernel density estimation is introduced, starting with the definition of a kernel.

Definition 5.7. (Kernel). A kernel is a non-negative, real-valued and integrable function k for which the following holds:

1. $\int_{-\infty}^{\infty} k(t)\,\mathrm{d}t = 1$ and;
2. $k(t) = k(-t)$ for all t in the domain of k.

The standard normal distribution $\mathcal{N}(0,1)$ is an example of a kernel, and one that is in fact spherical.

Definition 5.8. (Spherical Function). A function $g : \mathbb{R}^n \to \mathbb{R}$ is spherical if for every orthogonal matrix $H \in O(n)$ and $x \in \mathbb{R}^n$, $g(Hx) = g(x)$.

Spherical kernels are functions of the radial distance to the origin and can be thought of as probability densities that assign a probability of 1 to the unit sphere. Although kernel density estimation does not require the kernel used to be spherical, they are among other reasons chosen because their marginal and conditional distributions can be computed analytically.

For a general kernel k and an i.i.d. sample $X = \{x_i\}_{1 \leq i \leq m} \subset \mathbb{R}^n$ taken from an unobserved probability density $f : \mathbb{R}^n \to \mathbb{R}$, kernel density estimation is the approach of reconstructing f from the sample through use of the estimator:

$$\hat{f}_{\epsilon,m}(x) = \tfrac{1}{m\epsilon^n} \sum_{i=1}^{m} k(\tfrac{x - x_i}{\epsilon}).$$

The free parameter ϵ is a smoothing parameter called the *bandwidth*. After overlaying each datapoint in the sample with an adjusted copy of the kernel, the result is smoothed with a factor dictated by the bandwidth ϵ. Rather than choosing one bandwidth it is common to use a bandwidth selector, a sequence $\{\epsilon_i\}_{i \in \mathbb{N}}$ of bandwidths for every possible sample size. The optimal choice of bandwidth is an open problem, and one for which persistent homology may be employed. Before discussing how, the introduction of kernel density estimation is concluded by discussing how the performance of bandwidths is measured. In order to evaluate the quality of a candidate ϵ_i the mean integrated square error can be used, defined as:

$$\text{MISE}(\epsilon_i) = \mathbb{E}\left[\int (\hat{f}_{\epsilon_i,m}(x) - f(x))^2 \, \mathrm{d}x\right].$$

If the kernel is spherical, $\lim_{i\to\infty} \epsilon_i = 0$ and $\lim_{i\to\infty} i\epsilon_i^n = \infty$, this error can be asymptotically approximated [20] leading to an approximation of the optimal bandwidth value with respect to these error measures. The authors use these values as a benchmark for the bandwidths found using the topologically infused method.

The setting in which this method is applicable is as follows. Assume the sample $X = \{x_i\}_{1\leq i\leq m}$ is taken from a bounded subspace $\mathbb{X}$ of $\mathbb{R}^n$, and that a probability density f on $\mathbb{X}$ exists so that the sample is drawn from its distribution. If the chosen kernel has the unit sphere as its support, the support of $\hat{f}_{\epsilon,m}$ is X_ϵ. With the conditions on the Betti numbers known prior, the persistent homology of this point cloud sample can be computed and admissible values for ϵ can be found through for instance inspection of the barcodes. For all found admissible bandwidths ϵ the support of $\hat{f}_{\epsilon,m}$ then has the desired properties.

Experimentally this method always leads to an interval $[\epsilon_i^1, \epsilon_i^2]$ of admissible values, determined up to some fixed precision. Define:

$$\epsilon_i^* = \epsilon_i^1 + \left[m^{-1/(4+d)}\right] \tfrac{\epsilon_i^2 - \epsilon_i^1}{2}.$$

The authors show using several spherical kernels that in practice the bandwidth selector $\{\epsilon_i^*\}_{i\in\mathbb{N}}$ performs close to the approximated optimum, while guaranteeing the topological properties of the support. Furthermore they demonstrate that the proposed method can be used in combination with regression, applying this to a learning by demonstration scenario in which a racetrack is inferred from geospatial data provided by a car.

Controlling the geometry of the support of an estimated density can be useful for a variety of reasons. One is that it can incorporate prior knowledge of the density. If it is for instance known that there exist regions on which the density should not be supported, or the number of connected components is given, this method can reduce the number of densities that are considered eligible. The problem of determining a race track is an example of this. Because a racetrack is necessarily homeomorphic to S^1 when viewed topologically, the zeroth and first Betti numbers are restricted.

5.3 Neural Representation in the Visual Cortex

Information in the cortex is thought to be represented by the joint activity of groups of neurons, referred to as population activity. This paper aims to find topological structure in the activity of the neurons reflecting internal states and external stimuli. In this process the authors discriminate between spontaneous and evoked activity. It is easy to produce proof of the effect of stimulation in the case of evoked activity, but until recently it was an accepted theory that spontaneous cortical activity was random and unstructered. In the years leading up to this paper several new results were unearthed. First, it was found that spontaneous activity in individual cells is related to that of a specific neural population. Second, the patterns found in spontaneous activity appear to be related to the intrinsic connectivity of the cortex. Lastly, theoretical models of the primary visual cortex have shown how this activity could arise. These findings lead to the hypothesis that facts about the neuronal connectivity in the primary visual cortex can be learned by observing the spontaneous patterns of activity. In addition it was found through experimentation that spontaneous cortical activity resembles that of evoked activity brought forth by natural image stimulation. Altogether these results suggest the possibility of a computational model for the cortical activity that is in some regard shaped by natural signals. The goal is therefore to study basic aspects of both spontaneous and evoked activity, the latter by exposure to natural images.

Topology becomes involved through a specific characterization of the population activity. Earlier results show that spontaneous cortical states reproduce the patterns evoked by spatially oriented stimuli. Because this orientation is circular and the aforementioned computational model is expected to take the shape of a manifold, the activity patterns exhibit topological structure. By using persistent homology the experiments test if there are more relevant topological signatures to be found in neurological data.

The dataset used in the experiments is obtained from the primary visual cortex of macaques through an electrode array in a stereotaxic frame. After registering the signals from each electrode the five neurons with the highest firing rates were selected. The experiment concentrates on this subset because the spontaneous firing rate is low in general. By selecting neurons with a high firing rate the comparison between spontaneous and evoked activity becomes more reasonable. The point cloud in $\mathbb{R}^5$ is then generated from this dataset by means of binning, a data pre-processing technique that assigns each datapoint in a small subset to a value representative of that subset. Binning serves the purpose of both reducing the effect of measuring errors and normalisation of the data.

After the data collection and pre-processing was completed the barcodes of the point cloud were computed, and the persistent Betti numbers for certain tresholds on their persistence were taken as topological signatures of the dataset. The distributions of the persistent Betti numbers that remained constant over an interval of scale parameters of at least the length of the treshold were calculated. These distributions were for both types dominated by sequences starting with $(1, 1, 0)$ and $(1, 0, 1)$ at high tresholds. However, while the dominating signatures for both types were alike, the relative frequencies of observation were different. The likelihood of observing $(1, 1, 0)$ was found to be significantly higher in the case of evoked activity. For lower tresholds the signatures were more diverse and though still dominated by the same sequences of persistent Betti numbers, their likeliness of observation was closer.

In the conclusions the authors offer possible explanations for these observations from a neuroscientifical point of view. Spurred by these explanations they recommend further study using the techniques of topological data analysis, for instance by studying the correlation between the topological information of the specific stimulus and the cortical response.

Understanding the potential topological properties of population activity is useful for several reasons, for instance because it may aid in the design of decoding methods for brain to machine interfaces. If for example it is known that the activity has a structure with likeliness to a 2-sphere, the entire activity can be collapsed to a description of two variables. Further study is required to warrant such conclusions, as having the Betti numbers of a sphere does not conversely imply the geometric shape of a sphere. Even if this conclusion did not follow from the paper it has established the possibility of persistent homology being applicable to neuroscience.

6 Discussion

This final chapter sees a short contextual discussion of topological data analysis, and the description of a possibly novel algorithm that clusters a dataset in such a way that an algebraic structure present in the dataset is preserved in the clustering. In order to describe the algorithm some elementary notions from representation theory are included.

6.1 Others Results

Beside persistent homology, another major result in topological data analysis is that of the **Mapper** algorithm. This algorithm was introduced in [21] and is meant for direct visualization of datasets as a simplicial complex, achieved by use of point-set topological theory. The paper introduces two versions. The first is dubbed the topological version and is meant to be as general as possible. The second, meant to be used in practice, is called the statistical version. This version is a specification of the topological version applicable to finite datasets.

The main idea of the topological version is to define a continuous map f from a topological space into a parameter space Z that highlights certain properties of the domain. This function is called a *filter*. This filter is used to translate a provided finite covering of the target parameter space into a finite covering of the domain, which is then further decomposed into connected components. The connected components make up a set of vertices that is completed to a simplicial complex by taking the nerve of the covering by connected components.

The less general statistical version is explicitly defined on finite datasets paired with an $\mathbb{R}$-valued filter and uses arbitrary methods to find clusters in the sets that compose the dataset's covering. These clusters are pairwise connected in the complex only if their intersection is non-empty, leading to a 1-dimensional simplicial complex. In order to obtain a higher dimensional visualization multiple filters are combined into one real vector valued filter. Because the individual filters do not necessarily have to be the same, there is a large degree of freedom for the component filters, meaning the data can be studied in different ways from one visualization.

Other developments pertain to persistent homology more directly, with one example being the study of **zigzag persistence**. Zigzag persistence was introduced in [22] with the goal of extending the reach of persistent homology to a larger class of situations. The basis for this extended theory is that of zigzag diagrams. All sequences in traditional homological theories are monotone, meaning that for example all arrows in a horizontal chain complex are unidirectional. In a zigzag diagram this condition is non-existent.

$$\begin{tikzcd}
\cdots & & X_{i-1} \cup X_i & & X_i \cup X_{i+1} & & \cdots \\
& X_{i-1} \arrow[ul] \arrow[ur] & & X_i \arrow[ul] \arrow[ur] & & X_{i+1} \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

Figure 8: Part of an example of a zigzag diagram.

This sort of diagram is of use when the dimensions are not directly of interest, but a correspondence between the elements represented in the diagram is. If the X_i in the above zigzag diagram are samples, the theory of zigzag persistence could aid in analysing the homological consistency between them. Because a graphical representation of this information in the form of a barcode is still possible there is an easy to work with summary that carries over from general persistent homology to zigzag persistence.

Summaries like the barcode and persistence diagram are not guaranteed to be available for all additional persistent homology theories. Another theory, that of **multidimensional persistence**, had to develop a new summary instead. The initial thought of multidimensional persistence arises quite naturally. Instead of a filtration of simplicial complexes a multifiltration of structures is taken. This multifiltration is parametrized along multiple axes, meaning it's multidimensional. An example of a multifiltration that's closest to the theory discussed in this work is a filtration of simplicial complexes over a point cloud that is itself parametrized. Each object in this multifiltration is dependent on two parameters: the point cloud's parameter, and the scale parameter for the simplicial complex.

The paper that introduces multidimensional persistence [23] compiles the steps for computing the regular persistent homology of a point cloud into a recipe and generalises the steps to incorporate multifiltrations. Although the idea behind this generalisation is relatively easy it pulls a lot of abstract theory into its development, and in the process loses the completeness of the barcode invariant. In order to remedy this a new summary is introduced. This summary is called the rank invariant, and the paper shows that for one dimension it is equivalent to the barcode.

Other results have involved the succesful use of, or deepening of the knowledge surrounding, several types of simplicial complexes. Among these are the alpha complex, witness complex and tangent complex. Because there is always a trade-off between approximating qualities, ease of use and computational efficiency involved, different complexes are and will continue to be explored.

Besides studying different simplicial complexes, current research draws from several schools in mathematics such as Morse and sheaf theory in order to advance the current base of knowledge and to increase the efficiency of computation. Alongside this, work has been done in the categorification of persistent homology, and the combination of topological data analysis with machine learning is highly prominent. The next section explores the involvement of representation theory.

6.2 Structure-Preserving Clustering

Clustering is the task of grouping alike elements into clusters so that the similarity in the clusters is maximised with respect to some measure. Many different algorithms implement this task and every one of them implements a different process. This section outlines a process using representation theory and persistent homology that has the following fortes.

- It is capable of clustering qualitative data;
- If an algebraic structure is present in the dataset, it is respected;
- The number of clusters is not required to be given as a parameter;
- It has a high level of flexibility as the choice of metric is arbitrary and;
- Because it utilises persistent homology, it has a robustness to noise.

Before giving the definition of a representation the overarching plan is laid out.

The idea that is fundamental to the preservation of structure is that by adding a set of intermediate steps to a translation of the data into a point cloud the additional information is easily transferred, where it would be ignored in a naive translation. Following this procedure the persistent Betti numbers can be computed. As the zeroth Betti numbers measure the number of connected components, the longest surviving Betti number corresponds to the longest persisting division of the complexes into connected components. These components are precisely the clusters. The connected components are easily extracted from any complex in the subfiltration defined by the lifespan of the determined Betti number. The result is a clustering of the data dictated by the choice of metric and simplicial complex.

In a general setting, the most common problem with clustering of qualitative data is that it's usually not evident how similarity should be measured or how the datapoints should be represented in the first place. If an algebraic structure can be detected this problem is eliminated through this process, as there exists a natural translation to a point cloud via the matrices of a representation. Although this resolution is the most interesting possible result the clustering is fit for quantitative data as well, and the proposed algorithm is of interest even when the steps that translate the structure are omitted. Persistent homology can add a new resistance to noise that is not incorporated by other known clustering algorithms.

There is one drawback that keeps this algorithm from being fully fledged and widely deployable. In order to determine an interface between the dataset and a group, the dataset's structure needs to be manually examined on a case-by-case basis due to the absence of a canonical choice. Finding an automated algorithm to this end is not within the scope of this project. In addition the complexity of the algorithm is not computed. Without careful restraints on the process with regard to translation, choice of complex and datastructures, the complexity can vary too wildly. It is however a useful observation that only the 1-skeletons of the simplicial complexes are required, which makes some complexes better suited than others.

As a final note before proceeding, the algorithm is described for finite groups for brevity. Even if representation theory extends to other algebraic structures, groups are the most ubiquitous and lend themselves best to this illustration. Seeing as real world datasets are finite, the restriction to finite groups is a natural follow-up to this choice. The notation $\mathcal{X}$ will be used for the dataset that has been translated to a finite group.

The theorem that motivates the statement that the group structure of $\mathcal{X}$ can be preserved, and therefore justifies this endeavour, is Cayley's theorem.

Theorem 6.1. *(Cayley's Theorem). Every group G is isomorphic to a subgroup of the symmetric group acting on G.*

The elements of the symmetric group are easily translated to matrices and are therefore a prime candidate to aid in the translation from $\mathcal{X}$ to a point cloud. Representation theory comes into play when this translation is specified.

Definition 6.2. (Representation). Let G be a finite group and $\mathbb{F}$ a field. A representation of G over $\mathbb{F}$ of degree n is a homomorphism $\rho : G \to \mathrm{GL}(n, \mathbb{F})$ for $n \in \mathbb{N}$.

If im ρ is isomorphic to G the representation ρ is said to be *faithful.* This is precisely the case when ker $\rho = \{e\}$. Faithful representations are important because they are effectively an embedding of the group into the general linear group as matrices. Not coincidentally, this is precisely the property required for a succesful translation.

The plan is therefore as follows. First, set $n = |\mathcal{X}|$ and identify $\mathcal{X}$ with a subgroup $\mathcal{X}_* < S_n$ of the symmetric group on n symbols. Then take an n-th degree faithful representation σ of S_n over $\mathbb{R}$ and restrict the map σ to $\mathcal{X}_*$. Because group homomorphisms preserve subgroups, the image im $\sigma|_{\mathcal{X}_*}$ is a matrix subgroup of $\mathrm{GL}(n, \mathbb{R})$. Finally, this subgroup can be embedded as a point cloud in a real space through the natural identification of $\mathrm{M}(n, \mathbb{R})$ with $\mathbb{R}^{n^2}$.

The first step is to determine the subgroup $\mathcal{X}_*$. Left multiplication in $\mathcal{X}$ is a group action of the group on itself so that every element of $\mathcal{X}$ defines a permutation of the group elements. Because $\mathcal{X}$ is assumed to be finite it is enumerable, and by keeping track of the permutation the element x induces, each x can be uniquely identified with a permutation $\tau(x) \in S_n$. The claim is that $\tau(\mathcal{X})$ is the required subgroup $\mathcal{X}_*$.

Proposition 6.3. *The subgroup $\tau(\mathcal{X})$ of S_n is isomorphic to $\mathcal{X}$.*

Proof. Per the first isomorphism theorem it suffices to show that τ is a homomorphism with trivial kernel. It is shown through the group action that τ is a homomorphism. Let x, y, g be arbitrary elements of $\mathcal{X}$. Then:

$$\tau(xy)g = xyg = x(yg) = \tau(x)(yg) = \tau(x)\tau(y)g,$$

so that $\tau(xy) = \tau(x)\tau(y)$ as required. Next, assume ker τ contains more than one unique element. Per construction the identity element e of $\mathcal{X}$ is in the kernel and per assumption there exists a non-identity element $v \in \ker \tau$. As both e and v are in the pre-image of the trivial permutation both leave all elements of $\mathcal{X}$ fixed. This is in contradiction with the uniqueness of the identity element, hence the kernel must be trivial. □

Next take $\{e_1, ..., e_n\}$ to be the standard basis of $\mathbb{R}^n$. Every permutation s in S_n can be represented by a permutation matrix in which the columns are the permutated basis vectors $\{e_{s(1)}, ..., e_{s(n)}\}$, and this is the final key step in the translation process.

Proposition 6.4. *The map $\sigma : S_n \to \mathrm{GL}(n, \mathbb{R})$ defined by:*

$$\sigma : s \mapsto \{e_{s(1)}, ..., e_{s(n)}\}$$

is a faithful representation.

Following the definitions this too boils down to showing σ is a homomorphism with trivial kernel. This should come as no surprise, as the definition of a faithful representation was chosen in this way to guarantee the representation is an isomorphism from the group to its image.

Proof. Demonstrating the kernel of σ is trivial is easy, as the trivial permutation is mapped to the identity matrix, and the trivial permutation is the only permutation that keeps all symbols fixed. It is left to show that σ is a homomorphism, and this is done through multiplication with the basis vectors. Let e_i for $1 \leq i \leq n$ be a vector in the standard basis and let $x, y \in S_n$ be two arbitrary permutations. That σ is a homomorphism follows from:

$$\sigma(xy)e_i = e_{xy(i)} = e_{x(y(i))} = \sigma(x)e_{y(i)} = \sigma(x)\sigma(y)e_i.$$

□

For those familiar with representation theory it is apparent that this approach to the construction of a translation shows similarity with the regular representation. The main difference is that the regular representation of S_n is an $n!$-dimensional module, whereas the result of the current approach is at most n^2-dimensional. For virtually every dataset this is a huge difference in space complexity of the algorithm. For a further explanation of representations the reader is referred to [24].

For the final step, if ϕ denotes the natural map:

$$\mathrm{M}(n, \mathbb{R}) \to \mathbb{R}^{n^2},\ \{v_1, ..., v_n\} \mapsto (v_1, ..., v_n),$$

the group $\mathcal{X}$ is represented as a point cloud by its image under the composition $r = \phi \circ \sigma \circ \tau$, or in formula, $X = r(\mathcal{X})$. With the data pre-processed, the search for clusters can begin.

In order to do this the persistent homology and the zeroth persistent Betti numbers need to be computed. After this is done, the Betti numbers can be compared and the longest persisting one can be isolated. Assuming this Betti number is given by $\beta_0^{i,j}$, consider the ϵ-complex in the filtration for any $\epsilon \in [\epsilon_i, \epsilon_j]$. The connected components in the 1-skeleton of this complex form the clusters. These components give rise to an equivalence relation $\sim$ on the vertices where $x \sim y$ if and only if x and y lie in the same connected component. Equivalently, this relation is defined by the condition that $x \sim y$ if and only if there exists a 1-simplex in the complex with x and y as its vertices. Because the vertices stand in one-to-one correspondence with the point cloud, the clusters are given by the partitioning of X into equivalence classes. Formally the clusters are therefore given by $X/\sim$.

The algorithm is given in pseudocode below. Note that the number of clusters is not required to be given, but that the algorithm is easily modified to look for a predefined number of clusters.

Algorithm 1 The Clustering Algorithm

```
function HomologicalClustering(dataset D)
    $X \leftarrow$ The pre-processed dataset D as a point cloud        ▷ Group structure is optional
    $\mathcal{F} \leftarrow$ The filtration with respect to the chosen metric
    $\mathcal{B}_0 \leftarrow$ The set of zeroth persistent Betti numbers

    $\epsilon \leftarrow 0$                                               ▷ Placeholder values
    maxlen $\leftarrow 0$

    for all $\beta_0^{i,j}$ in $\mathcal{B}_0$ do
        if $(j - i) >$ maxlen then
            maxlen $\leftarrow (j - i)$
            $\epsilon \leftarrow \epsilon_i$

    F $\leftarrow$ The complex $K_\epsilon(X)$ in the filtration $\mathcal{F}$
    $\mathcal{C} \leftarrow$ A set of empty sets, one for every connected component of $F$    ▷ Constructing $X/\sim$

    for all vertices $v$ in $F$ do
        Assign the vertex to its equivalence class in $\mathcal{C}$

    return $\mathcal{C}$                                              ▷ Return the clusters
```

The choice to select ϵ_i for the scale ϵ is flexible, as any value in the interval $[\epsilon_i, \epsilon_j]$ will do. The complexes in the filtration form an increasing sequence with respect to inclusion so that the ϵ_i-complex has the least number of 1-simplices over the complexes in the subfiltration. This makes the choice for ϵ_i a natural one with a view to computation, unless more information is required. Finding the connected components in this complex can be done through the use of any graph traversal algorithm, for instance a breadth first search. When the procedure is finished the point cloud is returned, partitioned into clusters.

6.3 Conclusion

Topological data analysis is a viable area of research in the intersection of mathematics, computing science and data science. With data growing ever more important and with the sparked interest of the currently highly active machine learning community, many of the ideas from topological data analysis are bound to see more application in the future. In particular the ability to discriminate between noise and actual features of the data could be of great use in big data.

There is much left unsaid about the proposed clustering algorithm. Before a full implementation can be realised a general interface between datasets and groups based on the datasets would need to be designed. This interface itself raises new questions, especially if the presence of a complete group structure is required. Under certain conditions it might be possible to approximate or extrapolate to a group. No structures other than groups have been explored, and neither have specific types of groups such as topological groups. On the topic of time and space complexity, no analysis was done, but the space complexity could be further bounded if a lower dimensional point cloud representation were possible. Finally, there may be other topological concepts useful for clustering. Branching out from homology, (higher) homotopy groups could find applications in clustering.

In conclusion, even after this work there is more than enough left to explore.

References

[1] S. Mac Lane, *Categories for the Working Mathematician*, Springer Graduate Texts in Mathematics, 1978.

[2] D.S. Dummit, R.M. Foote, *Abstract Algebra*, Wiley, 2003.

[3] A. Hatcher, *Algebraic Topology*, Cambridge University Press, 2001.

[4] G.E. Bredon, *Topology and Geometry*, Springer Graduate Texts in Mathematics, 1993.

[5] L. Lovász, *Topological Methods in Combinatorics*, Eötvös Loránd University Lecture Notes, 2013. `http://web.cs.elte.hu/~lovasz/kurzusok/topol13.pdf`.

[6] V. de Silva, R. Ghrist, *Coverage in Sensor Networks via Persistent Homology*, Algebraic and Geometric Topology, 7:1 (2007), 339-358.

[7] G. Carlsson, *Topology and Data*, Bulletin of the American Mathematical Society, 46:2 (2009), 255-308.

[8] A. Zomorodian, G. Carlsson, *Computing Persistent Homology*, Discrete Computational Geometry, 33:2 (2005), 249-274.

[9] R.L. Schilling, *Measures, Integrals and Martingales*, Cambridge University Press, 2017.

[10] P. Bubenik, *Statistical Topological Data Analysis using Persistence Landscapes*, Journal of Machine Learning Research, 16:1 (2015), 77-102.

[11] B.T. Fasy, J. Kim, F. Lecci, C. Maria, V. Rouvreau, `https://cran.r-project.org/web/packages/TDA/index.html`.

[12] G. Carlsson, T. Ishkhanov, V. de Silva, A. Zomorodian, *On the Local Behavior of Spaces of Natural Images*, International Journal of Computer Vision, 76:1 (2008), 1-12.

[13] G. Carlsson, A. Zomorodian, A. Collins, L. Guibas, *Persistence Barcodes for Shapes*, Eurographics Symposium on Geometry Processing (2006).

[14] V. de Silva, G. Carlsson, *Topological Estimation using Witness Complexes*, Symposium on Point-Based Graphics (2004).

[15] F.T. Pokorny, M. Hawasly, S. Ramamoorthy, *Topological Trajectory Classification with Filtrations of Simplicial Complexes and Persistent Homology*, The International Journal of Robotics Research, 35:1 (2016), 204-223.

[16] F.T. Pokorny, C.H. Ek, H. Kjellström, D. Kragic, *Persistent Homology for Learning Densities with Bounded Support*, Advances in Neural Information Processing Systems, 25 (2012), 1817-1825.

[17] G. Singh, F. Memoli, T. Ishkhanov, G. Sapiro, G. Carlsson, D.L. Ringach, *Topological Analysis of Population Activity in Visual Cortex*, Journal of Vision, 8:8 (2008), 1-18.

[18] U. Bauer, H. Edelsbrunner, *The Morse Theory of Čech and Delaunay Filtrations*, Proceedings of the Thirtieth Annual Symposium on Computer Geometry, 2014, 484-490.

[19] J.M. Lee, *Introduction to Smooth Manifolds*, Springer Graduate Texts in Mathematics, 2013.

[20] M.P. Wand, M.C. Jones, *Kernel Smoothing*, Monographs on Statistics and Applied Probability, 60 (1995).

[21] G. Singh, F. Mémoli, G. Carlsson, *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*, Eurographics Symposium on Point-Based Graphics, 2007.

[22] G. Carlsson, V. de Silva, *Zigzag Persistence*, Foundations of Computational Mathematics, 10:4 (2010), 367-405.

[23] G. Carlsson, A. Zomorodian, *The Theory of Multidimensional Persistence*, Discrete Computational Geometry, 42:1 (2009), 71-93.

[24] W. Fulton, J. Harris, *Representation Theory A First Course*, Springer Graduate Texts in Mathematics, 2004.