Topological Data Analysis

Avik Laha

Columbia University

April 12, 2019

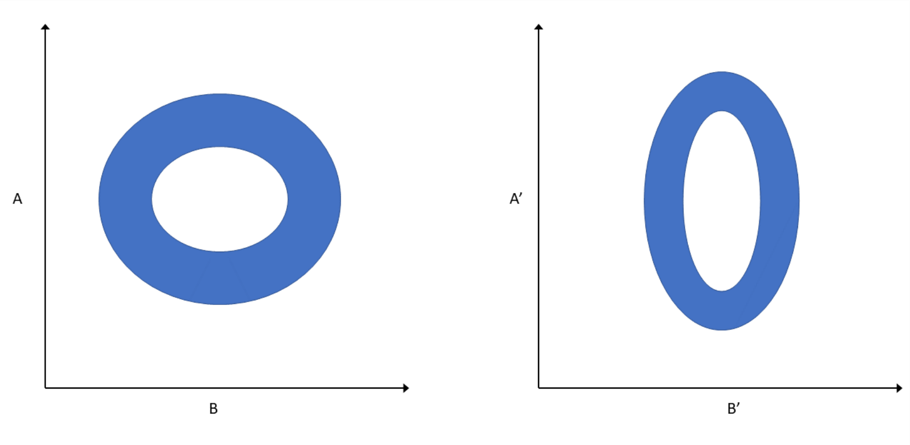
Motivation

- A somewhat different presentation than others – look at general method rather than paper
- Overall, the idea is that many interesting characteristics of data should not depend on certain details of the representation, i.e. they are topological
- Will largely make use of Chazal and Michel's `An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists`

Overview

- First, we will look at what it means for a feature in data to be "topological", and topological invariants
- Then, we will discuss persistent homology in particular as a realization of TDA
- Finally, we will briefly touch on applications

Topological features



- Toy example – for data obtained by different measurement schemes, interesting feature (hole) is preserved

What is topology?

Definition (Topological Space)

A pair $X = (S, \mathcal{T})$ where S is a set and $\mathcal{T}$ a set of its subsets such that:

1. $\emptyset, S \in \mathcal{T}$
2. $\mathcal{T}$ *is closed under arbitrary unions of its elements*
3. $\mathcal{T}$ *is closed under finite intersections of its elements*

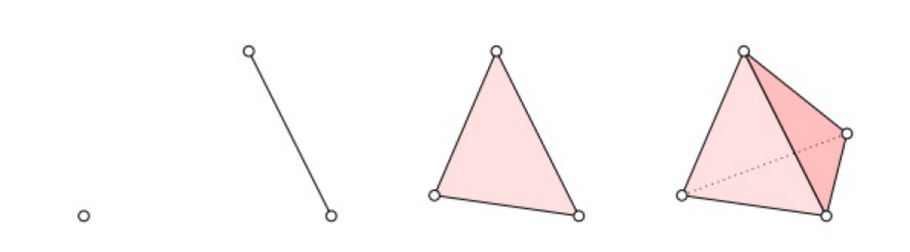
- Interpret elements of $\mathcal{T}$ as open sets
- Gives a notion of a continuous map (preimage of any open set is open) – topology is the study of such spaces and continuous maps between them
- For X, Y topological spaces, if $f : X \to Y$ is a continuous map with continuous inverse, it is a **homeomorphism**, and $X \cong Y$ are **homeomorphic**

Topology and learning

- Sensible to consider the sample space as a topological space, as any metric space has a natural topology
- Collection of data is application of some measurement map $f : X \to Y$ to elements of viable domain $A \subset X$
- Question (for future): how do we recover A or $f^{-1}(B)$ for $B \in Y$, given we only have finitely many samples?

Simplices

- First, need a way to encode topology which we can work with
- An n-simplex is intuitively a basic n-dimensional object, i.e. the convex hull of $n + 1$ affinely independent points



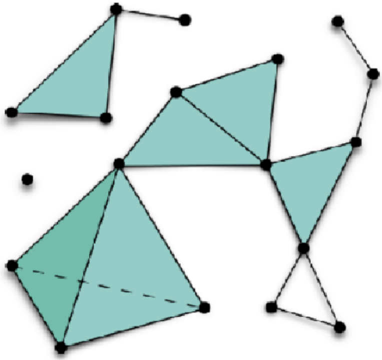
Simplicial complexes

- Abstractly, a generalization of a graph: a 0-simplicial complex is a set of points, a 1-simplicial complex is a graph...
- An n-simplicial complex contains up to n-dimensional simplices (but also all lower dimensions)
- Geometrically, just a set of simplices

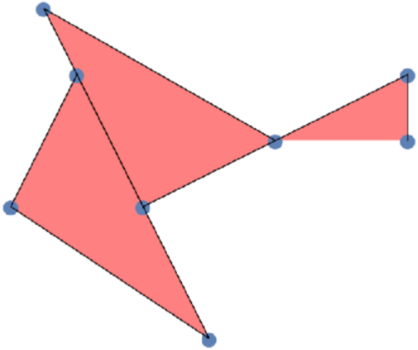
Definition (Simplicial complex)

A pair (V, K) where V consists of "vertices", K is a collection of finite subsets of V which contains all vertices, and obeys $\sigma \in K \implies$ any subset $\varsigma \subset \sigma \in K$ has $\varsigma \in K$

Simplicial complexes, cont.



A Simplicial Complex



Not a Simplicial Complex

Simplicial complexes from data

- For now, assume that X is a finite set of points in (M, ρ) a metric space, d is the inherited metric on X, and $\alpha \in \mathbb{R}^+$:

Definition (Vietoris-Rips Complex)

$\mathrm{Rips}_\alpha(X) :=$ *the set of simplices* $\sigma = [x_0, \ldots, x_n]$ *such that* $d(x_i, x_j) \leq \alpha$

Definition (Cech Complex)

$\mathrm{Cech}_\alpha(X) :=$ *the set of simplices* $\sigma = [x_0, \ldots, x_n]$ *such that* $\bigcap_{i=0}^{n} \overline{B_\alpha(x_i)} \neq \emptyset$

- Note that $\overline{B_\alpha(x_i)}$ is the (closed) ball of radius α centered on x_i
- Related by $\mathrm{Rips}_\alpha(X) \subset \mathrm{Cech}_\alpha(X) \subset \mathrm{Rips}_{2\alpha}(X)$

Rips and Cech complexes

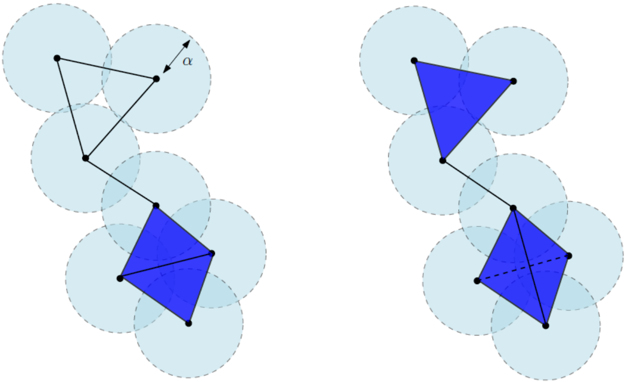


Figure 2: The Čech complex $\mathrm{Cech}_\alpha(\mathbb{X})$ (left) and the and Vietoris-Rips $\mathrm{Rips}_{2\alpha}(\mathbb{X})$ (right) of a finite point cloud in the plane $\mathbb{R}^2$. The bottom part of $\mathrm{Cech}_\alpha(\mathbb{X})$ is the union of two adjacent triangles, while the bottom part of $\mathrm{Rips}_{2\alpha}(\mathbb{X})$ is the tetrahedron spanned by the four vertices and all its faces. The dimension of the Čech complex is 2. The dimension of the Vietoris-Rips complex is 3. Notice that this later is thus not embedded in $\mathbb{R}^2$.

Summary so far

- The topology of data is potentially interesting, so we decided to look into it
- But actual datasets are just finite samples, and in any case topological spaces generally have infinite descriptions
- Introduced simplicial complexes and found a way to build them from finite sets of points, but does this actually help us understand the topology of data?

Nerve theorem

- In short, yes (given satisfaction of certain conditions)

Definition (Nerve)

For a cover $\mathcal{U} = \{U_i\}$ of M, the simplicial complex $C(\mathcal{U}) :=$ the set of simplices $\sigma = [U_{i_0}, \ldots, U_{i_n}]$ such that $\bigcap_{j=0}^{n} U_{i_j} \neq \emptyset$

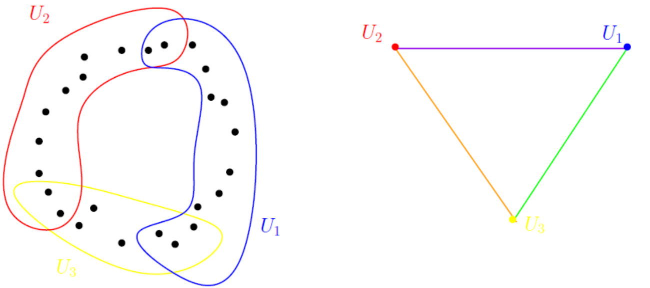


Figure 3: The nerve of a cover of a set of sampled points in the plane.

Nerve theorem, cont.

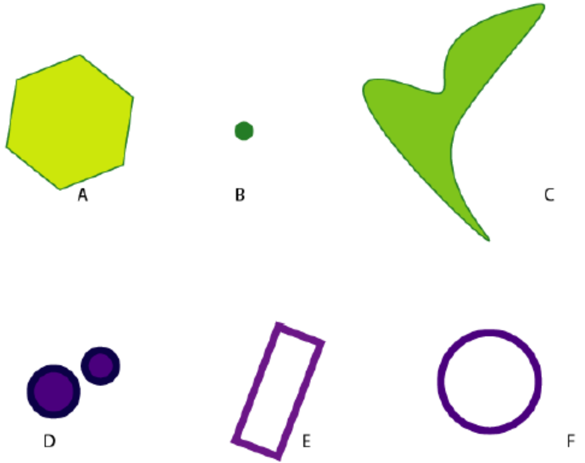
Definition (Homotopy, etc.)

For continuous $f, f' : X \to Y$, a continuous map $h : X \times [0,1] \to Y$ such that $h(x,0) = f(x)$ and $h(x,1) = f'(x)$. If f, f' permit a homotopy, they are **homotopic***, and if there exists $g : Y \to X$ such that $f \circ g$ and $g \circ f$ are homotopic to the identity maps, X and Y are* **homotopy-equivalent**

- Roughly, X can be continuously deformed into Y $\iff$ they are homotopy-equivalent
- If $X \cong Y$ then they are homotopy-equivalent, but the converse is not necessarily true

Nerve theorem, cont.

- If a space is homotopy-equivalent to a point, it is **contractible** – the top row is contractible while the bottom row is not:



Nerve theorem, cont.

Proposition (Nerve Theorem)

Let $\mathcal{U} = \{U_i\}_{i \in I}$ be a cover of M such that for any subset $A \subset I$, the intersection $U_A := \bigcap_{i \in A} U_i$ is empty or contractible. Then M is homotopy-equivalent to the nerve $C(\mathcal{U})$

- Note that as balls in $\mathbb{R}^n$ are convex (hence contractible), and the Cech complex is the nerve of such balls of fixed radius around a set of points, it is homotopy equivalent to the union of those balls

- Our previous observation might make us hope that the Cech complex can summarize the topological data of some space X, and the **Reconstruction Theorem** tells us that this is indeed true under certain (technical) conditions

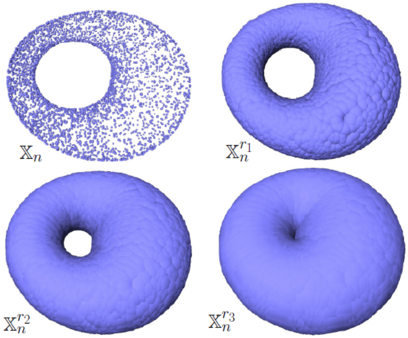
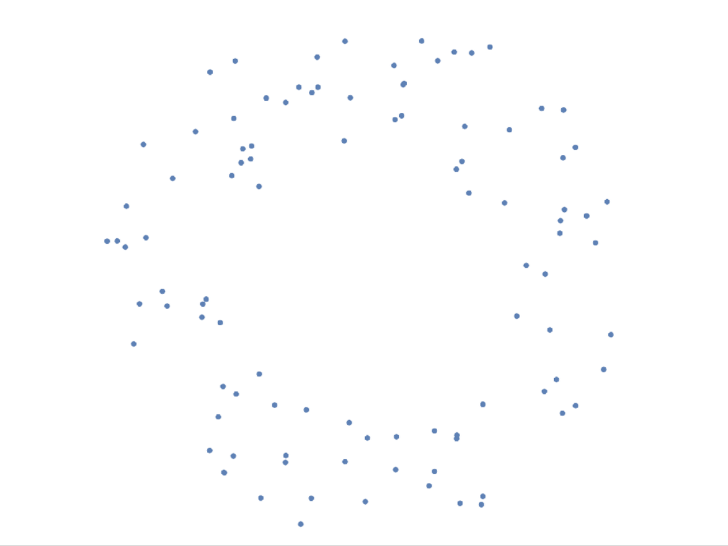
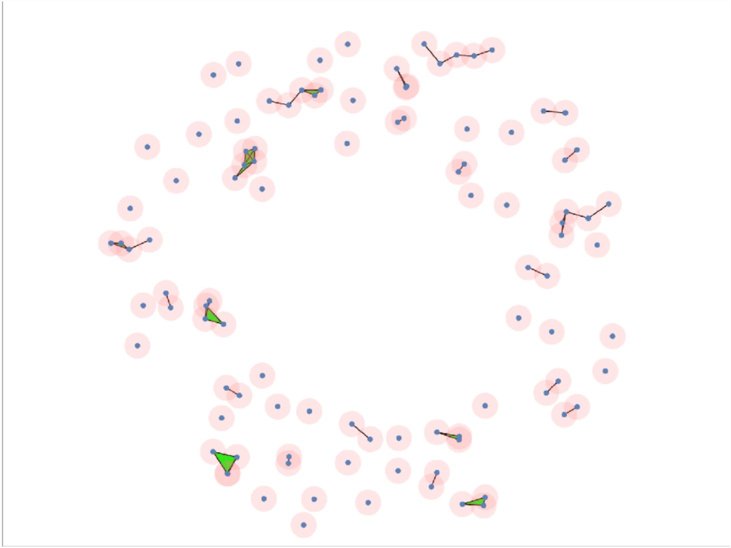


Figure 7: The example of a point cloud $\mathbb{X}_n$ sampled on the surface of a torus in $\mathbb{R}^3$ (top left) and its offsets for different values of radii $r_1 < r_2 < r_3$. For well chosen values of the radius (e.g. r_1 and r_2), the offsets are clearly homotopy equivalent to a torus.

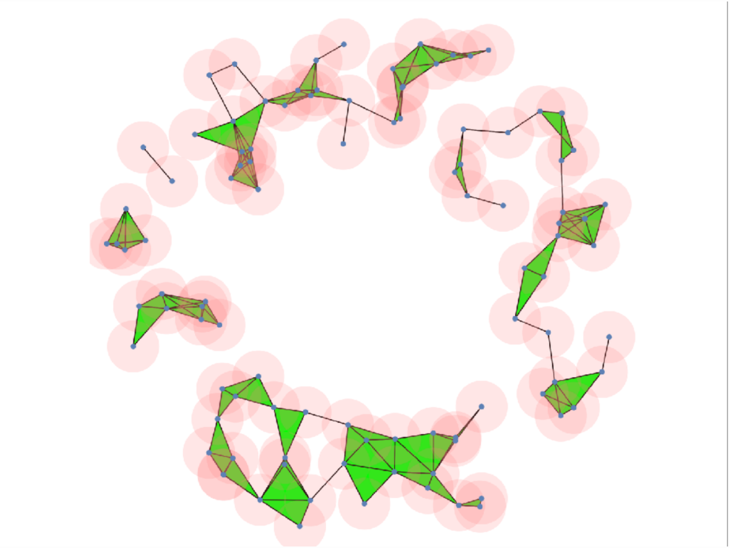
Another example



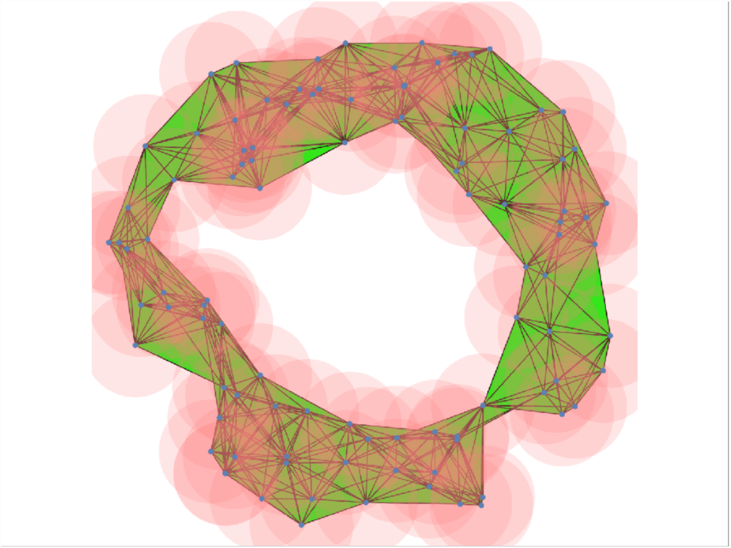
Another example, cont.



Another example, cont.



Another example, cont.

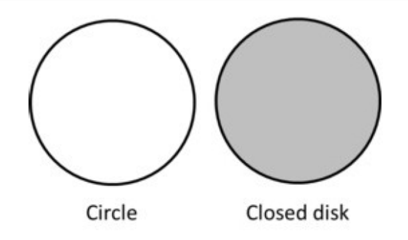


Homology

- We want a concise way of summarizing the topological characteristics of an object: homology provides a set of invariants which do just that
- Associates a set of groups (which will indeed be vector spaces for simplicial homology) to a topological space
- Does not uniquely identify a topological space: if X, Y are homotopy-equivalent, they have the same homology groups, but converse not necessarily true and certainly they are not necessarily homeomorphic (see link: pseudocircle)

Betti numbers

- The k-th **Betti number** of a topological space X is the dimension of its k-th homology group
- Roughly, β_0 corresponds to the number of connected components, β_1 to the number of punctures, β_2 to the number of "voids"...



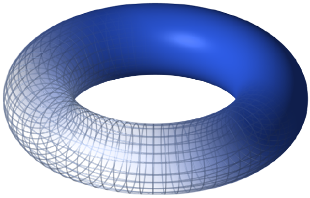


Figure 28: The torus has $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$.

Persistent homology

- Our primary issue remaining is that in general it is not obvious what the correct radius is for construction of our simplicial complex
- Persistent homology attempts to remedy this problem by highlighting the topological features which persist while growing the radii
- Use persistence diagrams: keeps track of increase/decrease of each Betti number, i.e. birth/death of features as radii increase

Toy example

- Can consider union of balls of radius r around $X \subset \mathbb{R}^n$ as sublevel set of the natural function $f_X : \mathbb{R}^n \to \mathbb{R}$, so let's look at persistence for a general function:

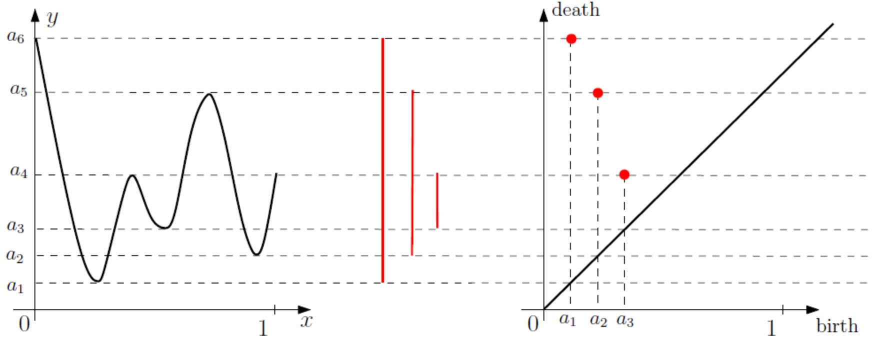
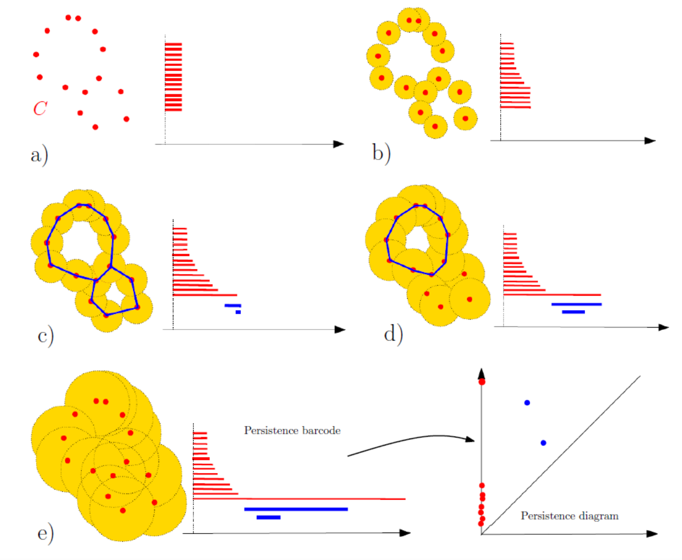


Figure 11: The persistence barcode and the persistence diagram of a function $f : [0, 1] \to \mathbb{R}$.

More complex example



Some good and bad things

- Persistence diagrams are fairly stable under certain perturbations of data, as desired from a topological learning method
- Care must be taken to deal with outliers – there are methods to mitigate this problem, but that is beyond the scope of this presentation

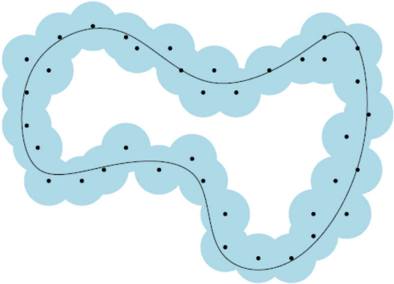


Figure 10: The effect of outliers on the sublevel sets of distance functions. Adding just a few outliers to a point cloud may dramatically change its distance function and the topology of its offsets.

Applications with machine learning

- TDA has found application in a number of fields, including biology, chemistry, sensor networks, shape analysis, materials science, and cosmology
- The method has done well with data which has some natural representation as a graph or complex, for example in genetics or cosmology, suggesting it may lend itself well to program analysis
- Often used with other learning methods, ex. an embedding of the initial data may be used to find the topological characteristics, or a CNN can be used to extract data from persistence diagrams

References I

[1] Chazal, F., and Michel, B. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists.

[2] Humphreys, D. P., McGuirl, M. R., Miyagi, M., and Blumberg, A. J. Fast Estimation of Recombination Rates Using Topological Data Analysis. *Genetics* (February 2019).

[3] Shiu, G. Topological Data Analysis for Cosmology and String Theory.

[4] So, G. Topological Data Analysis.

[5] Umeda, Y. Time Series Classification via Topological Data Analysis. *Transactions of the Japanese Society for Artifical Intelligence* (2017).