Topological Data Analysis

Anson Lai

1 Introduction

It is important in data analysis to understand the shape that our data takes but as the dimension increases many statistical methods like linear regression won't be precise because the shape of our data is much more complicated and no longer represented by a straight line. So we use topological data analysis in which the theory underlying it is simplicial homology. Using simplicial homology to study topological features of our data we won't have to worry about the coordinates of the data, we can keep track of noises or features that are insignificant and most importantly we are able to do this by representing our data in a simpler manner which is called a simplicial complex. Most of this report will just explain the underlying theory which is simplicial homology instead of describing the computational methods and how to actually get our data into a simplicial complex representation since they all basically rely on simplicial homology. So we start by defining simplicial homology and then in the end briefly describe persistent homology which is a method to see which topological features are persistent. The article used to learn about the steps in simplicial homology is [3].

2 Simplicial homology

If we have a point cloud and we wish to compute the homology of it in order to study the shape of the data, this may be a complicated task. We may want to use simiplicial homology since the underlying objects which are simplicial complexes are combinatorial data. Simplicial homology provides computational techniques to search for topological features. To use simplicial homology on our data, we need to find a simplicial complex representation of it with similar homology and this is part of persistent homology to be discussed in 3. Before getting there, we first define simiplicial homology.

2.1 Defining simplices and simplicial complexes

A simplex is a generalization of a triangle in higher dimensions. More formally, let $k \geq 0$ and suppose we have $k+1$ points $v_0, v_1, \ldots, v_k \in \mathbb{R}^m$ such that $v_1 - v_0, v_2 - v_0, \ldots, v_k - v_0$ are linearly independent. The convex hull of the points $v_0, v_1, \ldots, v_k$ is called a k-simplex, which is the set

$$\sigma = \{a_0 v_0 + a_1 v_1 + \cdots + a_k v_k \mid \sum_{i=0}^{k} a_i = 1, a_i \geq 0\}$$

We denote the k-simplex as $\sigma = [v_0, v_1, \ldots, v_k]$ and $v_0, \ldots, v_k$ are called the vertices of the simplex. From this definition, a 0-simplex is a point, 1-simplex a line segment, 2-simplex a filled triangle, 3-

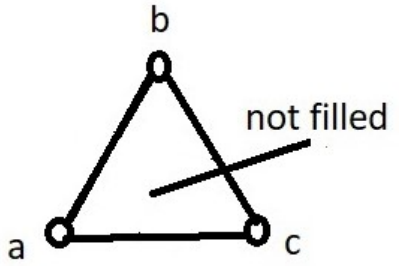

Figure 1: This is a simplicial 1-complex created by connecting 3 edges

simplex a solid tetrahedron. A k-simplex can be thought of as a k dimensional generalized triangle. A few properties of simplices are:

1. Given a k-simplex, the convex hull of any of its $m+1$ vertices is called an m-face of it. For example, given $\sigma = [v_0, v_1, \ldots, v_{10}]$, a 10-simplex, a 4-face of it is $[v_1, v_3, v_9, v_2, v_5]$. There are $\binom{11}{5} = 462$ 4-faces of a 10-simplex. Since $v_1 - v_0, \ldots, v_{10} - v_0$ are linearly independent by definition, $v_3 - v_1, v_9 - v_1, v_2 - v_1, v_5 - v_1$ are also linearly independent which means $[v_1, v_3, v_9, v_2, v_5]$ is a 4-simplex leading to the second property.
2. The m-face is a m-simplex, so the 0-face is a 0-simplex (vertex) and 1-face is a 1-simplex (edge).

Now we define a simplicial complex. It is a glueing of the simplices we defined above together. This is useful because we can represent shapes by a glueing of simplices which are most of the time a homotopically equivalent representation. Then we can apply algebraic calculations on the simplicial complex to study features of the original shape.

A simplicial complex is a collection $\mathcal{K}$ of simplices in $\mathbb{R}^n$ such that

1. If $\sigma \in \mathcal{K}$, then every face of σ is in $\mathcal{K}$
2. A non-empty intersection of any two simplices $\sigma_1, \sigma_2 \in \mathcal{K}$ is a face of both σ_1 and σ_2

We define the dimension of $\mathcal{K}$ to be $\dim(\mathcal{K}) = \max\{\dim(\sigma) \mid \sigma \in \mathcal{K}\}$ where $\dim(\sigma)$ is the dimension of the simplex σ. If it is a r-simplex, then it has dimension r.

For example, the simplicial complex given in figure 1 is the collection $\{[a], [b], [c], [a, b], [b, c], [c, a]\}$ and it is one dimensional since the highest dimensional simplex are edges.

2.2 Orientation

So far it doesn't matter in what order we write the vertices within the bracket in the k-simplex notation. $[a, b, c]$ is the same as $[b, c, a]$ and any other reordering. But in order to properly define the boundary operator and other things leading to simplicial homology we need to define an orientation for simplices. We define an orientation of a k-simplex $[v_0, v_1, \ldots, v_k]$ for $k \geq 1$ to be an equivalence class of reorderings of the vertices where two reorderings are equivalent if and only if they differ by an even permutation. Thus, there are exactly two orientations. Reorderings of $[v_0, v_1, \ldots, v_k]$ by an even permutation forms one class and reorderings by odd permutation forms the other class. From now on, a k-simplex σ is one that is oriented and we denote the simplex with opposite orientation to be $-\sigma$. For example, $[a, b, c] = [b, c, a] = [c, a, b]$ and $-[a, b, c] = [b, a, c] = [a, c, b] = [c, b, a]$. For

a 0-simplex $[v_0]$, there is only 1 reordering but we still consider it to have an opposite orientation denoted $-[v_0]$. That is, $[v_0]$ is different from $-[v_0]$, we need this so that the 0-chain group to be defined is actually a group (i.e. $[v_0]$ has inverse -$[v_0]$).

2.3 Cycle and boundary group

We would like to quantify topologically important features in our data that are basic such as components and holes but we will do this on the simplicial complex representation of our data which preserves topological features. So now we will identify cycles and holes in the simplicial complex. We want the identification to be systematic and algebraic so we will need to define chains, cycles, boundary group, etc to say exactly what a hole is and count them.

Let $\mathcal{K}$ be a simplicial complex. A k-chain is a finite formal sum $\sum_{i=1}^{N} c_i\sigma_i$ where $c_i \in \mathbb{Z}$ and $\sigma_i \in \mathcal{K}$ is an oriented k-simplex. We also treat σ_i to be the same as $-\tau_i$ where τ_i is σ_i with the opposite orientation. So we can replace σ_i with $-\tau_i$ or $-\sigma_i$ with τ_i in the summation whenever we want. The set of all k-chains is a free abelian group and is denoted C_k. We choose an orientation for each k-simplex in $\mathcal{K}$ and the set of them will be a basis for the free abelian group. There is a 0-chain group C_0 all the way up to a $\dim(\mathcal{K})$-chain group, $C_{\dim(\mathcal{K})}$. For $r < 0$ or $r > \dim(\mathcal{K})$, we have $C_r = 0$.

To motivate how to detect holes, in figure 1 the simplicial complex has a hole but to express this idea we can only use objects that we have which are 1-simplices. Perhaps we can say the 1-chain $[a, b] + [b, c] + [c, a]$ represents a hole but then we need a computational way to output this 1-chain without knowing there is a hole in the first place. It turns out we can define a boundary operator which has a nice property that says the boundary of a boundary is 0. So in the context of figure 1, the boundary operator applied to the 3 edges with a certain orientation will output 0. Thus we can just compute the kernel of the boundary operator and those are potential candidates that represent a hole which are called cycles. Before getting into detail, we define the boundary operator.

Let $\sigma = [v_0, v_1, \ldots, v_k] \in \mathcal{K}$ be a basis element of C_k which is an oriented k-simplex, the boundary operator is $\partial_k : C_k \to C_{k-1}$ which is a group homomorphism on the basis elements of C_k defined by

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

and then we extend linearly. $\hat{v}_i$ indicates that this vertex should be removed. For $k \leq 0$ or $k > \dim(\mathcal{K})$, we define ∂_k to be the zero map. ∂_k is also a well defined map which means every ordering of σ in the same orientation is mapped to the same thing. One reason the operator is defined like this with alternating sign is because we get that $\partial_k \partial_{k+1} = 0$ for all k and this leads to a way to detect holes by solving kernel of boundary operator but first we show this holds. Using the definition of boundary operator,

$$\partial_k \partial_{k+1}([v_0, v_1, \ldots, v_{k+1}]) = \partial_k \left(\sum_{i=0}^{k+1} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_{k+1}] \right) = \sum_{i=0}^{k+1} (-1)^i \partial_k [v_0, \ldots, \hat{v}_i, \ldots, v_{k+1}]$$

$$= \sum_{i=0}^{k+1} (-1)^i \left(\sum_{j=0}^{i-1} (-1)^j [v_0, \ldots, \hat{v}_j, \ldots, \hat{v}_i, \ldots, v_{k+1}] + \sum_{j=i}^{k} (-1)^j [v_0, \ldots, \hat{v}_i, \ldots, \hat{v}_{j+1}, \ldots, v_{k+1}] \right) = 0$$

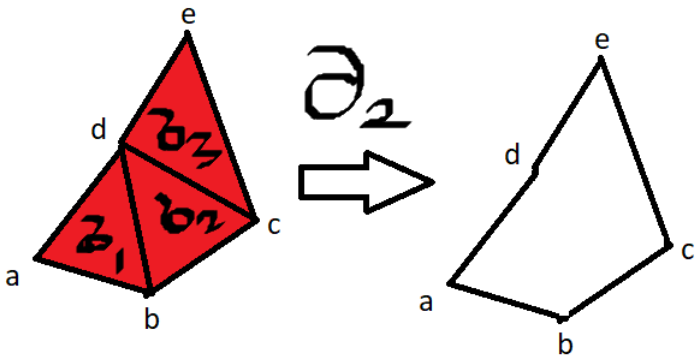

Figure 2:

since the sign of $[v_0, \ldots, \hat{v}_a, \ldots, \hat{v}_b, \ldots, v_{k+1}]$ is $(-1)^{a+b} + (-1)^{a+b-1} = 0$.

Now we can formally define a k-cycle group for the simplicial complex $\mathcal{K}$. The kernel of the boundary map $\partial_k : C_k \to C_{k-1}$ is the k-cycle group of $\mathcal{K}$ and is denoted

$$Z_k = ker(\partial_k) = \{x \in C_k \mid \partial_k(x) = 0\}$$

An element of Z_k is called a k-cycle. The k-cycle group contains k-chains which tell us where a hole is making it possible to count them. We could have multiple k-cycles representing the same hole so we want to get rid of this redundant information which will be done by defining the homology group. It could also be that a k-cycle is not representing a hole because it is the boundary of higher dimensional simplices but the point is that the k-cycle group will not miss information on any hole if there are any. For example on the right of figure 2, we have a hole which we can fill with 2-simplices $\sigma_1, \sigma_2, \sigma_3$. Taking the boundary of $[a, d, b] + [c, b, d] + [d, e, c]$, we get $[d, b] - [a, b] + [a, d] + [b, d] - [c, d] + [c, b] + [e, c] - [d, c] + [d, e] = [b, a] + [a, d] + [d, e] + [e, c] + [c, b]$ and taking boundary again will give 0 since $\partial_1 \partial_2 = 0$. Thus, $[b, a] + [a, d] + [d, e] + [e, c] + [c, b] \in Z_1$ and this is exactly what we want since $[b, a] + [a, d] + [d, e] + [e, c] + [c, b]$ is a 1-cycle which represents a hole and we can compute it by solving $ker(\partial_1)$. Solving for $ker(\partial_1)$ doesn't use the assumption that there is a hole as shown in figure 2.

We define the k-boundary group to be the image of the boundary operator $\partial_{k+1} : C_{k+1} \to C_k$ which is denoted

$$B_k = im(\partial_{k+1}) = \{\partial_{k+1}(x) \mid x \in C_{k+1}\}$$

An element of B_k is called a k-boundary. If we have a k-cycle and it is also a k-boundary then this cycle can not represent a hole and we call it a bounded cycle. If the k-cycle is not in the k-boundary group then this cycle represents a hole and it is called a bounded cycle or k-hole. To summarize this information we have to define the k-homology group which contains the unique k-holes. Before getting there, since $\partial_k \partial_{k+1} = 0$ we have B_k is a subgroup of Z_k. Also, since C_k is an abelian group every subgroup is a normal subgroup. So B_k is a normal subgroup of Z_k.

2.4 Homology group

For a simplicial complex $\mathcal{K}$, the *kth* homology group is defined to be the quotient group Z_k/B_k and is denoted H_k. An element of H_k is called a k dimensional homology class of $\mathcal{K}$. The elements of

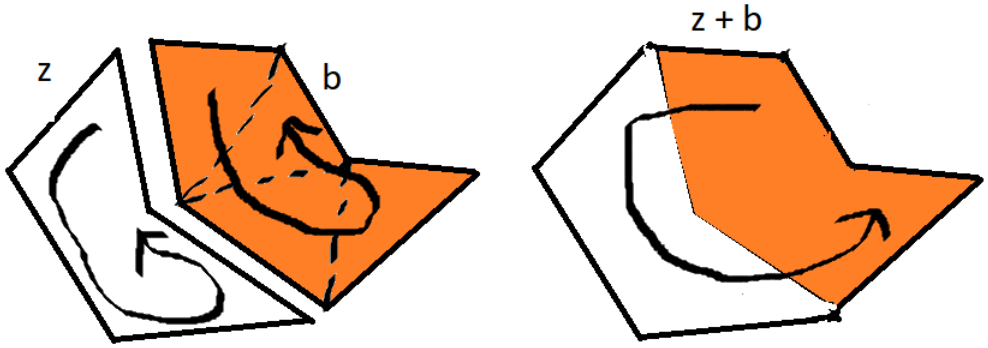

Figure 3: a non-bounded 1-cycle z and bounded 1-cycle b with their sum

the quotient group H_k is given by the equivalence classes $[z] = z + B_k$ and the equivalence relation is $z_1 \sim z_2 \Leftrightarrow z_1 - z_2 \in B_k$ which is called homology. If $z_1, z_2 \in Z_k$ and $z_1 \sim z_2$, then the cycles are called homologous. So it follows that we have a non-zero k dimensional homology class if and only if it is represented by a k-cycle which is not a k-boundary. Having non-zero k dimensional homology classes means we have a hole in the complex.

If we have 2 distinct non-bounded k-cycles and they represent the same hole, then we have redundant information. However, if non bounded cycles represent the same hole they must differ by a boundary and by the above equivalence relation they are in the same homology class. So H_k actually contains the unique k-holes within the complex. To illustrate this, consider figure 3. We have $z \in Z_1$, $z \notin B_1$ and $b \in B_1$. When we add them, the shared boundaries cancel out and we get $z + b$ shown by the figure. Both $z + b$ and z represent the same hole so they should be in the same homology class. The homology group formally defines this as both z and $z + b$ are in the homology class $[z] = z + B_1$. So z and $z + b$ are not independent elements of the homology group H_1 which is what we want in order to not count the same hole multiple times.

2.5 Betti number

The betti number will count the number of unique k-holes using the k-homology group. The k-th betti number is the rank of H_k which is the cardinality of a maximal linearly independent subset of H_k. If we have two non bounded cycles z_1, z_2 with no face in common and $z_1 + z_2$ is a non-bounded cycle we don't want to count $z_1 + z_2$ as a hole since it is the "union of two disjoint holes". $[z_1]$ and $[2z_1]$ are two distinct homology class representing the same hole so we don't want to count twice. Thus we require maximal linearly independent subset of H_k. The k-th betti number β_k is given by

$$\beta_k = \text{rank}(H_k) = \text{rank}(Z_k) - \text{rank}(B_k)$$

Consider the simplicial complex in figure 1. It is the collection of oriented simplices $\{[a], [b], [c], [a, b], [b, c], [c, a]\}$. We have $C_0 \cong \mathbb{Z}^3$ with basis $[a], [b], [c]$ and $C_1 \cong \mathbb{Z}^3$ with basis $[a, b], [b, c], [c, a]$. The other chain groups are trivial. From the boundary operator $\partial_1 : C_1 \to C_0$, $ker(\partial_1)$ is obtained by solving $k_1([b] - [a]) + k_2([c] - [b]) + k_3([a] - [c]) = 0$ for k_1, k_2, k_3. In which we get the solution to be span of $(1, 1, 1)$. Thus, $ker(\partial_1) = Z_1 \cong \mathbb{Z}$. Since $C_2 = 0$, we have B_1, the boundary group, is the trivial group. Hence, $H_1 = Z_1/B_1 \cong \mathbb{Z}$. Then it follows that $\beta_1 = rank(H_1) = 1$ which formalizes the idea

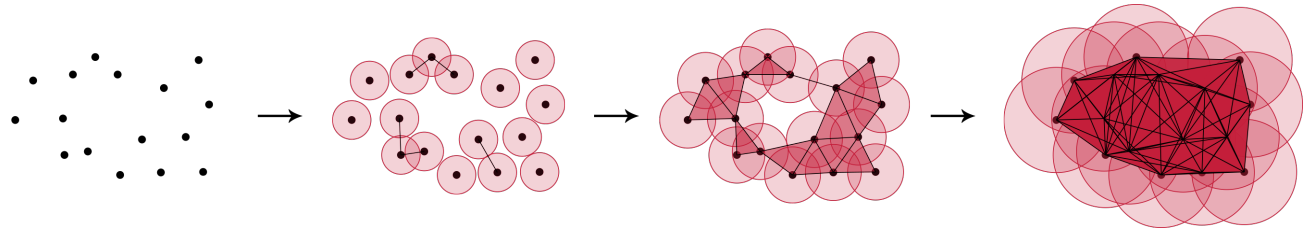

Figure 4:

that there is one 1-hole.

Furthermore, the 0 cycle group Z_0 is C_0 since $C_{-1} = 0$. The 0 boundary group is generated by the elements $[b] - [a], [a] - [c], [c] - [b]$ which is a 2 dimensional subgroup of C_0. Thus, the 0 homology group is $H_0 = Z_0/B_0 \cong \mathbb{Z}$. The H_0 group also measures the number of connected components in the complex, which in this case we have 1 connected component. There is a path made of 1 simplices to go from vertex to vertex. The betti number is $\beta_0 = rank(H_0) = 1$ showing that there is one connected component.

For complicated simplicial complex, usually a simplification is made like defining the k-chain so that the coefficients are in $\mathbb{Z}_2$. This way we can remove the need for defining an orientation.

3 Persistent homology

We have defined simplicial homology groups but we need to get our data into a simplicial complex representation in order to apply simplicial homology. Because data is usually discrete points we need to find the connection and continuity between the points. The point cloud doesn't have a topological structure at first and we have to define the Čech complex to get an interesting structure relating the points. As shown in 4, taken from the website by Christian Bock[1], we start with a point cloud consisting of 16 points. Then define a ball around each point in which the diameter is described by the parameter $t > 0$ and we work with euclidean distance. As t increases and 2 balls intersect, a 1-simplex is formed between them. When 3 balls intersect a 2-simplex is formed between them and so on. By doing this we get a simplicial complex for each value of t. We can also apply simplicial homology to each value of the parameter. For example, the first diagram which is just a simplicial complex with vertices has 16 connected components so the 0 homology group has rank 16 or betti number 16. As betti number is the number of connected components. Then in the second diagram we have 11 connected components as we start to form 1-simplices. As the parameter increases the number of connected components decreases. We can also count the number of 1-holes and how the value changes as we increase the parameter t. For example in the 3rd phase there is clearly one 1-hole and then it disappears when we increase the parameter. So we obtain different homology groups and betti numbers as the parameter increases. Doing this we can see how persistent the topological features are for this space. We also have the values of t for which the hole first appears and when it disappears since when t is large enough all the balls will intersect leaving no holes. Using this we know when a feature is noise since a hole that appears and disappears quick is insignificant to the data. This information is usually described in a persistence diagram.[1][2]

References

[1] Christian Bock. *A gentle introduction to persistent homology*. URL: https://christian.bock.ml/posts/persistent_homology/.

[2] JDPORRAS. *Understanding the shape of data*. URL: https://quantdare.com/understanding-the-shape-of-data/.

[3] Alexander D. Smith, Paweł Dłotko, and Victor M. Zavala. "Topological data analysis: Concepts, computation, and applications in chemical engineering". In: *Computers and Chemical Engineering* 146 (2021), p. 107202. ISSN: 0098-1354. DOI: https://doi.org/10.1016/j.compchemeng.2020.107202. URL: https://www.sciencedirect.com/science/article/pii/S009813542031245X.

[4] Hendrik Suess. *Simplicial homology notes from Hendrik Suess*.

[5] Wikipedia. *Simplicial complex*. URL: https://en.wikipedia.org/wiki/Simplicial_complex.

[6] Wikipedia. *Simplicial homology*. URL: https://en.wikipedia.org/wiki/Simplicial_homology.