TDA lecture

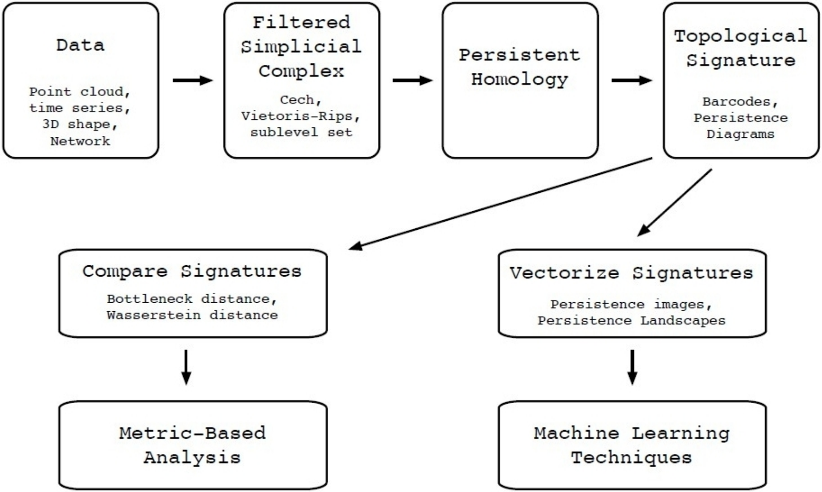
Priyavrat Deshpande

Chennai Mathematical Institute

February 20, 2022

Outline

The TDA pipeline



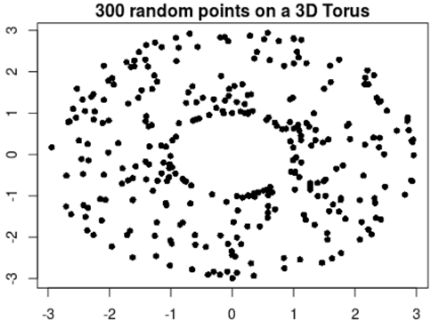
Visualizing persistence

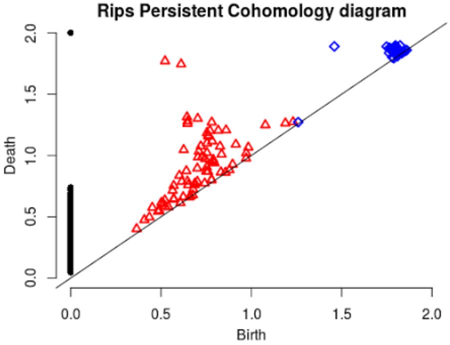
Definition (Persistence diagram)

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.

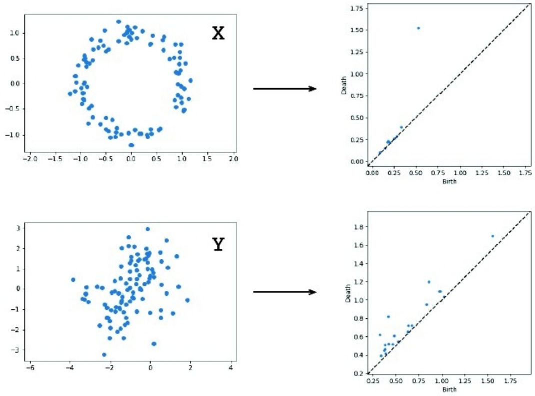
1. The lifetime of a cycle x_i is called persistence; $\mathrm{pers}(x_i) = d_i - b_i$.
2. The space of all PDs supports various metrics.
3. Advantage: points are grouped by scale similarity.
4. Stable w.r.t. perturbation in the data.
5. Sensitive to "*small/big*" holes.
6. Possible to track holes, record size/scale of the feature.
7. Not sensitive to outliers.
8. Computable in practice.

Persistence diagrmas





Persistence diagrmas



The bottleneck distance

Definition

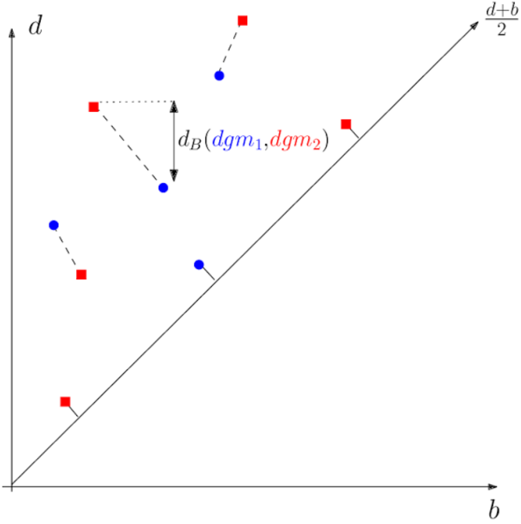
For two PDs X, Y the bottleneck distance (∞-Wasserstein metric) is defined as

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

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

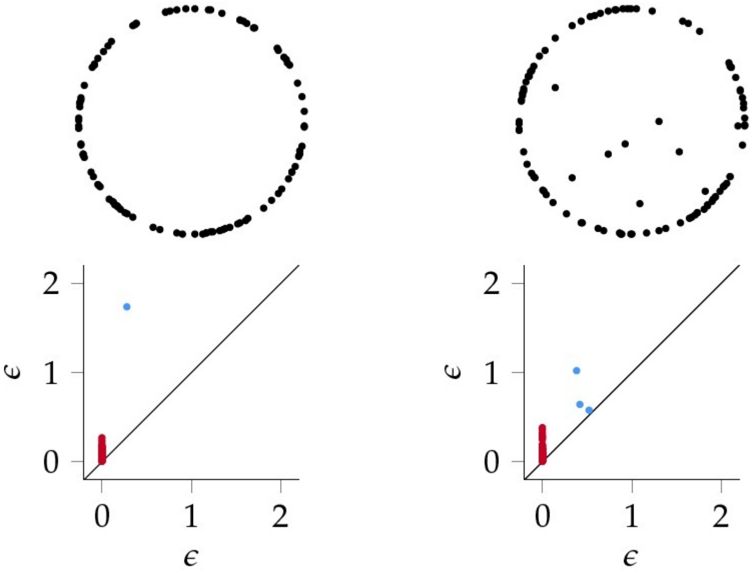
$$d_B(D_p(\mathbb{X}_1), D_p(\mathbb{X}_2)) \leq d_H(\mathbb{X}_1, \mathbb{X}_2),$$

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



Outline

Statistics on PDs

Question

Given a PD $D_p(\mathbb{X}) =: X$. Does X behave like a random variable of the data?

Consider independent random variables $X_1, \ldots, X_k$ with the same distribution as the PD (Does this even make sense?). We want good interpretation for

- the mean μ of X.
- the mean $\overline{X}_n$ of the samples.

In order

- to say $\lim \overline{X}_n = \mu$ (law of large numbers),
- hypothesis testing $(\mu_x = \mu_y)$,
- confidence interval on $\overline{X}_n - \mu$?

The bad news

Suppose (M, d) is a metric space. The Frechet mean of $a_1, \ldots, a_n \in M$ is the unique $p \in M$ which minimizes

$$\sum_i d(a_i, p)^2.$$

The stats doesn't make sense

For PDs with the bottleneck distance the Frechet mean is never unique.

Outline

Persistence landscapes

First described in

P. Bubenik, Statistical topological data analysis using persistence landscapes. J. Machine Learning Research, (2015).

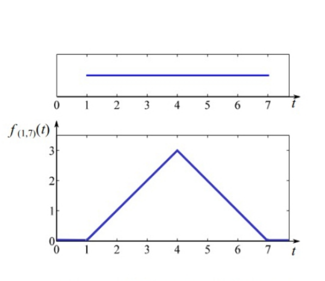
- They quantify ‘covered’ topological features.
- The idea is to ‘peel off’ layers iteratively.
- A landscape can be sampled at regular intervals to obtain a fixed-size feature vector.
- There is a built-in hierarchy.
- No information is lost.
- Recently it has been used a neural net layer.

Persistence landscapes

The single interval case

For $J := [b, d]$ consider the $\mathbb{R}$-function

$$f_J(t) := \begin{cases} 0 & \text{if } t \notin J, \\ t - b & \text{if } b \leq t \leq \frac{b+d}{2} \\ d - t & \text{if } \frac{b+d}{2} \leq t \leq d \end{cases}$$



- Switch to $(m = \frac{b+d}{2}, h = \frac{d-b}{2}$ coordinates (diagonal becomes $h = 0$).
- Construct 'peak' functions for each persistence cycle.

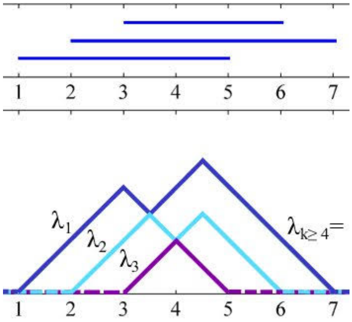
Persistence landscapes

Given an interval $J := [b, d]$ consider the real valued function

$$f_J(t) := \begin{cases} 0 & \text{if } t \notin J, \\ t - b & \text{if } b \leq t \leq \frac{b+d}{2} \\ d - t & \text{if } \frac{b+d}{2} \leq t \leq d \end{cases}$$

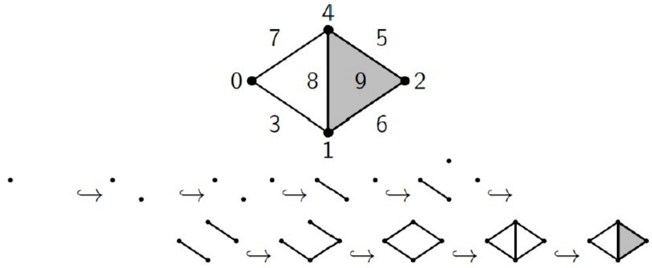
Given a collection of intervals J_i in a barcode B, we get a sequence λ_k of functions, for $k \in \mathbb{N}$:

$$\lambda_k(x) := k \max\{f_{J_i}(x)\}.$$



1. The sequence $\{\lambda_k\} \in L^p(\mathbb{N} \times \mathbb{R})$, a Banach space.
2. The norm measures *how much homology* there is (quantifies long and many barcodes).
3. The distance compares shapes of point clouds.

Consider the following simplicial complex



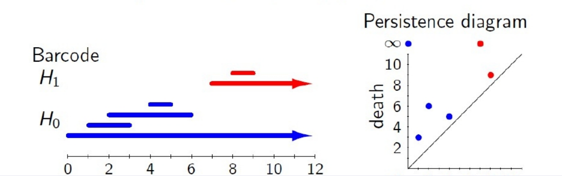
Time	0	1	2	3	4	5	6	7	8	9
Betti number	β_0	β_0	β_0	β_0	β_0	β_0	β_0	β_1	β_1	β_1
effect	$+$	$+$	$+$	$-$	$+$	$-$	$-$	$+$	$+$	$-$

Birth–Death pairs for H_0: $(0, \infty)$, $(1, 3)$, $(2, 6)$, $(4, 5)$

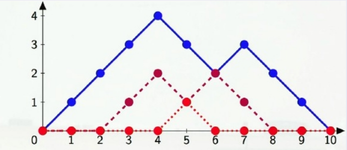
Birth–Death pairs for H_1: $(7, \infty)$, $(8, 9)$

Example

The barcode



The 0-landscape



λ_1	0	1	2	3	4	3	2	3	2	1	0
λ_2	0	0	0	1	2	1	2	1	0	0	0
λ_3	0	0	0	0	0	1	0	0	0	0	0

A Statistical landscape

If $f, g \colon \mathbb{R} \to \mathbb{R}$ are real-valued functions, their mean is defined as

$$\mu_{f,g}(x) := \frac{1}{2}(f(x) + g(x)).$$

A well-defined mean for landscapes

If $\Lambda = \{\lambda_i\}$ and $\Xi = \{\xi_i\}$ are two landscape functions then their mean is

$$\frac{1}{2}(\Lambda + \Xi) = \{\frac{1}{2}(\lambda_1 + \xi_1)\}.$$

Theorem (The stability theorem)

For any $t \in R$ and any $k \in \mathbb{N}$,

$$|\lambda_k(t) - \lambda'_k(t)\ |< d_B(D_i(\mathbb{X}), D_i(\mathbb{X}')).$$

More on landscapes

Definition (The p-norm)

Let $D_i(\mathbb{X})$ be i-dimensional PD, its p-norm is:

$$||D_i(\mathbb{X})||_p = \left(\sum_{k=1}^{\infty} \int_{\mathbb{R}} |\lambda_k(t)|^p dt \right)^{\frac{1}{2}}.$$

- The main advantage is that landscapes form a vector space.
- The notions of distance and norm are present.
- Equipped with these tools one can compare shapes of PCDs.

Persistence Entropy

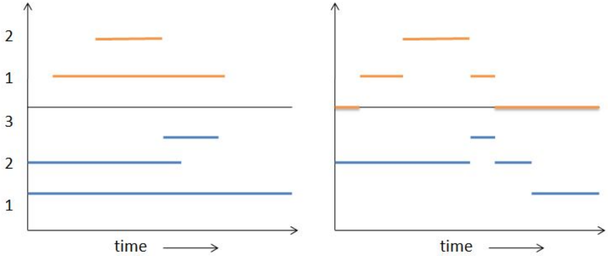
Let the persistence diagram be represented by $D = \{(b_j, d_j)\}_{j \in I}$, where I is the set of all points. The length of each bar is $l_i = d_i - b_i$. Let $L = \sum_i l_i$ denote the total length. Persistent Entropy: The persistent entropy of the barcode is the Shannon entropy of the lengths of the bars.

$$PE(D) = \frac{1}{L} \sum l_i \log(\frac{l_i}{L}) \tag{1}$$

This gives a measure of how similar the length of the barcodes are with the maximum entropy of persistent diagram achieved when all bars are equal.

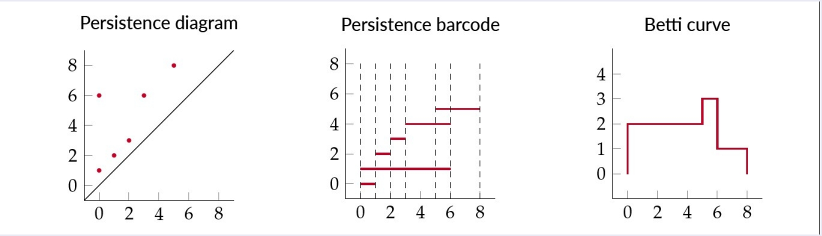
The Betti curve

The Betti curve is a real valued function defined on the set of parameter values. At each point, its value is the number of bars that contain this point. The L^p norm of these curves are considered.



(Left) Persistent Barcode; (Right) Betti Curve

The Betti curve

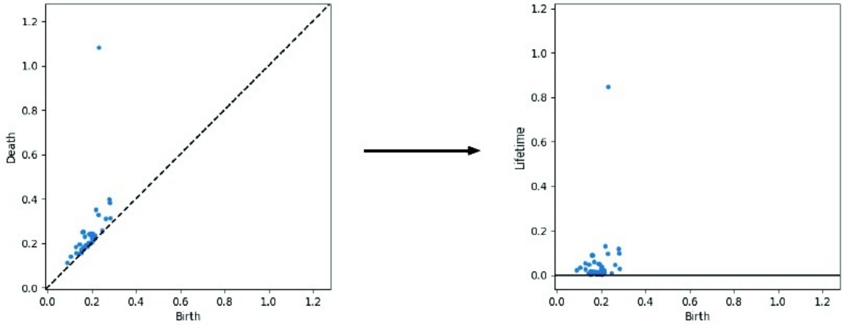


- Easy to calculate.
- Simple representation: a piecewise linear function.

Persistence images

Step 1: The lifetime representation

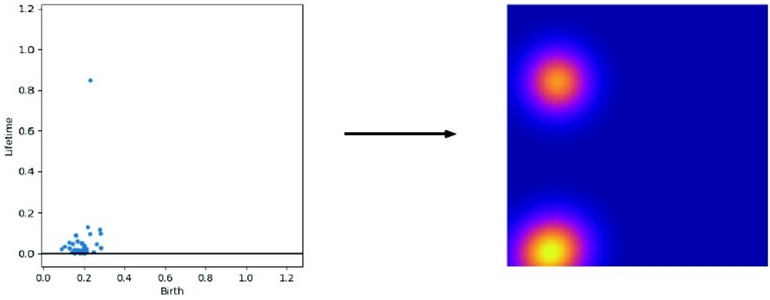
Change the coordinates from $(b, d) \mapsto (b, d - b)$.



Persistence images

Step 2: Heat map

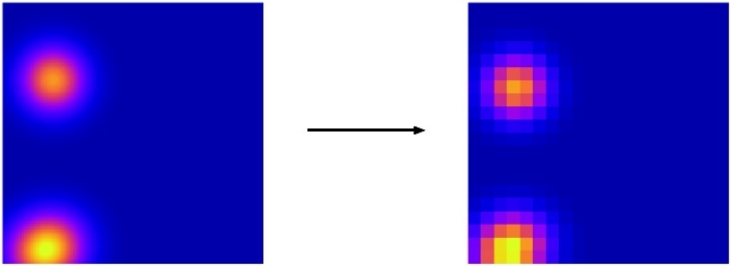
1. Each cycle in the PD is the center of a symmetric Gaussian.
2. Sum the Gaussians to get a real-valued function.
3. Multiply by a weight function, say $w(x, y) = y$.



Persistence images

Step 3: vectorization

Pixelate the image: slice the domain into a grid, then take the average of the function over each square.



PI details

1. For each point (u, v) in the original PD the Gaussian distribution is

$$g(x, y) = \frac{1}{2\pi\sigma^2} e^{-[(x-u)^2+(y-v)^2]/2\sigma^2}$$

2. Now the persistence surface from the transformed PD:

$$\rho(x, y) := \sum_{u \in T(B)} w(u) g_u(x, y).$$

3. The image value at pixel p is:

$$I(\rho)_p := \int \int_p \rho(x, y) dy dx$$

Where is the vector?

1. A persistence diagram B is mapped to an integrable function $\rho_B : \mathbb{R}^2 \to \mathbb{R}$.
2. The function ρ_B is called the persistence surface.
3. Discretize a subdomain of ρ_B to define a grid.
4. Create a matrix of pixel values by computing the integral of ρ_B on each grid.
5. This matrix is the desired vector; it is called the persistence image.

Advantages of PI

- PI is stable w.r.t. input noise.
- Computationally efficient.
- PI maintains an interpretable connection to the original PD.
- PI allows one to adjust the relative importance of points in different regions of the PD.
- PI is an intuitive description in terms of density estimates.
- Easy to use in a classification setting.
- However, parameter choices are hard.
- Not necessarily a sparse representation.

Other approaches

- Wasserstein amplitude of order p is the L_p norm of the vector of point distances to the diagonal.
- A vector obtained by rearranging the entries of the distance matrix between points in a PD.
- A vector obtained by superimposing a grid over PD and counting the number of points in each bin.
- First produce a surface from a PD by taking sum of a positive Gaussian centered at each point together with negative Gaussian centered on its reflection below the diagonal.

Outline

Stats for non-vector data

- Let Ω be a data set from which certain finite data points are obtained.
- To calculate statistical summaries, the set Ω is desired to have structures of addition, scalar multiplication and even inner product.
- The space of PDs is not an inner product space.
- If we can define a 'nice' map

$$\phi : \Omega \to \mathcal{H}$$

where $\mathcal{H}$ is a Hilbert space then we can calculate summaries and ML models from the inner product

$$< \phi(x_i), \phi(x_j) > .$$

The kernel method: basics

Definition

Let Ω be a set, a function $k : \Omega \times \Omega \to \mathbb{R}$ is called a positive definite kernel if:

- $k(x, y) = k(y, x)$,
- for any $x_1, \ldots, x_n \in \Omega$, the matrix (called the Gram matrix) $[k(x_i, x_j)]$ is positive semi-definite.

Example

Let $\Omega = \mathbb{R}^n$:

- Linear kernel: $< x, y >$.
- Polynomial kernel: $(< x, y > +c)^n$.
- Gaussian kernel: $e^{-\frac{||x-y||^2}{2\sigma^2}}$.

Theorem (Reporducing kernel Hilbert space)

A p.s-d. kernel uniquely defines a Hilbert space $\mathcal{H}$ satisfying

- *for any $x \in \Omega$ the function $k(\cdot, x) : \Omega \to \mathbb{R}$ is an element of $\mathcal{H}$,*
- *the span of $\{k(\cdot, x) : x \in \Omega\}$ is dense,*
- *for $x \in \Omega$ and $f \in \mathcal{H}$, $< f, k(\cdot, x) >= f(x)$.*

Given a data set Ω and a kernel k use the Gram matrix to construct the corresponding RKHS. If k has additional differentiable properties then the RKHS embeds in the space of signed Radon measures.
Conclusion: One can talk about probability distributions on Ω.

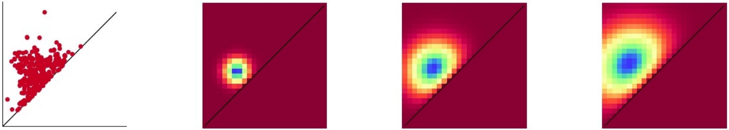
Example

Stable multi-scale kernel of Reininghaus et al.

For two PDs B, D we have

$$k(B,D) := \frac{1}{8\pi\sigma} \sum_{p\in B, q\in D} \exp\left(\frac{||p-q||^2}{8\sigma}\right) - \exp\left(\frac{||p-\overline{q}||^2}{8\sigma}\right).$$

Gaussians of standard deviation σ are placed over every point of B and a -ve Gaussian of σ over the mirror image of the point across the diagonal. The output of this operation is a real-valued function on $\mathbb{R}^2$.



Persistence diagram $\quad \sigma = 0.1 \quad \sigma = 0.5 \quad \sigma = 1.0$

Other kernels

- Kernel based on sliced Wasserstein distance by by Carriere et al. (PMRL 2017)
- Kernel embeddings method by Kusano et al. (JMLR 2018)
- Kernel based on Riemannian geometry by Le et al. (ANIPS 2018)
- Kernels on Betti curves by Rieck et al. (arXiv 1907.13496)