Topological Data Analysis: History and Challenges

David Letscher

Saint Louis University

SLU Topology Seminar

Motivation

Original Question?

Given a finite set of points $X_0 \in \mathbb{R}^d$ what can you say about the topology of the shape they represent.

Motivation

Original Question?

Given a finite set of points $X_0 \in \mathbb{R}^d$ what can you say about the topology of the shape they represent.

Discrete point set = no interesting topology ... Right?

Motivation

Original Question?

Given a finite set of points $X_0 \in \mathbb{R}^d$ what can you say about the topology of the shape they represent.

Discrete point set = no interesting topology ... Right?

Followup

Can we detect geometric information using topological information about the points?

Filtrations

Input $X_0 \subset \mathbb{R}^d$. (Assumed to be generic)

α-shape

$$X_\alpha = \bigcup_{x \in X_0} B(x, \alpha)$$

α-filtration

$$X_{\alpha_1} \subset X_{\alpha_2} \subset \cdots \subset X_{\alpha_k}$$

More generally

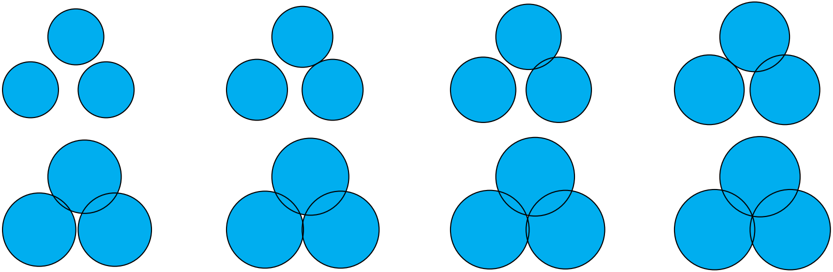
For continuous $f : \mathbb{R}^d \to \mathbb{R}$, define $X_\alpha = \{x \mid f(x) \leq \alpha\}$.
For example, f might measure density (medical imaging, ...).

Topological Critical Points

Definition

α's where $X_{\alpha-\epsilon} \ncong X_\alpha$ for all sufficiently small α.

If X_0 is generic or f is a (generic) Morse function then there are finitely many critical point and at each there is a single change in topology.



Homology

Suppose that X is a triangulated space.

Chains

$C_k(X)$ if the vector space (with base field $\mathbb{Z}/2\mathbb{Z}$) generated by the k-simplices of X (vertices, edges, triangles, ...).

Boundary

For a k-simplex σ, $\partial\sigma$ is the formal sum of its $(k-1)$-simplices. For chains, the boundary operator can be extended linearly

$$\partial\left(\sum \sigma_i\right) = \sum \partial(\sigma_i)$$

Homology

Cycles

$Z_k(X) =$ subspace of chains c with $\partial c = 0$

Boundaries

$B_k(X) =$ spaces of boundaries of $(k+1)$-chains

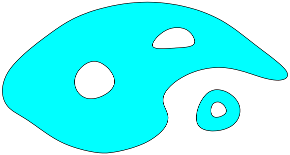
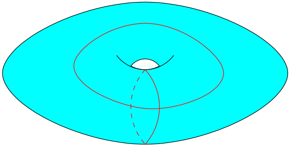
Homology

$H_k(X) = Z_k(X)/B_k(X)$

If $X \subset \mathbb{R}^3$,

Rank of $H_0(X) =$ number of components of X
Rank of $H_1(X) =$ number of independent loops of X
Rank of $H_2(X) =$ number of voids of X

Homology, examples

X	Rank of $H_0(X)$	Rank of $H_1(X)$	Rank of $H_2(X)$
	2	3	0
	1	2	1

Persistent Homology

[L-Edelsbrunner-Zomordian, 2000]

$$H_k^p(X_\alpha) = Z_k(X_\alpha)/\left(Z_k(X_\alpha) \cap B_k(X_{\alpha+p}\right)$$

Equivalently,

$$H_k^p(X_\alpha) = image(H_k(X_\alpha) \to H_k(X_{\alpha+p}))$$

$H_0^p(X_\alpha)$	counts the number of components of X_α that are still separate in $X_{\alpha+p}$.
$H_1^p(X_\alpha)$	measures the number of (independent) loops in X_α that are not "filled-in" in $X_{\alpha+p}$.
$H_2^p(X_\alpha)$	measures the number of voids in X_α that are not "filled-in" in $X_{\alpha+p}$.

Challenges

- Representing persistence
- (Efficiently) calculating persistence
- Isolating signal from noise
- Stability
- Robustness
- Can you simplify a shape to remove topological noise?
- Sample means and variances

Persistence Diagrams

Every topological critical point corresponds to a "birth" or "death" of a cycle. (We assume these critical points happen at distinct times.)

Birth time

An α where $H_k(X_\alpha) \cong H_k(X_{\alpha-\epsilon})$ for all sufficiently small $\epsilon > 0$.

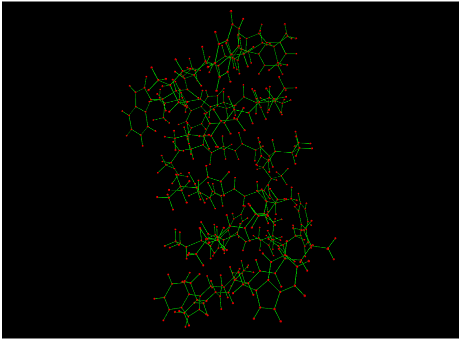
Death time

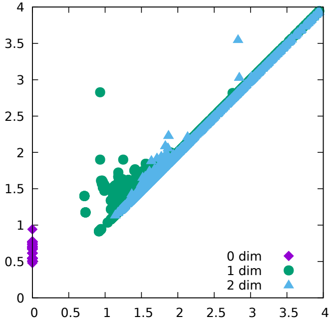
If c is some cycle born at time α, then it's death time is the smallest β such that there exists a cycle $c' \in H_k(X_\alpha)$ such that $c + c'$ bounds a cycle in X_β.

Persistence Diagrams

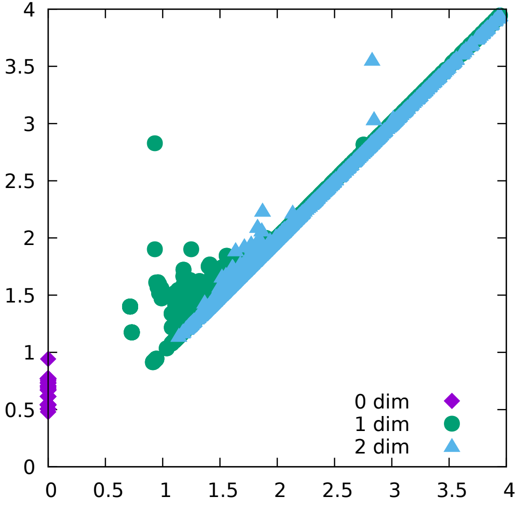
Persistence Diagrams

A persistence diagram [Cohen-Steiner-Edelsbrunner-Harer, 2005] is a plot of the birth-death pairs of the plane.

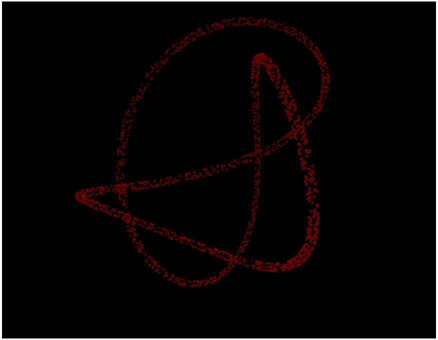


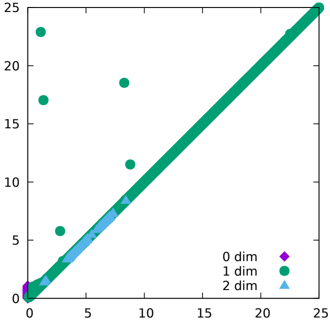


Persistence Diagrams

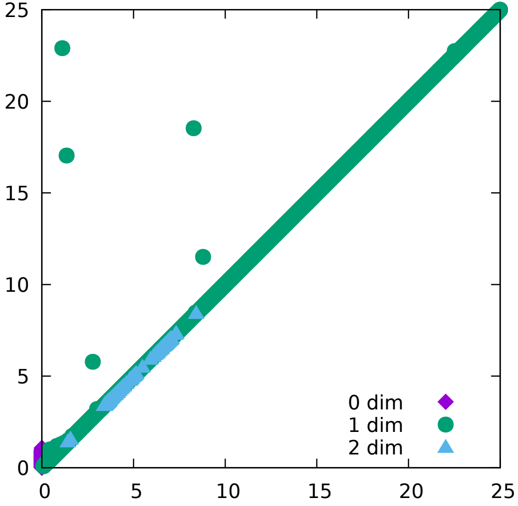


Persistence Diagrams





Persistence Diagrams



Stability of Persistence Diagrams

We will augment the persistence diagrams by adding the diagonal with infinite multiplicity.

Bottleneck distance

If D_k and D'_k are two persistence diagrams the bottleneck distance

$$d_B(D_k, D'_k) = \inf_{\text{bijections } f:D_k \to D'_k} \sup_{x \in D_k} ||x - f(x)||_\infty$$

Bottle neck distance make the space of persistence diagrams a complete separable metric space.

Stability [Cohen-Steiner-Edelsbrunner-Harer, 2005]

If $f, g : \mathbb{R}^d \to \mathbb{R}$ have k-dimensional persistence diagram D_f and D_g, respectively, then

$$d_B(D_f, D_g) \leq ||f - g||_\infty$$

Robustness of Persistence Diagrams

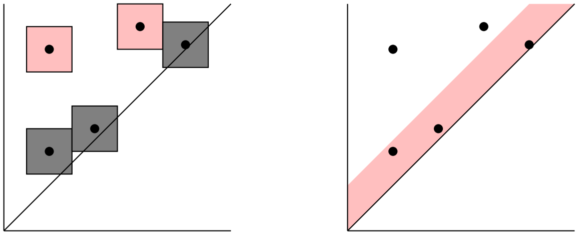
Breakdown point

The minimum fraction of data points that need to be changed in order to change a statistic arbitrarily.
e.g. For a sample of n points, the breakdown point for the mean is $1/n$ and is $1/2$ for the median.

Breakdown points for persistence diagrams [L]

For n points in $\mathbb{R}^d$ filtered using α-shapes, the break down points for the k-dimensional persistence diagram is $\frac{k+1}{n}$.

Isolating Features from Noise



Suppose that D is the diagram for space being sampled and D_n is a diagram for an n point subsample (chosen uniformly for some distribution).

Goal

For any p find functions $c(n)$ and $f(n)$ such that

$$\mathbb{P}\left(d_B(D, D_n) > c(n)\right) < p + f(n)$$

where $c(n) \to 0$ and $f(n) \to 0$.

Confidence Intervals

[Fasy-Lecci-Rinaldo-Wasserman-Balakrishan-Singh]

Subsampling $c(n) = 2/p(n)$, where $p(n)$ is the probability of a random sample of n points being within Hausdorff distance α from the given point sample and $f(n) = O\left(\frac{1}{(\log n)^{1/4}}\right)$.

Concentration of Measure $c(n) = O\left(\left(\frac{\log n}{n}\right)^{1/d}\right)$ and $f(n) = O\left(\frac{1}{n \log n}\right)$.

Method of Shells More complicated $f(n)$ for an constant $c(n)$ that is sufficiently small.

Density Estimation Can have $f(n) = 0$ for a more complication $c(n)$.

Note all functions depend on invariants of the space the points are sampled from and cannot be estimated (yet) for non-trivial spaces.

Sample Means and Variances

Given diagrams $D_1, \ldots, D_n$.

[Turner-Mileyko-Mukherjee-Harer]

A Frechet mean is a diagram D that minimizes

$$\sum_i (d_B(D, D_i))^2$$

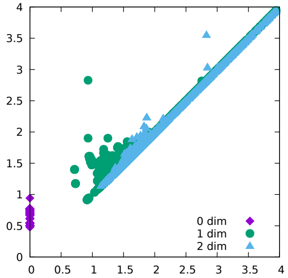
and the sum is the Frechet variance.

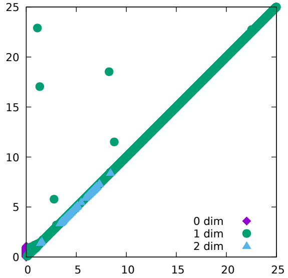
- Frechet means are conjectured to be biased estimators.
- As the sample density for the D_i goes to infinity, is the Frechet mean asymptotically unbiased?
- Are there alternative sample "averages" that are unbiased? Asymptotically unbiased?

Simplification

Question

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ has persistence diagrams D_k. For a fixed $\epsilon > 0$ let D'_k be D_k with all points within a distance ϵ removed.
Does there exists f' with $||f - f'||_\infty < \epsilon$ and the k-dimensional persistence diagrams of f' equal to D'_k.





Simplification

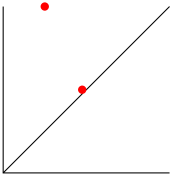
[Bauer-Lange-Wardetzky]

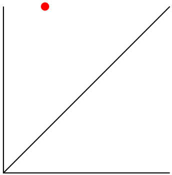
If $f : \mathbb{R}^2 \to \mathbb{R}$ is a tame Morse function then for any $\epsilon > 0$ there exits $f' : \mathbb{R}^2 \to \mathbb{R}$ such that

- $||f - f'||_\infty < \epsilon$
- The persistence diagram for f' are the persistence diagrams for f with every point within ϵ of the boundary removed.

Simplification

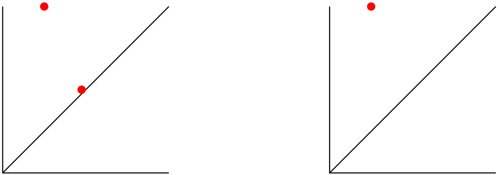
Similar simplification is not possible in $\mathbb{R}^3$ using persistent homology.

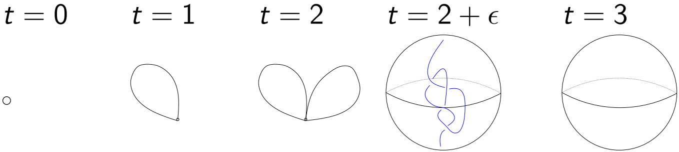




Simplification

Similar simplification is not possible in $\mathbb{R}^3$ using persistent homology.





If f' is any function with the second persistence diagram then $||f - f'||_\infty \geq 1$ but $d_B(D, D') \leq \epsilon$.

Thanks

Questions?