An Introduction to Topological Data Analysis

Michael Catanzaro

NCS MAA Fall 2018
Southwest Minnesota State University

Iowa State University

Outline

Crash course in algebraic topology

Persistent Homology and examples

Mapper and examples

Implementation and resources

Algebraic Topology

What is algebraic topology?

Topology is a branch of mathematics which is good at extracting global qualitative features from complicated geometric structures.

Algebraic topology provides a set of *algebraic* descriptors to topological objects.

What is algebraic topology?

Topology is a branch of mathematics which is good at extracting global qualitative features from complicated geometric structures.

Algebraic topology provides a set of *algebraic* descriptors to topological objects.

Questions and scope

Topological questions surround different notions of connectedness: connected components, loops, voids, etc.

Two topological spaces are equivalent through the lens of topology if one can be *continuously* deformed to the other.

A B C D

Algebraic Topology

What is algebraic topology?

Topology is a branch of mathematics which is good at extracting global qualitative features from complicated geometric structures.

Algebraic topology provides a set of *algebraic* descriptors to topological objects.

Questions and scope

Topological questions surround different notions of connectedness: connected components, loops, voids, etc.

Two topological spaces are equivalent through the lens of topology if one can be *continuously* deformed to the other.

A B C D

Invariants of topological spaces

- Algebraic Topology assigns *invariants* to topological spaces. These take the form of groups, rings, fields, vector spaces, etc.
- Our computations will be over the field $\mathbb{F}_2$, so it suffices to record only the *dimensions* of vector spaces.
- If two spaces are the same, then the invariants must be the same.
  If the invariants are not the same, then the two spaces are not the same.

Topological Data Analysis

Goal

Topological data analysis uses topology to summarize and study the 'shape' of data.

Topological Data Analysis

Goal

Topological data analysis uses topology to summarize and study the 'shape' of data.

Examples:

- motion tracking problems,
- analysis of brain arteries,
- analysis of social and spatial networks, including neurons, genens, Twitter, co-authorship,
- study of viral evolution,
- measurement of protein compressibility,
- analysis of phase transitions,
- financial crash analysis,
- piecewise constant signal analysis,
- study of cosmic web and its filamentary structure,
- identification of breast cancer subtypes,
- study of plant root systems,
- discrimination of EEG signals before and during epileptic seizures,
- steganalysis of images,
- sphere packing and colloid,
- population activity in the visual cortex,
- fMRI data

Overview of TDA

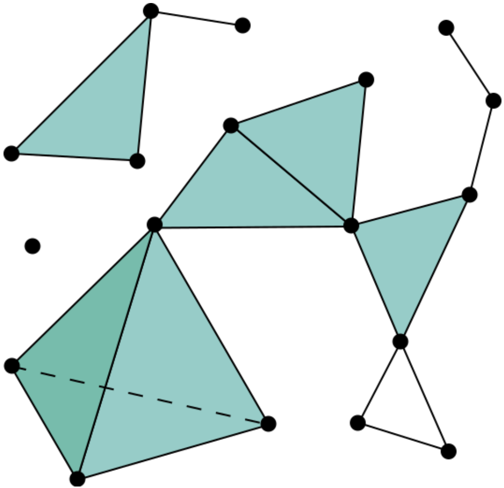
Topological data analysis comes in a variety of flavors.

The two most popular methods in TDA are

1. Persistent Homology
2. Mapper

Simplicial Complexes

A *simplicial complex* is a combinatorial object, generalizing the notion of a graph. Each simplicial complex is built out of *simplices* of varying dimensions.

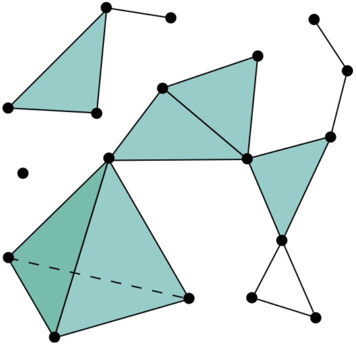


Betti numbers of simplicial complexes

$\beta_0 = \#$ of connected components

$\beta_1 = \#$ of holes

$\beta_2 = \#$ of voids



$\beta_0 \quad =$

$\beta_1 \quad =$

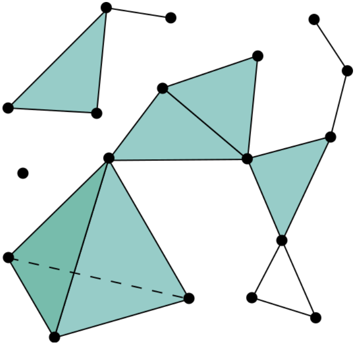
$\beta_2 \quad =$

Betti numbers of simplicial complexes

$\beta_0 = \#$ of connected components

$\beta_1 = \#$ of holes

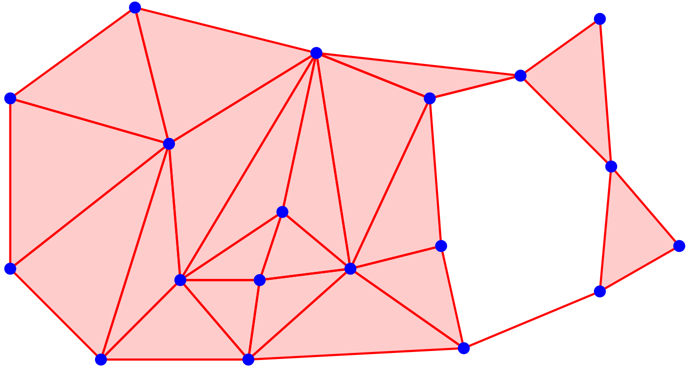
$\beta_2 = \#$ of voids



$$\beta_0 = 3$$
$$\beta_1 = 1$$
$$\beta_2 = 1$$

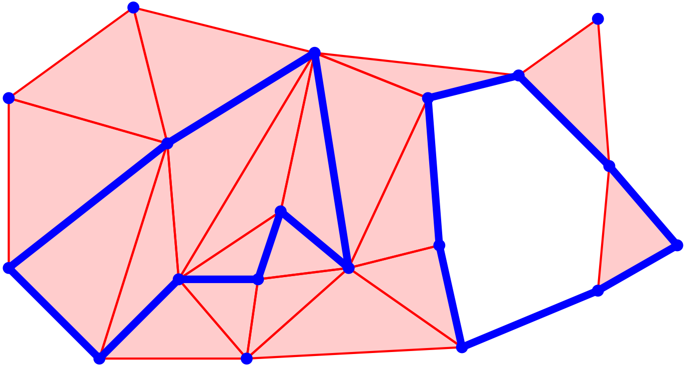
Definition

Homology in degree k is given by k-cycles modulo the k-boundaries.



Definition

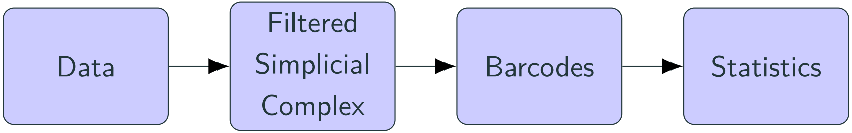
Homology in degree k is given by k-cycles modulo the k-boundaries.



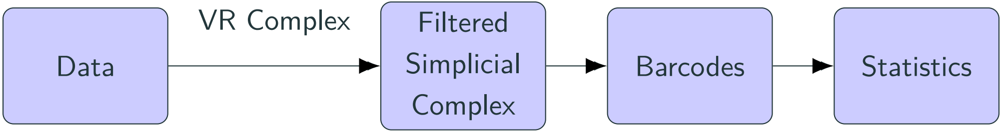
β_k = rank of homology in degree k

Overview of PH

Persistent homology consists of the following pipeline:



Overview of PH



Simplicial Complexes from Point data

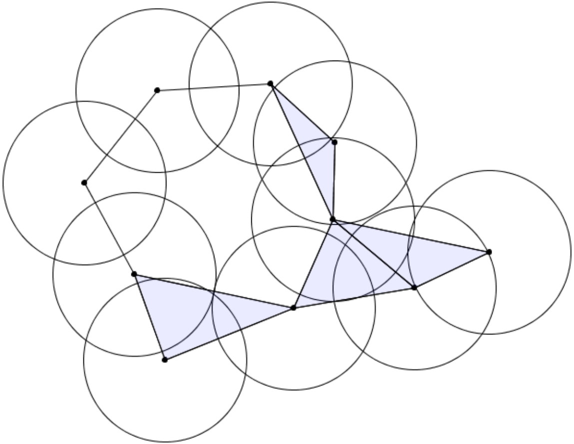
Definition

A *point cloud* P is a finite metric space.

Simplicial Complexes from Point data

Definition

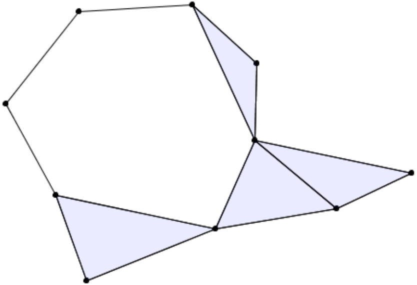
The Vietoris-Rips complex is a simplicial complex built out of a point cloud. Put a circle of radius r around each point. Add an edge whenever two circles overlap. Add a triangle whenever three circles overlap.



Simplicial Complexes from Point data

Definition

The Vietoris-Rips complex is a simplicial complex built out of a point cloud. Put a circle of radius r around each point. Add an edge whenever two circles overlap. Add a triangle whenever three circles overlap.

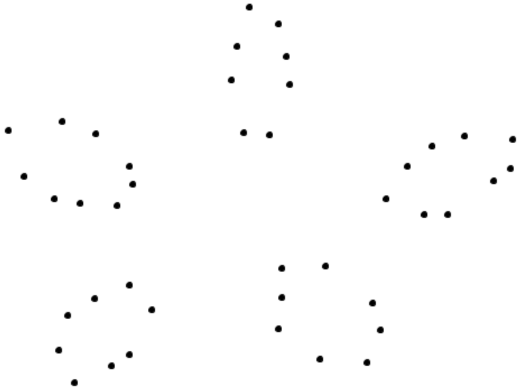


Vietoris-Rips parameter

Question

How do we choose the correct radius for the Vietoris-Rips construction?

Often, there is no one "right" choice.

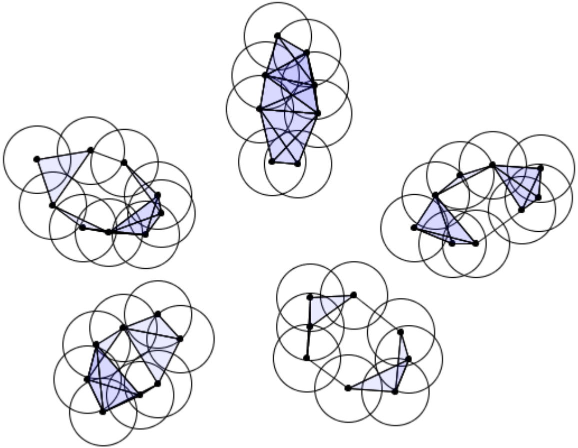


Vietoris-Rips parameter

Question

How do we choose the correct radius for the Vietoris-Rips construction?

Often, there is no one "right" choice.

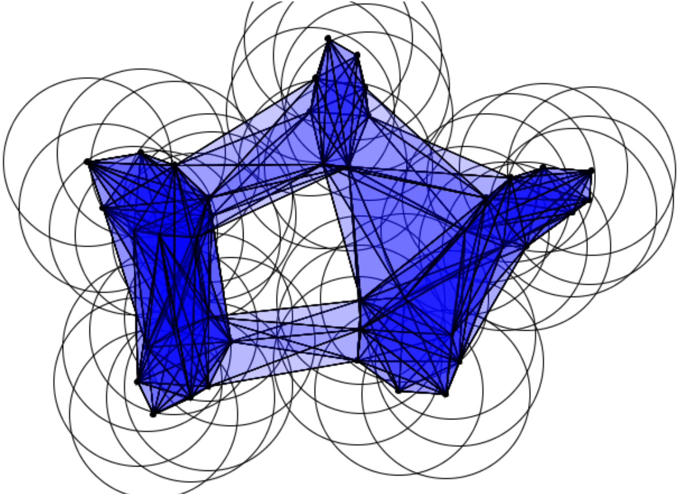


Vietoris-Rips parameter

Question

How do we choose the correct radius for the Vietoris-Rips construction?

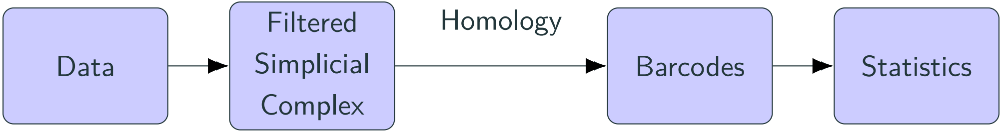
Often, there is no one “right” choice.



Processing demo

Processing demo

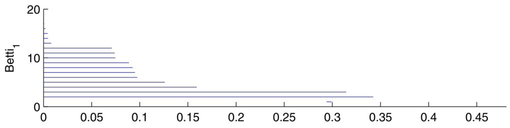
Overview of PH

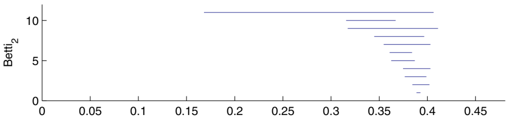


Barcodes

- The barcode provides a summary of how the homology changes as the radius varies in the Vietoris-Rips construction.
- We look for topological features which 'persist' over many values of radii.

Barcodes typically look like:

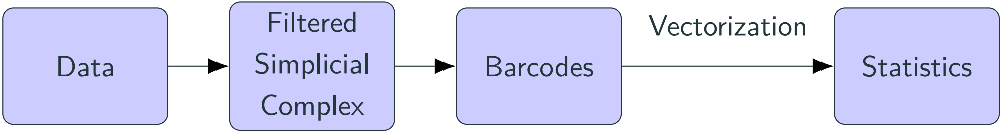




Processing demo

Processing demo

Overview of PH

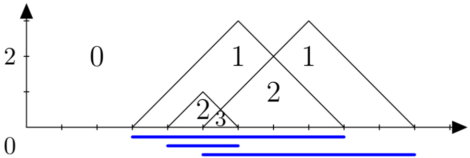


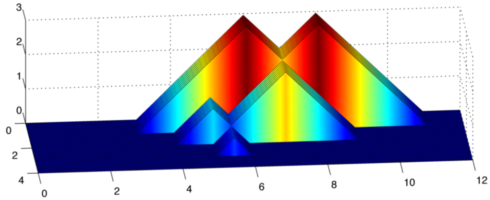
Vectorization

- The barcode provides a convenient visualization of persistent topological features of potentially high-dimensional data sets. With barcodes:
  - Clustering, certain hypothesis testing are easy,
  - Calculating averages, understanding variances, and classification are hard.
  - Reason: No good metric space structure on barcodes directly.
- We need a way of *vectorizing* the output. If we can map the barcodes into a vector space, we can add, take differences, averages, etc.
- We can implement more advanceed statistical methods, e.g., machine learning techniques like SVM.

Persistence Landscapes

Relatively simple, yet powerful method of vectorization.





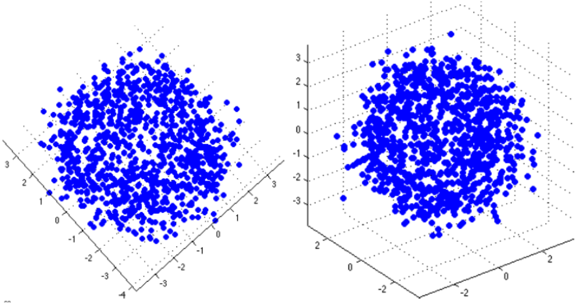
Each

$$\lambda_k : \mathbb{R} \to \mathbb{R} .$$

Functions can be added, subtracted, averaged, etc.

PH example: mathematical data

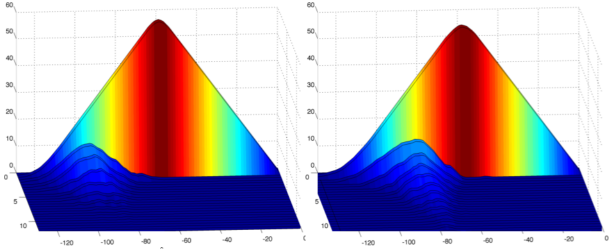
- We sample 1000 points from a noisy sphere and a noisy torus.



- Can we use persistent homology to distinguish these spaces?

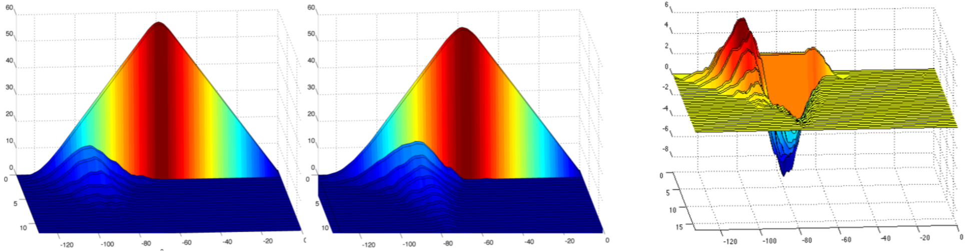
PH example: mathematical data

- Randomly choose 10 points from each space. Build the VR complex on those 10 points, compute β_0 barcodes, and build landscapes.
- Repeat this 10,000 times. Average all the sphere landscapes and average all the torus landscapes.



PH example: mathematical data

- Randomly choose 10 points from each space. Build the VR complex on those 10 points, compute β_0 barcodes, and build landscapes.
- Repeat this 10,000 times. Average all the sphere landscapes and average all the torus landscapes.



Doing a permutation test with 10,000 repititions gives a p-value of 0.0111!

Peter Bubenik, Statistical Topological Data Analysis using persistence landscapes, JoMLR, 2015.

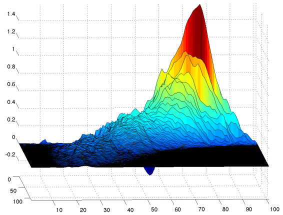
PH example: fMRI data

- An fMRI patient has a screen in front of them. They tap a pad every time a stimulus flashes on the screen. The stimuli flash both periodically, randomly for 200 seconds. There are also rest periods.
- We focus on a region of the brain known as the Anterior Cingulate Cortex (ACC).
- **Can persistent homology tell the difference between these periods based on the fMRI signal?**

PH example: fMRI data

- The fMRI machine treats the brain as a 3-dimensional grid, so the data is 5-dimensional: $(x, y, z, t, \mathrm{BOLD})$.
- For each time slice, compute the VR complex, and then the barcodes and landscapes.
- Average the periodic time periods, random time periods, and rest time periods.
- Doing a permutation test with 10,000 repetitions gives:

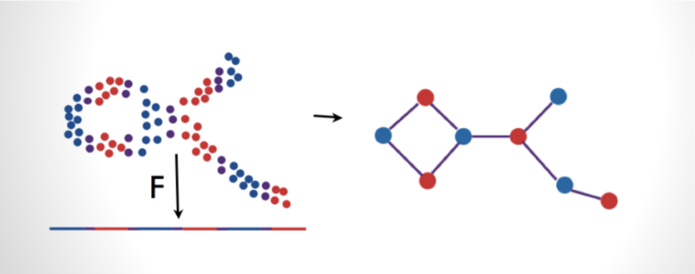
p-values for	Periodic-Random	Periodic-Rest	Random-Rest
H_0	0	0	0
H_1	.0007	.0007	0
H_2	0	.002	.1307



Mapper

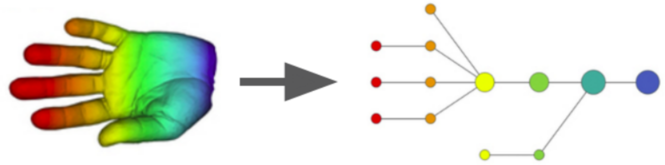
Originally developed by Carlsson and Singh, Mapper provides a different approach to classification of data.

1. Choose a 'filter' function on the point cloud $f : P \to \mathbb{R}$.
2. Cover $\mathbb{R}$ and pull back to cover the point cloud P using f.
3. Within each open set, run single-linkage clustering
4. Draw a node for each cluster. Connect two nodes from different covers with an edge if they share linked points.



Mapper properties

- Mapper provides a different form of visualization of high dimensional data compared to persistent homology.
- Complimentary method to persistent homology, as well other statistical methods.
- There are several parameters to be chosen. In particular, the filter function f needs to be chosen carefully!



Mapper examples: Breast cancer

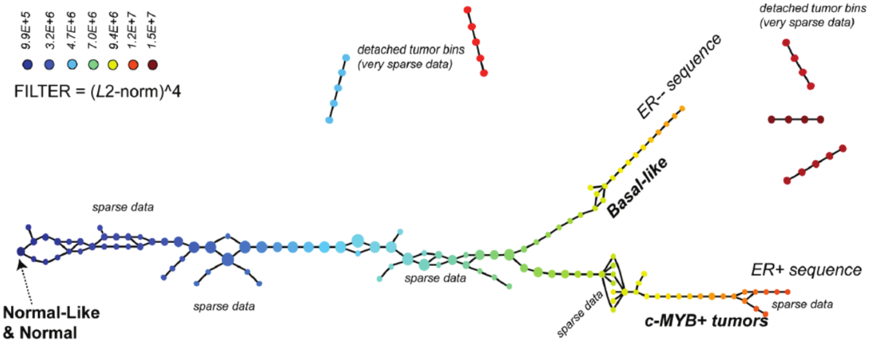


Diagram of gene expression profiles for breast cancer
M. Nicolau, A. Levine, and G. Carlsson, PNAS 2011

Algorithms

There are lots of software packages implementing the algorithms of persistent homology:

Persistent Homology:

- Javaplex
- Dionysus
- Perseus
- Ripser
- PHAT
- GUDHI
- CHOMP
- SimBa
- SimPers
- Eirene
- R-TDA

Vectorizations:

- Persistence landscapes
- Persistence images
- Persistence silhouettes

Mapper:

- Pymapper
- TDAmapper

References

General Overviews:

More technical introduction: