Informal Introduction to Topological Data Analysis

Christophe Ange Napoléon Biscio

Overview

Motivation:

- Topological Data Analysis (TDA) is a new field at the intersection of several mathematical fields.
- Various approaches depending on your scientific field.
- Many new concepts: Topology, Homology, Persistence, Quiver, cycle, Reeb graph, mapper, Morse Theory ...
- Require background in field traditionally unknown by most statisticians.

Aim of this talk:

- To provide the basic concepts and vocabulary appearing in TDA.
- To provide an introduction to Anne Marie's talk.

TDA – History and Showcases

Where is TDA coming from ?

The theory and main objects can be trace back to the 90s:

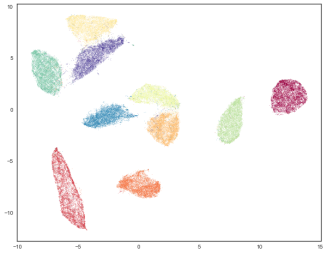
- Early concept of persistence, Frosini (1992)
- New descriptor: Persistent Betti numbers, Robins (1999)
- The currently most used object: Persistence Diagram, Edelsbrunner et al. (2000)

Other approaches have been developed since then

- Mapper: Singh et al. (2007)
- UMAP – Uniform Manifold APproximation: McInnes et al. (2018)

All of them have been used for different applications.

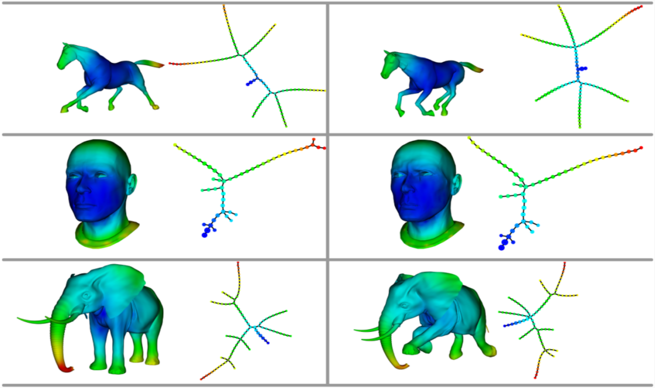
Applications – Dimension reduction



Encoding of each image:

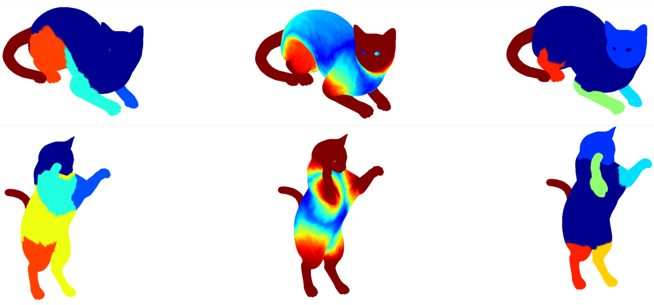
- a vector of dimension 28x28 ”the number of pixels”.
- The value on each coordinate is the gray level of the pixel.

Applications – Shape Classification



Motivation: Since topology is invariant by continuous deformation, then the same shape in different position will be well identified, Singh et al. (2007).

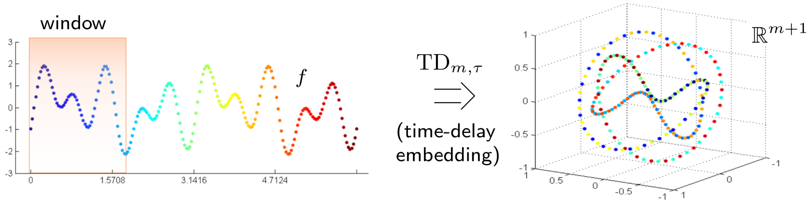
Applications – Image Segmentation



Skraba et al. (2010)

Applications – Shape of Data

Data have shapes and their topology can be used as a descriptor.



TDA – What is topology?

Definition (Topology)

A topological space is a set E equipped with a family of subsets $\mathcal{O}$ such that, and such that

- $\emptyset, E \in \mathcal{O}$,
- $\mathcal{O}$ is stable under union,
- $\mathcal{O}$ is stable under finite intersection.

Warning: This is the first error actually.

In TDA: we do algebraic topology, not general topology.

- Roughly, this is the study of invariant quantities via continuous deformation of a shape, i.e. no tearing.
- Example: from an ”algebraic topologist” point of view, a mug is a donut.

What are the topological features? (informally)

Topogical features: Every feature that is invariant under continuous deformation.

- The (path) connected components – 0-dimensional features
- The loops – 1-dimensional features
- The voids – 2-dimensional features
- In higher dimension, n-dimensional holes.

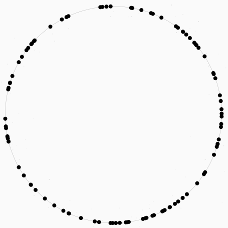
Example: The torus has

- 1 connected component,
- 1 loop,
- 1 void – the inside of the donut.

TDA – Persistent Homology – What is it?

Original motivation

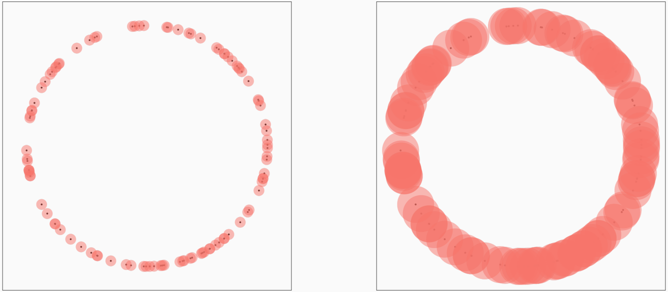
- We observe points sampled on an (unobserved) shape.
- Original motivation: How can we find the original shape only from the points?



Here comes Topological Data Analysis (TDA).

Topology of Points?

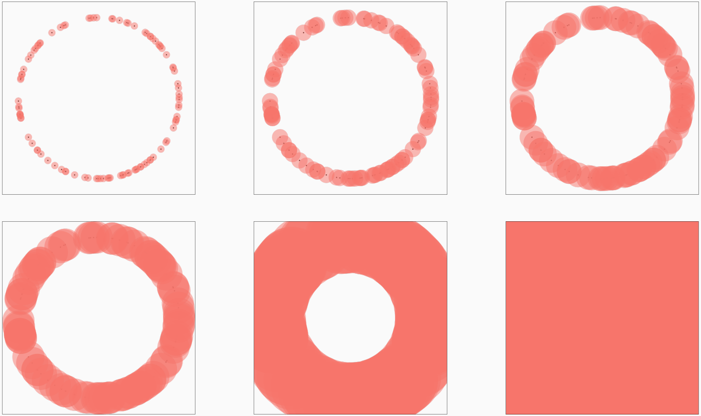
- We replace each point with a ball of radius $r > 0$.
- For a r large enough, we find indeed the loop.



How to choose r?

Here comes Persistence

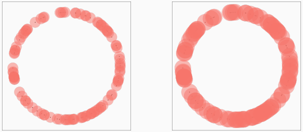
We let r growing from 0 up to ∞.



Original Idea: Important features will be the ones that ”persist” a long ”time” when r increases.

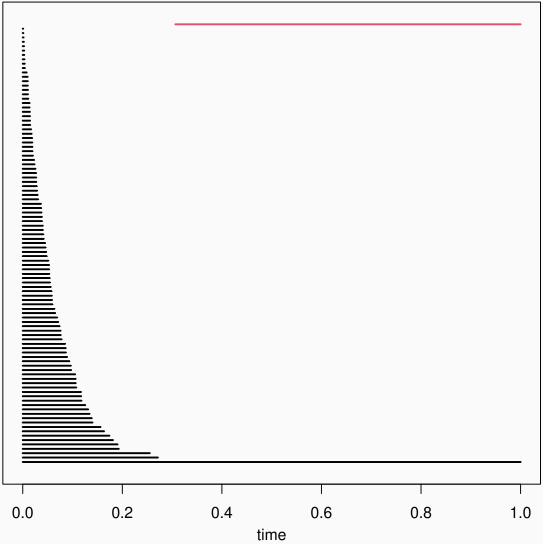
How do we record information?

- We record each ”radius/time” where change in the topology of the union of balls happens.
- Each time two balls connect: one connected component disappear – it dies.
- When a loop appears for a radius r_{loop} we say it is the birth time of the loop.

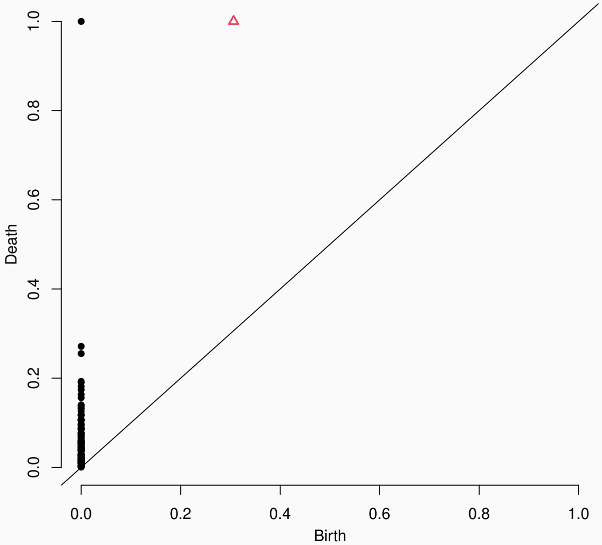


- When the loop is completely covered we say it is its death.
- There is two common ways to display this information.

The Barcode



Persistent Diagram



Remarks

- This is the most standard way in which TDA is performed.
- This is the so-called **persistent homology** approach
- Although these two representations are equivalents, the persistence diagrams appears to be the most used for applications.

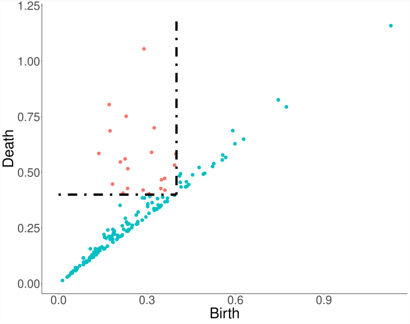
From this ”mathematically complicated” object, we define other statistics.

Betti Numbers

Definition: Let $b, d > 0$ with $b < d$ and D be a PD. The persistent Betti number is

$$\beta_{b,d}^{D} = \#\{(x, y) \in D,\ x \leq b, y \geq d\}.$$

Example: The number of point in red is $\beta_{0.4,0.4}^{D}$.



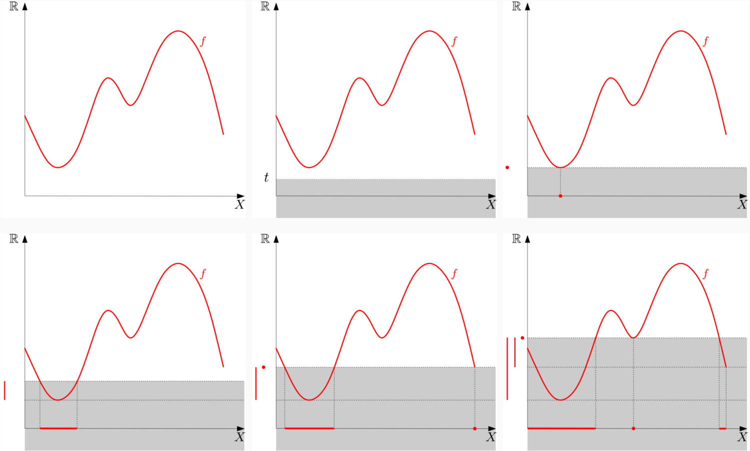
Important: The knowledge of $\beta_{b,d}^{D}$ for all b, d defines completely D.

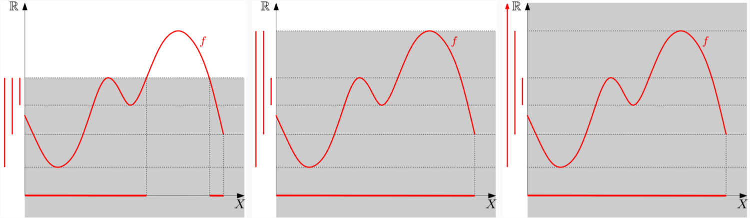
Filtrations

But, even within the persistent homology approach, there may be some variations.

- You do not need the balls.
- Building the union of balls $\rightarrow$ constructing an increasing sequence of topological spaces called a **filtration**.
- Analytically the union of balls corresponds to the sublevel sets of the distance function to the data points.
- Any sublevel sets of a function (smooth enough) actually works and may be used for analysis.

Example





ℝ
f
X
ℝ
f
X
ℝ
f
X

Back to the balls

- The union of balls is just the interpretation
- The true mathematical objects are the distance function to the set of data points X:

$$d_X(u) := d(u, X) = \min\{|u - x|,\ x \in X\}$$

- and its level sets at level $r > 0$:

$$\{u \in \mathbb{R}^d,\ d_X(u) \leq r\} = \bigcup_{x \in X} B(x, r).$$

Stability with respect to perturbation

The space of persistence diagram is a metric space: the Bottleneck distance on persistence diagram d_B.

Theorem (Stability, Cohen-Steiner et al. (2005))

Let X and Y be two sets of points in $\mathbb{R}^d$ with d_X and d_Y being the distance function to X and Y, respectively. Let further $PD(X), PD(Y)$ be the persistence diagrams obtained from the points X and Y, respectively. Then

$$d_B(PD(X), PD(Y)) \leq |d_X - d_Y|_\infty.$$

Main idea: If I perturb my data X by ϵ to get Y, $|d_X - d_Y|_\infty$ is small and the persistence diagrams are similar.

Remark on Stability Theorem

- There exists others stability theorems
- For example, assume the points X to be sampled on a manifold M.
- There is stability theorem to bound the Bottleneck distance $d_B(PD(X), PD(M))$ in function of the number of points.
- This is useful for shape analysis, to prove that you can recover the topology of the shape from points sampled on it.

What about PD computations? – Simplices

Simplicial Complexes

On the union of balls:

- We do not know how to define and compute easily the connected components, loops and other topological features of higher dimensions.
- Solution – Using another mathematical objects easier to work with:

Simplicial Complexes

Simplex

Definition (k-simplex)

Given a set of $k+1$ points $\{x_0, \dots, x_k\} \subset \mathbb{R}^d$, the k-dimensional simplex $[x_0, \dots, x_k]$ is the convex hull of the $k+1$ points.

Remark: the dimension depends on the number of points, not the dimension of the space.

Simplicial Complexes

A simplicial complex is a (valid) union of simplices.

They can be build from the data in many ways:

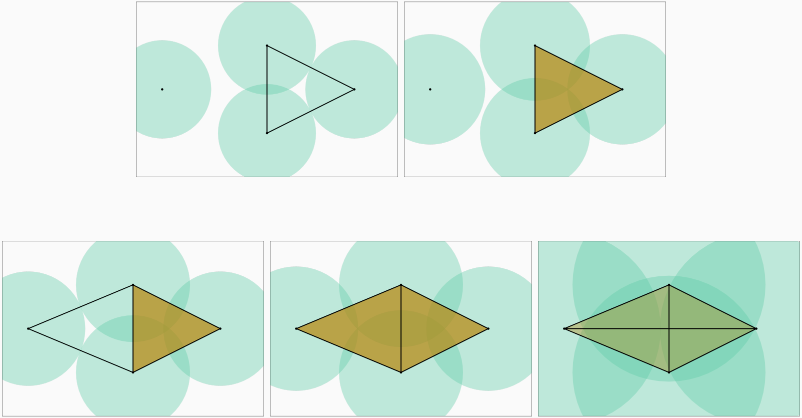
- Vietoris-Rips complexes
- Cech complexes
- α-complexes.
- Cubical complexes (suitable for images)

Cech complex – Definition

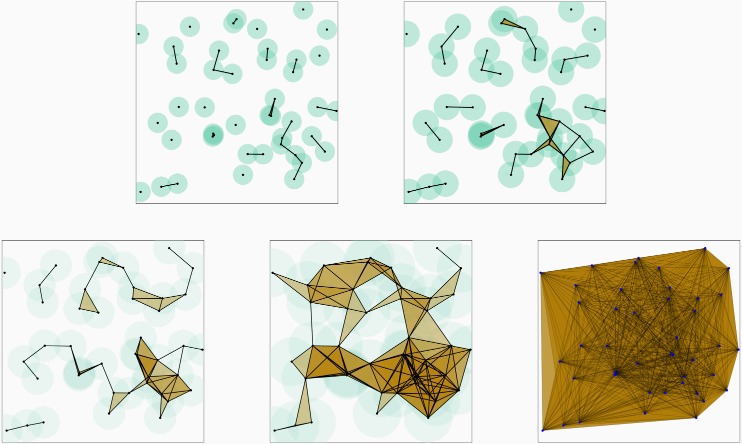
- Let's us consider the point pattern: $\mathbf{x} = \{x_1, \ldots, x_n\}$.
- The Cech complex at radius $r > 0$ of $\mathbf{x}$ is an union of simplices noted $C_r(\mathbf{x})$.
- For $k \in \mathbb{N}$, a k-simplex $[y_0, \ldots, y_k]$ belongs to $C_r(\mathbf{x})$ if and only if $\{y_0, \ldots, y_k\} \subset \mathbf{x}$ and

$$\cap_{j=0}^{k} B(y_j, r) \neq \emptyset.$$

Cech complex – Toy example



Cech complex – Poisson



Cech complexes – Advantages, Inconvenients

Pros:

- From a theoretical point of view: easy to study
- It verifies a Nerve Lemma: at each radius r, $C_r(\mathbf{x})$ is homotopic to the union of balls of radius r.
- **Main message:** Nerve Lemma $\Rightarrow$ Studying the union of balls or simplicial complexes is the same

Cons:

- Contains simplices of very high dimensions.
- Computationally hard to handle when lot of points.
- Slow to compute.

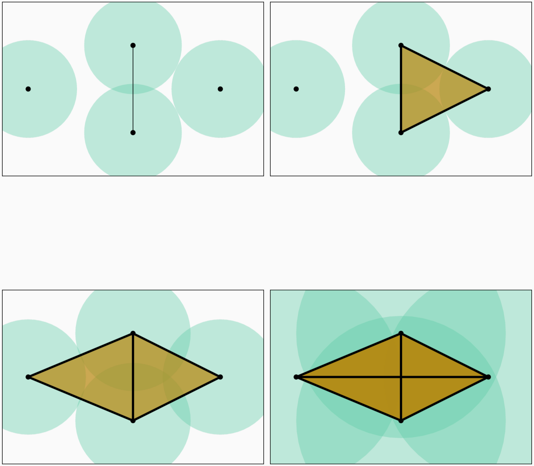
Rips complexes – Definition

- Let's us consider the point pattern: $\mathbf{x} = \{x_1, \ldots, x_n\}$.
- The (Vietoris-)Rips complex at radius $r > 0$ of $\mathbf{x}$ is an union of simplices noted $R_r(\mathbf{x})$.
- For $k \in \mathbb{N}$, a k-dimensional simplex $[y_0, \ldots, y_k]$ belongs to $R_r(\mathbf{x})$ if and only if $\{y_0, \ldots, y_k\} \subset \mathbf{x}$ and for all $i, j \in \{0, \ldots, k\}$:

$$B(y_i, r) \cap B(y_j, r) \neq \emptyset.$$

To compare with the Cech complex: $\cap_{j=0}^{k} B(y_j, r) \neq \emptyset$.

Rips complex – Toy example



Rips complexes

- Cech complex: good but slow and hard to compute.
- Vietoris-Rips complex: the quickest to compute but no Nerve Lemma.
- However, it the persistence diagram is still a good approximation in some sense.

In conclusion: For applications, no major differences.

The End – Thank you