Statistical topological data analysis using persistence landscapes

Peter Bubenik

Cleveland State University

February 3, 2014

funded by AFOSR
TDA attempts to recover topological and geometric information from sampled data.
radius = 1
radius = 2
radius = 3
radius = 4
radius = 5
Motivation

Persistence landscape

radius = 6
radius = 7
Motivation

Persistence landscape

radius = 8
radius = 9
radius = 11
Motivation

Persistence landscape

Degree 1 persistent homology: $\{(3, 9), (4, 6), (5, 11)\}$
Mathematical viewpoint

For each radius $r$, have
- a simplicial complex $S_r(X)$
- a vector space $H(S_r(X))$

For $r \leq r'$, have
- the inclusion $S_r(X) \subseteq S_{r'}(X)$
- a linear map $H(S_r(X)) \to H(S_{r'}(X))$

**Persistent homology** is the image of this map.

This set of vector spaces and linear maps is called a **persistence module**.

It has a complete discrete invariant: $\{(\text{birth}_j, \text{death}_j)\}$.

There exist good algorithms. This summary is stable.
The topological summary as a random variable:

\[(\Omega, \mathcal{F}, \mathcal{P}) \xrightarrow{TS} (SS, A, \mathcal{P}_*)\]

\[X \xrightarrow{TS} TS(X)\]
Motivation
Persistence landscape

Challenges

Goal: use topological summaries to make inferences.

We want to:
- construct summaries
- compare summaries
- average summaries
- use summaries for hypothesis testing

and do so efficiently.

My approach: the persistence landscape
Recall that the persistence module consisted of linear maps

\[ H(S_r(X)) \rightarrow H(S_{r'}(X)), \text{ for } r \leq r'. \]

The ranks of these maps gives us a function from \( \mathbb{R}^2 \) to \( \mathbb{R} \).
Persistence landscape

Peter Bubenik

Persistence landscapes
Persistence landscape
Persistence landscape
Persistence landscape
Properties

Lemma

- $\lambda_k(t) \geq 0$
- $\lambda_k(t) \geq \lambda_{k+1}(t)$
- $\lambda_k$ is 1-Lipschitz
Consider $\lambda_1, \lambda_2, \lambda_3, \ldots$ as

$$\lambda : \mathbb{N} \times \mathbb{R} \rightarrow \mathbb{R}.$$ 

Then

- $\|\lambda\|_{\infty} = \|\lambda_1\|_{\infty}$, and
- for $1 \leq p < \infty$, $\|\lambda\|_p = \left(\sum_k \int \lambda_k^p\right)^{\frac{1}{p}}$. 

![Diagram of persistence landscapes](image-url)
For a persistence landscape $\lambda$, let $(b_j, d_j)$ be the corresponding birth-death pairs.

**Lemma**

1. $\|\lambda\|_{\infty} = \frac{1}{2} \max_j (d_j - b_j)$, and
2. $\|\lambda\|_1 = \frac{1}{4} \sum_j (d_j - b_j)^2$. 
Persistence landscapes, \( \lambda^{(1)}, \ldots, \lambda^{(n)} \), have mean, \( \overline{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \lambda^{(i)} \).

That is,
\[
\overline{\lambda}_k(t) = \frac{1}{n} \sum_{i=1}^{n} \lambda^{(i)}_k(t)
\]

Interpretation: This is the average value of the largest \( h \) such that
\[
H(S_{t-h}(X)) \rightarrow H(S_{t+h}(X))
\]
has rank at least \( k \).
Mean diagram vs mean landscape
Linked annuli
Linked annuli
Linked annuli

Persistence landscape

Motivation
Definition
Properties
Mean
Hypothesis testing
Stability

Peter Bubenik
Persistence landscapes
Recall \( \| \lambda \|_p = \left( \sum_k \int \lambda_k^p \right)^{\frac{1}{p}} \).

Let \( 1 \leq p < \infty \). We assume \( \| \lambda \| := \| \lambda \|_p < \infty \). That is, \( \lambda \in L^p(\mathbb{N} \times \mathbb{R}) \).

So \( \lambda \) is a random variable with values in a separable Banach space.
\( \lambda \in L^p(\mathbb{N} \times \mathbb{R}), \quad \|\lambda\| \) is a real random variable.

If \( E\|\lambda\| < \infty \) then there exists \( E(\lambda) \in L^p(\mathbb{N} \times \mathbb{R}) \) such that \( E(f(\lambda)) = f(E(\lambda)) \) for all continuous linear functionals \( f \).

For \( X_1, \ldots, X_n \) be an iid sample, and let \( \lambda^{(1)}, \ldots, \lambda^{(n)} \) be the corresponding persistence landscapes.

**Theorem (Strong Law of Large Numbers)**

\[ \bar{\lambda}^{(n)} \rightarrow E(\lambda) \text{ almost surely if and only if } E\|\lambda\| < \infty. \]
Central limit theorems

Theorem (Central Limit Theorem in $L^p(\mathbb{N} \times \mathbb{R})$)

Assume $p \geq 2$. If $E\|\lambda\| < \infty$ and $E(\|\lambda\|^2) < \infty$ then

$$\sqrt{n}[\lambda^{(n)} - E(\lambda)] \text{ converges weakly to a Gaussian random variable with the same covariance structure as } \lambda.$$  

Corollary (Practical Central Limit Theorem)

For any $f \in L^q(\mathbb{N} \times \mathbb{R})$ with $\frac{1}{p} + \frac{1}{q} = 1$, let

$$Y = \int_{\mathbb{N} \times \mathbb{R}} f \lambda. \quad (1)$$

Then

$$\sqrt{n}[Y_n - E(Y)] \xrightarrow{d} N(0, \text{Var}(Y)). \quad (2)$$
Mean landscapes for Gaussian Random Fields

Persistence landscapes

Peter Bubenik

Motivation
Persistence landscape
Definition
Properties
Mean
Hypothesis testing
Stability
Mean landscapes for Gaussian Random Fields

Peter Bubenik

Persistence landscapes
Topological hypothesis testing

Motivation
Persistence landscape
Definition
Persistence landscapes
Topological hypothesis testing

Points $\rightarrow$ kernel density estimator $\rightarrow$ filtered simplicial complex
Topological hypothesis testing
Topological hypothesis testing

Null hypothesis: \( \| \bar{\lambda}_S \|_1 = \| \bar{\lambda}_T \|_1 \).

Student’s t-test:

<table>
<thead>
<tr>
<th>dim</th>
<th>decision</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>cannot reject</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>reject</td>
<td>( 3 \times 10^{-6} )</td>
</tr>
<tr>
<td>2</td>
<td>cannot reject</td>
<td></td>
</tr>
</tbody>
</table>
Topological hypothesis testing, noisy
Topological hypothesis testing, noisy
Topological hypothesis testing, noisy

Null hypothesis: \( \| \bar{\lambda}_S - \bar{\lambda}_T \|_2 = 0 \).

Permutation test:

<table>
<thead>
<tr>
<th>dim</th>
<th>decision</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>reject</td>
<td>0.0111</td>
</tr>
<tr>
<td>1</td>
<td>reject</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>reject</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Topological hypothesis testing, noisy

Topological hypothesis testing, noisy
Stability

Given \( f : X \rightarrow \mathbb{R} \),
let \( \lambda(f) \) the persistence landscape of sublevel sets of \( f \).

**Theorem (Landscape stability theorem)**

Let \( f, g : X \rightarrow \mathbb{R} \).

\[
\|\lambda(f) - \lambda(g)\|_{\infty} \leq \|f - g\|_{\infty}.
\]

If \( X \) is nice and \( f \) and \( g \) are tame and Lipschitz then

\[
\|\lambda(f) - \lambda(g)\|_p \leq C \|f - g\|_{\infty}^{p-k}.
\]