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

Question

How close is $\hat{\mathcal{P}}$ to $\mathcal{P}$?
Let $\mathcal{P}$ be an unknown persistence diagram and $\hat{\mathcal{P}}$ be an estimate of $\mathcal{P}$. 

**Question**

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

**Answer with Statistics**

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

$$P(\mathcal{P} \in \{\mathcal{P}_* : W_\infty(\mathcal{P}_*, \hat{\mathcal{P}}) < \delta_\alpha\}) \leq 1 - \alpha.$$
Let $\mathcal{P}$ be an unknown persistence diagram and $\hat{\mathcal{P}}$ be an estimate of $\mathcal{P}$.

**Question**

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

**Answer with Statistics**

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

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

\( \mathcal{M} \) is a manifold.
\( P \) is a probability distribution supported on \( \mathcal{M} \).
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 \( \mathbb{E}(\Theta_n) = \Theta_n(\mathcal{M}) \)?
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

\( \mathcal{M} \) is a manifold.

\( P \) is a probability distribution supported on \( \mathcal{M} \).

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 \( \mathbb{E}(\Theta_n) = \Theta_n(\mathcal{M}) \)?

**Answer**

Find \( C \) such that \( \mathbb{P}(\Theta_n(\mathcal{M}) \in C) \geq 1 - \alpha \). How?
Computing a Confidence Interval
With Infinite Resources

Repeatedly sample $n$ data points, obtaining:

Confidence Intervals

$$\mathbb{P}(\Theta_n(M) \in [0, q^\alpha]) \geq 1 - \alpha.$$
Computing a Confidence Interval
With Infinite Resources

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

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

$$\Pr(\Theta_n(M) \in [0, q^\alpha]) \geq 1 - \alpha.$$
Bootstrapping
When We Can Only Take One Sample

We have one sample:
\[ S_n = \{ X_1, \ldots, X_n \} \]
**Bootstrapping**

When We Can Only Take One Sample

We have one sample:

\[ S_n = \{X_1, \ldots, X_n\} \]

Subsample (with replacement), obtaining:

\[ \{X_1^*, \ldots, X_n^*\} \]
Bootstrapping
When We Can Only Take One Sample

We have one sample:
\[ S_n = \{ X_1, \ldots, X_n \} \]

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:
\[ S_n = \{ X_1, \ldots, X_n \} \]

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: 
\[ S_n = \{ X_1, \ldots, X_n \} \]

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: \( p_h = p \ast K_h \)

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: $p_h = p \ast K_h$

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

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

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

Bootstrap Theorem [F. LRWBS]
Bootstrapping Example

Estimating Densities

$P$ has density $p$.

**Smoothed Density:** $p_h = p \ast K_h$

**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.

\[ \mathcal{P} = \text{Dgm}_p^+ (f) \]
Persistent Homology
A Pairing of Critical Values.

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

$\mathcal{P} = \text{Dgm}_p^+(f)$
Persistent Homology
A Pairing of Critical Values.

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

$\mathcal{P} = \text{Dgm}_p^+(f)$
Persistent Homology
A Pairing of Critical Values.

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

\[ \mathcal{P} = Dgm^+_P(f) \]
Given two persistence diagrams $\mathcal{P}$ and $\mathcal{P}^\wedge$, 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]

\[ \| p - \hat{p} \|_{\infty} \geq W_{\infty}(\mathcal{P}, \hat{\mathcal{P}}) \]
Bottleneck Stability Theorem

\[ \| p - \hat{p} \|_\infty \geq W_\infty(\mathcal{P}, \hat{\mathcal{P}}) \]
Putting It All Together

Bottleneck Stability Theorem

\[ \|p - \hat{p}\|_\infty \geq W_\infty(\mathcal{P}, \hat{\mathcal{P}}) \]

Bootstrap Theorem

\[ \mathbb{P}(\sqrt{nhD} \|\hat{p}_h - p_h\|_\infty > q_\alpha) = \alpha + O\left(\sqrt{1/n}\right) \]
Putting It All Together

Bottleneck Stability Theorem

\[ \| p - \hat{p} \|_\infty \geq W_\infty(\mathcal{P}, \hat{\mathcal{P}}) \]

Bootstrap Theorem

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

Confidence Sets for Persistence Diagrams

\[ \mathbb{P}(W_\infty(\mathcal{P}, \hat{\mathcal{P}}) \leq \frac{q_\alpha^*}{\sqrt{nh^D}}) \geq 1 - \alpha - O\left(\sqrt{1/n}\right) \]
Putting It All Together

**Bottleneck Stability Theorem**

\[ \|p - \hat{p}\|_\infty \geq W_\infty(\mathcal{P}, \hat{\mathcal{P}}) \]

**Bootstrap Theorem**

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

**Confidence Sets for Persistence Diagrams**

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

**Asymptotic Confidence Sets for Persistence Diagrams**

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

\[ d_M(a) = \inf_{x \in M} ||x - a|| \]
\[ P_1 = \text{Dgm}_p(d_X) \]
Distance Function

Distance to a Subset

\[ \text{support}(P) = \mathbb{M} \]

\[ S_n = \{X_1, \ldots, X_n\} \sim P \]

\[ \hat{P}_1 = \text{Dgm}_p^-(d_{S_n}) \]

\[ P_1 = \text{Dgm}_p^-(d_X) \]

\[ d_{\mathbb{M}}(a) = \inf_{x \in \mathbb{M}} |x - a| \]
Subsampling

Confidence Interval from Subsampling [FLRWBS]

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

\[
\mathbb{P} \left( W_\infty(P_1, \hat{P}_1) > c_n \right) \leq \alpha + \frac{2^d}{n \log n} + O \left( \sqrt{\frac{b_n \log n}{n}} \right)
\]
Varying $\alpha$
Varying $\alpha$

$\alpha = 0.001, 0.05, 0.25$
Two More Methods

\[ S_n = S_{1,n} \cup S_{2,n}. \]

**Theorem (Concentration of Measure)**

There exists \( \hat{t}_{cm} = \hat{t}_{cm}(\alpha, d, n, S_{1,n}) \) such that

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

**Theorem (Method of Shells)**

There exists \( \hat{t}_s = \hat{t}_s(\alpha, d, n, K, S_{1,n}) \) such that

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

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

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

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

