Topological Inference

http://geometrica.saclay.inria.fr/team/Steve.Oudot/courses/EMA/
Context: The data deluge

Modern data sets are ever more massive and complex:

- academia
- industry
- general public
Context: The data deluge

Modern data sets are ever more massive and complex:

- academia
- industry
- general public

Need scalable and robust methods to analyze and classify these data
**Exploratory analysis of geometric data**

**Input:** point cloud equipped with a metric or (dis-)similarity measure

**Data point** ≡ image/patch, geometric shape, protein conformation, patient, LinkedIn user...
Exploratory analysis of geometric data

**Input:** point cloud equipped with a metric or (dis-)similarity measure

**data point** ≡ image/patch, geometric shape, protein conformation, patient, LinkedIn user...

**Goal:** describe the structure of the geometry underlying the data, for interpretation or summary
Challenges

Scale

Noise

Dimensionality
Challenges

4 million data points in $\mathbb{R}^9$

(source: [Lee, Pederson, Mumford 2003])

Motivation: study cognitive representation of space of images

Topology
Challenges

4 million data points in $\mathbb{R}^9$
(source: [Lee, Pederson, Mumford 2003])

Motivation: study cognitive representation of space of images

This is a lie!
The topology of data (TDA)

topological invariants for classification

$$\beta_0 = \beta_2 = 1$$

$$\beta_1 = 2$$

Algebraic topology in the 20th century

Algebraic topology in the 21st century

topological descriptors for inference and comparison
Topology from Data

Input: point cloud \( P \subset \mathbb{R}^d \)

→ uncover the topological structure of the space(s) underlying the data
→ inspect data at all scales and see what ‘persists’
Approach: Compute persistence of distance function

\[ d_P : \mathbb{R}^2 \to \mathbb{R} \]

\[ x \mapsto \min_{p \in P} \| x - p \|_2 \]
Approach: Compute persistence of distance function

\[ d_P : \mathbb{R}^2 \rightarrow \mathbb{R} \]

\[ x \mapsto \min_{p \in P} \| x - p \|_2 \]
Approach: Compute persistence of distance function

d_P : \mathbb{R}^2 \rightarrow \mathbb{R}
\quad x \mapsto \min_{p \in P} \|x - p\|_2
Approach: Compute persistence of distance function

\[ d_P : \mathbb{R}^2 \rightarrow \mathbb{R} \]
\[ x \mapsto \min_{p \in P} \| x - p \|_2 \]
Approach: Compute persistence of distance function

\[ d_P : \mathbb{R}^2 \rightarrow \mathbb{R} \]

\[ x \mapsto \min_{p \in P} \| x - p \|_2 \]
Approach: Compute persistence of distance function

\[ d_P : \mathbb{R}^2 \to \mathbb{R} \]
\[ x \mapsto \min_{p \in P} \|x - p\|_2 \]
Approach: Compute persistence of distance function

\[ d_P : \mathbb{R}^2 \rightarrow \mathbb{R} \]
\[ x \mapsto \min_{p \in P} \| x - p \|_2 \]
Approach: Compute persistence of distance function

\[ d_P : \mathbb{R}^2 \rightarrow \mathbb{R} \]

\[ x \mapsto \min_{p \in P} \| x - p \|_2 \]
Approach: Compute persistence of distance function

\( d_P : \mathbb{R}^2 \rightarrow \mathbb{R} \)

\( x \mapsto \min_{p \in P} \| x - p \|_2 \)

Challenge: provide theoretical guarantees

(sufficient sampling conditions under which the barcode of \( d_P \) reveals the homology of the underlying space)
In practice: The inference pipeline

- point cloud
- proximity rule
- simplicial filtration
- homology
- barcode (signature)
In practice: The inference pipeline

Challenges:

- scaling with input size and dimensionality
- theoretical guarantees (*sweet range*, SNR)
Motivating example (manufactured data)

\((\mathbb{R} \mod \mathbb{Z})^2\) \(\mapsto\) \(\frac{1}{\sqrt{2}}(\cos(2\pi u), \sin(2\pi u), \cos(2\pi v), \sin(2\pi v))\) \(\subset\mathbb{S}^3 \subset \mathbb{R}^4\)

source: http://en.wikipedia.org/wiki/Clifford_torus
Motivating example (manufactured data)

\[(u, v) \mapsto \frac{1}{\sqrt{2}} (\cos(2\pi u), \sin(2\pi u), \cos(2\pi v), \sin(2\pi v))\]

\(n = 2000\) data points

ambient dimension \(d = 4\)

intrinsic dimension \(k = 1, 2, 3\)

Motivating example (manufactured data)

$n = 2000$ data points
ambient dimension $d = 4$
intrinsic dimension $k = 1, 2, 3$

$\text{size} \sim 2^n \mid n^{d+1}$
This has been one of the most commonly used filtrations since the introduction of persistence in the early 2000’s. ... on the existence of a sweet range for it. Before that, people were using it in practice with no theoretical guarantees.

Motivating example (manufactured data)

$n = 2000$ data points
ambient dimension $d = 4$
intrinsic dimension $k = 1, 2, 3$

$3$-sphere $(37 \cdot 10^9$ simplices)
Motivating example (manufactured data)

$n = 2000$ data points
ambient dimension $d = 4$
intrinsic dimension $k = 1, 2, 3$

$\mathsf{size} \sim 2^{d^2} n$

\(12 \cdot 10^6\) simplices

$\mathsf{mesh-based\ filtration}$

---

Motivating example (manufactured data)
Motivating example (manufactured data)

$n = 2000$ data points
ambient dimension $d = 4$
intrinsic dimension $k = 1, 2, 3$

$(200 \cdot 10^3$ simplices)

Rips zigzags
size $\sim 2^{k^2} n$
Natural images data

4 million data points in $\mathbb{R}^9$
(source: [Lee, Pederson, Mumford 2003])

Motivation: study cognitive representation of space of images

Topology
Natural Images Data

Preprocessing: 
- select bottom $x\%$ of data points according to $k$-NN distance
- sample 5000 points uniformly at random from filtered point set

\[ k = 1200, \ x = 10 \]
\[ k = 1200, \ x = 20 \]
\[ k = 1200, \ x = 30 \]

\[ k = 8000, \ x = 10 \]
\[ k = 8000, \ x = 20 \]
\[ k = 8000, \ x = 30 \]

\[ k = 24000, \ x = 10 \]
\[ k = 24000, \ x = 20 \]
\[ k = 24000, \ x = 30 \]

(source: [de Silva, Carlsson 04])
Natural Images Data

Preprocessing: - select bottom $x\%$ of data points according to $k$-NN distance
- sample 5000 points uniformly at random from filtered point set

(source: [O., Sheehy 13])
Natural Images Data

**Preprocessing:**
- select bottom $x\%$ of data points according to $k$-NN distance
- sample 5000 points uniformly at random from filtered point set

5000 landmarks

$k = 1200, x = 30$

(source: [O., Sheehy 13])
Natural Images Data

Preprocessing:  
- select bottom $x\%$ of data points according to $k$-NN distance  
- sample 5000 points uniformly at random from filtered point set

5000 landmarks

$k = 24000$, $x = 30$

(source: [O., Sheehy 13])
Natural Images Data

Preprocessing: - select bottom $x\%$ of data points according to $k$-NN distance
- sample 5000 points uniformly at random from filtered point set

5000 landmarks

$k = 24000, x = 30$

(source: [O., Sheehy 13])
Natural Images Data

Preprocessing: - select bottom $x\%$ of data points according to $k$-NN distance
- sample 5000 points uniformly at random from filtered point set

(source: [Carlsson, Ishkhanov, de Silva, Zomorodian 2008])