Persistent homology and statistical inference: Peristence Landscapes

Peter Bubenik

Cleveland State University

January 5, 2012

Joint Mathematical Meetings 2012
Boston, MA

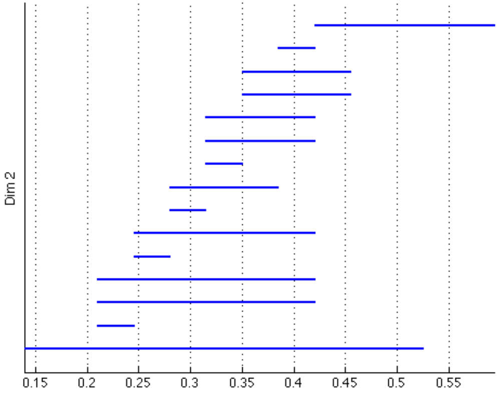
An applied topology pipeline

$$\text{Raw data} \xrightarrow[\text{preprocess}]{\text{Encode and}} \text{Cleaned data}$$

$$\xrightarrow{\text{Geometry}} \text{Geometric object} \xrightarrow[\text{topology}]{\text{Algebraic}} \text{Algebraic diagram}$$

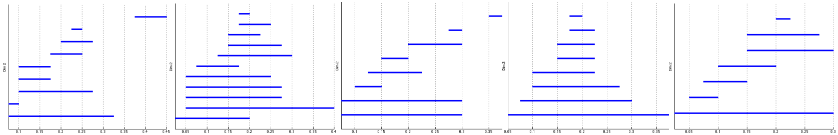
$$\xrightarrow[\text{homology}]{\text{Persistent}} \text{Descriptor}$$

The usual descriptor



Statistical applied topology

Suppose we have calculated a sequence of descriptors:

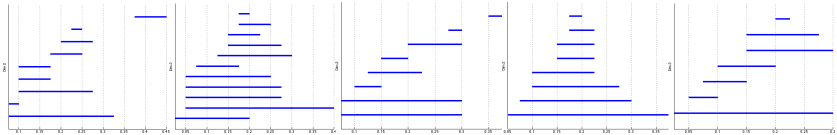


Question

What is the mean? What is the standard deviation? Can we use it for hypothesis testing?

Statistical applied topology

Suppose we have calculated a sequence of descriptors:



Question

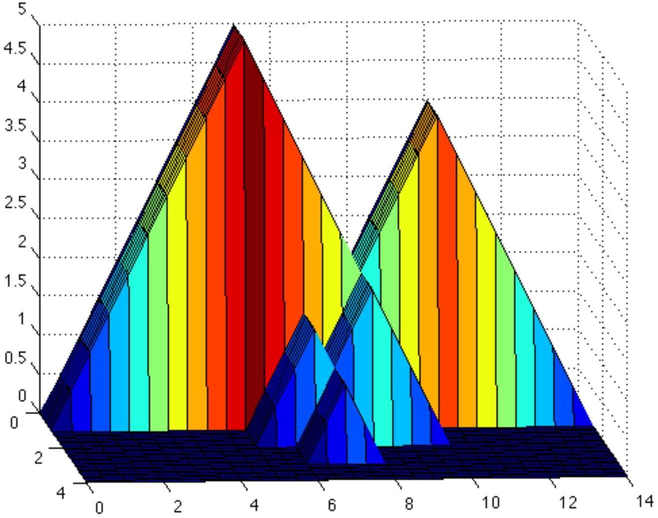
What is the mean? What is the standard deviation? Can we use it for hypothesis testing?

A solution

We define a new descriptor: *the peristence landscape*.

It is a function from $\mathbb{R}^2$ to $\mathbb{R}$.

Persistence Landscape example



Persistence Landscape

Let M be a vector space diagram indexed by $\mathbb{R}$.
For all $a \leq b$, we have a linear transformation $M(a) \to M(b)$.
Define $\beta^{a,b}(M) = \dim(\mathrm{im}(M(a) \to M(b)))$.

Definition

$$\lambda(M)(x,t) = \begin{cases} \sup(s \mid \beta^{t-s,t+s}(M) \geq x) & \text{if } 0 < x \leq \dim M(t) \\ 0 & \text{otherwise.} \end{cases}$$

Persistence Landscape

Let M be a vector space diagram indexed by $\mathbb{R}$.
For all $a \leq b$, we have a linear transformation $M(a) \to M(b)$.
Define $\beta^{a,b}(M) = \dim(\mathrm{im}(M(a) \to M(b)))$.

Definition

$$\lambda(M)(x,t) = \begin{cases} \sup(s \mid \beta^{t-s,t+s}(M) \geq x) & \text{if } 0 < x \leq \dim M(t) \\ 0 & \text{otherwise.} \end{cases}$$

Theorem

Let $\lambda_{\mathcal{B}}$ denote the persistence landscape corresponding to a barcode $\mathcal{B}$. Then,

$$||\lambda_{\mathcal{B}}||_2^2 = \frac{1}{12}\,\mathrm{pers}_3(\mathcal{B}).$$

The space of persistence landscapes

Definition

Let PL be the set of functions that "look like persistence landscapes" (a "decreasing stack of 1-Lipschitz functions"). Let $\mathrm{PL}^2 = \mathrm{PL} \cap L^2$ together with the metric from L^2.

The space of persistence landscapes

Definition

Let PL be the set of functions that "look like persistence landscapes" (a "decreasing stack of 1-Lipschitz functions"). Let $\mathrm{PL}^2 = \mathrm{PL} \cap L^2$ together with the metric from L^2.

Theorem

PL^2 *is complete and seperable, i.e. it is a Polish space.*

Fréchet mean, variance

Definition

Given $x_1, \ldots, x_n$ in a metric space. The Fréchet mean is the point y that minimizes

$$\sum_{i=1}^{n} d(y, x_i)^2.$$

The corresponding sum is called the Fréchet variance.

Replacing summation with integration we define the Fréchet mean and variance of a probability measure.

Fréchet mean, variance

Definition

Given $x_1, \ldots, x_n$ in a metric space. The Fréchet mean is the point y that minimizes

$$\sum_{i=1}^{n} d(y, x_i)^2.$$

The corresponding sum is called the Fréchet variance.

Replacing summation with integration we define the Fréchet mean and variance of a probability measure.

Theorem

In PL^2, *the Fréchet mean is given by the pointwise mean, and the Fréchet variance is the integral of the pointwise variances.*

Estimator for the Fréchet mean

Let $\lambda_1, \ldots, \lambda_n$ be a sample of persistence landscapes drawn from a probability measure with Fréchet mean h.
Let $\bar{\lambda}_n$ be the pointwise mean of the sample.

Estimator for the Fréchet mean

Let $\lambda_1, \ldots, \lambda_n$ be a sample of persistence landscapes drawn from a probability measure with Fréchet mean h.
Let $\bar{\lambda}_n$ be the pointwise mean of the sample.

For all x, t,

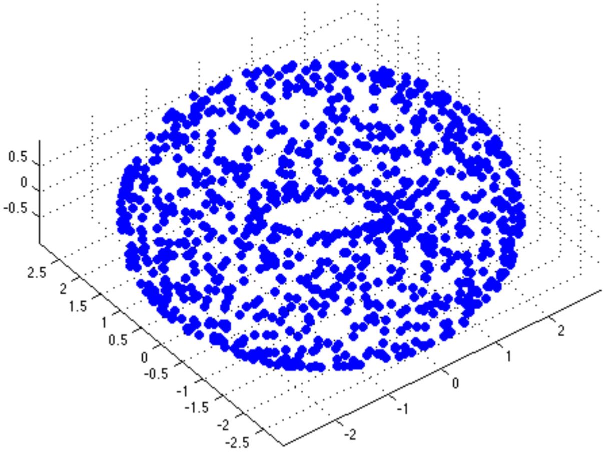
Strong Law of Large Numbers

$$\bar{\lambda}_n(x,t) \xrightarrow{a.s.} h(x,t).$$

Central Limit Theorem

$$\sqrt{n}\left(\bar{\lambda}_n(x,t) - h(x,t)\right)/\sigma \xrightarrow{d} N(0,1).$$

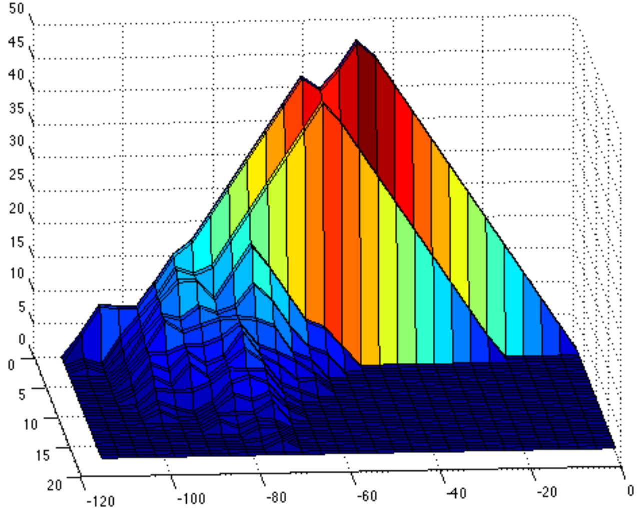
Torus: point cloud



Torus: filtered simplicial complex

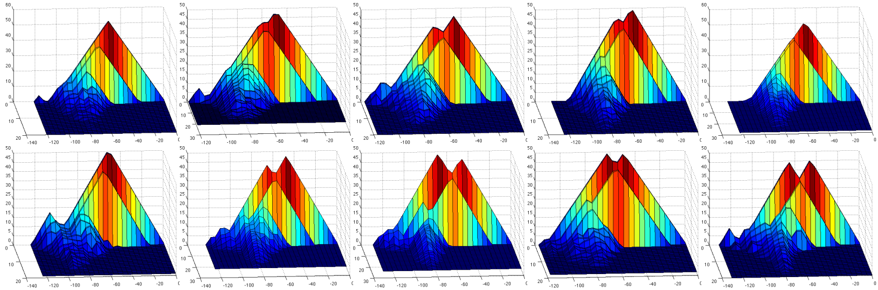
see rendering from `plex_viewer`

Torus: persistence landscape in dimension 1



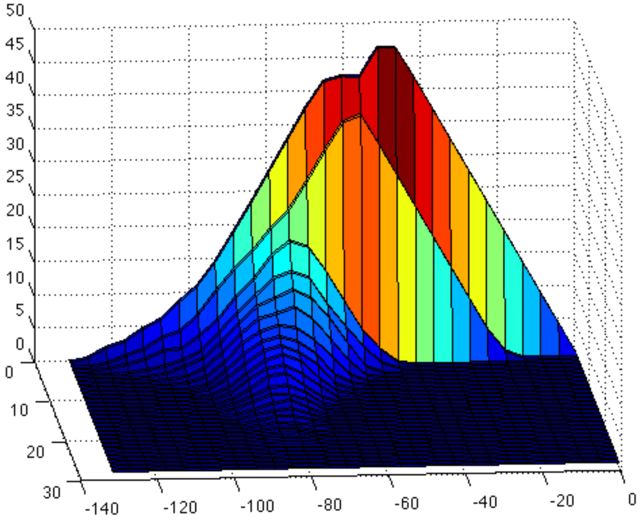
Torus: persistence landscape in dimension 1

Now repeat 10 times.



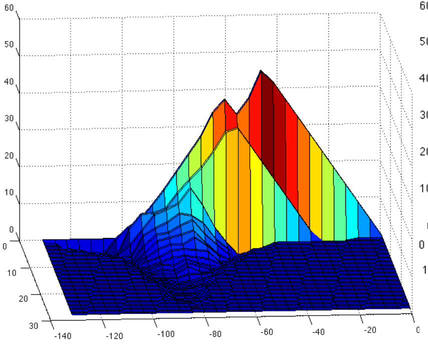
Average pointwise.

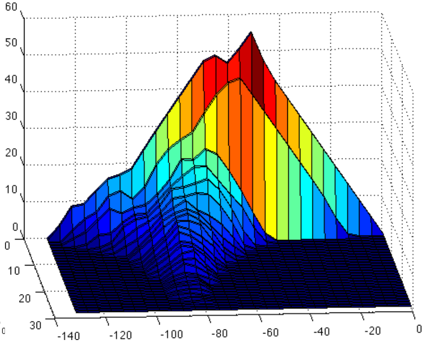
Torus: average persistence landscape in dimension 1



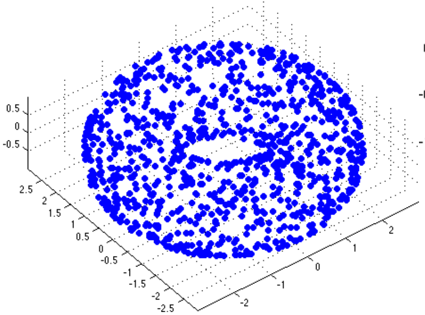
Torus: average persistence landscape in dimension 1

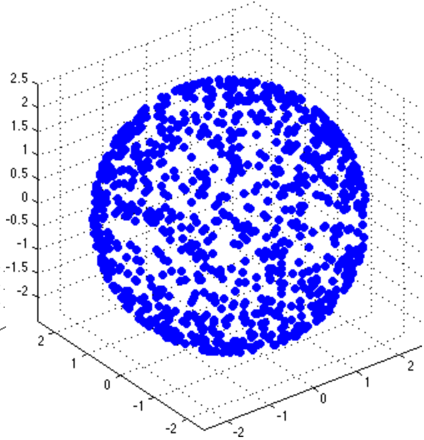
± 2 standard deviations



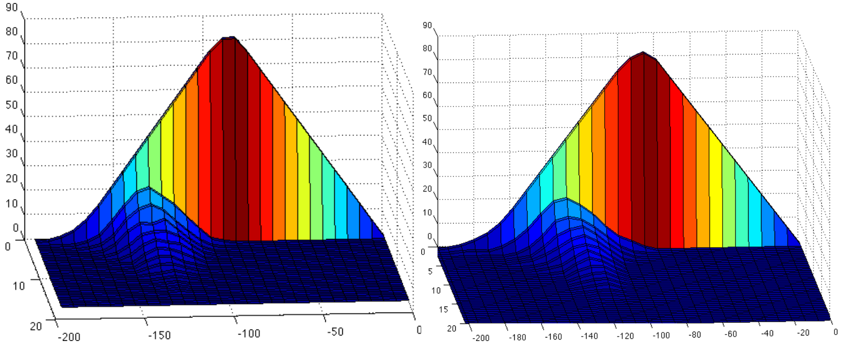


Torus vs Sphere: point clouds

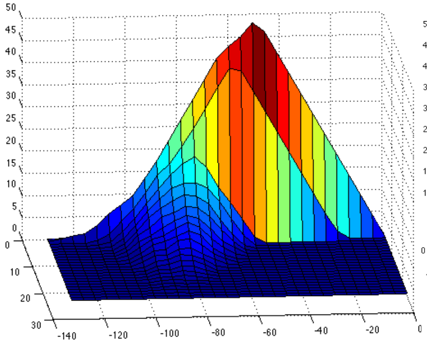


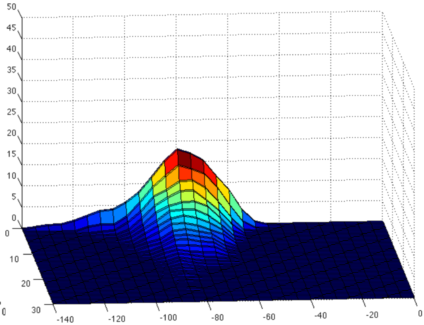


Torus vs Sphere: average persistence landscapes dim 0

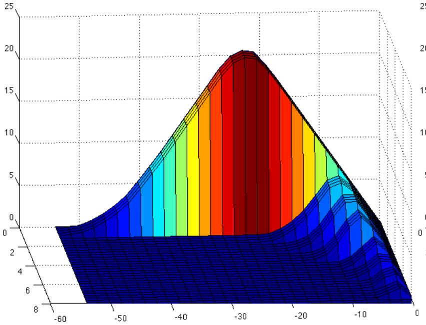


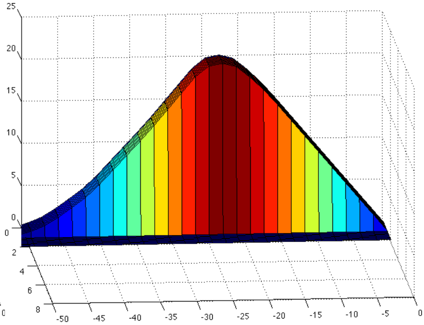
Torus vs Sphere: average persistence landscapes dim 1





Torus vs Sphere: average persistence landscapes dim 2





Torus vs Sphere

Question

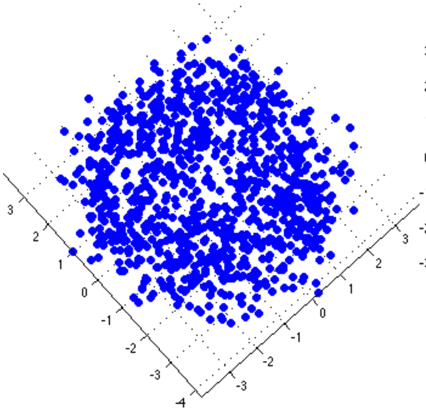
Is there a statistically significant difference?

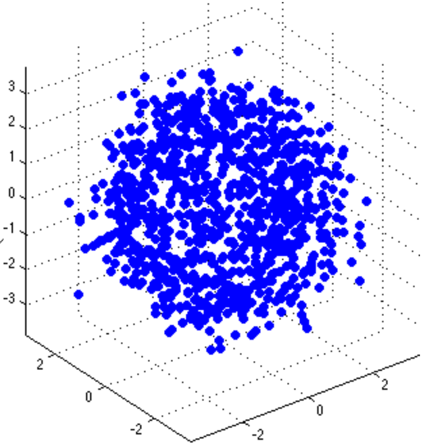
Answer

Using a two sample t test at the 0.01 significance level:

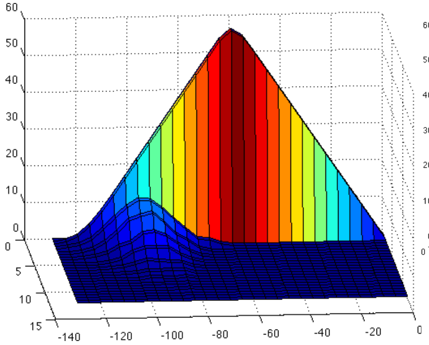
	dim 0	*dim 1*	*dim 2*
peak 1	$\boxed{\textit{no}}$	$\boxed{\textit{yes}}$	$\boxed{\textit{no}}$
peak 2	*no*	$\boxed{\textit{yes}}$	*yes*

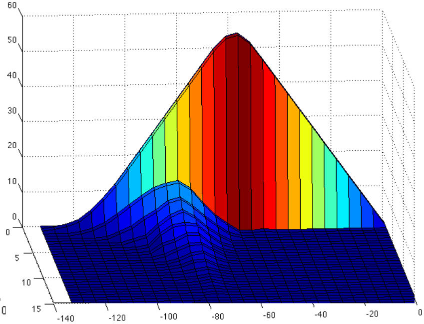
Noisy Torus vs Sphere: point clouds



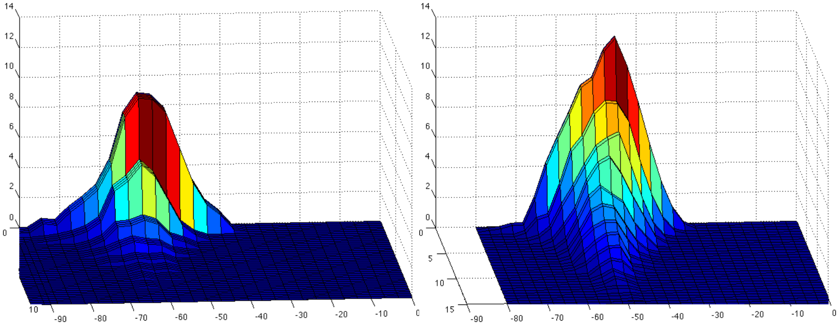


Noisy Torus vs Sphere: avg persistence landscapes dim 0

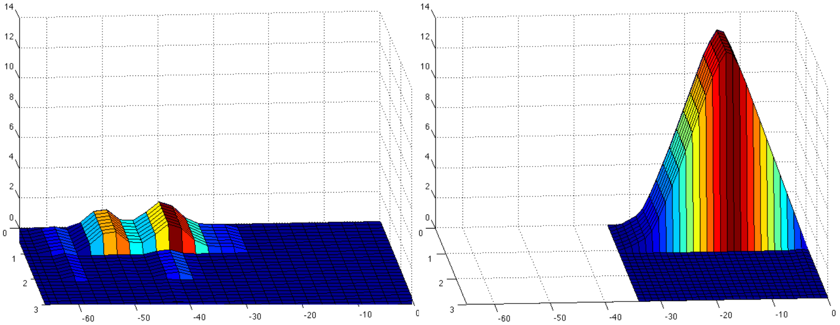




Noisy Torus vs Sphere: avg persistence landscapes dim 1



Noisy Torus vs Sphere: avg persistence landscapes dim 2



Noisy Torus vs Noisy Sphere

Question

Is there a statistically significant difference?

Answer

Using a two sample t test at the 0.01 significance level:

	dim 0	dim 1	dim 2
peak 1	$\boxed{no}$	$\boxed{yes}$	$\boxed{yes}$
peak 2	no	$\boxed{yes}$	no

Future directions

- Stability: $||\lambda(M) - \lambda(N)||_2 \leq K \ d(M, N)$?
- Fréchet median
- Uniform convergence, bootstrap
- Asymptotic persistence landscapes
- Applications!

Torus vs Sphere: p values

	dim 0	dim 1	dim 2
peak 1	0.8807	0.0000	0.7357
peak 2	0.7547	0.0000	0.0000
peak 3	0.4038	0.0002	0.0000
peak 4	0.7163	0.0032	0.0000
peak 5	0.2837	0.0000	0.0000

Noisy Torus vs Noisy Sphere: p values

	dim 0	dim 1	dim 2
peak 1	0.3246	0.0077	0.0000
peak 2	0.4967	0.0032	0.3217
peak 3	0.8132	0.0000	0.3306