Persistence paths and signature features in topological data analysis (TDA)

by Chevyrev, Nanda and Oberhauser

$$\text{TDA} \xrightarrow{\text{vectorization}} \text{ML}$$

More specifically:

$$\text{Met} \xrightarrow{①} \text{PD} \xrightarrow{②} \text{Path} \xrightarrow{③} \text{Banach}$$

(persistence diagrams)

(complete metric v.s)

good for feature vectors !

Outline:

1. Persistent Homology
2. Path embeddings
3. Path signatures
4. Results

① Let X_t $t \in \mathbb{R}$ be a family of top.sp. such that

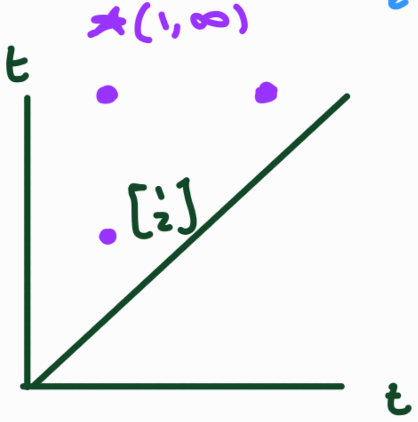
$$X_i \hookrightarrow X_j \quad \text{if} \quad i<j$$

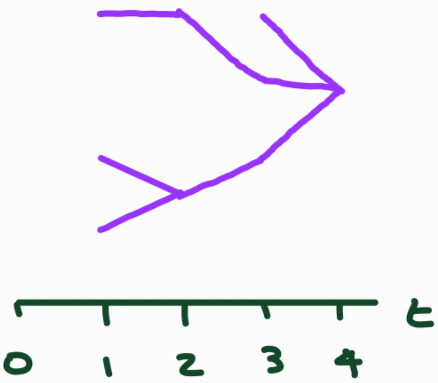
Let $H(-)$ be your favorite topological invariant (e.g. connected components)

Persistent homology consists on keeping track of $H(X_t)$ as t varies.

Example:

Hmm... the farther from diagonal the 'longer they persist'.





Persistence Diagram (PD)

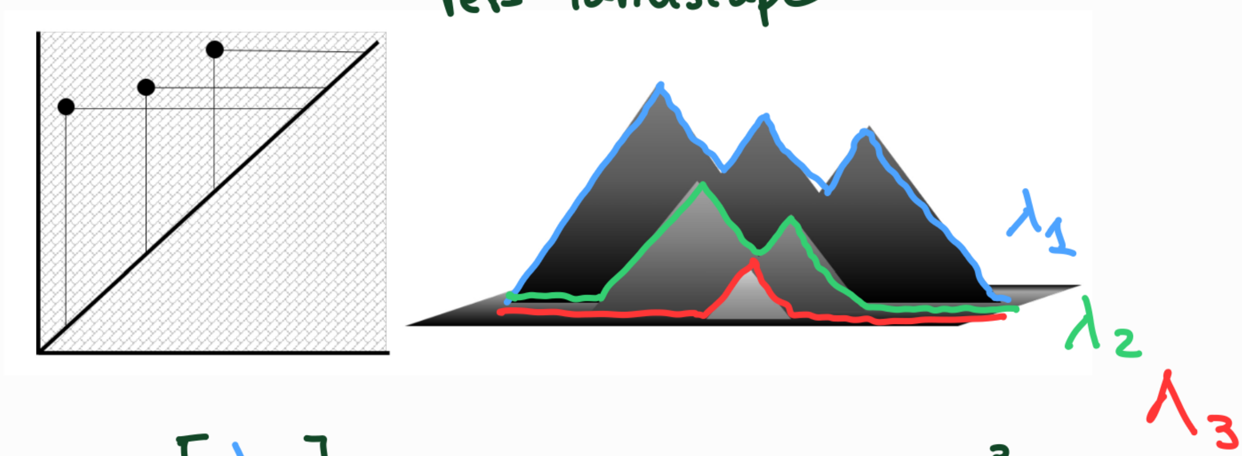
Rmk: PD are bad for stats/ML!

integrated landscape emb.

② From PD to Path^{BV}

* They describe 4 options, I will focus on 1.

Pers landscape



$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} : [0, T] \longrightarrow \mathbb{R}^3$$

aha! a (piecewise linear) path!

$$z_{iL}(t) = \int_0^t \lambda$$

$$\boxed{\text{Met} \xrightarrow{①} \text{PD} \xrightarrow{②} \text{Path} \xrightarrow{③} \text{Banach}}$$

③ Path signatures

Recall: Tensor algebra $(\mathbb{R}, v, \text{mat}, \text{rubiks}, \dots)$

$$T(V) = \prod_{i=0}^{\infty} \underbrace{V \otimes \cdots \otimes V}_{i\text{-times}}$$

V ↳ vector space

$$\begin{array}{c} \vdots \\ V \otimes V \otimes V \\ V \otimes V \\ V \\ V^0 := \mathbb{R} \end{array}$$

Def: Let $V = \mathbb{R}^n$ and $\{e_1, \dots, e_n\}$ basis.

The path sig of $\lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix}$ is

$$S(\lambda) = (S_0(\lambda), S_1(\lambda), \dots)$$

where

$S_0(\lambda) = 1$.

$$S_1(\lambda) = \int_0^{\text{end}} d\lambda_1 \; e_1 + \dots + \int_0^{\text{end}} d\lambda_n \; e_n$$

$$S_2(\lambda) = \sum_{i_1, i_2 \in \{1, \dots, n\}} \int_0^{\text{end}} \int_0^{t} d\lambda_{i_1}(x) \, d\lambda_{i_2}(y) \; e_{i_1} \otimes e_{i_2}$$

$\vdots$

Properties

I. $S(\lambda) = S(\gamma)$ iff λ and γ are tree-like equivalent

$\cong_{te}$

II. $S(-)$ is reparametrization invariant

III. If λ is 'a segment' then $\lambda = t \cdot v$

$$S(\lambda) = \left(1, v, \frac{v^{\otimes 2}}{2!}, \frac{v^{\otimes 3}}{3!}, \dots\right)$$

[think 'moments']

IV. Behaves well with concat.

$$S(\gamma * \lambda) = S(\gamma) \otimes S(\lambda).$$

AlgTop (Chen 60's) $\rightsquigarrow$ SDE (Lyons 90's) $\rightsquigarrow$ Data (now)

4 Results

Bar = PD

THEOREM 5. *Define*

$$\Phi : \mathbf{Bar}/\iota \to \mathbf{T}(V), \quad B \mapsto \mathrm{S} \circ \iota(B).$$

On each compact subset $K \subset \mathbf{Bar}/\iota$, the map Φ has the following properties.

(1) **(Universal)** *Let $f : K \to \mathbb{R}$ be continuous. For each $\epsilon > 0$, there exists ℓ in $\bigoplus_{m\geq 0}(V')^{\otimes m}$ (the dual space of the tensor algebra) such that*

$$\sup_{B\in K} |f(B) - \langle \Phi(B), \ell \rangle| < \epsilon.$$

(2) **(Characteristic)** *Denoting by $\mathcal{M}$ the set of Borel probability measures on K, the map*

$$\mathcal{M} \to \mathbf{T}(V), \quad \mu \mapsto \mathbb{E}_{B\sim\mu}[\Phi(B)]$$

is injective.

(3) **(Kernelized)** *Suppose further that V is a Hilbert space. Then the map*

$$k : K \times K \to \mathbb{R}, \quad k(B, B') = \langle \Phi(B), \Phi(B') \rangle$$

defines a bounded, continuous kernel[6] *which is universal for the space of continuous functions $C(K, \mathbb{R})$ and characteristic for Borel probability measures on K.*

$$\text{Met} \xrightarrow{\;1\;} \text{PD} \underset{*}{\xrightarrow{\;2\;}} \text{Path} \xrightarrow{\;3\;} \underset{*}{\text{Banach}}$$

embed

trunc

kernel Y/N.

TABLE 1. Mean accuracy ($\pm$ standard deviation).

Method	Textures	Orbits	Shapes
k_{SW}	96.8 ± 1.0	94.6 ± 1.3	95.8 ± 1.6
Φ_{PI}	93.7 ± 1.0	$\mathbf{99.86 \pm 0.21}$	90.3 ± 2.3
k_{E}	90.4 ± 1.5	96.6 ± 0.9	92.7 ± 1.5
k_{χ}	94.9 ± 0.6	NA	92.4 ± 3.0
k_{β}	$\mathbf{97.8 \pm 0.2}$	NA	93.0 ± 3.0
Φ_{E}	88.1 ± 0.8	98.1 ± 1.0	95.0 ± 0.9
Φ_{χ}	92.9 ± 0.7	98.8 ± 0.6	98.0 ± 1.1
Φ_{β}	96.6 ± 0.6	97.7 ± 0.8	$\mathbf{98.1 \pm 0.7}$

TABLE 2. Best truncation level M.

M	Textures	Orbits	Shapes
k_{E}	2	2	2
k_{χ}	3	NA	2
k_{β}	3	NA	2
Φ_{E}	5	8	4
Φ_{χ}	8	7	7
Φ_{β}	6	8	5

Shameless plug:

Discrete signature tensors for persistence landscapes

Vincenzo Galgano, Heather A. Harrington, Daniel Tolosa

Abstract

Signature tensors of paths are a versatile tool for data analysis and machine learning. Recently, they have been applied to persistent homology, by embedding barcodes into spaces of paths. Among the different path embeddings, the persistence landscape embedding is injective and stable, however it loses injectivity when composed with the signature map. Here we address this by proposing a discrete alternative. The critical points of a persistence landscape form a time-series, of which we compute the discrete signature. We call this association the *discrete landscape feature map* (DLFM). We give results on the injectivity, stability and computability of the DLFM. We apply it to a knotted protein dataset, capturing sequence similarity and knot depth with statistical significance.

Keywords: persistent homology, barcodes, persistence landscapes, feature maps, time-series, time warping, path signatures, tensors, knotted proteins, vectorisation.

MSC2020 codes: 55N31, 68T09, 46B85.

Acknowledgements. We thank Agnese Barbensi, Darrick Lee, Vidit Nanda and Leon Renkin for helpful discussions. We also thank MPI-MiS Leipzig for hosting the conference "MEGA 2024", which informed this research and led to new connections with Francesco Galuppi and Pierpaola Santarsiero. We are grateful to MPI-CBG Dresden and CSBD for the excellent working conditions. HAH gratefully acknowledges funding from the Royal Society RGF/EA/201074, UF150238 and EPSRC EP/R018472/1, EP/Y028872/1 and EP/Z531224/1. VG is member of the Italian national group GNSAGA-INdAM.

Introduction

As the scale and complexity of biological data increases, problems such as interpretation, classification and quantification require advanced mathematical methods for investigation. Applying persistent homology from topological data analysis to biological data provides a geometric interpretation of the shapes of data. Here, we propose an alternative approach combining persistent homology and nonlinear algebra.