Topological Data Analysis

Michael Catanzaro

Midwest Big Data Summer School 2019

Iowa State University

`github.com/mjcatanz/MBDS19_TDA`

Outline

Quick course in algebraic topology

Persistent Homology and examples

Mapper and examples

Implementation and resources

Outline

Quick course in algebraic topology

Persistent Homology and examples

Mapper and examples

Implementation and resources

Topology

What is topology?

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

Topology is what remains from geometry after stripping away angles and distances.

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

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

[coffee cup vs doughnut]

Algebraic Topology

What is algebraic topology?

Algebraic Topology is the study of topological spaces through the lens of algebra.

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

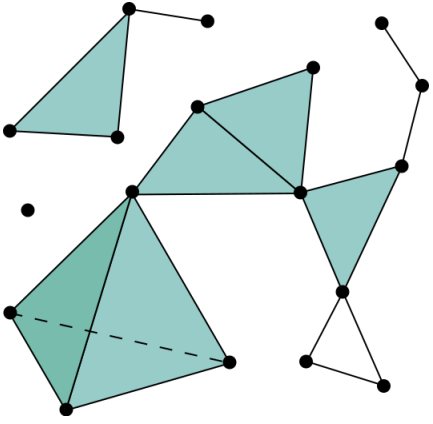
Invariants of topological spaces

- Algebraic Topology assigns *invariants* to topological spaces. These take the form of groups, rings, fields, vector spaces, etc.
- In applied algebraic topology, we assign the simplest invariants: **a list of numbers**.
- If two spaces are the 'same', then the list of numbers must be the same.
  If the list of numbers are not the same, then the two spaces are not the 'same'.

Simplicial Complexes

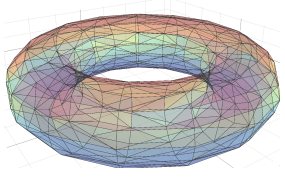
We focus our attention on *simplicial complexes*, a simple class of topological spaces.

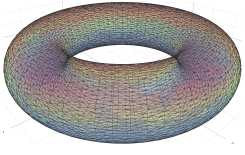
A simplicial complex is a combinatorial object, generalizing the notion of a network or graph. Each simplicial complex is built out of *simplices* of varying dimensions.

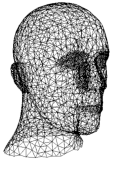


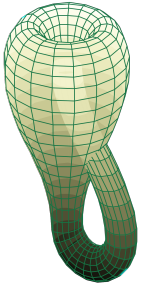
Simplicial Complexes

One way of obtaining a simplicial complex is to triangulate a surface.









Mederos, Velho, & Figueiredo. Smooth surface reconstruction from noisy clouds. Journal of the Brazilian Computer Society, 9(3), 52-66. https://dx.doi.org/10.1590/S0104-65002004000100005

Betti numbers of simplicial complexes

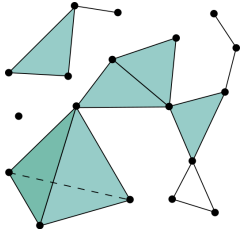
The list of numbers we assign to a topological space consists of the following.

$$\beta_0 = \#\text{ of connected components}$$
$$\beta_1 = \#\text{ of holes}$$
$$\beta_2 = \#\text{ of voids}$$
$$\vdots \qquad\qquad \vdots$$
$$\beta_k = \#\text{ of k-dimensional holes}$$

Betti numbers of simplicial complexes

The list of numbers we assign to a topological space consists of the following.

$$\beta_0 = \#\text{ of connected components}$$
$$\beta_1 = \#\text{ of holes}$$
$$\beta_2 = \#\text{ of voids}$$
$$\vdots \qquad\qquad \vdots$$
$$\beta_k = \#\text{ of k-dimensional holes}$$

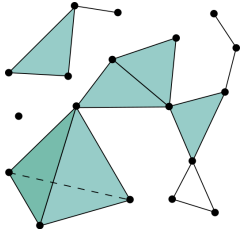


$$\beta_0 =$$
$$\beta_1 =$$
$$\beta_2 =$$

Betti numbers of simplicial complexes

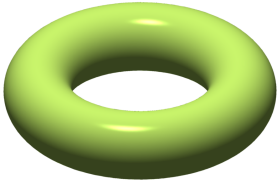
The list of numbers we assign to a topological space consists of the following.

$$\beta_0 = \#\text{ of connected components}$$
$$\beta_1 = \#\text{ of holes}$$
$$\beta_2 = \#\text{ of voids}$$
$$\vdots \qquad\qquad \vdots$$
$$\beta_k = \#\text{ of k-dimensional holes}$$



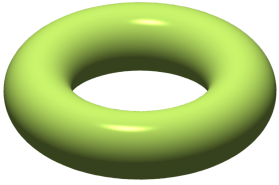
$$\begin{aligned} \beta_0 &= 3 \\ \beta_1 &= 1 \\ \beta_2 &= 1 \end{aligned}$$

More Betti numbers

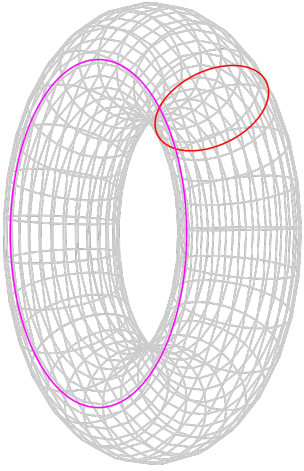


$$\beta_0 =$$
$$\beta_1 =$$
$$\beta_2 =$$

More Betti numbers



$$\beta_0 = 1$$
$$\beta_1 = 2$$
$$\beta_2 = 1$$



Topological Data Analysis

Main Idea

Given a data set, build a topological space in such a way that the topological properties of the constructed space reflect the geometry/statistics of the data.

Topological Data Analysis

Main Idea

Given a data set, build a topological space in such a way that the topological properties of the constructed space reflect the geometry/statistics of the data.

Slogan

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

Topological Data Analysis

Main Idea

Given a data set, build a topological space in such a way that the topological properties of the constructed space reflect the geometry/statistics of the data.

Slogan

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

Examples:

- analysis of brain arteries,
- analysis of social and spatial networks, including neuronal networks, Twitter, co-authorship,
- study of viral evolution,
- measurement of protein compressibility,
- financial crash analysis,
- identification of breast cancer subtypes,
- study of plant root systems,
- discrimination of EEG signals before and during epileptic seizures,
- population activity in the visual cortex,
- fMRI data

Why TDA?

TDA provides a set of tools with very useful features.

- TDA provides a **multiscale summary of data**, encoding geometric and topological features of data. Practitioners need not make an a priori choice of scale (micro to macro at once).

Why TDA?

TDA provides a set of tools with very useful features.

- TDA provides a **multiscale summary of data**, encoding geometric and topological features of data. Practitioners need not make an a priori choice of scale (micro to macro at once).
- The output of TDA is **robust with respect to noise**. Small perturbations of input data yield small changes in the output descriptors.

Why TDA?

TDA provides a set of tools with very useful features.

- TDA provides a **multiscale summary of data**, encoding geometric and topological features of data. Practitioners need not make an a priori choice of scale (micro to macro at once).
- The output of TDA is **robust with respect to noise**. Small perturbations of input data yield small changes in the output descriptors.
- Topological methods are coordinate-free and non-parametric. This allows us to study data intrinsically without worrying about parameter tuning.
- It is efficiently computable, especially with recent advances in algorithms.

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

Outline

Simplicial Complexes from Point data

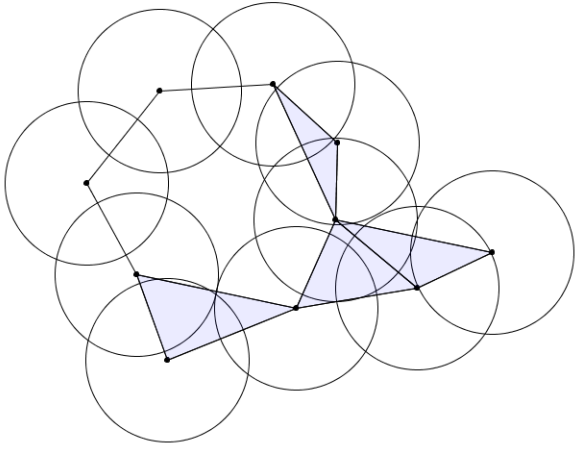
Definition

A *point cloud* P is a finite data set in some Euclidean space $\mathbb{R}^n$.

Simplicial Complexes from Point data

Definition

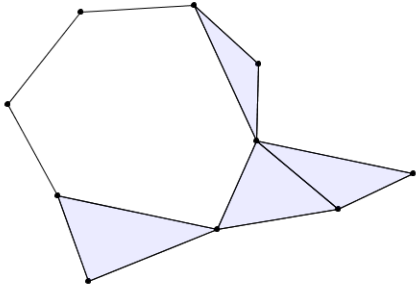
The Vietoris-Rips complex is a simplicial complex built out of a point cloud. Put a circle of radius r around each point. Add an edge whenever two circles overlap. Add a triangle whenever three circles overlap, and so on.



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, and so on.

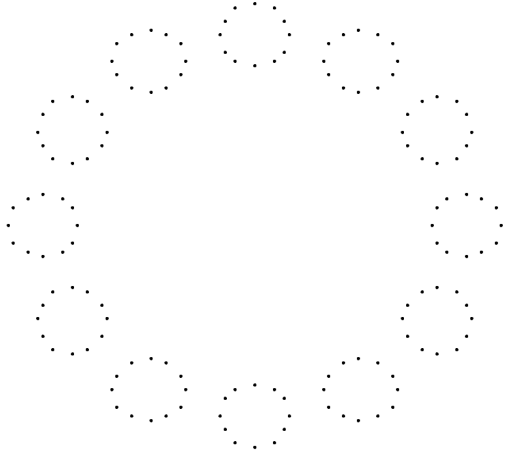


Vietoris-Rips parameter

Question

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

Often, there is no one “right” choice.

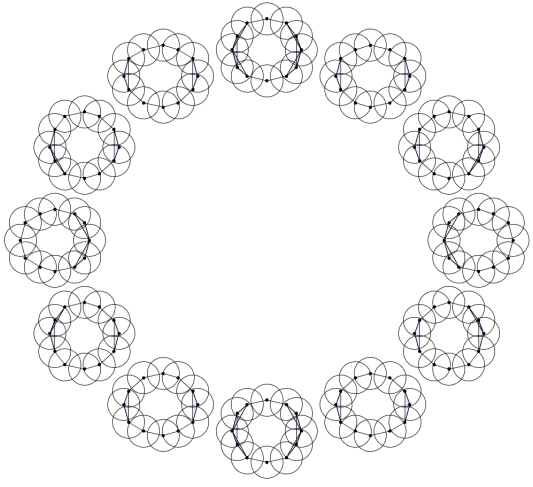


Vietoris-Rips parameter

Question

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

Often, there is no one “right” choice.

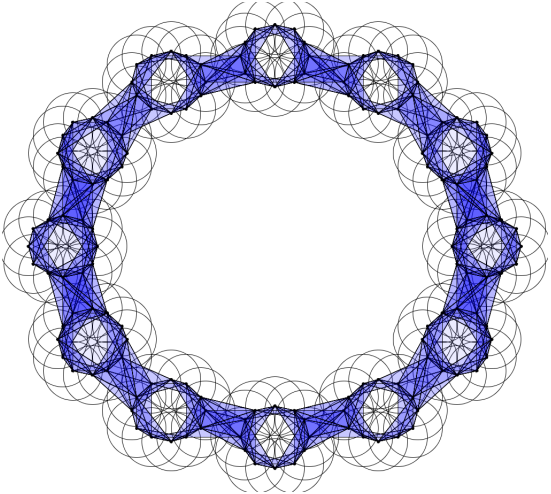


Vietoris-Rips parameter

Question

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

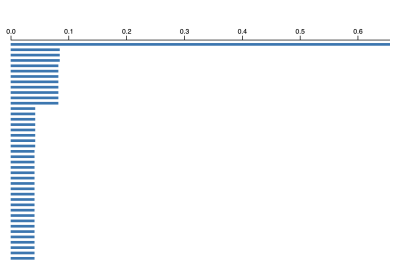
Often, there is no one "right" choice.

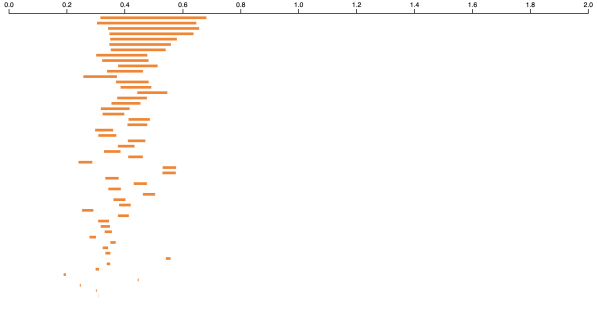


Barcodes

- The output of persistent homology is known as a **barcode**.
- 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

Examples

- Let's apply these tools to some data sets.
- Go to `github.com/mjcatanz/MBDS19_TDA` and clone the repo.
- To perform the computations, go to `live.ripser.org`.

Examples

- We'll start with some mathematical examples.
- Select 'point cloud data' from the dropdown menu, and set the dimensions to 0 to 2, and distance to 2.
- Load the first three synthetic data sets from the `data` directory
  - `synth_data1.txt`,
  - `synth_data2.txt`,
  - `synth_data3.txt`
- What can you say about these spaces?
- With the max distance set to 1, compute the persistent homology of `synth_data4.txt`.

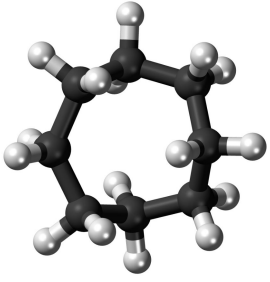
Examples

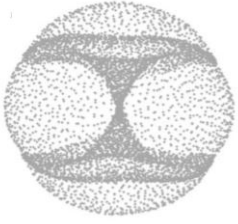
- Set the max distance to 1.3, and compute the persistent homology of `synth_data5.txt`.

Examples

- Set the max distance to 1.3, and compute the persistent homology of `synth_data5.txt`.

This data set is the set of conformations (physically possible arrangement modulo rotations and translations) of cyclooctane. It has 8 carbon atoms, each bonded to a pair of hydrogen atoms. Each conformation gives a point in $\mathbb{R}^{24}$.

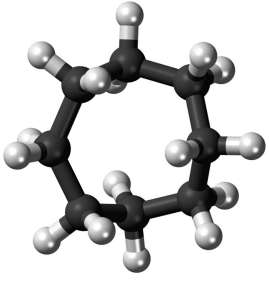


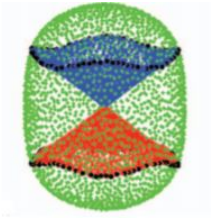


Examples

- Set the max distance to 1.3, and compute the persistent homology of `synth_data5.txt`.

This data set is the set of conformations (physically possible arrangement modulo rotations and translations) of cyclooctane. It has 8 carbon atoms, each bonded to a pair of hydrogen atoms. Each conformation gives a point in $\mathbb{R}^{24}$.

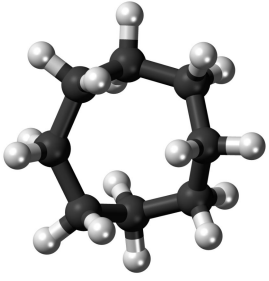


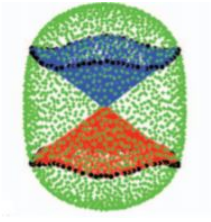


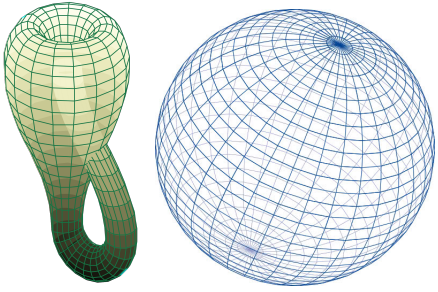
Examples

- Set the max distance to 1.3, and compute the persistent homology of `synth_data5.txt`.

This data set is the set of conformations (physically possible arrangement modulo rotations and translations) of cyclooctane. It has 8 carbon atoms, each bonded to a pair of hydrogen atoms. Each conformation gives a point in $\mathbb{R}^{24}$.







Outline

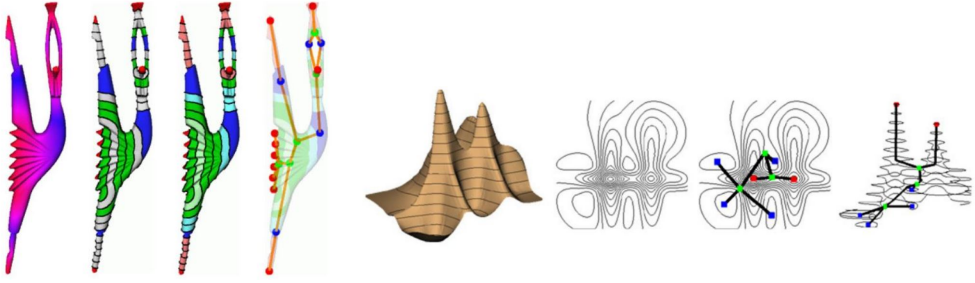
Mapper

Mapper is motivated by practical motivations:

- Imaging, shape analysis, animation, surface parameterizations.
- Computer graphics.

Idea

Provide a **quantitative** representation/visualization of a data set instead of a **qualitative** one.



Mapper

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

1. Choose a 'filter' function on the point cloud $f : P \to \mathbb{R}$.

Mapper

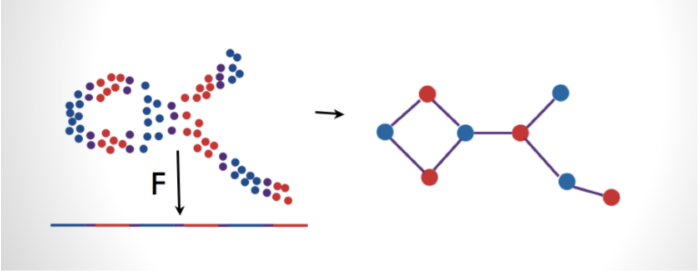
Originally developed by Carlsson, Singh, and Memoli, Mapper provides a different approach to classification of data.

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 a clustering method.

Mapper

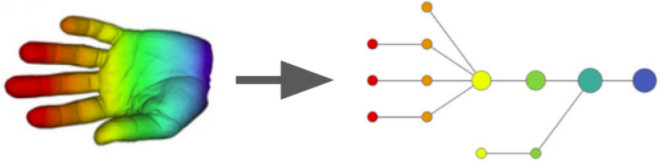
Originally developed by Carlsson, Singh, and Memoli, Mapper provides a different approach to classification of data.

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 a clustering method.
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

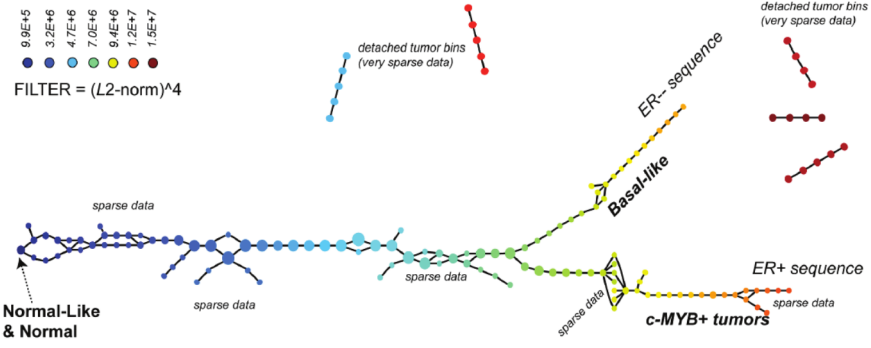


Diagram of gene expression profiles for breast cancer
M. Nicolau, A. Levine, and G. Carlsson, PNAS 2011

Mapper examples

Let's apply Mapper to some data sets.

Outline

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:

- Kepler-Mapper
- Pymapper
- TDAmapper

References

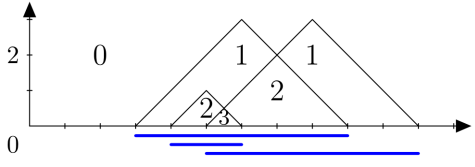
- Peter Bubenik. "Statistical topological data analysis using persistence landscapes". *J. Mach. Lear. R.* 16.1 (2015).
- Gunnar Carlsson. "Topology and data". *Bull. Amer. Math. Soc.* 46.2 (2009), pp. 255–308.
- Robert Ghrist. "Homological algebra and data". (2017). URL: https://www.math.upenn.edu/~ghrist/preprints/HAD.pdf.
- Steve Y Oudot. *Persistence theory: from quiver representations to data analysis*. Vol. 209. Amer. Math. Soc. Providence, RI, 2015.
- Jose A. Perea. "A Brief History of Persistence". (2018). URL: http://arxiv.org/abs/1809.03624.
- Matthew Wright. "Introduction to Persistent Homology - YouTube". (2016). URL: https://www.youtube.com/watch?v=h0bnG1Wavag.

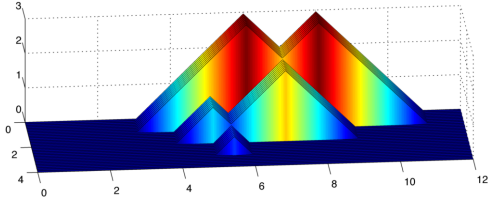
Vectorization

- 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 advanced statistical methods, e.g., machine learning techniques like SVM.

Persistence Landscapes

Relatively simple, yet powerful method of vectorization.





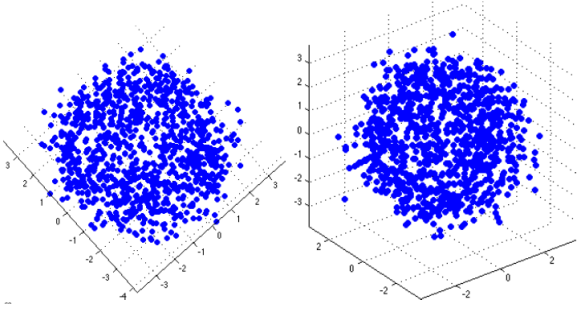
For every $k \geq 1$,

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

Functions can be added, subtracted, averaged, etc.

PH example: mathematical data

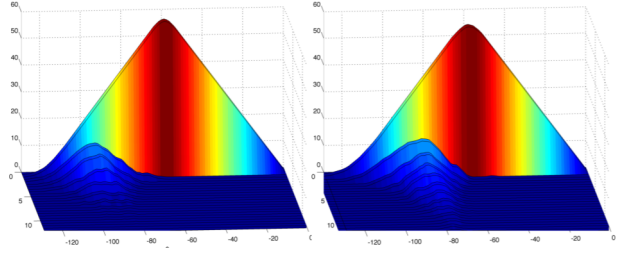
- We sample 1000 points from a noisy sphere and a noisy torus.



- Can we use persistent homology to distinguish these spaces?

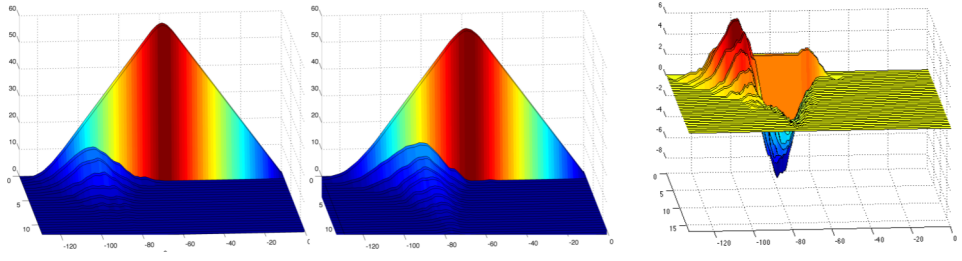
PH example: mathematical data

- 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 repetitions gives a p-value of 0.0111!

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

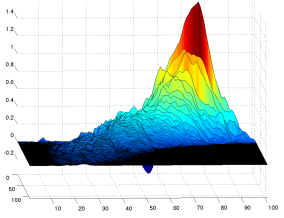
PH example: fMRI data

- 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:

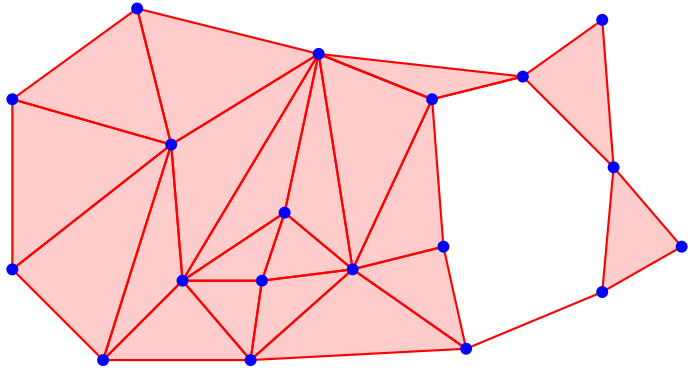
p-values for	Periodic-Random	Periodic-Rest	Random-Rest
H_0	0	0	0
H_1	.0007	.0007	0
H_2	0	.002	.1307



Homology of simplicial complexes

Definition

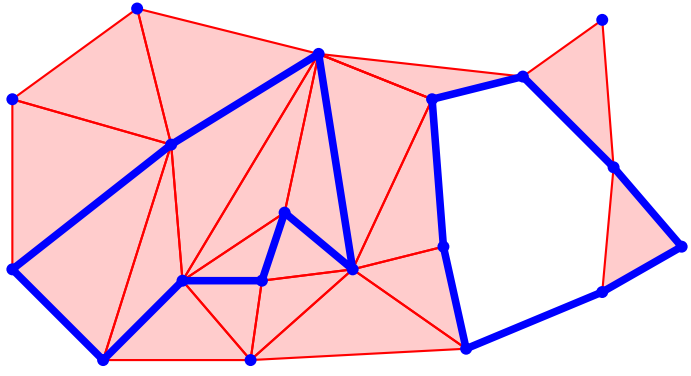
Homology in degree k is given by k-cycles modulo the k-boundaries.



Homology of simplicial complexes

Definition

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



β_k = rank of homology in degree k