Topological Data Analysis

Priyangshu Ghosh[1]*, guided by Dr. Kuntal Roy[2]

Abstract
This report focuses on applications of Topology in Data Analysis(TDA) and Machine Learning. Over the last 15 years, I have witnessed that topology and algorithms form a potent mix of disciplines with many real-world applications. Here, we will try to expose the central ideas using simple means and minimize the amount of technical detail.

Keywords
Topological Data Analysis — Metric Space Data — Persistent Homology —

[1] *Chennai Mathematical Institute*
[2] *Department of EECS, Indian Institute of Science Education and Research, Bhopal*
***Corresponding author**: priyangshu@cmi.ac.in

Contents

Introduction 1
1 Some background in Topology 1
1.1 Simplicial Complexes and Homology 1
1.2 Working examples 2
2 Persistent Homology 2
2.1 Čech and Vietoris-Rips complex 3
2.2 Computing persistence 3
2.3 Persistence barcodes and diagrams 3
3 Stability theory 4
3.1 Formalism 4
Acknowledgments 4

Introduction

Topology informally refers to encoding features of the shape of data. Hence, it provides a complementary approach to localised and rigid geometric features. Topological features can capture large-scale and intrinsic properties of data sets effectively. All data can be considered as vectors in a space and, therefore have corresponding shapes. TDA borrows from algebraic topology, computational geometry, and algorithms to study data's higher dimensional topological properties.

In the forthcoming section, I will summarize the most the common path is taken in applications of TDA to transform data into measures of topology, and subsequently to machine learning.

1. Some background in Topology

Our primary focus in this introduction is to understand homology. Homology is a technique of studying topological spaces using algebraic objects. We will first be constructing homology groups, which are topological invariants(invariant properties under homeomorphisms). Homological invariants have a convenient hierarchical characterization that is expressed through a homology number. These homology numbers are just one type of topological invariant. Whereas geometric invariants typically require many detailed calculations, homology is a summary of geometric features that can be formulated less computationally intensively. The crossover from homology to statistics is natural in some sense since both disciplines' motivation lies in summarizing information. A very important class of algebraic invariants are the homology groups, which encode a a great deal of information while still being efficiently computable

In many cases. Homology groups arise from combinatorial representations of the manifold, the chain complexes.

1.1 Simplicial Complexes and Homology

The n-simplex, Δ^n, is the simplest geometric figure determined by a collection of $n+1$ points in Euclidean space $\mathbb{R}^n$. Geometrically, it can be thought of as the complete graph on $n+1$ vertices, which is solid in n dimensions. $\Delta^n := \{(x_0,..,x_n) \in \mathbb{R}^{n+1} | \sum x_i = 1, x_i \geq 0 \forall i\}$. A simplicial complex K is a finite set of simplices satisfying the following conditions:

1. For all simplices $A \in K$ with α a face of A, we have $\alpha \in K$.

2. If $A, B \in K$, then A and B are properly situated.

Given a set $A_1^n, \ldots, A_k^n$ of n-simplices of a complex K and the group $\mathbb{Z}/2\mathbb{Z}$, we define an n-chain x as a formal sum: $x = g_1 A_1^n + g_2 A_2^n + \cdots + g_k A_k^n$, where $g_i \in \{0,1\}$. The set of n-chains forms an abelian group over addition: for $x = \sum_{i=1}^k g_i A_i^n$ and $y = \sum_{i=1}^k h_i A_i^n$, we have $x + y = \sum_{i=1}^k (g_i + h_i) A_i^n$. We denote the group of n-chains by L_n. Let A^n be an oriented n-simplex in a complex K. The boundary of A^n is defined as the $(n-1)$-chain of K over $\mathbb{Z}/2\mathbb{Z}$ given by

$$\partial_n(A^n) = A_0^{n-1} + A_1^{n-1} + \cdots + A_n^{n-1},$$

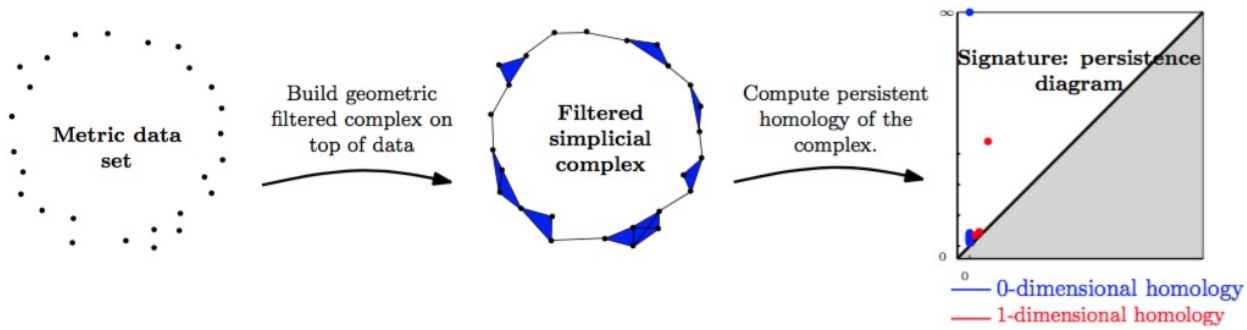

Figure 1. Procedure for abstracting persistence features

where A_i^{n-1} is an $(n-1)$-face of A^n. If $n=0$, we define $\partial_n(\Delta^0)=0$. We can extend the definition of boundary linearly to all of L_n: for an n-chain $x=\sum_{i=1}^k g_i A_i^n$, define

$$\partial_n(x)=\sum_{i=1}^{k} g_i \partial_n(A_i^n),$$

where A_i^n are the n-simplices of K. Therefore, the boundary operator ∂_n is a homomorphism $\partial_n : L_n \to L_{n-1}$. It is important to note that, $\partial_d(\partial_{d+1}(x))=0$. We call an n-chain a cycle if its boundary is zero and denote the set of n-cycles of K over $\mathbb{Z}$ by Z_n. Z_n is a subgroup of L_n, and can also be written as $Z_n=\ker(\partial_n)$. The subgroup of boundaries $B_n=\mathrm{im}(\partial_{n+1})$

The group $H_n(K):=Z_n(K)/B_n(K)$ is the n-dimensional homology group of the complex K over $\mathbb{Z}/2\mathbb{Z}$. Simpler invariants derived from these homology groups, such as betti numbers $\beta_n:=\mathrm{rank}(H_n(K))$. Betti numbers are useful classifiers. $beta_0$ counts the number of connected components, β_1 counts number of holes, and more generally β_k counts number of k dimensional holes.

1.2 Working examples

Consider the simplicial complex in Figure 2

$X=\{a,b,c,d,e,ab,bc,ca,bd,de,ec,abc\}$.

Notice that the boundary of any 1 cell is sum of even 1 simplexes, and via combinatorial identity that $\sum_{k=0}^{n}\binom{n}{k}/\sum_{i=0}^{n//2}\binom{n}{2i}=2$

we see that $|L_0|/|B_0|=2$ Hence $H_0=\mathbb{Z}/2\mathbb{Z}$, $\beta_0=1$

$Z_1=\{ab+bc+ca, bc+ce+ed+db\}$

and $B_1=ab+bc+ca$

$|H_1|=2$, hence $\beta_1=1$, Thus we have a loop.

Since there are no 2 cycles, H_2 will be trivial.

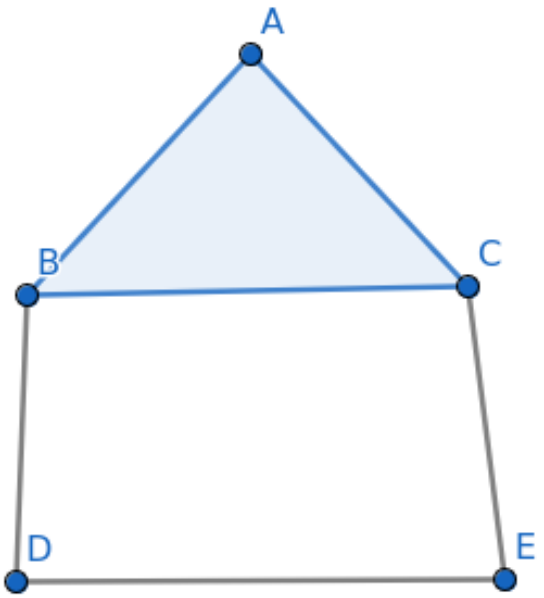

Figure 2. Simplicial complex

2. Persistent Homology

The concept of persistence is motivated by the practical need to cope with noise in data. This includes defining, recognizing, and possibly eliminating noise. Because of the loose definitions of noise and features, we will focus on a range of scales and try to attain a point of view. Persistent homology is a measure of the structure of a filtered simplicial complex. It helps us address homology's major weakness, instability under small changes.

Consider a simplicial complex, K, and a function $\varphi : K \to \mathbb{R}$. We require that φ be monotonic, by which we mean it is non-decreasing along increasing chains of faces, that is, $\varphi(\sigma)\le\varphi(\tau)$ whenever σ is a face of τ. Monotonicity implies that the sublevel set, $K(a)=\varphi^{-1}(-\infty,a]$, is a subcomplex of K for every $a\in\mathbb{R}$. Letting m be the number of simplices in K, we get $n+1\le m+1$ different subcomplexes, which we arrange as an increasing sequence:

$$\emptyset=K_0\subseteq K_1\subseteq\ldots\subseteq K_n=K.$$

This sequence is called a filtered simplicial complex.

2.1 Čech and Vietoris-Rips complex

Čech Complex and Vietoris-Rips complex are abstract simplicial complices defined on metric spaces with a distance parameter over a point cloud.

Let S be a finite set of points in $\mathbb{R}^n$, let $r \geq 0$ be a real number,then

$$\check{\text{C}}\text{ech}(S,r) := \{\sigma \subseteq S | \bigcap_{x\in\sigma} B_x(r) \neq \varnothing\}$$

To construct the Čech complex, we need to test whether a collection of disks has a non-empty intersection, which can be difficult or impossible in some metric spaces. Hence we introduce Rips complex.

The Rips complex of S and r, consists of all abstract simplices in 2^S. whose vertices are at most a distance 2r from one another. In other words, we connect any two vertices at a distance at most 2r from each other by an edge, and we add a triangle or higher-dimensional simplex to the complex if all its edges are in the complex.

$$VR(S,r) = \{\sigma \subseteq S | \text{diam}(\sigma) \leq 2r\}$$

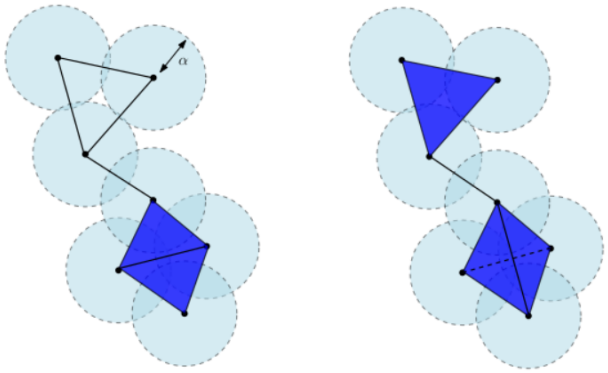

Figure 3. Čech complex(left) and VR complex(right)

2.2 Computing persistence

For a filtered complex $K_0, K_1, .., K_n$ we define For $0 \leq i \leq j \leq n$, the inclusion $K_i \hookrightarrow K_j$ induces a homomorphism $\varphi_p^{i,j}$: $H_p(K_i) \to H_p(K_j)$.The filtration thus corresponds to a sequence of homology groups connected by homomorphisms,

$$0 = H_p(K_0) \to H_p(K_1) \to \ldots \to H_p(K_n) = H_p(K),$$

The *p-th persistent homology groups* are the images of the homomorphisms induced by inclusion, $H_{i,j}^p = \text{im}\,\varphi_{i,j}^p$, for $0 \leq i \leq j \leq n$. The corresponding *p-th persistent Betti numbers* are the ranks of these groups, $\beta_{i,j}^p = \text{rank}\, H_{i,j}^p$. Similarly, we define reduced persistent homology groups and reduced persistent Betti numbers. Note that $H_{i,i}^p = H_p(K_i)$.

The persistent homology groups consist of the homology classes of K_i that are still alive at K_j or, more formally, $H_{i,j}^p = Z_p(K_i)/(B_p(K_j) \cap Z_p(K_i))$.

We can be more concrete about the classes in which the persistent homology groups count. Letting γ be a class in $H_p(K_i)$, we say it is born at K_i if $\gamma \notin \text{im}\,\varphi_{i-1,i}^p$. Furthermore, if γ is born at K_i, then it dies entering K_j if it merges with an older class as we go from K_{j-1} to K_j, that is, $\varphi_p^{i,j-1}(\alpha) \notin \text{im}\,\varphi_p^{i-1,j-1}$ but $\varphi_p^{i,j}(\alpha) \in \text{im}\,\varphi_p^{i-1,j}$.

The index persistence of γ is $j-i+1$. In most applications, we have a function that governs the construction of the filtration, and we call the difference between the function values at the birth and the death the persistence of the class.

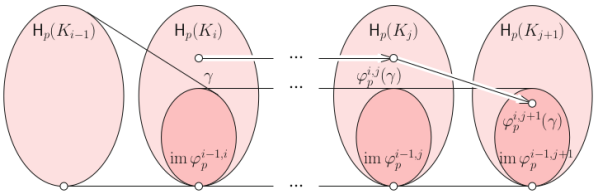

Figure 4. Birth and death of feature γ

2.3 Persistence barcodes and diagrams

In TDA, we define persistence more rigorously as the death birth times of a homology feature. In the union-of-spheres algorithm described previously, these times would be radii at which simplices are born (connected components are born at ε= 0; higher-dimensional features are born later) or die (are absorbed into more significant connected components). Persistence information can be recorded/represented in many ways, including the two plots in Figure 6. In these persistence diagrams (PDs), the black dots represent the connected components of 1-dimension, and the red triangles represent loops of 2-dimension. The y-axis of the graph on the right of Figure 6 has had the transformation persistence = T (birth, death) = death - birth, so its coordinates are (birth, persistence), whereas the graph on the left has coordinates (birth, death). It is common to use this rotation for later applications. These graphs seem to show a very persistent loop in a noisy set. Despite the loop's persistence, in empirical application, these loops would be challenging to observe in the raw data. This result will be discussed more in the applications section. I also provide a persistence barcode example in Figure 6, since it has recently become a commonly used filtration in ML applications of TDA. A persistence barcode is simply another parameterization of the betti numbers of a persistence diagram, wherein data are discrete line segments (1-dimensional signals) over the radii at which a found and stratified by the hierarchy of Betti numbers. Each simplex in the union-of-spheres algorithm has a horizontal line in the barcode, which begins at birth and terminates at death. The persistence barcode was not frequently used in application because of its hierarchical, 1-dimensional nature.

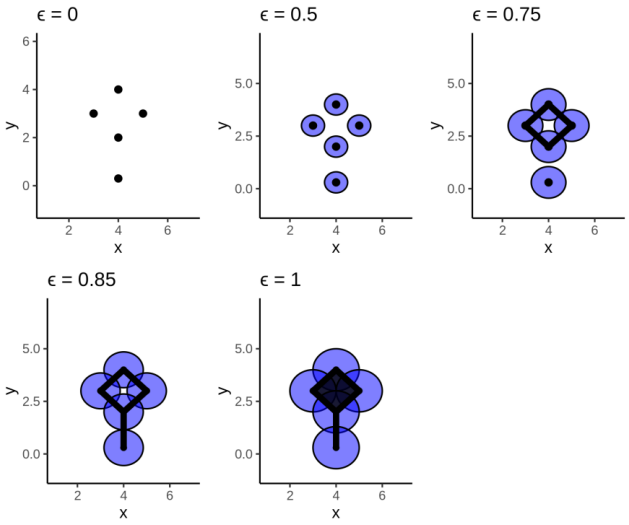

Figure 5. VR complex with ϵ as scaling parameter

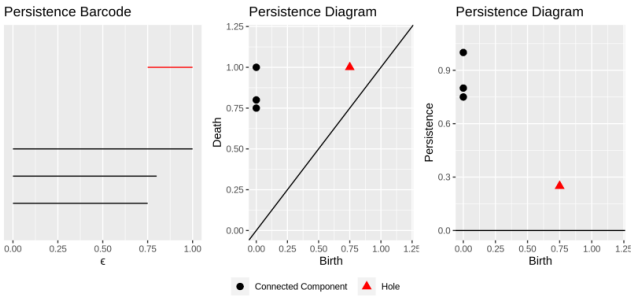

Figure 6. Persistence diagrams from the filtered complex in Figure 6

3. Stability theory

One reason we care about Persistence is its stability under data perturbations. Small changes in the data imply only small changes in the measured persistence. If X and Y are two persistent diagrams, then we define the bottleneck distance

$$W_\infty(X,Y) = \inf_{f:X\to Y} \sup_{x\in X} ||x - f(x)||_\infty$$

where f is a bijection.

A drawback of the bottleneck distance is that it is receptive to noise, as we only look at the furthest pair of the corresponding points. Hence we introduce the degree q Wasserstein distance between two persistence diagrams X and Y

$$W_q(X,Y) = \Big[\inf_{\eta:X\to Y} \sum_{x\in X} ||x - \eta(x)||_\infty^q\Big]^{\frac{1}{q}}$$

Since two persistence diagrams don't always have the same number of points, we add infinitely many points on the diagonal with infinite multiplicities, so for the leftover points, we can project those onto the diagonal.

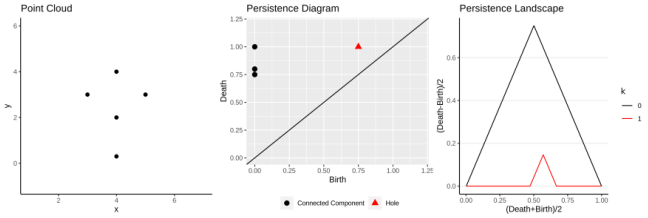

Figure 7. Corresponding persistence landscape

3.1 Formalism

Let K be a simplicial complex and $f, g : K \to \mathbb{R}$ are two monotonic functions. For each dimension p, the bottleneck distance between the diagrams $X = \mathrm{Dgm}_p(f)$ and $Y = \mathrm{Dgm}_p(g)$, we have $W_\infty(X,Y) \le ||f - g||_\infty$.

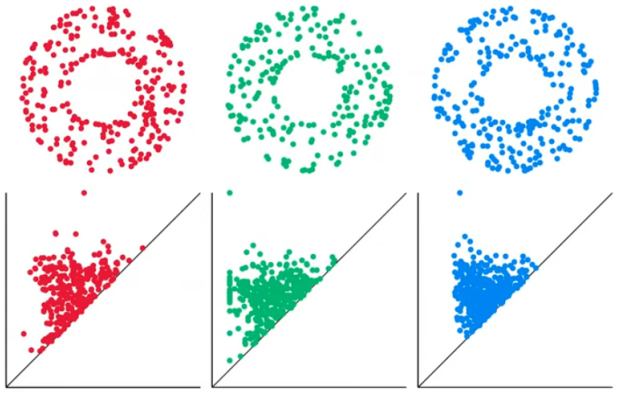

Figure 8. Figure depicting how persistence diagrams behave under small perturbations

Acknowledgments

We want to acknowledge all the researchers and scholars whose work paved the way for advancements in TDA. We thank everyone who contributed to the knowledge pool, making this journey possible.

References

[1] Gunnar Carlsson et al. "Persistence Barcodes for Shapes". In: *International Journal of Shape Modeling* 11.02 (2005), pp. 149–187.

[2] Frédéric Chazal and Bertrand Michel. "An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists". In: *arXiv preprint arXiv:1710.04019* (2017). URL: https://arxiv.org/abs/1710.04019.

[3] Herbert Edelsbrunner. *A Short Course in Computational Geometry and Topology*. Cham, Switzerland: Springer, 2014. URL: https://www.springer.com/gp/book/9783319059560.

[4] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. Providence, Rhode Island: American Mathematical Society, 2010. URL: https://www.ams.org/books/surv/082/.

[5] Ephraim Robert Love. "Machine Learning with Topological Data Analysis". PhD diss. University of Tennessee, 2021. URL: https://trace.tennessee.edu/utk_graddiss/6170/.

[6] J. P. May. *A Concise Course in Algebraic Topology*. Chicago, IL: University of Chicago Press, 1999.

[7] Wikipedia. *Persistent Homology*. https://en.wikipedia.org/wiki/Persistent_homology. Accessed: 2024-07-01. 2023.