The Intersection of Statistics and Topology:

Confidence Sets

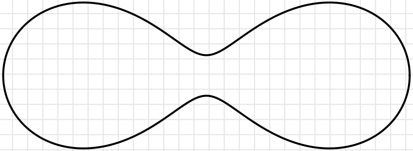
Brittany Terese Fasy

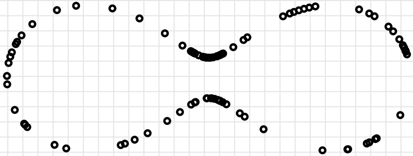
joint work with S. Balakrishnan, F. Chazal, F. Lecci,
A. Rinaldo, A. Singh, L. Wasserman

18 January 2014

How do we Interpret Data?

Data can be a fininte subset of $\mathbb{R}^D$.

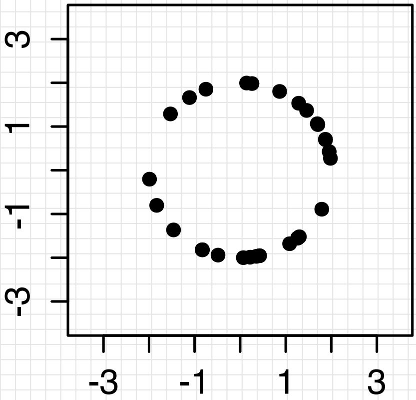




What is the homology / the structure of the underlying space?

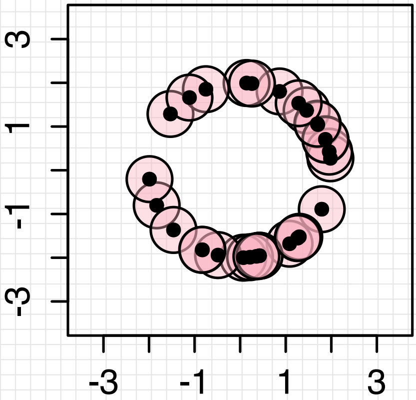
How do we Interpret Data?

Induced Topological Space



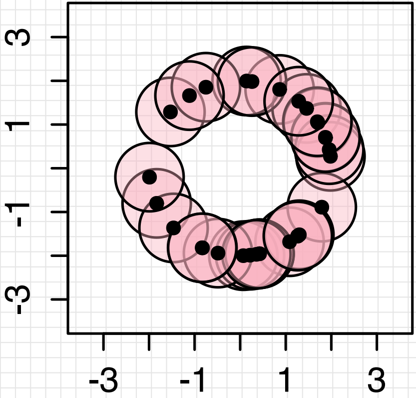
How do we Interpret Data?

Induced Topological Space



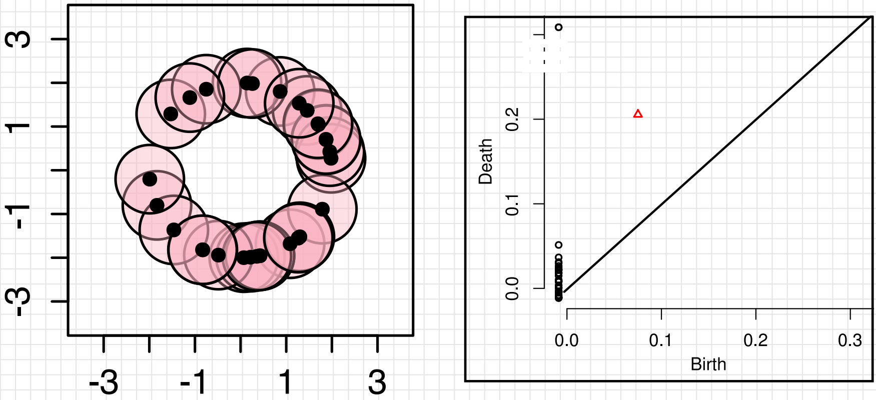
How do we Interpret Data?

Induced Topological Space



How do we Interpret Data?

Induced Topological Space



Objective

Let $\mathcal{P}$ be an unknown persistence diagram and $\widehat{\mathcal{P}}$ be an estimate of $\mathcal{P}$.

Objective

Let $\mathcal{P}$ be an unknown persistence diagram and $\widehat{\mathcal{P}}$ be an estimate of $\mathcal{P}$.

Question

How close is $\widehat{\mathcal{P}}$ to $\mathcal{P}$?

Objective

Let $\mathcal{P}$ be an unknown persistence diagram and $\widehat{\mathcal{P}}$ be an estimate of $\mathcal{P}$.

Question

How close is $\widehat{\mathcal{P}}$ to $\mathcal{P}$?

Answer with Statistics

Given $\alpha \in (0,1)$, we want δ_α such that

$$\mathbb{P}(\mathcal{P} \in \{\mathcal{P}_* : W_\infty(\mathcal{P}_*, \widehat{\mathcal{P}}) < \delta_\alpha\}) \leq 1 - \alpha.$$

Objective

Let $\mathcal{P}$ be an unknown persistence diagram and $\widehat{\mathcal{P}}$ be an estimate of $\mathcal{P}$.

Question

How close is $\widehat{\mathcal{P}}$ to $\mathcal{P}$?

Answer with Statistics

Given $\alpha \in (0, 1)$, we want δ_α such that

$$\mathbb{P}(\mathcal{P} \in \{\mathcal{P}_* : W_\infty(\mathcal{P}_*, \widehat{\mathcal{P}}) < \delta_\alpha\}) \leq 1 - \alpha.$$

Statistical Model

$\mathbb{M}$ is a manifold.
P is a probability distribution supported on $\mathbb{M}$.
Observe data $X_1, X_2, \ldots, X_n \sim P$.
Compute $\hat{\Theta}_n = \Theta(X_1, \ldots, X_n)$

Statistical Model

$\mathbb{M}$ is a manifold.
P is a probability distribution supported on $\mathbb{M}$.
Observe data $X_1, X_2, \ldots, X_n \sim P$.
Compute $\hat{\Theta}_n = \Theta(X_1, \ldots, X_n)$

Question

How does $\hat{\Theta}_n$ compare to $\mathbb{E}(\Theta_n) = \Theta_n(\mathbb{M})$?

Statistical Model

$\mathbb{M}$ is a manifold.
P is a probability distribution supported on $\mathbb{M}$.
Observe data $X_1, X_2, \ldots, X_n \sim P$.
Compute $\hat{\Theta}_n = \Theta(X_1, \ldots, X_n)$

Question

How does $\hat{\Theta}_n$ compare to $\mathbb{E}(\Theta_n) = \Theta_n(\mathbb{M})$?

Answer

Find C such that $\mathbb{P}(\Theta_n(\mathbb{M}) \in C) \geq 1 - \alpha$.

Statistical Model

$\mathbb{M}$ is a manifold.
P is a probability distribution supported on $\mathbb{M}$.
Observe data $X_1, X_2, \ldots, X_n \sim P$.
Compute $\hat{\Theta}_n = \Theta(X_1, \ldots, X_n)$

Question

How does $\hat{\Theta}_n$ compare to $\mathbb{E}(\Theta_n) = \Theta_n(\mathbb{M})$?

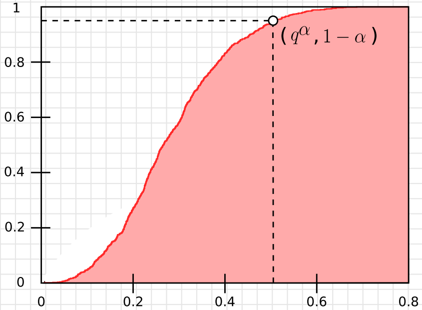
Answer

Find C such that $\mathbb{P}(\Theta_n(\mathbb{M}) \in C) \geq 1 - \alpha$. How?

Computing a Confidence Interval

With Infinite Resources

Repeatedly sample *n* data points, obtaining:



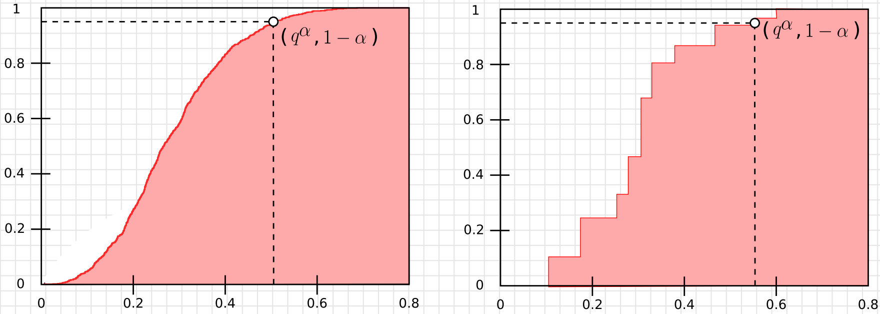
Confidence Intervals

$$\mathbb{P}(\Theta_n(\mathbb{M}) \in [0, q^\alpha]) \geq 1 - \alpha.$$

Computing a Confidence Interval

With Infinite Resources

Repeatedly sample n data points, obtaining: $\hat{\Theta}_{n,1}, \ldots, \hat{\Theta}_{n,N}$



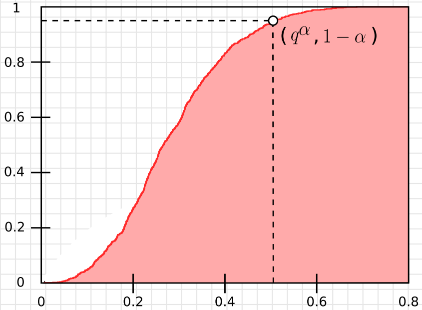
Confidence Intervals

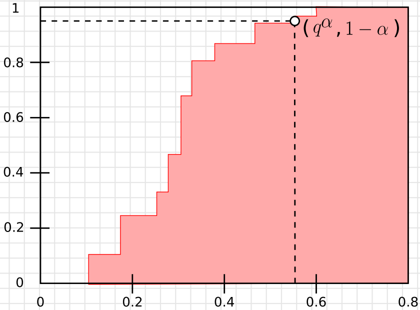
$$\mathbb{P}(\Theta_n(\mathbb{M}) \in [0, q^\alpha]) \geq 1 - \alpha.$$

Computing a Confidence Interval

With Infinite Resources

Repeatedly sample n data points, obtaining: $\hat{\Theta}_{n,1}, \ldots, \hat{\Theta}_{n,N}$ via simulation.





Confidence Intervals

$$\mathbb{P}(\Theta_n(\mathbb{M}) \in [0, q^\alpha]) \geq 1 - \alpha.$$

Bootstrapping

When We Can Only Take One Sample

We have one sample:
$\mathcal{S}_n = \{X_1, \ldots, X_n\}$

Bootstrapping

When We Can Only Take One Sample

We have one sample:
$\mathcal{S}_n = \{X_1, \ldots, X_n\}$

Subsample (with replacement), obtaining: $\{X_1^*, \ldots, X_n^*\}$

Bootstrapping

When We Can Only Take One Sample

We have one sample:
$\mathcal{S}_n = \{X_1, \ldots, X_n\}$

Subsample (with replacement), obtaining: $\{X_1^*, \ldots, X_n^*\}$

Compute $\hat{\Theta}_n^* = \Theta(X_1^*, \ldots, X_n^*)$.

Bootstrapping

When We Can Only Take One Sample

We have one sample: $\mathcal{S}_n = \{X_1, \ldots, X_n\}$

Subsample (with replacement), obtaining: $\{X_1^*, \ldots, X_n^*\}$

Compute $\hat{\Theta}_n^* = \Theta(X_1^*, \ldots, X_n^*)$.

Repeat N times, obtaining: $\hat{\Theta}_{n,1}^*, \ldots, \hat{\Theta}_{n,N}^*$.

Bootstrapping

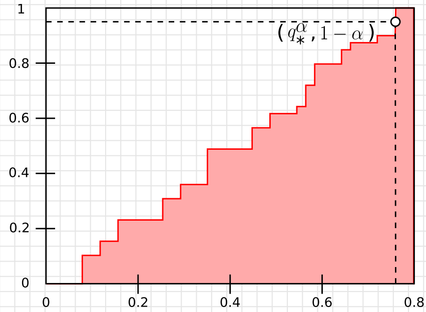
When We Can Only Take One Sample

We have one sample:
$\mathcal{S}_n = \{X_1, \ldots, X_n\}$

Subsample (with replacement), obtaining: $\{X_1^*, \ldots, X_n^*\}$

Compute $\hat{\Theta}_n^* = \Theta(X_1^*, \ldots, X_n^*)$.

Repeat N times, obtaining:
$\hat{\Theta}_{n,1}^*, \ldots, \hat{\Theta}_{n,N}^*$.



Bootstrapping Example

Estimating Densities

P has density p.

Smoothed Density: $p_h = p \star K_h$

KDE: $\hat{p}_h(x) = \frac{1}{n} \sum_1^n \frac{1}{h^D} K\left(\frac{||x - X_i||}{h}\right)$.

Bootstrapping Example

Estimating Densities

P has density p.
Smoothed Density: $p_h = p \star K_h$
KDE: $\hat{p}_h(x) = \frac{1}{n} \sum_1^n \frac{1}{h^D} K\left(\frac{||x - X_i||}{h}\right)$.

$\Theta_n = (\sqrt{nh^D} ||\hat{p}_h - p_h||_\infty)$.
$\Theta_n^* = (\sqrt{nh^D} ||\hat{p}_h^* - \hat{p}_h||_\infty)$.

Bootstrapping Example

Estimating Densities

P has density p.
Smoothed Density: $p_h = p \star K_h$
KDE: $\hat{p}_h(x) = \frac{1}{n} \sum_1^n \frac{1}{h^D} K\left(\frac{||x - X_i||}{h}\right)$.

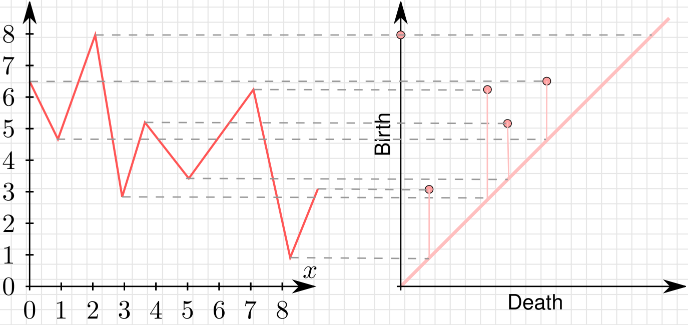
$\Theta_n = (\sqrt{nh^D}||\hat{p}_h - p_h||_\infty)$.
$\Theta_n^* = (\sqrt{nh^D}||\hat{p}_h^* - \hat{p}_h||_\infty)$.

Bootstrap Theorem [FLRWBS]

$$\mathbb{P}(\sqrt{nh^D}||\hat{p}_h - p_h||_\infty > q_*^\alpha \mid X_1, \ldots, X_n) = \alpha + O\left(\sqrt{1/n}\right)$$

Persistent Homology

A Pairing of Critical Values.

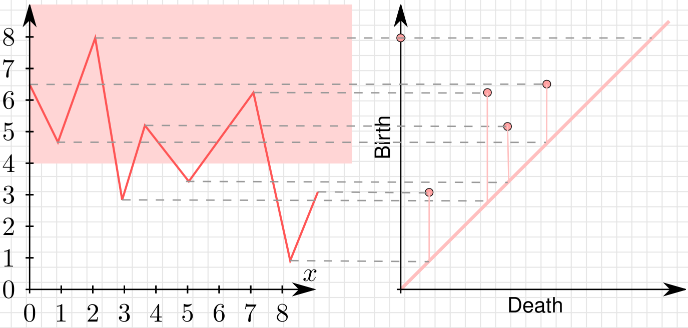


$$\mathcal{P} = \mathrm{Dgm}_p^+(f)$$

Persistent Homology

A Pairing of Critical Values.

Tracking $H\left(f^{-1}([t,\infty))\right)$.

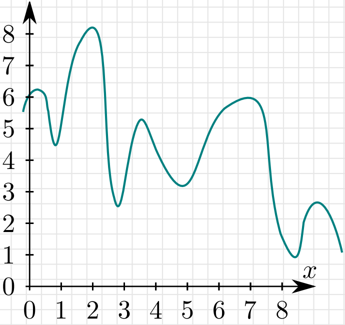


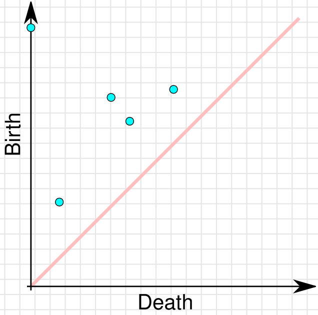
$$\mathcal{P} = \mathrm{Dgm}_p^+(f)$$

Persistent Homology

A Pairing of Critical Values.

Tracking $H\left(f^{-1}([t,\infty))\right)$.



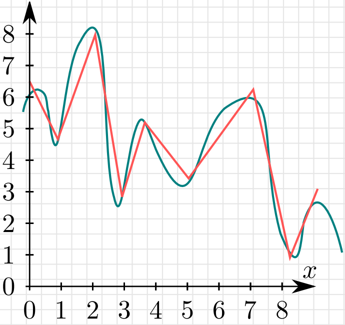


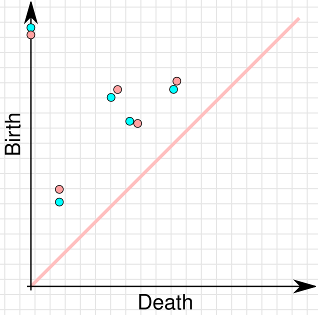
$$\mathcal{P} = \mathrm{Dgm}_p^+(f)$$

Persistent Homology

A Pairing of Critical Values.

Tracking $H\left(f^{-1}([t,\infty))\right)$.





$$\mathcal{P} = \mathsf{Dgm}_p^+(f)$$

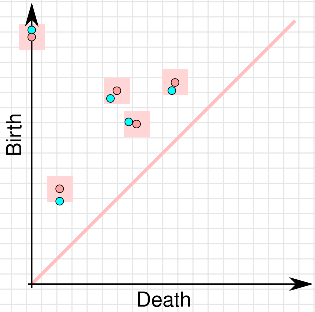
Bottleneck Distance

Given two persistence diagrams $\mathcal{P}$ and $\hat{\mathcal{P}}$, find the best *perfect matching* between the point sets.

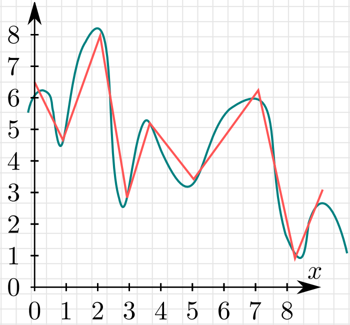
Minimize Cost

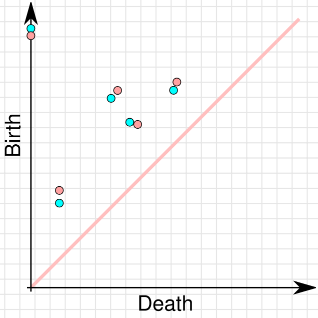
We wish to find

$$W_\infty = \min_M \{ \max_{(p,q)\in M} ||p - q||_\infty \}.$$



Stability of Matchings





Bottleneck Stability Theorem [CDGO]

$$||p - \hat{p}||_\infty \quad \geq \quad W_\infty(\mathcal{P}, \widehat{\mathcal{P}})$$

Putting It All Together

Bottleneck Stability Theorem

$$||p - \hat{p}||_\infty \quad \geq \quad W_\infty(\mathcal{P}, \widehat{\mathcal{P}})$$

Putting It All Together

Bottleneck Stability Theorem

$$||p - \hat{p}||_\infty \quad \geq \quad W_\infty(\mathcal{P}, \widehat{\mathcal{P}})$$

Bootstrap Theorem

$$\mathbb{P}(\sqrt{nh^D}||\hat{p}_h - p_h||_\infty > q_*^\alpha) \quad = \quad \alpha + O\left(\sqrt{1/n}\right)$$

Putting It All Together

Bottleneck Stability Theorem

$$||p - \hat{p}||_\infty \quad \geq \quad W_\infty(\mathcal{P}, \widehat{\mathcal{P}})$$

Bootstrap Theorem

$$\mathbb{P}(\sqrt{nh^D}||\hat{p}_h - p_h||_\infty > q_*^\alpha) \quad = \quad \alpha + O\left(\sqrt{1/n}\right)$$

Confidence Sets for Persistence Diagrams

$$\mathbb{P}(W_\infty(\mathcal{P}, \widehat{\mathcal{P}}) \leq \frac{q_*^\alpha}{\sqrt{nh^D}}) \quad \geq \quad 1 - \alpha - O\left(\sqrt{1/n}\right)$$

Putting It All Together

Bottleneck Stability Theorem

$$||p - \hat{p}||_\infty \quad \geq \quad W_\infty(\mathcal{P}, \widehat{\mathcal{P}})$$

Bootstrap Theorem

$$\mathbb{P}(\sqrt{nh^D}||\hat{p}_h - p_h||_\infty > q_*^\alpha) \quad = \quad \alpha + O\left(\sqrt{1/n}\right)$$

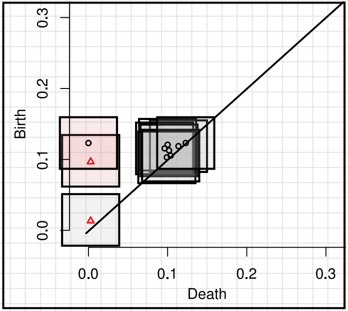
Confidence Sets for Persistence Diagrams

$$\mathbb{P}(W_\infty(\mathcal{P}, \widehat{\mathcal{P}}) \leq \frac{q_*^\alpha}{\sqrt{nh^D}}) \quad \geq \quad 1 - \alpha - O\left(\sqrt{1/n}\right)$$

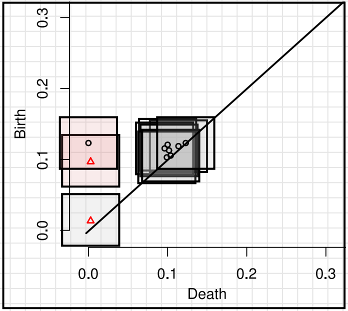
Asymptotic Confidence Sets for Persistence Diagrams

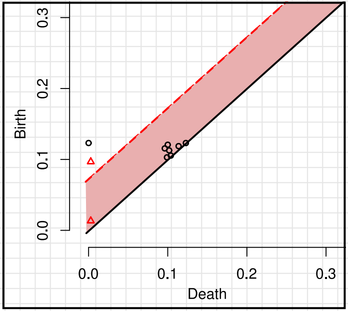
$$\lim_{n\to\infty} \mathbb{P}(W_\infty(\mathcal{P}, \widehat{\mathcal{P}}) \leq \frac{q_*^\alpha}{\sqrt{nh^D}}) \quad \geq \quad 1 - \alpha$$

Visualizing Confidence Intervals



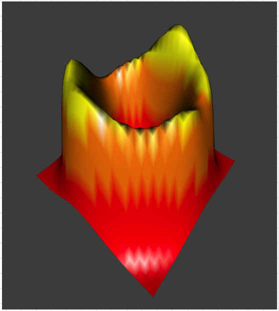
Visualizing Confidence Intervals





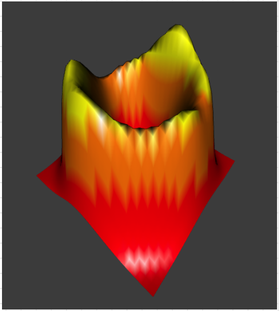
Density Function Examples

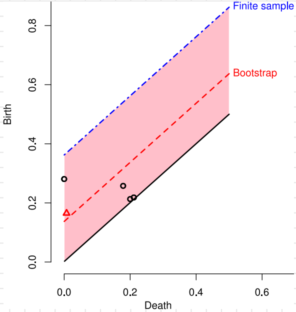
Uniform Distribution on Unit Circle



Density Function Examples

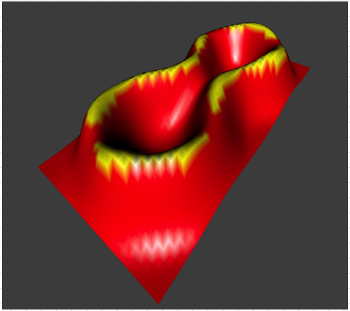
Uniform Distribution on Unit Circle





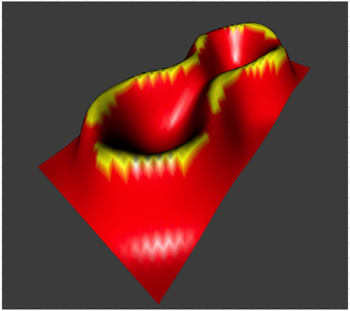
Density Function Examples

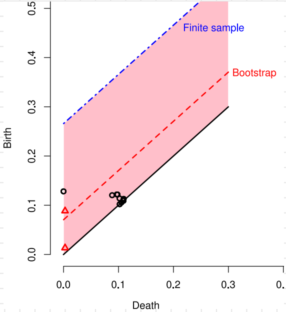
Uniform Distribution on Cassini Curve



Density Function Examples

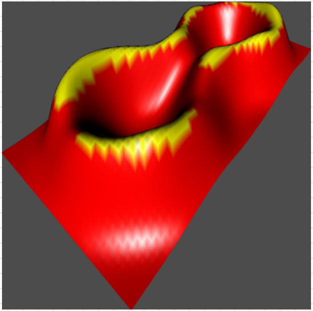
Uniform Distribution on Cassini Curve





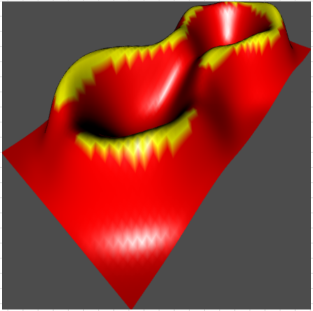
Density Function Examples

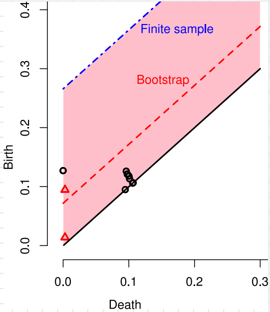
Cassini Curve with Outliers



Density Function Examples

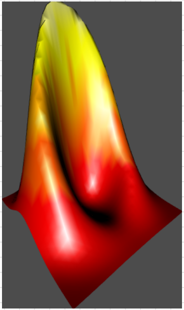
Cassini Curve with Outliers





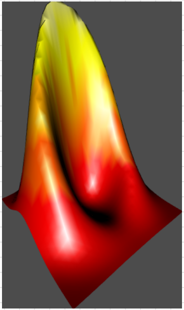
Density Function Examples

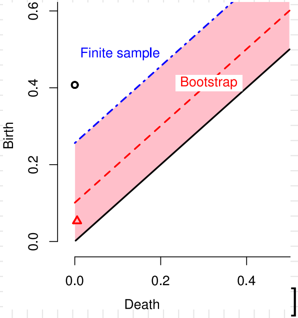
Normal Distribution on Unit Circle



Density Function Examples

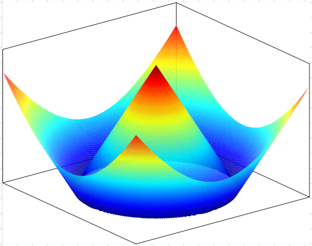
Normal Distribution on Unit Circle





Distance to a Subset

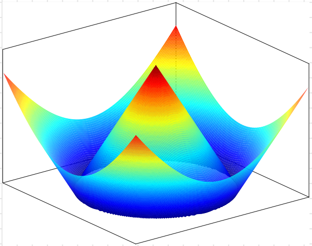
$d_{\mathbb{M}}(a) = \inf_{x \in \mathbb{M}} ||x - a||$
$\mathcal{P}_1 = \mathrm{Dgm}_p^-(d_{\mathbb{X}})$



Distance to a Subset

$d_{\mathbb{M}}(a) = \inf_{x \in \mathbb{M}} ||x - a||$
$\mathcal{P}_1 = \mathrm{Dgm}_p^{-}(d_{\mathbb{X}})$

P has continuous density p.
support$(P) = \mathbb{M}$.
$\mathcal{S}_n = \{X_1, \ldots, X_n\} \sim P$
$\widehat{\mathcal{P}}_1 = \mathrm{Dgm}_p^{-}(d_{\mathcal{S}_n})$



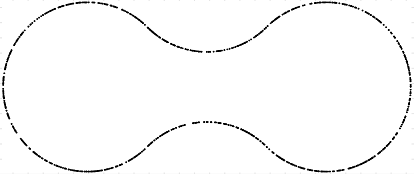
Subsampling

Confidence Interval from Subsampling [**F**LRWBS]

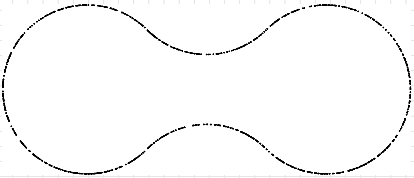
Assume that $p(x)$ is bounded away from zero.
Then, almost surely, for all large n,

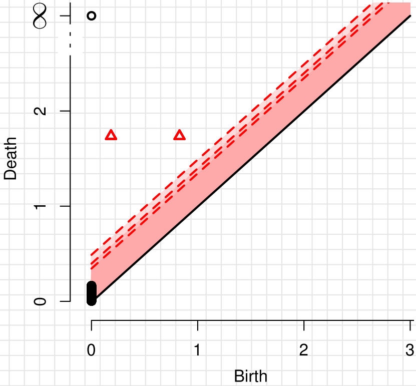
$$\mathbb{P}\left(W_\infty(\mathcal{P}_1, \widehat{\mathcal{P}}_1) > c_n\right) \leq \alpha + \frac{2^d}{n \log n} + O\left(\sqrt{\frac{b_n \log n}{n}}\right)$$

Varying α



Varying α





$$\alpha = 0.001, 0.05, 0.25$$

Two More Methods

$\mathcal{S}_n = \mathcal{S}_{1,n} \bigsqcup \mathcal{S}_{2,n}$.

Theorem (Concentration of Measure)

There exists $\hat{t}_{cm} = \hat{t}_{cm}(\alpha, d, n, \mathcal{S}_{1,n})$ *such that*

$$\mathbb{P}\left(W_\infty(\mathcal{P}_1, \widehat{\mathcal{P}}_1) > \hat{t}_{cm}\right) \leq \alpha + O\left(\left(\frac{\log n}{n}\right)^{1/d+2}\right).$$

Theorem (Method of Shells)

There exists $\hat{t}_s = \hat{t}_s(\alpha, d, n, K, \mathcal{S}_{1,n})$ *such that*

$$\mathbb{P}\left(W_\infty(\mathcal{P}_1, \widehat{\mathcal{P}}_1) > \hat{t}_s\right) \leq \alpha + O\left(\left(\frac{\log n}{n}\right)^{1/d+2}\right).$$

These Methods are Different

Concentration of Measure

$\hat{t}_{cm}$ is found by solving the following for t:

$$\frac{2^{d+1}}{t^d \hat{\rho}_{1,n}} \exp\left(-\frac{nt^d \hat{\rho}_{1,n}}{2}\right) = \alpha.$$

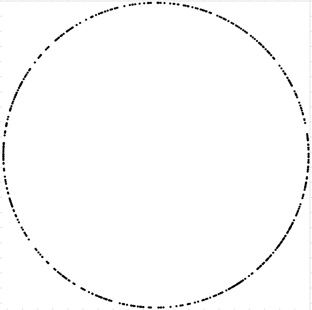
Shells

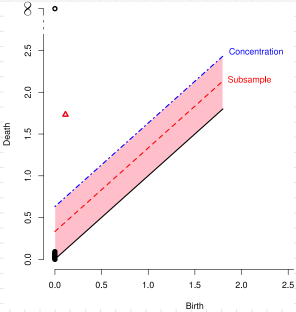
$\hat{t}_s$ is found by solving the following for t:

$$\frac{2^{d+1}}{t^d} \int_{\hat{\rho}_n}^{\infty} \frac{\hat{g}(v)}{v} \exp\left(-\frac{nvt^d \hat{\rho}_{1,n}}{2}\right) dv = \alpha.$$

Distance Function Examples

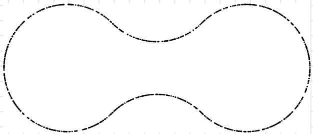
Uniform Distribution on Unit Circle

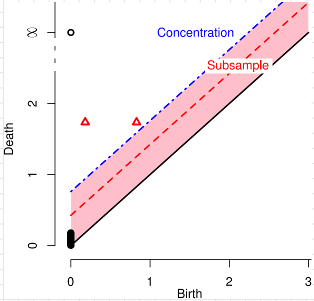




Distance Function Examples

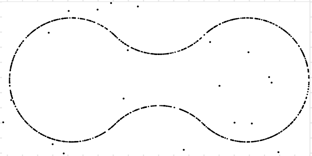
Uniform Distribution on Cassini Curve

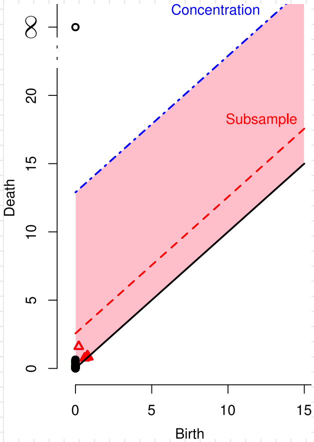




Distance Function Examples

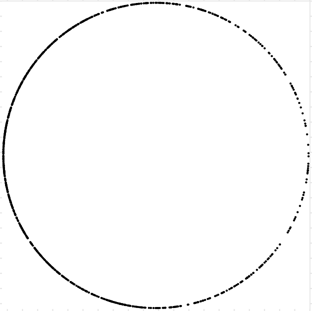
Cassini Curve with Outliers

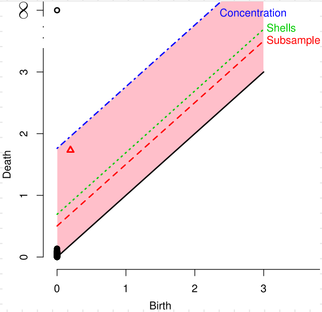




Distance Function Examples

Normal Distribution on Unit Circle





Summary

Recalling the Problem

- Sample from a distribution on a manifold.

Summary

Recalling the Problem

- Sample from a distribution on a manifold.
- Create sample function (distance or density).

Summary

Recalling the Problem

- Sample from a distribution on a manifold.
- Create sample function (distance or density).
- Now, we have (unknown) $\mathcal{P}$ and (known) $\widehat{\mathcal{P}_n}$.

Summary

Recalling the Problem

- Sample from a distribution on a manifold.
- Create sample function (distance or density).
- Now, we have (unknown) $\mathcal{P}$ and (known) $\widehat{\mathcal{P}_n}$.
- **Find c_n such that $\mathbb{P}\left(W_\infty(\mathcal{P}, \widehat{\mathcal{P}_n}) > c_n\right) \leq \alpha$.**

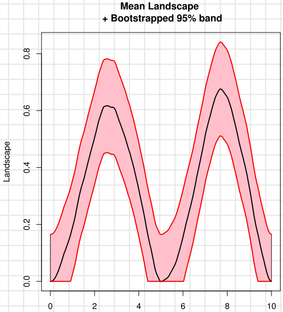
Summary

Recalling the Problem

- Sample from a distribution on a manifold.
- Create sample function (distance or density).
- Now, we have (unknown) $\mathcal{P}$ and (known) $\widehat{\mathcal{P}_n}$.
- **Find c_n such that $\mathbb{P}\left(W_\infty(\mathcal{P}, \widehat{\mathcal{P}_n}) > c_n\right) \leq \alpha$.**
- The pair $\widehat{\mathcal{P}_n}$ and $[0, c_n]$ define a confidence set for $\mathcal{P}$.

Ongoing Research

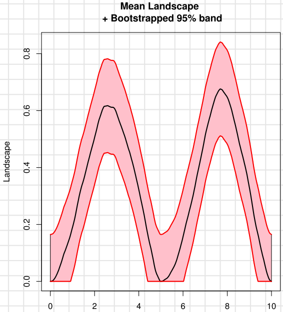
Ongoing Research

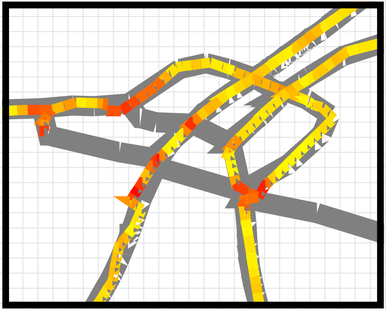


Functional Analysis

Confidence Bands for Landscapes
joint w/ F. Chazal, F. Lecci,
A. Rinaldo, L. Wasserman

Ongoing Research





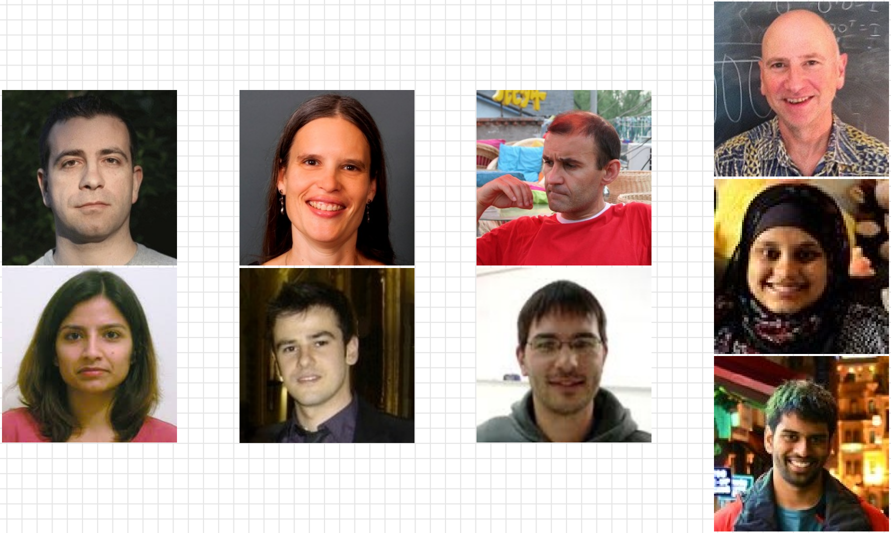
Functional Analysis

Confidence Bands for Landscapes
joint w/ F. Chazal, F. Lecci,
A. Rinaldo, L. Wasserman

Really Great Upcoming Talk

Carola Wenk
Map Construction & Comparison
3:30 Here!

Collaborator Collage



Thank you!

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

References

[CDGO] The Structure and Stability of Persistence Modules. ArXiv 1207.3674.

[C**F**LRSW] On the Bootstrap for Persistence Diagrams and Landscapes. Modeling and Analysis of Information Systems, **20**:6 (Dec. 2013), 96–105.

[**F**LRWBS] Statistical Inference for Persistent Homology: Confidence Sets for Persistence Diagrams. ArXiv 1303.7117. Tentatively accepted, Annals of Statistics.