INTRODUCTION TO TOPOLOGICAL DATA ANALYSIS

JASON LI

1. INTRODUCTION

Topological data analysis allows us to precisely determine the geometric properties of a finite point cloud data set of a metric space. The overall structure of the data pipeline is given a data set, we first construct various simplicial complexes on the data set using different parameters to find which parameters best capture the important features of the data. This idea is made rigorous by persistent homology. For the simplicial complexes that best capture the features we want to analyze, we then analyze the simplicial complex using the tools of algebraic topology. Topological data helps solve two difficult problems that frequently arise in data analysis. Topological data analysis, inter alia, revealing geometric aspects of noisy and potentially incomplete data even if the data is of very high dimension.

2. SIMPLICIAL COMPLEXES

We will assume the data sets we are working with are finite subsets of $\mathbb{R}^n$. However, the theory discussed can be generalized to any metric space in place of $\mathbb{R}^n$.

Simplicial complexes are a way to give the data set an easy to work with topology. They are generally easy to construct on a finite set of a metric space. Furthermore, homology groups, the main tool used to analyze the simplicial complexes, are relatively easy to compute on simplicial complexes because of their purely combinatorial nature. To define what simplicial complexes are, we first begin with defining a K-simplex.

Definition 2.1. *K-Simplex: Let $K+1$ be a set of points $\{x_1, \ldots, x_{K+1}\} \subseteq \mathbb{R}^n$, such that the points are all affinely independent. Then the K-simple determined by those set of points, denoted by σ, is*

$$\sigma = \{\theta_1 x_1 + \cdots + \theta_{k+1} x_{k+1} | \sum_{i=1}^{k+1} \theta_i = 1 \text{ and } \theta_i \geq 0 \text{ for all } i\}$$

*We also define the **face** of a simplex to be any sub-simplex generated by a non-empty subset of the $K+1$ points.*

We can think of a K-simplex as a K dimensional generalization of a 2-d triangle or a 3-d pyramid with the vertex set being the set of K+1 points. We require the points to be affinely independent so that no more than 2 points are in the same 1-d subspace.

Definition 2.2. *Simplicial Complex: Let $\mathcal{K}$ be a set of simplices of potentially varying dimension. $\mathcal{K}$ is a simplicial complex if it satisfies the following conditions:*

(1) *every face of a simplex $\sigma \in \mathcal{K}$ is also in $\mathcal{K}$*

(2) *given $\sigma_1, \sigma_2 \in \mathcal{K}$ and $\sigma_1 \cap \sigma_2 \neq \emptyset$, the intersection is a face of both σ_1 and σ_2*

Now we give some examples of different constructions of simplicial complexes.

Definition 2.3. *Vietoris-Rips Complex: Given a set of finite points $\mathbb{X} \subseteq \mathbb{R}$ and an $\epsilon > 0$, define the simplicial complex $Rips_\epsilon(\mathbb{X})$ as follows. K-simplex $\sigma \in Rips_\epsilon(\mathbb{X})$ if and only if the distance between two furthest points in σ is less or equal to ϵ. This distance is sometimes also called the **diameter** of the simplex.*

An important property of Vietoris-Rips Complexes, and many simplicial complexes in general, is that they preserve the overall geometric features of the set they are constructed from. This is explored in greater detail in [1].

Definition 2.4. *Čech Complex: Given a finite set of points $\mathbb{X} \subseteq \mathbb{R}$ and an $\epsilon > 0$, define the simplicial complex $\check{C}ech_\epsilon(\mathbb{X})$ as follows. Let $\mathbb{X}$ be the vertex set. For $\sigma \subseteq \mathbb{X}$, where σ is a simplicial complex, $\sigma \in \check{C}ech_\epsilon(\mathbb{X})$ if and only if the set of $\mathcal{B}_\epsilon(x)$ as $x \in \sigma$ varies has a non-empty intersection.*

Proposition 2.5. *Let $\mathbb{X}$ be a finite point cloud of a metric space and $\epsilon > 0$. Then $Rips_\epsilon(\mathbb{X})$ and $\check{C}ech_\epsilon(\mathbb{X})$ are simplicial complexes.*

Proposition 2.6. *Let $\epsilon > 0$ and $\mathbb{X} \subseteq \mathbb{R}$ be a finite set of points, then*

$$Rips_\epsilon(\mathbb{X}) \subseteq \check{C}ech_\epsilon(\mathbb{X}) \subseteq Rips_{2\epsilon}(\mathbb{X})$$

Proof. To show $\text{Rips}_\epsilon(\mathbb{X}) \subseteq \check{\text{C}}\text{ech}_\epsilon(\mathbb{X})$, let $\sigma \in \text{Rips}_\epsilon(\mathbb{X})$ be a simplex. Then for any $x_1, x_2 \in \sigma$ we have and $|x_1 - x_2| < \epsilon$.

Consider balls $\mathcal{B}_\epsilon(x_1)$ and $\mathcal{B}_\epsilon(x_2)$. Since $|x_1 - x_2| < \epsilon$ then $\mathcal{B}_\epsilon(x_1) \cap \mathcal{B}_\epsilon(x_2) \neq \emptyset$. Since x_1 and x_2 are arbitrary, it follows by induction that $\bigcap_{x_i \in \sigma} \mathcal{B}_\epsilon(x_i) \neq \emptyset$ since σ is finite. Hence the first inclusion is proven.

To show $\check{\text{C}}\text{ech}_\epsilon(\mathbb{X}) \subseteq \text{Rips}_{2\epsilon}(\mathbb{X})$, let $\sigma' \in \check{\text{C}}\text{ech}_\epsilon(\mathbb{X})$. Then the intersection of all the ϵ-balls around the points in σ is non-empty. Thus none of the balls have an empty intersection. So given any two points $y_1, y_2 \in \sigma'$ and $y_3 \in \mathcal{B}_\epsilon(y_1) \cap \mathcal{B}_\epsilon(y_2)$, we have $|y_1 - y_3| + |y_3 - y_2| < 2\epsilon$ and by the triangle inequity, $|y_1 - y_2| < 2\epsilon$. Since the y_1 and y_2 are arbitrary, it follows that $\sigma \in \text{Rips}_{2\epsilon}(\mathbb{X})$. Hence the second inclusion is proven. □

The Vietoris-Rips complex and the Čech complex construct simplicial complexes from points in $\mathbb{R}^n$. Now we will introduce a more flexible way to construct simplicial complexes. A nerve is a simplicial complex determined uniquely after being given a topological space and a cover of the topological space. It turns out a Čech complex is a special case of this more general construction with the topological space being some subset of $\mathbb{R}^n$ contain the data and the cover being the ϵ-balls around each data point.

Definition 2.7. *Nerve: Given an open cover $\mathcal{U} = \{U_i\}_{i \in \Lambda}$ of a topological space X, the nerve of $\mathcal{U}$ denoted $\mathcal{C}(\mathcal{U})$, is the simplicial complex with vertices U_i and*

$$\sigma = [U_{i_1}, \ldots, U_{i_k}] \in \mathcal{C}(U) \text{ if and only if } \bigcap_{j=0}^{k} U_i \neq \emptyset$$

Proposition 2.8. *A Čech complex is a nerve*

Proof. Give $\check{\text{C}}\text{ech}_\epsilon(\mathbb{X}) \subseteq \mathbb{R}^n$ the subspace topology. We know that $\mathcal{U} = \{\mathcal{B}_\epsilon(x_i)\}_{x_i \in \mathbb{X}}$ is a cover of the Čech complex. It is indeed a cover is because the Čech complex is defined in a way such that a simplex

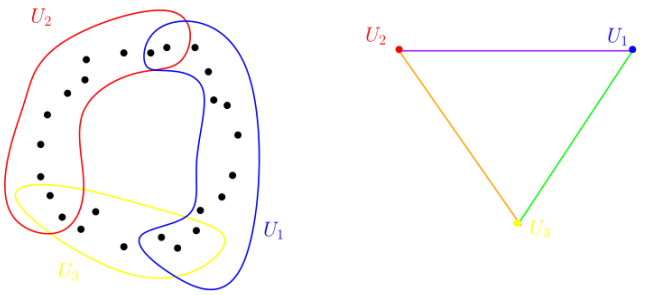

FIGURE 1. Nerve of a cover of points in $\mathbb{R}^2$. Image taken from [2]

is in the complex if and only if it is in one of the balls (since otherwise it would not be in a non-empty intersection). It follows that $\mathcal{C}(\mathcal{U}) = \check{\text{C}}\text{ech}_\epsilon(\mathbb{X})$. □

The nerve theorem is a important result connecting the simplicial complexes to the data sets.To state the theorem we first have to introduce the notion of contractible spaces. Contractible spaces are some of the simplest spaces because they are "almost" a single point, where "almost" means homotopy equivalent.

Definition 2.9. *Contractible: Let X be a topological space. X is said to be contractible if it is homotopy equivalent to a point. Equivalently, X is said to be contractible if the identity map on X is null homotopic.*

Proposition 2.10. *The two definitions of a contractible space are indeed equivalent.*

Proof. Suppose X is homotopic equivalent to $p = \{x\}$. Let $f : X \to p$, $g : p \to X$ such that $f \circ g \simeq id_p$ and $g \circ f \simeq id_X$. Since $g \circ f(x) = g(f(x)) = g(p) = c_{x_o}$ where c_{x_o} the constant map to a x_o. Hence $id_X \simeq x_o$ so the identity on X is null homotopic.

Suppose that the identity on X is null homotopic so $id_X \simeq c_{x_o}$. Then if $f : X \to p$, $g : p \to X$, we have $g \circ f = c_{x_1}$ for some x_1. Since $c_{x_o} \simeq (c_{x_1} = g \circ f)$ since constant loops are homotopic, we have $id_X \simeq g \circ f$.

We $f \circ g = c_p$ so it follows $f \circ g \simeq (c_p = id_p)$. Thus X is homotopy equivalent to a point and the two definitions are indeed equivalent.

□

Definition 2.11. *Good Open Cover: An open cover $\mathcal{U} = \{U_i\}_{i\in\Lambda}$ of topological space X is called a good open cover if every U_i is contractible and every non-empty finite intersection of U_i is contractible.*

Theorem 2.12. *Nerve Theorem: Let $\mathcal{U}$ be a good open cover of topological space X. Then it's nerve $\mathcal{C}(\mathcal{U})$ and X are homotopy equivalent.*

Corollary 2.13. *Let $\mathbb{X} \subseteq \mathbb{R}^n$ be a finite set of points, and fix $\epsilon > 0$. Then there exists the following homotopy equivalence*

$$\check{C}ech_\epsilon(\mathbb{X}) \simeq \bigcup_{x\in\mathbb{X}} \mathcal{B}_\epsilon(x)$$

Proof. $\check{\text{C}}\text{ech}_\epsilon(\mathbb{X})$ is a nerve with the cover being the ϵ-balls around each $x \in \mathbb{X}$. Balls in $\mathbb{R}^n$ are homotopic equivalent to a point thus they are contractible as well as their finite interaction. Thus, the balls are a good cover of their union. Apply the nerve theorem and the result immediately follows. □

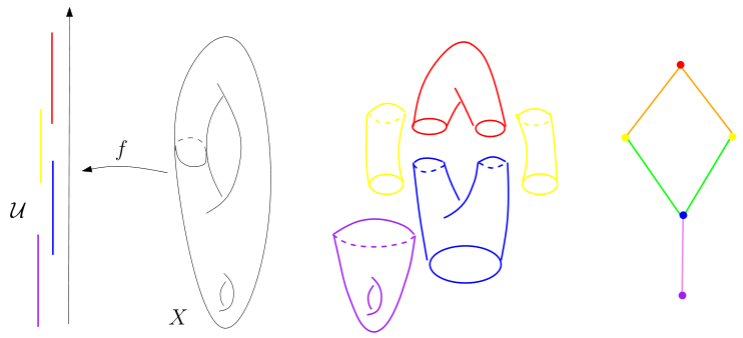

FIGURE 2. Visual depiction of the mapper algorithm. Image taken from [2]

Due to their homotopy equivalence, the Čech complex behaves exactly like the union of ϵ-balls, a subset of $\mathbb{R}^n$. In general, the nerve theorem guarantees us that all that the tools of algebraic topology we are going to use, can indeed be used. This is because the topological features we discover from analyzing the simplicial complex of a data set can be transferred back to the actual data set via homotopy. This is because many concepts in algebraic topology, are homotopy invariants. Thus, they are preserved under homotopy equivalence. For example, homotopic equivalent topological spaces have isomorphic homology groups. Hence, a simplicial complex of a data set being the homotopic invariant to the data set is important because homotopy preserves many algebraic topology concepts.

3. THE MAPPER ALGORITHM

The mapper algorithm was originally formulated by Gurjeet Singh, Facundo Mémoli and Gunnar Carlsson in 2007 in their joint paper. The mapper algorithm is another way to construct an simplicial complex out of a data set $\mathbb{X}$ of a metric space. Specifically, the mapper algorithm will take a cover of a $\mathbb{R}^n$ and create a nerve out of the refined pull back cover of some continuous map $f : \mathbb{X} \to \mathbb{R}^n$, sometimes called the filter or lens function. Following the earlier motif, we apply the nerve theorem realizing that the nerve of the pullback cover and $\mathbb{X}$ are homotopy equivalent.

Definition 3.1. *Pullback Cover: Let X be a topological space and $f : X \to \mathbb{R}^n$ a continuous map. If $\mathcal{U} = \{U_i\}_{i\in\Lambda}$ a cover of $\mathbb{R}^n$. Then the pullback cover of X induced by $(f,\mathcal{U})$ is $\{f^{-1}(U_i)\}_{i\in\Lambda}$. The refined pullback is $\{f^{-1}U_i)'\}_{i\in\Lambda'}$ where $f^{-1}(U_i)' \subseteq f^{-1}(U_j)$, for some j, is a connected component of the pullback cover.*

Proposition 3.2. *The refined pullback cover is a refinement of the regular pullback cover.*

Now we have the prerequisites needed to define the mapper algorithm.

Definition 3.3. *Mapper algorithm: Let $\mathbb{X}$ be a finite set of data points of a metric space, and let $f : \mathbb{X} \to \mathbb{R}^n$ a continuous map with $\mathcal{U} = \{U_i\}_{i\in\Lambda}$ being a cover of $f(\mathbb{X})$.*

Then compute the refined pullback cover $\mathcal{U}' = \{f^-1(U_i)'\}_{i\in\Lambda'}$ induced by $(f,\mathcal{U})$. Then construct the output of the mapper algorithm, which is the nerve $\mathcal{C}(\mathcal{U})$.

Natural questions arise regarding how to choose a function $f : \mathbb{X} \to \mathbb{R}^n$ and how to choose a cover $\mathcal{U}$ of $f(\mathbb{X})$. The answers are discussed in detail in [2] but here is a summary.

Choice of f: The choice of f greatly changes the output of the function. The appropriate choice of f depends on the features we want to highlight in the data.

If we do not have any specific knowledge of the data, standard choices of f include the centrality function $f(x) = \sum_{x \in \mathbb{X}} d(x, y)$ and the eccentricity function $f(x) = \max_{y \in \mathbb{X}} d(x, y)$, where d is the metric of the space the data is in.

Choice of $\mathcal{U}$: The standard choice of cover $\mathcal{U}$ of $f(\mathbb{X})$ is of regularly spaced intervals of some length $r > 0$. r is sometimes called the **resolution** of the cover. The percentage of overlap of these spaced intervals is sometimes called the **gain** of the cover. An interesting result is that when the gain is greater than 50%, the nerve produced by the mapper algorithm is a graph. Resolution and gain are the two of the most important parameters regarding this standard cover of $f(\mathbb{X})$.

In general, the output of the mapper algorithm varies greatly with choice of $\mathcal{U}$. The strategy employed to counteract this problem is to consider the various outputs of the mapper algorithm as the parameters of the cover change. When the output stays constant during a large change in the parameters, we can consider nerve to be capturing the main structure of the data. This strategy helps us distinguish between when a nerve is capturing the noise of the data and when it is capturing the main features of the data. This idea is further explored in the application of persistent homology.

4. Persistent Homology and Applications

Homology groups, loosely speaking, are a way to detect and classify holes in a topological space similar to homotopy groups. The major relevant difference is, however, that homology groups are typically much easier to compute than homotopy groups. Homology groups of a simplicial complex are especially easy to compute. In the context of this paper, we will assume that the computation of homology is already understood.

To motivate persistent homology, a specific way of applying homology, consider finite data set $\mathbb{X} \subseteq \mathbb{S}^1$. When we construct the $\check{\text{C}}\text{ech}_\epsilon(\mathbb{X})$ we want to choose ϵ large enough so that the complex is connected. However, we also want to choose ϵ small enough so that there are not any 1-simplicies "crisscrossing" hole in the center of $\mathbb{S}^1$. In essence, we want to choose ϵ such that the topological features of the simplicial complex are equivalent to the topological features of the data set. A priori, how do we choose such an ϵ? The answer is persistent homology.

Steps in Persistent Homology:

(1) Choose a kth-homology group to analyze.
(2) Choose some interval $[a, b] \subseteq \mathbb{R}$
(3) Vary ϵ from a to b constructing a simplicial complex with that ϵ on the data set. Record the kth-homology group as ϵ varies. This change in the homology groups is usually visualized by a **barcode diagram**. A barcode diagram is a plot with the horizontal axis being the changing parameters and the vertical axis being homology groups or homology generators.
(4) Conclude that the homology groups that stayed constant for the large ranges of ϵ encode the major features of the data.

Using persistent homology we can choose appropriate ϵ to construct our simplicial complexes. Now we know the homology group(s) that best capture the features of the data set.

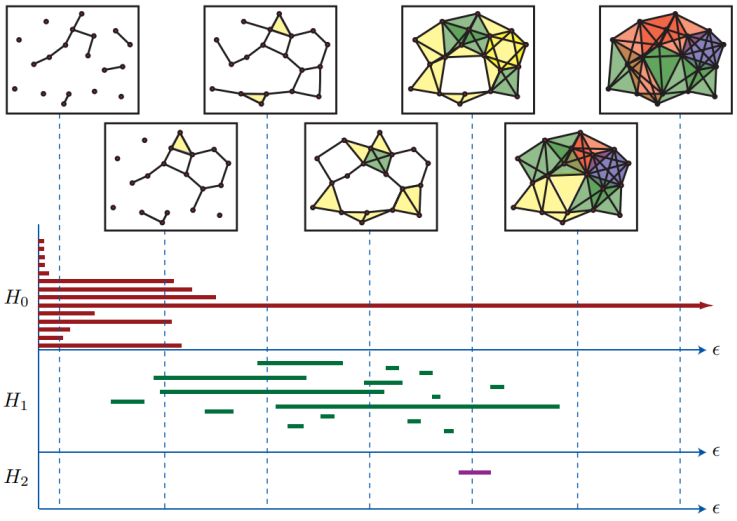

FIGURE 3. Barcode diagram of the first to third homology group for a data set. Image taken from [4]

As a final note, we will give an example of topological feature we can analyze from the homology group of a simplex with a well choosen ϵ. Geometrically, Betti numbers, with the i-th Betti number being denoted b_i, are the number of n-dimension holes in a topological space. Formally, b_i is defined as the rank of the ith-homology group.

For example, b_0 is the number of connected components of the topological space. b_1 is the number of one-dimensional "circular holes" in the topological space. b_2 is the number of two-dimensional holes in the topological space. i.e the number of "voids" or "cavities".

Betti numbers are just one example of a concept from algebraic topology that be applied to reveal the geometric and topological features of a finite data set.

5. CONCLUSION

The data pipeline explored is that once given a data set, we first construct the appropriate simplicial complex(exs) using persistent homology. Then using the computed homology groups, we can analyze the geometric and topological features of a simplicial complex, and in turn, reveal the geometric and topological features of the data set, which through other methods, may be difficult to detect.

REFERENCES

[1] Erin W Chambers, Vin De Silva, Jeff Erickson, and Robert Ghrist. Vietoris–rips complexes of planar point sets. *Discrete & Computational Geometry*, 44(1):75–90, 2010.

[2] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists, 2017.

[3] R. Ghrist. *Elementary Applied Topology*. CreateSpace Independent Publishing Platform, 2014.

[4] Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[5] A. Hatcher, Cambridge University Press, and Cornell University. Dept. of Mathematics. *Algebraic Topology*. Algebraic Topology. Cambridge University Press, 2002.

[6] John Lee. *Introduction to topological manifolds*, volume 202. Springer Science & Business Media, 2010.

[7] Raúl Rabadán and Andrew J Blumberg. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, 2019.