TDA lecture

Priyavrat Deshpande

Chennai Mathematical Institute

February 16, 2022

Outline

1. Complexes from Data
2. Persistent homology
3. Homology representation

The Cech complex

Point cloud data

A point cloud $\mathbb{X}$ is a set of vectors $\{x_1, \dots, x_N\}$ in $\mathbb{R}^d$.

Definition (Cech Complex)

Given a PCD $\mathbb{X}$ and $r > 0$ the Cech complex $\mathrm{Ch}_r(\mathbb{X})$ is an (abstract) simplicial complex whose simplices are those subsets $\sigma \in \mathbb{X}$ such that

$$\bigcap_{x \in \sigma} B(x; r) \neq \emptyset.$$

The Cech complex

- If $r < r'$ then $\mathrm{Ch}_r(\mathbb{X}) \subseteq \mathrm{Ch}_{r'}(\mathbb{X})$.
- For a radius r, a subset σ of $\mathbb{X}$ a simplex if and only if the corresponding set of points is contained in a ball of radius r.
- Checking whether a set of points is contained in a ball of given radius is a well studied problem in computational geometry.

Definition (Diameter)

The diameter of $\mathbb{X}$ is the upper bound of the set of all pairwise distances, i.e.,

$$\operatorname{diam}(\mathbb{X}) := \sup\{d(x,y) \mid x, y \in \mathbb{X}\}.$$

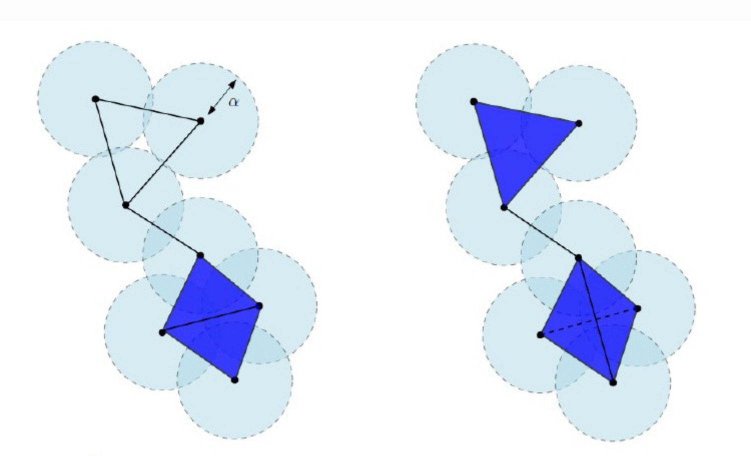
Definition

Let $X \subset \mathbb{R}^N$ be a finite point cloud and $r > 0$ be the scale parameter. The Vietoris-Rips complex, $\mathrm{V}_r(X)$, is defined as the simplicial complex that contains all subsets whose diameter is at most r:

$$\mathrm{V}_r(X) := \{\sigma \subset \mathbb{X} \mid \operatorname{diam}(\sigma) \le r\}.$$

- In general $\mathrm{V}_r(X)$ does not embed in $\mathbb{R}^d$.
- If $r < r'$ then $\mathrm{V}_r(X) \subseteq \mathrm{V}_{r'}(X)$.
- It is a clique (or flag) complex, i.e., completely determined by its 1-skeleton.

An example



Left hand side we have Ch_α and on the right hand side we have $\mathrm{V}_{2\alpha}$.

The relationship

Proposition

$$\mathrm{Ch}_r(\mathbb{X}) \subseteq \mathrm{V}_{2r}(\mathbb{X}) \subseteq \mathrm{Ch}_{2r}(\mathbb{X}).$$

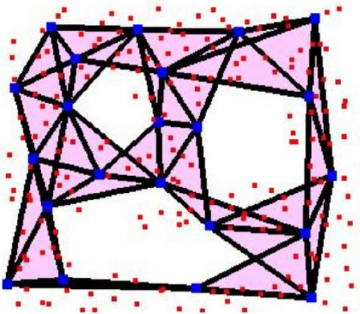
Proof

- Let $\sigma \in \mathrm{Ch}_r$. Then there is a ball of radius r that contains σ. The diameter of such a ball is at most $2r$, hence $\sigma \in \mathrm{V}_{2r}$.
- Let $\tau \in \mathrm{V}_{2r}$. Then $\mathrm{diam}(\tau) \leq r$. Hence there is a ball of radius $2r$ containing τ. Hence $\tau \in \mathrm{Ch}_{2r}$.

Other witness complex

Definition

Given a PCD $\mathbb{X}$ and a chosen subset L of landmark points the witness complex, $W(\mathbb{X}, L)$ is defined as follows: the vertices of W are the points in L. For each $x \in \mathbb{X}$, we find two points $l_1, l_2 \in L$ that are closest to x, and add the edge $\{l_1, l_2\}$. A higher simplex is added if and only if all its edges are present.



The witness complex

How to choose L?

- Randomly.
- Pick l_1 at random. Then choose l_2 such that $d(l_1, l_2)$ is maximum. Choose l_3 that maximizes $\min\{d(l_1, l_3), d(l_2, l_3)\}$ etc.
- Choose from denser regions.

The scale parameter

For a scale parameter $r > 0$ the witness complex $W_r(\mathbb{X}, L)$ has vertex set L and $\{l_1, l_2\}$ is an edge if there exists $x \in X$ such that

$$d(x, l_1), d(x, l_2) \leq \text{const.} + r.$$

Other complexes

1. The Alpha complex.
2. The flow complex.
3. The Delaunay triangulation.

Outline

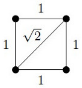
1. Complexes from Data
2. Persistent homology
3. Homology representation

Where is the correct shape?

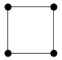
- From data one gets a filtered simplicial complex.
- How to pick the right r? (For which r we get the right shape from which the data is sampled?)
- There might not be just 'one correct' value of r.
- Matrix reduction(s) for every value of r must be expensive!!

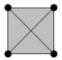
Just calculate topological features for all possible scales.

An example

Consider the PCD consisting of 4 points. 

For $r < 1$ the VR complex is just 4 points.

For $1 \leq r < \sqrt{2}$ the VR complex is 

For $r \geq \sqrt{2}$ the VR complex is a tetrahedron 

Question

How do we quantify and keep track of these changes?

Filtered complexes

An increasing sequence $\epsilon_{i_1} < \cdots < \epsilon_{i_n}$ induces a filtration

$$\emptyset \subset VR_{\epsilon_{i_1}}(X) \subseteq \cdots \subseteq VR_{\epsilon_{i_n}}(X).$$

For every $p \geq 0$ we have:

$$H_p(VR_1(X)) \xrightarrow{f_p^{0,1}} H_p(VR_2(X)) \xrightarrow{f_p^{0,2}} \cdots \xrightarrow{f_p^{n-1,n}} H_p(VR_n(X)).$$

In general for $i < j$

$$f_p^{i,j} : H_p(VR_i(X)) \to H_p(VR_j(X)).$$

Persistent homology

$$f_p^{i,j} : H_p(VR_i(X)) \to H_p(VR_j(X)).$$

Definitions

- p-**th persistent homology group**: $\mathcal{H}_p^{i,j} := \mathrm{Im}(f_p^{i,j})$.
- p-**th persistence Betti number**: $\beta_p^{i,j} := \mathrm{rank}(\mathcal{H}_p^{i,j})$.
- **Birth at i-th stage**: A class c such that

$$c \in H_p(VR_i(X)) \text{ but } c \notin \mathcal{H}_p^{i-1,i}.$$

- **Death of a class at j-th stage**: A class c such that

$$f_p^{i,j-1}(c) \notin \mathcal{H}_p^{i-1,j-1} \text{ and } f_p^{i,j}(c) \in \mathcal{H}_p^{i-1,j}.$$

- The number of cycles that are born at ϵ_i and are dead at ϵ_j is:

$$\mu_p^{i,j} := (\beta_p^{i,j-1} - \beta_p^{i-1,j-1}) - (\beta_p^{i,j} - \beta_p^{i-1,j})$$

Are there other filtrations?

Sublevel set filtration

Suppose K is a simplcial complex and f an $\mathbb{R}$-valued function defined on the vertices of K. Define following "weight function" on K

$$w(\sigma) := \begin{cases} f(v) & \text{if } \sigma = \{v\}, \\ \max_{\tau \subset \sigma} w(\tau) & \text{else.} \end{cases}$$

Then K can be filtered using ascending order of weights of simplices.

One can analogously define superlevel set filtrations.

Outline

1 Complexes from Data

2 Persistent homology

3 Homology representation

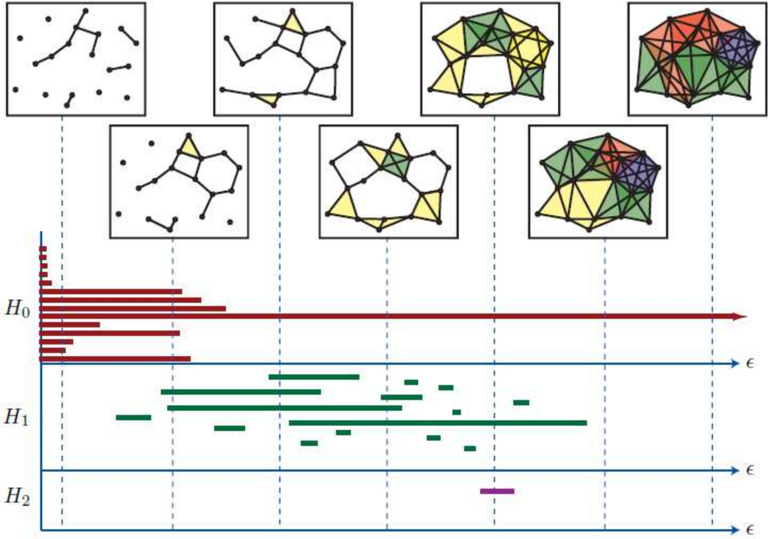
Visualizing persistence

Definition (Persistence barcodes)

For every $p \geq 0$ we draw a graph whose verticle axis corresponds to all possible p-homology generators and the horizontal axis is the time parameter.

1. Advantage: "judging lengths of lines"
2. However, topological features exist at different scales with large relative differences.
3. Pretty large data, unimportant details.
4. How does on order features?

Example



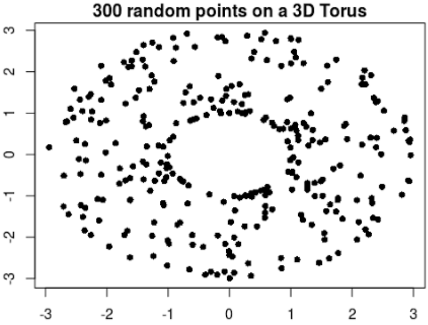
Visualizing persistence

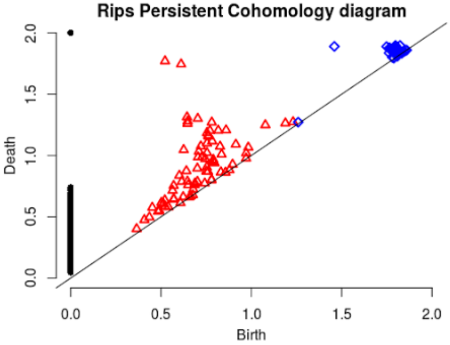
Definition (Persistence diagram)

The p-persistence diagram is a 2-d coordinate system where x is the birth coordinate and y is the death coordinate. For every p-homology class there is a point (b, d) representing its birth and death time.

1. The lifetime of a cycle x_i is called persistence; $\mathrm{pers}(x_i) = d_i - b_i$.
2. Persistence diagram is a type of topological summary.
3. The space of all PDs supports various metrics that differentiate topological features.
4. They appear very cluttered. Suffer from overplotting.
5. Advantage: points are grouped by scale similarity.
6. Stable w.r.t. perturbation in the data.
7. Sensitive to "*small/big*" holes.
8. Possible to track holes, record size/scale of the feature.
9. Not sensitive to outliers.
10. Computable in practice.
11. Provides a flexible framework (i.e., clusters/flares etc.)

Persistence diagrmas

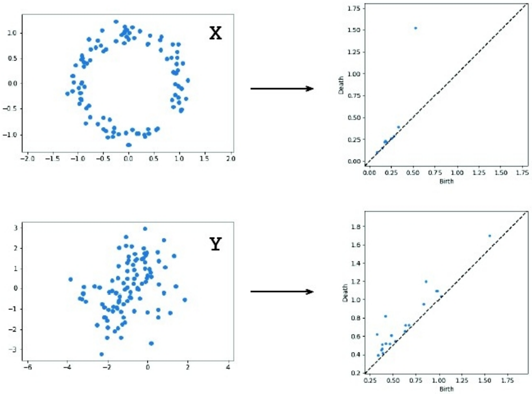




How to interprete?

1. High persistence implies existence of **robust** features.
2. Spurious topological features are short-lived, i.e., **noise**.
3. The *summary description* is always 2-dimensional.
4. Persistent diagrams are a similarity metirc.
5. β_0: number of connected components (clusters?).
6. β_1: number of cycles (periodic features?).
7. β_2: number of hollow spaces (?).

Persistence diagrmas



The bottleneck distance

Definition

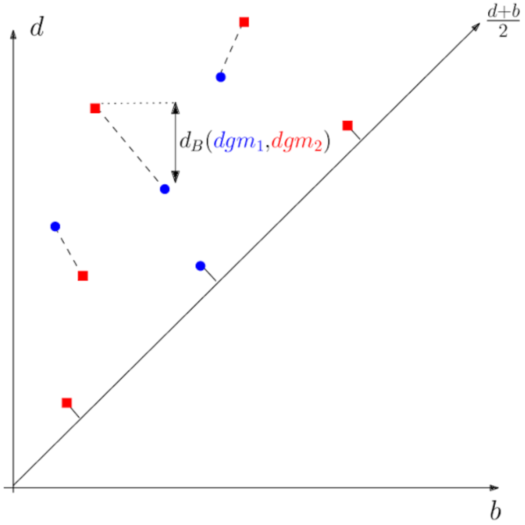
For two PDs X, Y the bottleneck distance (∞-Wasserstein metric) is defined as

$$d_B(X,Y) := \inf_{\gamma} \sup_{x \in X} ||x - \gamma(x)||_{\infty},$$

where γ runs over all the matchings (bijections) from X to Y.

1. The space of PDs with d_B is a metric space.
2. There are similar distance functions.
3. Proves stability of PH operation.
4. PD is not a vector.

Optimal transport



The stability theorem

Theorem

Denote by Let $\mathbb{X}_1, \mathbb{X}_2$ *be two PCDs and denote by* $D_p(\mathbb{X})$ *the persistence diagram correponding* p*-persistence homology. Then*

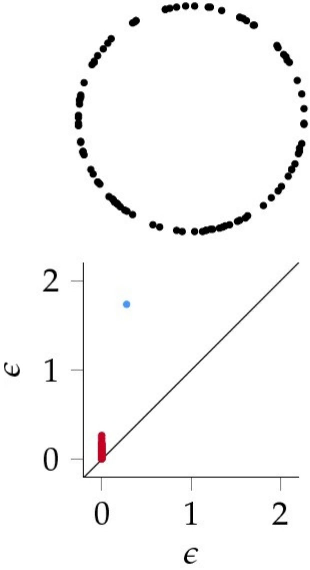
$$d_B(D_p(\mathbb{X}_1), D_p(\mathbb{X}_2)) \leq d_H(\mathbb{X}_1, \mathbb{X}_2),$$

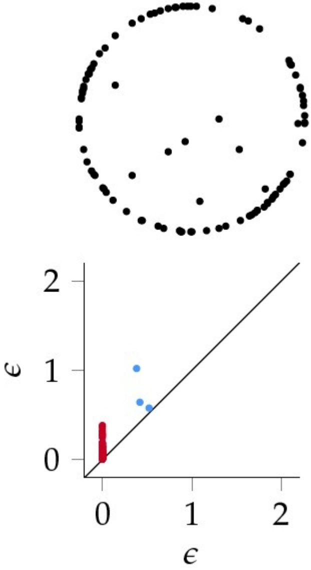
where $d_H(\ ,\ )$ *is the Hausdorff distance between the sets.*

Intuitive meaning

The persistent homology doesn't change under mild perturbation of the data.

An example





Digression

- PDs can be constructed for functions defined over point clouds.
- PDs are defined and stable for a large class of continuous functions defined over (pre-)compact metric spaces.
- Topological signatures have been used for shape classification and segmentation and clustering.

Topology to Statistics

Data → Persistence Diagram → **Vectorization** → Prediction, Classification, Testing, Estimation

Recently a lot of methods have been discovered that convert topological features into vectors that can be used for statistical analysis as well as input to ML algorithms.