A Practical Introduction to Topological Data Analysis

Michael Catanzaro

Midwest Big Data Summer School 2021

Iowa State University

`github.com/catanzaromj/MBDS21_TDA`

First things first

If you want to participate in the interactive part of this talk, now would be a good time to

- Clone the repo at `github.com/catanzaromj/MBDS21_TDA`
- Install scikit-tda and giotto-tda via

  `pip install scikit-tda, giotto-tda`

  and all of their dependencies.

Outline

Quick course in algebraic topology

Persistent Homology and examples

Mapper and examples

Implementation and resources

Outline

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.

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.

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.

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.

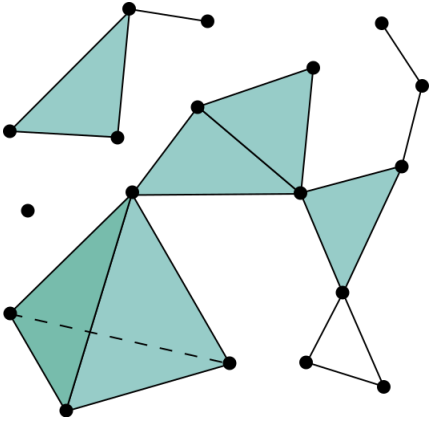
Invariants of topological spaces

- 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 topological 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 topological spaces are not the 'same'.

Combintorial Topology

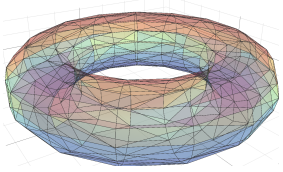
We focus our attention on combinatorial spaces, e.g., *simplicial complexes*, a simple class of topological spaces.

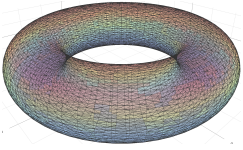
A simplicial complex is a combinatorial object, generalizing the notion of a network or graph. Each simplicial complex is built out of *simplices* of varying dimensions.

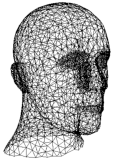


Simplicial Complexes

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



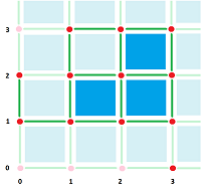


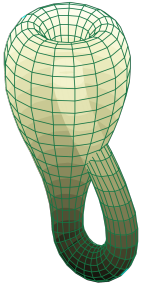


Mederos, Velho, & Figueiredo. Smooth surface reconstruction from noisy clouds. Journal of the Brazilian Computer Society, 9(3), 52-66. https://dx.doi.org/10.1590/S0104-65002004000100005

Cubical Complexes

A cubical complex is a similar generalization of a graph. Instead of using triangles and tetrahedra in higher dimensions, we use squares and cubes (and hypercubes).





Cubical complexes are common in all types of imaging data: pixels on a screen, voxels of a region in a medical scan, etc.

https://calculus123.com/index.php?curid=2514

Betti numbers of simplicial complexes

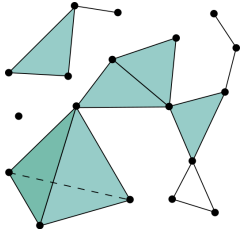
The list of numbers we assign to a topological space consists of the following.

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

Betti numbers of simplicial complexes

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

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

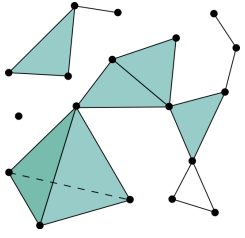


$$\beta_0 =$$
$$\beta_1 =$$
$$\beta_2 =$$

Betti numbers of simplicial complexes

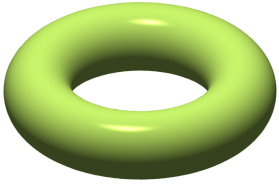
The list of numbers we assign to a topological space consists of the following.

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



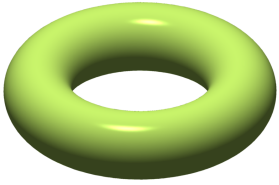
$$\begin{aligned} \beta_0 &= 3 \\ \beta_1 &= 1 \\ \beta_2 &= 1 \end{aligned}$$

More Betti numbers

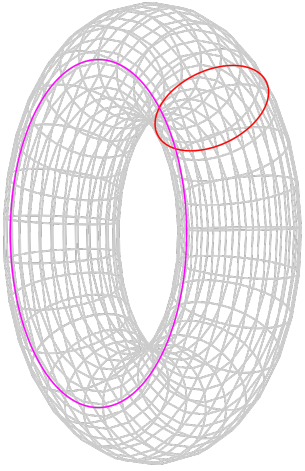


$$\beta_0 =$$
$$\beta_1 =$$
$$\beta_2 =$$

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.

Examples

Examples where TDA has proven useful:

- motion tracking problems,
- analysis of brain arteries,
- analysis of social and spatial networks, including neuronal networks, 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 packings,
- 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

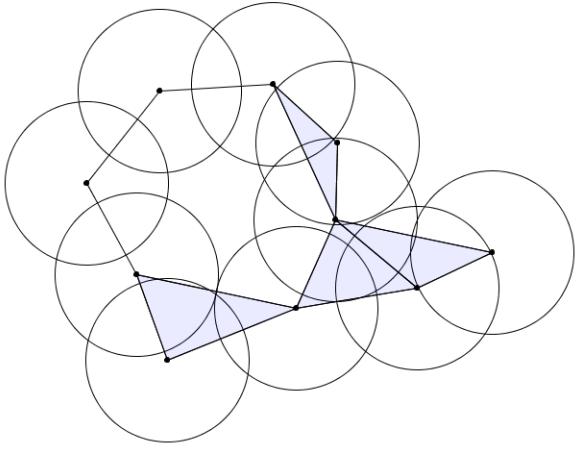
Definition

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

Simplicial Complexes from Point data

Definition

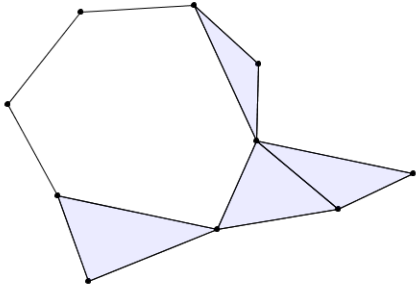
The Cech-Vietoris-Rips complex is a simplicial complex built out of a point cloud. Put a circle of radius r around each point. Add an edge whenever two circles overlap. Add a triangle whenever three circles overlap, and so on.



Simplicial Complexes from Point data

Definition

The Cech-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.

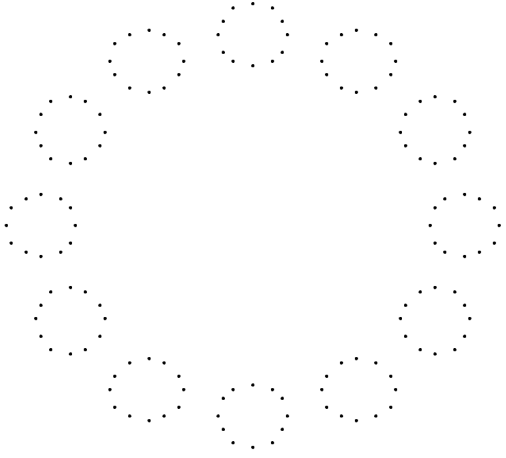


Vietoris-Rips parameter

Question

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

Often, there is no one “right” choice.

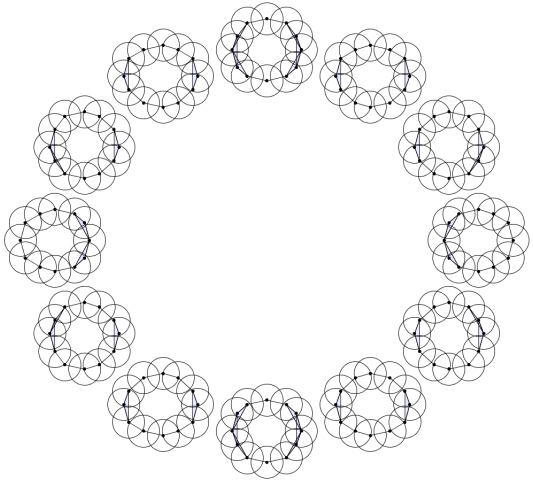


Vietoris-Rips parameter

Question

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

Often, there is no one “right” choice.

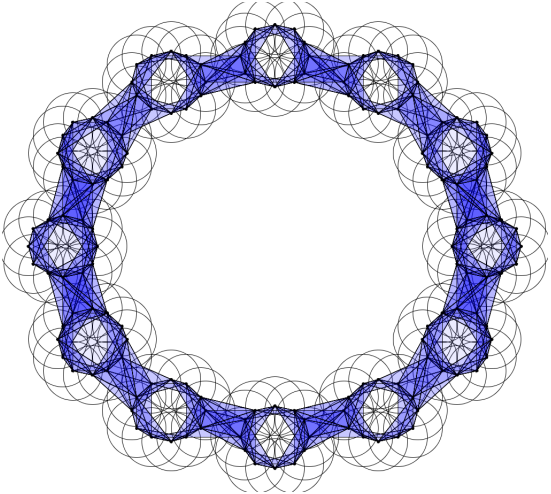


Vietoris-Rips parameter

Question

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

Often, there is no one "right" choice.



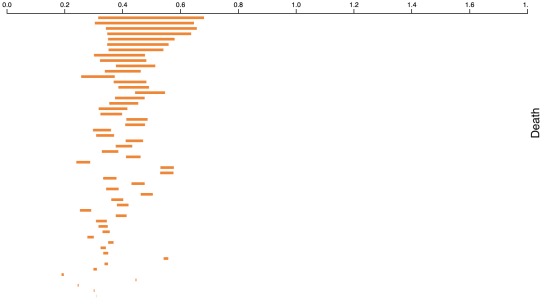
Other filtrations

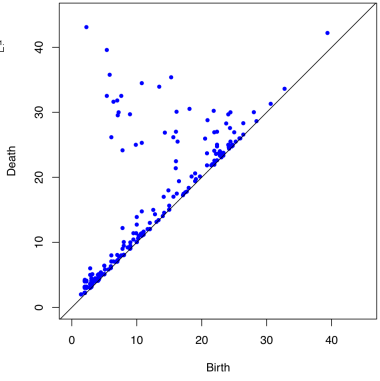
Persistent homology can be performed on any *filtration* of topological spaces: a nested sequence of spaces, each contained in the next.

The idea of persistence is to track how topological features vary in a family.

Barcodes

- The output of persistent homology is known as a **barcode** or a **persistence diagram**.
- Barcodes and persistence diagrams are completely equivalent encodings of the same information.
- They provide summaries of how the homology changes as the radius varies in the Vietoris-Rips construction.





Processing demo

Processing demo

Examples

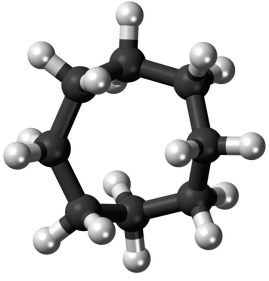
- Let's apply these tools to some data sets.
- Feel free to clone the repo at `github.com/catanzaromj/MBDS21_TDA`.
- If you don't want to install anything, you can upload the data sets to `live.ripser.org`.

Examples

- We'll start with some mathematical examples.
- First work through `Intro_to_PH.ipynb`.

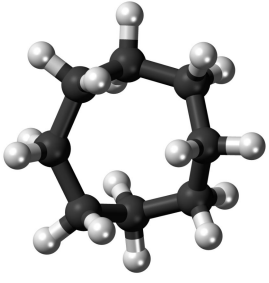
Cyclooctane

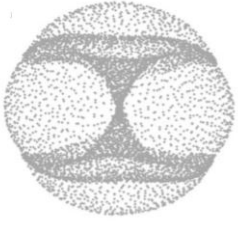
- Cyclooctane is a molecule with 8 carbon atoms, each bonded to a pair of Hydrogen atoms.
- `data/data6.txt` is a data set consisting of sampling the set of conformations (physically possible arrangement modulo rotations and translations) of cyclooctane. Each conformation gives a point in $\mathbb{R}^{24}$.



Cyclooctane

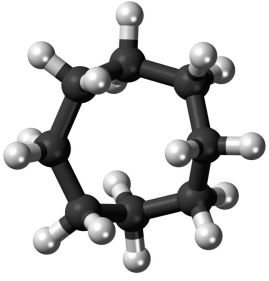
- Cyclooctane is a molecule with 8 carbon atoms, each bonded to a pair of Hydrogen atoms.
- `data/data6.txt` is a data set consisting of sampling the set of conformations (physically possible arrangement modulo rotations and translations) of cyclooctane. Each conformation gives a point in $\mathbb{R}^{24}$.

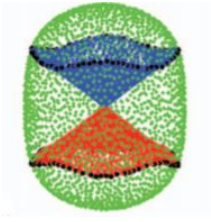




Cyclooctane

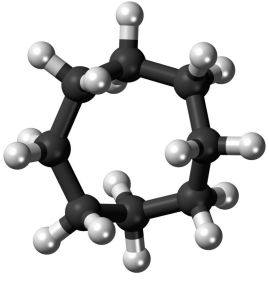
- Cyclooctane is a molecule with 8 carbon atoms, each bonded to a pair of Hydrogen atoms.
- `data/data6.txt` is a data set consisting of sampling the set of conformations (physically possible arrangement modulo rotations and translations) of cyclooctane. Each conformation gives a point in $\mathbb{R}^{24}$.

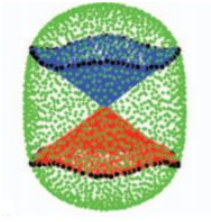


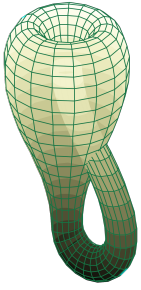


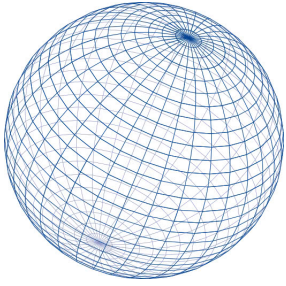
Cyclooctane

- Cyclooctane is a molecule with 8 carbon atoms, each bonded to a pair of Hydrogen atoms.
- `data/data6.txt` is a data set consisting of sampling the set of conformations (physically possible arrangement modulo rotations and translations) of cyclooctane. Each conformation gives a point in $\mathbb{R}^{24}$.









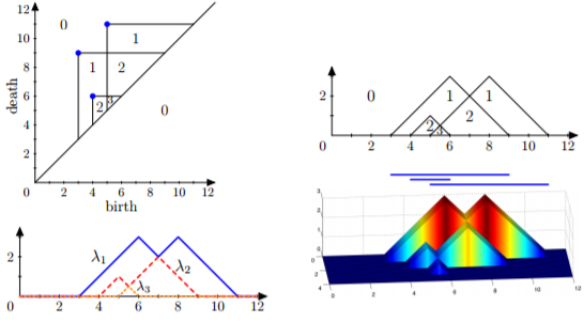
Vectorizations

- Barcodes and Persistence Diagrams provide a convenient visualization of persistent topological features of potentially high-dimensional data sets. With these:
  - 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, Random Forests, etc.
- Luckily there are several popular ways to do this.

Persistence Landscapes

Input: Persistence diagram.

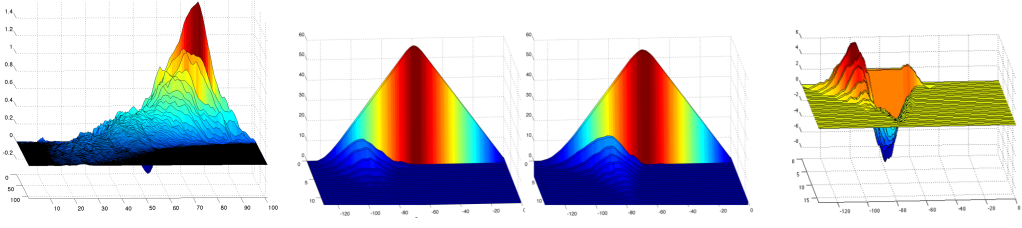
Output: A sequence of functions $\lambda_k : \mathbb{R} \to \mathbb{R}$ for each $k \geq 1$.



Persistence Landscapes

Persistence Landscapes allow one to perform statistical tests easily.

- Addition: add each layer as functions $\lambda_k + \lambda'_k$.
- Scalar multiplication applies to each layer: $c \cdot \{\lambda_k\} = \{c \cdot \lambda_k\}$.



Any vector operation can be performed now: inner products, norms, etc.

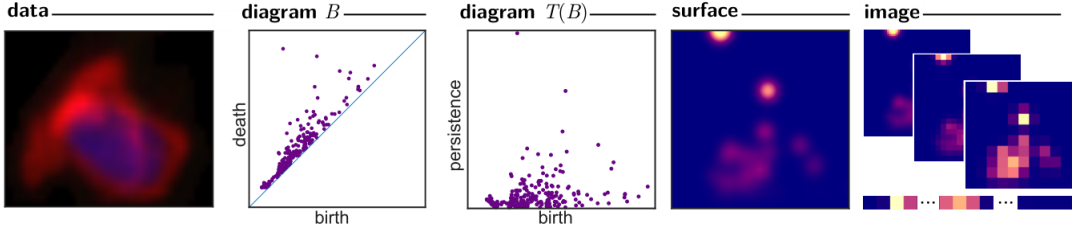
Let's see such a statistical test with landscapes now. Check out the `Differentiation_with_Persistence_Landscapes` notebook.

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

Persistence Images

Input: A persistence diagram, Gaussian parameters (height, width), weighting function, and a pixel size.

Output: A real number to each point in the grid determined by pixel size.



Again, any vector operation can be performed (sums, scalar multiplication, norms, etc.), making comparisons quantifiable.

Adams et. al., Persistence Images: A Stable Vector Representation of Persistent Homology, JoMLR, 2017.

Outline

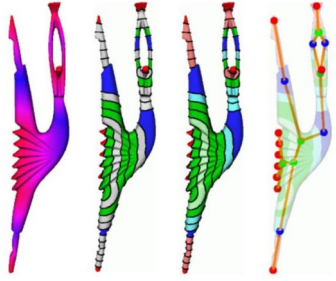
Mapper

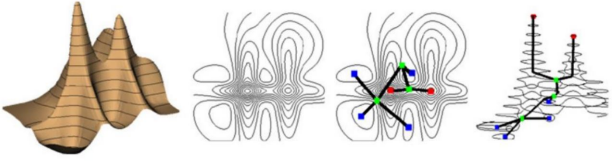
Mapper is motivated by practical motivations:

- 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

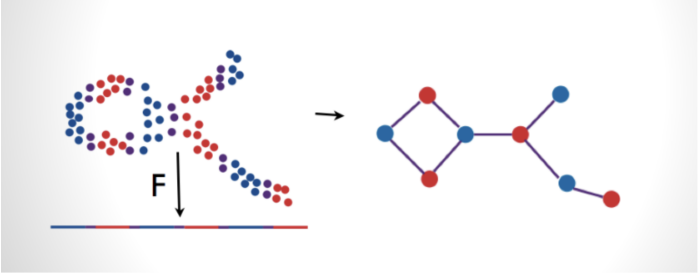
Originally developed by Carlsson, Singh, and Memoli, Mapper provides a different approach to classification of data.

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

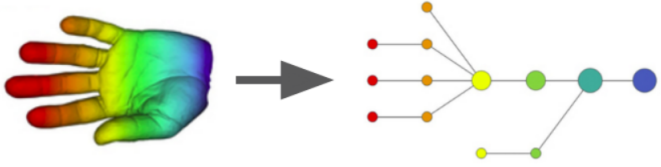
Originally developed by Carlsson, Singh, and Memoli, Mapper provides a different approach to classification of data.

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.
- Complementary 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

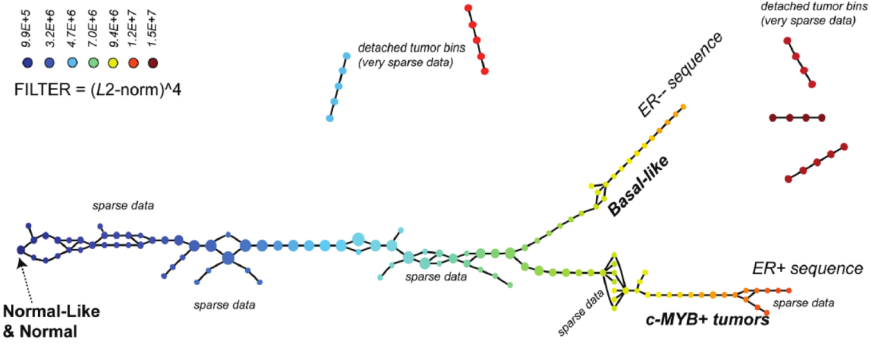


Diagram of gene expression profiles for breast cancer
M. Nicolau, A. Levine, and G. Carlsson, PNAS 2011

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
- scikit-tda
- giotto-tda

Vectorizations:

- Persistence landscapes
- Persistence images
- Persistence silhouettes
- Kernels

Mapper:

- Kepler Mapper
- Pymapper
- TDAmapper

References

[1] Peter Bubenik. "Statistical topological data analysis using persistence landscapes". *J. Mach. Lear. R.* 16.1 (2015).

[2] Gunnar Carlsson. "Topology and data". *Bull. Amer. Math. Soc.* 46.2 (2009), pp. 255–308.

[3] Robert Ghrist. "Homological algebra and data". (2017). URL: https://www.math.upenn.edu/~ghrist/preprints/HAD.pdf.

[4] Steve Y Oudot. *Persistence theory: from quiver representations to data analysis*. Vol. 209. Amer. Math. Soc. Providence, RI, 2015.

[5] Jose A. Perea. "A Brief History of Persistence". (2018). URL: http://arxiv.org/abs/1809.03624.

[6] Matthew Wright. "Introduction to Persistent Homology - YouTube". (2016). URL: https://www.youtube.com/watch?v=h0bnG1Wavag.