Stochastic Convergence of Persistence Landscapes and Silhouettes

Brittany Terese Fasy
brittany.fasy@alumni.duke.edu

joint work with F. Chazal, F. Lecci, A. Rinaldo, L. Wasserman

Postdoctoral Researcher, Tulane University

11 June 2014
Sample (Bounded) Persistence Diagrams
Sample (Bounded) Persistence Landscapes
Persistence Landscapes
Persistence Landscapes
Persistence Landscapes

\[
\frac{(\text{Death} - \text{Birth})}{2}
\]

\[
\frac{(\text{Death} + \text{Birth})}{2}
\]
Persistence Landscapes

Motivation
Where Do the Diagrams Come From?

Scenario 1

Draw $n$ functions from a probability distribution over the set of (Morse) functions. This induces a sample of persistence landscapes. Ex: each function is the distance to a compact set embedded in $\mathbb{R}^d$. 
Motivation
Where Do the Diagrams Come From?

**Scenario 1**

Draw \( n \) functions from a probability distribution over the set of (Morse) functions. This induces a sample of persistence landscapes. Ex: each function is the distance to a compact set embedded in \( \mathbb{R}^d \).

**Scenario 2**

Given a large dataset with \( N \) points, it is very expensive to compute the persistence landscape \( \lambda \) exactly. Instead, we use subsampling to compute approximations \( \lambda_1, \lambda_2, \ldots, \lambda_n \). Then, we can upper bound \( \mathbb{E}[\lambda_i - \lambda] \) with

\[
\mathbb{E}[\lambda_i - \mu] + \mathbb{E}[\mu - \lambda]
\]
Motivation

Where Do the Diagrams Come From?

**Scenario 1**

Draw $n$ functions from a probability distribution over the set of (Morse) functions. This induces a sample of persistence landscapes. Ex: each function is the distance to a compact set embedded in $\mathbb{R}^d$.

**Scenario 2**

Given a large dataset with $N$ points, it is very expensive to compute the persistence landscape $\lambda$ exactly. Instead, we use subsampling to compute approximations $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then, we can upper bound $\mathbb{E}[\lambda_i - \lambda]$ with $\mathbb{E}[\lambda_i - \mu]$. 
Topological Inference
Pointwise Convergence of Landscapes

Let \( \lambda_1, \ldots, \lambda_n \sim \mathcal{L}_T \).

Key Properties
Landscapes are \( \left( \frac{T}{2} \right) \)-bounded and one-Lipschitz!
Pointwise Convergence of Landscapes

Let $\lambda_1, \ldots, \lambda_n \sim \mathcal{L}_T$.  
$\mu = \mathbb{E}(\lambda_i)$
Let \( \lambda_1, \ldots, \lambda_n \overset{\text{iid}}{\sim} \mathcal{L}_T \).

\[
\mu = \mathbb{E}(\lambda_i)
\]

\( \bar{\lambda}_n \) : empirical average landscape
Let $\lambda_1, \ldots, \lambda_n \overset{\text{iid}}{\sim} \mathcal{L}_T$.

$\bar{\lambda}_n$ converges pointwise to $\mu$.

Pointwise Convergence [B-2012].
Pointwise Convergence of Landscapes

Let $\lambda_1, \ldots, \lambda_n \overset{\text{iid}}{\sim} \mathcal{L}_T.
\mu = \mathbb{E}(\lambda_i)
\bar{\lambda}_n : \text{empirical average landscape}

Pointwise Convergence \cite{B-2012}.

$\bar{\lambda}_n$ converges pointwise to $\mu$.

Key Properties

Landscapes are $(T/2)$-bounded and one-Lipschitz!
Gaussian Process

A GP over $I$ is a set of independent random variables associated to each $t \in I$ such that every finite collection of random variables has a multi-variate normal distribution.
Gaussian Process

A GP over \( I \) is a set of independent random variables associated to each \( t \in I \) such that every finite collection of random variables has a multi-variate normal distribution.

Brownian Bridge

A Brownian Bridge \( B \) defined over \( I \) is a continuous GP over \( I \) with a nice covariance structure such that \( B(0) = B(1) = \mathbb{E}[B(i)] = 0. \)
Emperical Process

For \( t \in [0, T] \), we define

\[
G_n(f_t) = G_n(t) := \frac{1}{\sqrt{n}} (f_t(\bar{\lambda}_n) - f_t(\mu)).
\]
Emperical Process

For $t \in [0, T]$, we define $G_n(f_t) = G_n(t) := \frac{1}{\sqrt{n}}(f_t(\bar{\lambda}_n) - f_t(\mu))$. 

$f_t : \mathcal{L}_T \rightarrow \mathbb{R}$

$\lambda \mapsto \lambda(t)$

$\bar{\lambda}_n(t) - \mu(t)$
Emperical Process on $[0, T]$

$$f_t : \mathcal{L}_T \rightarrow \mathbb{R}$$
$$\lambda \mapsto \lambda(t)$$

$$\bar{\lambda}_n(t) - \mu(t)$$

$$f_t(\bar{\lambda}_n) - f_t(\mu)$$
Emperical Process

For \( t \in [0, T] \), we define \( G_n(f_t) = G_n(t) := \frac{1}{\sqrt{n}} \left( f_t(\bar{\lambda}_n) - f_t(\mu) \right) \).
Weak Convergence

\( \mathbb{G}_n(t) = \frac{1}{\sqrt{n}}(\bar{\lambda}_n(t) - \mu(t)) \) converges weakly to the Brownian bridge \( \mathbb{G} \) with covariance function

\[
\kappa(f, g) = \int f(u)g(u)dP(u) - (\int f(u)dP(u))(\int g(u)dP(u)).
\]
Uniform Convergence

Let \( \sigma(t) = \sqrt{n \ Var \tilde{\lambda}_n(t)} \).
Assume \( \sigma(t) > 0 \) on \([t_*, t^*]\) \(\subset [0, T]\).

Uniform CLT

There exists a random variable \( W \overset{d}{=} \sup_{t \in [t_*, t^*]} |G(f_t)| \) such that

\[
\sup_{z \in \mathbb{R}} \left\| \mathbb{P} \left( \sup_{t \in [t_*, t^*]} |G_n(t)| \leq z \right) - \mathbb{P} (W \leq z) \right\| = O\left( \frac{(\log n)^{\frac{7}{8}}}{n^{\frac{1}{8}}} \right).
\]
Confidence Band

A $(1 - \alpha)$-confidence band for $\mu$ is a pair of functions $\ell_n, u_n : [0, T] \rightarrow \mathbb{R}$ such that

$$\mathbb{P}(\ell_n(t) \leq \mu(t) \leq u_n(t) \text{ for all } t) \geq 1 - \alpha.$$
The Multiplier Bootstrap

Let $\xi_1, \ldots, \xi_n \sim N(0, 1)$. Then,

$$\widetilde{G}_n(f_t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i(\lambda_i(t) - \bar{\lambda}_n(t))$$

is the multiplier bootstrap version of $G_n(f_t)$.

$\alpha$-Quantile

$\widetilde{Z}_\alpha$ is the unique value such that

$$\mathbb{P}\left( \sup_t |\widetilde{G}_n(f_t)| > \widetilde{Z}_\alpha \bigg| \{\lambda_i\} \right) = \alpha$$
The Multiplier Bootstrap

Let $\xi_1, \ldots, \xi_n \sim \mathcal{N}(0, 1)$. Then,

$$\tilde{G}_n(f_t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i (\lambda_i(t) - \bar{\lambda}_n(t))$$

is the multiplier bootstrap version of $G_n(f_t)$.

$\alpha$-Quantile

$\tilde{Z}_\alpha$ is the unique value such that

$$\mathbb{P} \left( \sup_t |\tilde{G}_n(f_t)| > \tilde{Z}_\alpha \left| \{\lambda_i\} \right. \right) = \alpha$$

*Approx. $\tilde{Z}_\alpha$ by MC simulation
Recalling
\[ G_n(t) = \frac{1}{\sqrt{n}}(\bar{\lambda}_n(t) - \mu(t)), \]
let
\[ \ell_n = \bar{\lambda}_n(t) - \frac{\tilde{Z}(\alpha)}{\sqrt{n}}, \]
\[ u_n = \bar{\lambda}_n(t) + \frac{\tilde{Z}(\alpha)}{\sqrt{n}}. \]

**Uniform Band**

\[
P(\ell_n(t) \leq \mu(t) \leq u_n(t) \text{ for all } t) \geq 1 - \alpha - O\left(\frac{(\log n)^{7/8}}{n^{1/8}}\right).
\]
Variable Width Confidence Bands

\[ \mathbb{H}_n(f_t) := \mathcal{G}_n(t)/\sigma(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\lambda_i(t) - \mu(t)}{\sigma(t)} \]

\[ \tilde{\mathbb{H}}_n(f_t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \frac{\lambda_i(t) - \bar{\lambda}_n(t)}{\hat{\sigma}_n(t)} \]

\( \tilde{Q}_\alpha \) is the unique value such that

\[ \mathbb{P} \left( \sup_t |\tilde{\mathbb{H}}_n(f_t)| > \tilde{Q}_\alpha \mid \{\lambda_i\} \right) = \alpha \]
Variable Width Confidence Bands

\[ \ell_n = \bar{\lambda}_n(t) - \frac{\tilde{Q}(\alpha) \hat{\sigma}_n(t)}{\sqrt{n}} \]

\[ u_n = \bar{\lambda}_n(t) + \frac{\tilde{Q}(\alpha) \hat{\sigma}_n(t)}{\sqrt{n}} \]

Variable Band

\[ \mathbb{P}(\ell_n(t) \leq \mu(t) \leq u_n(t) \text{ for all } t) \geq 1 - \alpha - O\left(\frac{(\log n)^{7}}{n^{1/8}}\right). \]
Relax, that was the most technical part.
Persistence Silhouettes

Definitions

\[ \frac{(\text{Death} - \text{Birth})}{2} \]

\[ \frac{\text{Death} + \text{Birth}}{2} \]
Persistence Silhouettes

Definitions

\( \Lambda : \mathbb{R} \times \mathbb{Z}_+ \rightarrow \mathbb{R} \)
Persistence Silhouettes

Definitions
Silhouettes

Persistence Silhouettes

Definitions

Weighted Silhouette

\[ \phi(t) = \frac{\sum_{i=1}^{n} w_i \Lambda_i(t)}{\sum_{j=1}^{n} w_j} \]
Persistence Silhouettes

Definitions

Weighted Silhouette
\[ \phi(t) = \frac{\sum_{i=1}^{n} w_i \Lambda_i(t)}{\sum_{j=1}^{n} w_j} \]

Power-Weighted Silhouette
\[ w_i = |d_i - b_i|^p \]
Power-Weighted Silhouettes

Two Examples

![Persistence Diagram](#)

- $(\text{Death-Birth})/2$ vs $(\text{Birth+Death})/2$

![Persistence Diagram](#)

- $(\text{Death-Birth})/2$ vs $(\text{Birth+Death})/2$
Power-Weighted Silhouettes

Two Examples

Persistence Diagram

Silhouette (p=0.1)

Persistence Diagram

Silhouette (p=0.1)
Power-Weighted Silhouettes

Two Examples

Persistence Diagram

Silhouette \( p=0.1 \)

Silhouette \( p=1 \)

Silhouette \( p=0.1 \)

Silhouette \( p=3 \)
Persistence Silhouettes

Results

Since $\phi$ is one-Lipschitz for non-negative weights $w_j$ ...

Convergence of Empirical Process

$$
\frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n} \phi_i(t) - \mathbb{E}[\phi(t)] \right)
$$

converges weakly to a Brownian bridge, with known rate of convergence.

Confidence Bands

We can use the multiplier bootstrap to create a uniform (or a variable width) confidence band defined by $\ell_n^{sil}$ and $u_n^{sil}$ such that

$$
\lim_{n \to \infty} \mathbb{P}\left( \ell_n^{sil}(t) \leq \mu(t) \leq u_n^{sil}(t) \text{ for all } t \right) = 1 - \alpha.
$$
Example I

A Toy Example

Sample Space

1 of 30 Diagrams

Mean 1st Landscape (n= 30 )
with Adaptive 95% band

Mean Silhouette (p= 4 )
with Adaptive 95% band

Mean 3rd Landscape (n= 30 )
with Adaptive 95% band

Mean Silhouette (p= 0.1 )
with Adaptive 95% band
Example II
Earthquake Epicenters

Earthquakes epicenters

Mean 1st Landscape (n=30) with 95% confidence band

Mean Silhouette (p= 0.01 ) with 95% confidence band

1 of 30 Diagrams

Mean 1st Landscape (n=30) with Adaptive 95% band

Mean Silhouette (p= 0.01 ) with Adaptive 95% band

Silhouettes
Summary

First real use of statistical data analysis in TDA. Not just theoretical: we've implemented these techniques! Develops the theory of subsampling techniques. (Did you see the ArXiv this week?)

Silhouette (p=0.1)
Summary

- First real use of statistical data analysis in TDA.
- Not just theoretical: we’ve implemented these techniques!
- Develops the theory of subsampling techniques. (Did you see the ArXiv this week?)
Coauthors
Stochastic Convergence of Persistence Landscapes and Silhouettes

Brittany Terese Fasy
brittany.fasy@alumni.duke.edu

joint work with F. Chazal, F. Lecci, A. Rinaldo, L. Wasserman

Postdoctoral Researcher, Tulane University

11 June 2014