Topological Data Analysis I

Applications in Spatial statistics

Christophe Ange Napoléon Biscio

Overview

Motivation of the course:

- Topological Data Analysis (TDA) is a relatively new field at the intersection of several mathematical fields.
- Consequently it may be hard to follow.
- There is various approaches depending on your scientific field.
- Tons of new concepts: Topology, Homology, Persistence, Quiver, cycle, Reeb graph, mapper, Morse Theory ...
- Many of theses concepts require background in field traditionally unknown by most statisticians.

Aim of this course:

- To provide the basic concepts and vocabulary appearing in TDA.
- To apply some of its basic ideas to spatial statistics for goodness-of-fit (deviation) test.
- Introduce the R package TDA and its use. Other programs works similarly.

What is not cover? A lot

What this course does not cover:

- Rigorous introduction to the theory.
- Connections with Morse theory and level-sets of functions.
- Mapper algorithm and UMAP algorithm.
- Applications of TDA with machine learning: kernel embedding technique, vectorisations, shape reconstruction, classification.
- Computational aspects of persistent homology: reduction algorithms, many properties of simplicial complexes, cubical complexes.

But: With today introduction and the mentioned references you will be better equipped to dive into TDA.

Topological Data Analysis – History

History

- Topology is the study of shapes.
- Topologist are interested in shapes invariant via continuous deformation of an object, i.e. no tearing.
- Example: from a ”topologist” point of view, a mug is a donut.

Important Questions:

- How do use it for applications?
- Spatial statistics, where?

What are the topological features? (informally)

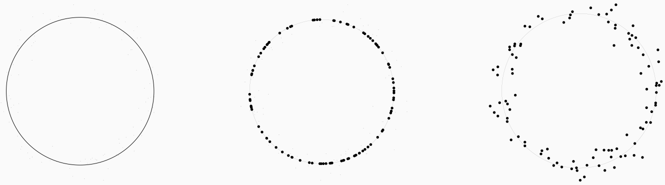
Topogical features: Every feature that is invariance under continuous deformation.

- The connected components.
- The loops
- In more that 2D: the voids ...

Example: The torus has 1 loop and 1 void (the inside of the donut).

Original motivation

- Assume we study a shape (circle)
- But we only observe points sampled on the shape $+$ noise.
- How can we find the original shape only from the points?



Here comes Topological Data Analysis (TDA).

There is several approaches in TDA:

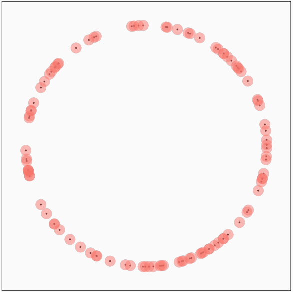
- Persistent Homology
- Mapper
- UMAP, Uniform Manifold Approximation

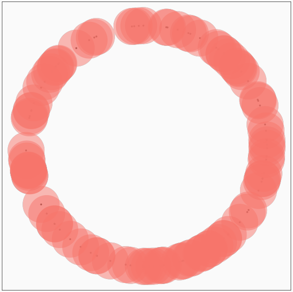
We will only cover Persistent Homology.

Persistent Homology

Topology of Points?

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

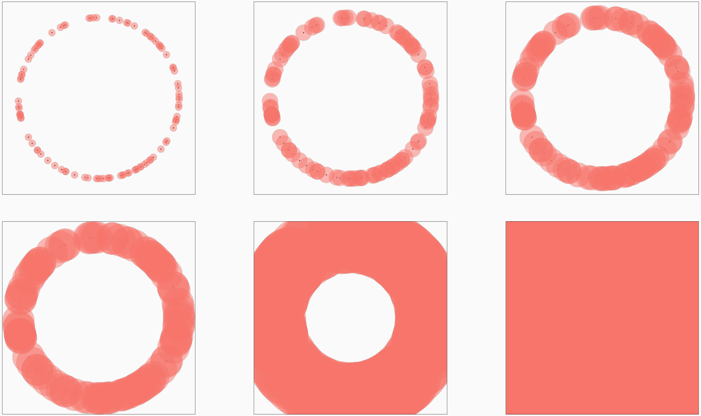




How to choose r?

Here comes Persistence

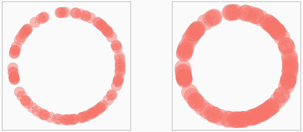
We let r growing from 0 up to ∞.



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

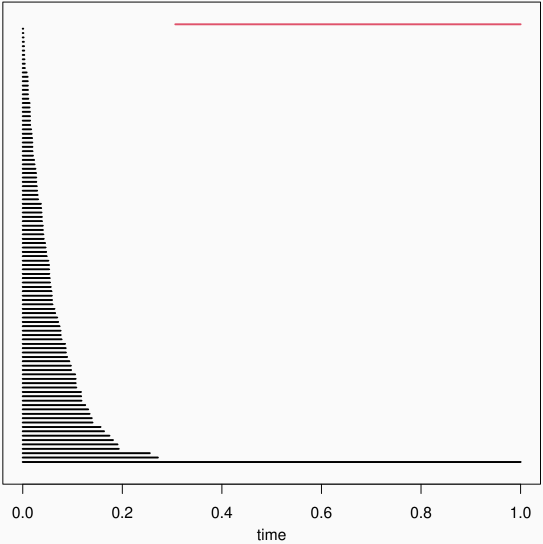
How do we record information?

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

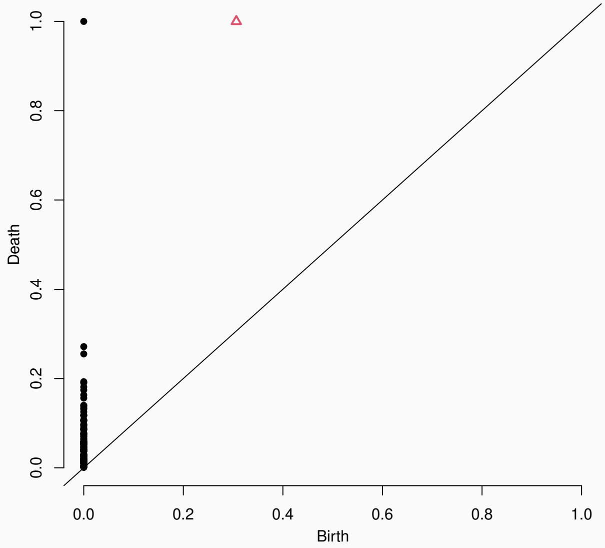


- When the loop is completely covered we say it is its death.
- There is two common ways to display this information.

The Barcode



Persistent Diagram



Computational aspects

Simplicial Complexes

- Both from a practical and theoretical point of view, we do not work with the union of balls directly.
- Why? 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

The union of balls is introduced mostly for pedagogical reasons.

Simplicial Complexes II

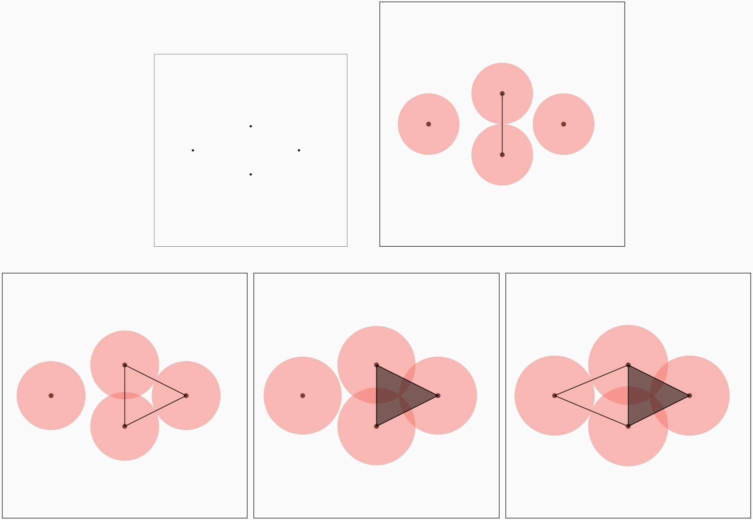
- Roughly, a simplicial complex is a generalisation of a graph in higher dimension.
- Meaning that is composed of vertices, edges, triangles (filled), tetrahedron ...
- Given a set of $k+1$ points $\{x_0, \ldots, x_k\}$, the k-dimensional simplex $[x_0, \ldots, x_k]$ is the convex hull of the $k+1$ points.
- Vertices $\Leftrightarrow$ 0-simplices
- Edges $\Leftrightarrow$ 1-simplices ...
- A ”good” simplicial complex will have, for a given radius r, the same topological features as the union of balls of radius r.

Simplicial Complexes III

Simplicial complexes may be build in many ways:

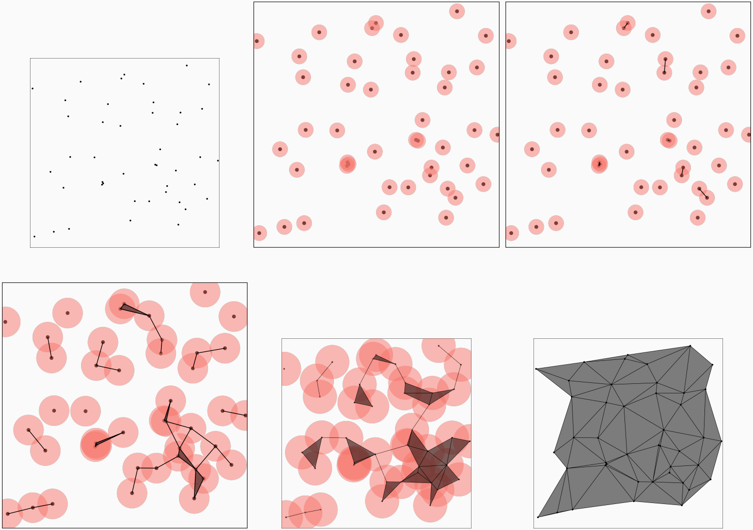
- Vietoris-Rips complexes
- Cech complexes
- α-complexes.
- Cubical complexes (suitable for images)

α-complexes – toy example



The sequence of simplicial complexes forms a sequence of topological spaces called a ”filtration”.

α-complexes – Poisson Process



Advantages of the α-complex

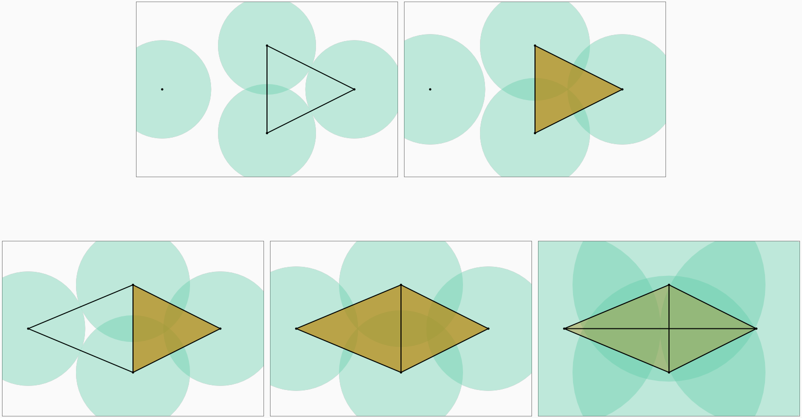
- The dimension of the simplices are bounded by the dimension of the ambient space.
- Easy to compute in low dimension with few thousands of points.
- For a each radius r the α-complex is homotopic equivalent to the union of balls of radius r. This is what we called a **Nerve lemma**.
- In other words, they share the same numbers of connected components, loops ...
- The α-complex is strongly connected to Voronoi tessellations, see Edelsbrunner and Harer (2010) for more details.

Cech complex – Definition

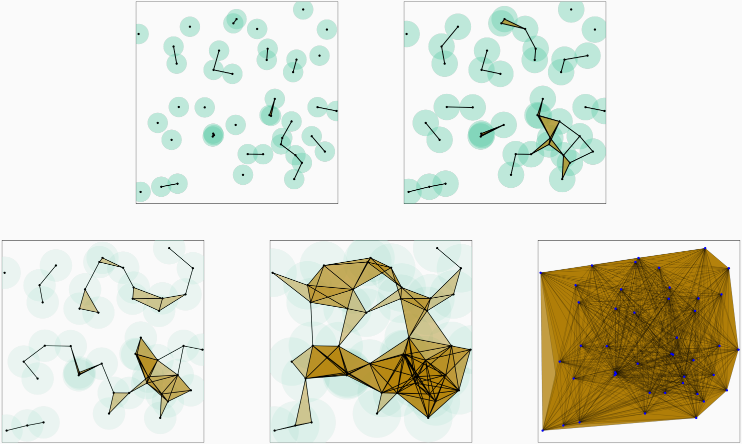
- Let's us consider the point pattern: $\mathbf{x} = \{x_1, \ldots, x_n\}$.
- The Cech complex at radius $r > 0$ of $\mathbf{x}$ is an union of simplices noted $C_r(\mathbf{x})$.
- For $k \in \mathbb{N}$, a k-dimensional simplex $[y_0, \ldots, y_k]$ belongs to $C_r(\mathbf{x})$ if and only if $\{y_0,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

Cons:

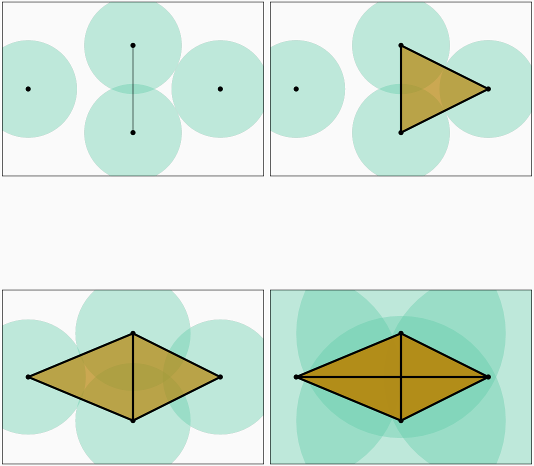
- Contains simplices of very high dimensions.
- Computationally hard to handle when lot of points.
- Slow to compute.

Rips complexes – Definition

- Let's us consider the point pattern: $\mathbf{x} = \{x_1, \dots, x_n\}$.
- 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 $[y_0, \dots, y_k]$ belongs to $R_r(\mathbf{x})$ if and only if $\{y_0, \dots, y_k\} \subset \mathbf{x}$ and for all $i, j \in \{0, \dots, k\}$:

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

Rips complex – Toy example



Summary

- Cech complex: good but slow and hard to compute.
- Vietoris-Rips complex: the quickest to compute but no Nerve Lemma.
- α-complex: easy to compute in low dimension.

In conclusion: For data analysis with few thousands hundreds points there is no major differences between this complexes and conclusions of a study should (in theory) be the same for each of them.

Coding Part I