The Intersection of Statistics and Topology:

Confidence Sets for Persistence Diagrams

Brittany Terese Fasy

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

3 June 2014

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

The Intersection of Statistics and Topology:

Stochastic Convergence of Persistence Landscapes and Silhouettes

Brittany Terese Fasy

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

3 June 2014

[C**F**LRW] Stochastic Convergence of Persistence Landscapes and Silhouettes. ArXiv. SoCG 2014 Proceedings, Kyoto. Journal version pending.

The Intersection of Statistics and Topology:

Confidence Sets for Persistence Diagrams

Brittany Terese Fasy

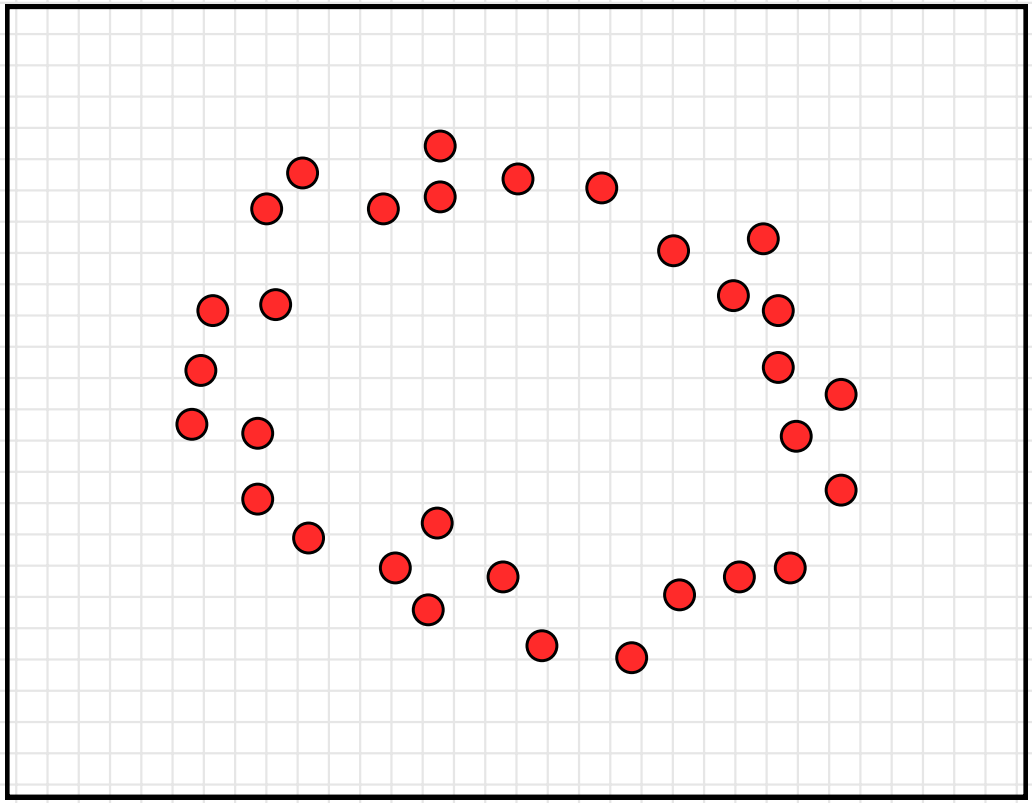
joint work with S. Balakrishnan, F. Lecci,
A. Rinaldo, A. Singh, L. Wasserman

3 June 2014

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

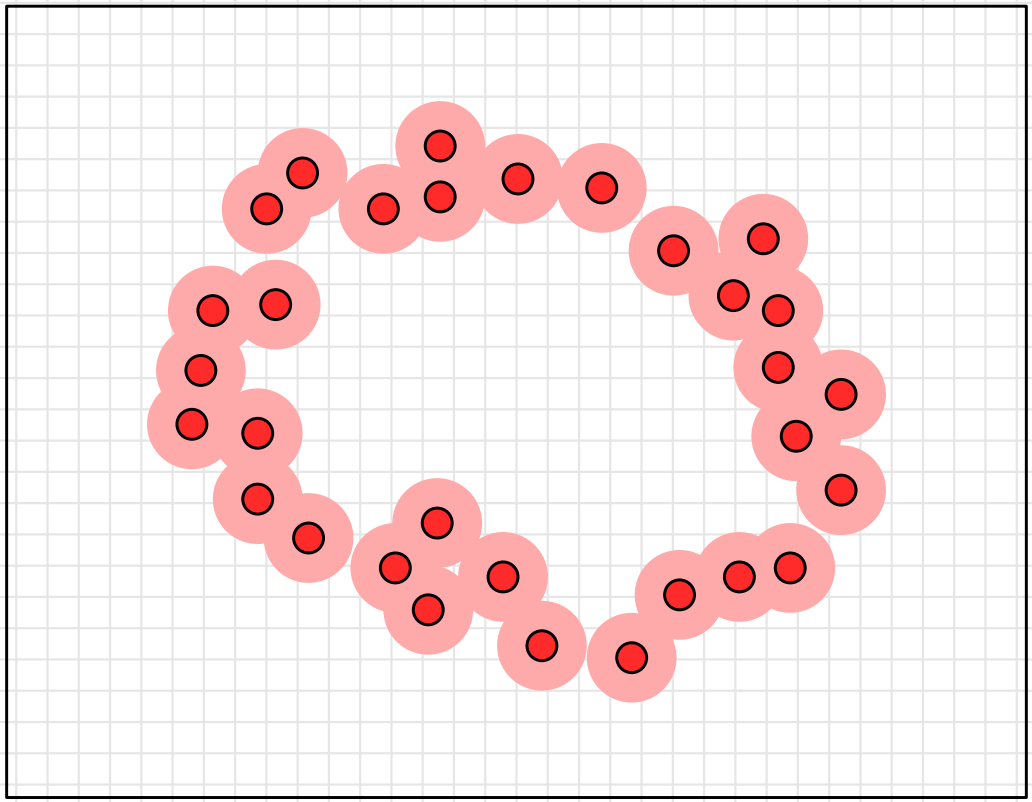
Motivation

Extracting Structure from Point Cloud Data



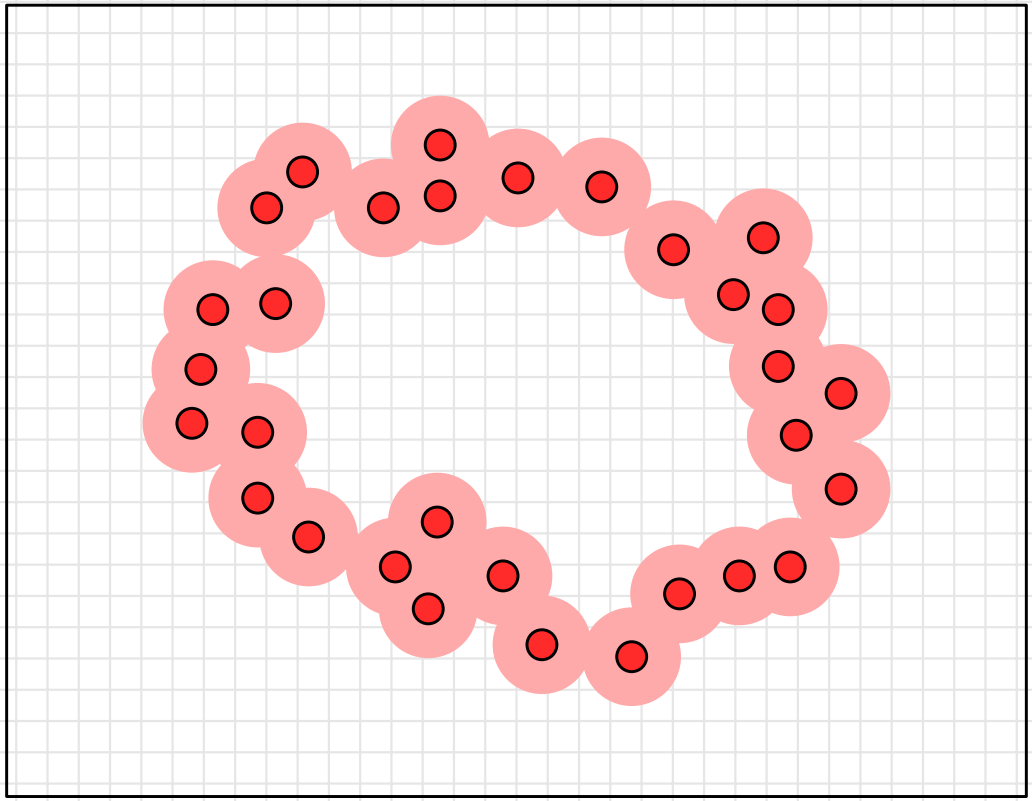
Motivation

Extracting Structure from Point Cloud Data



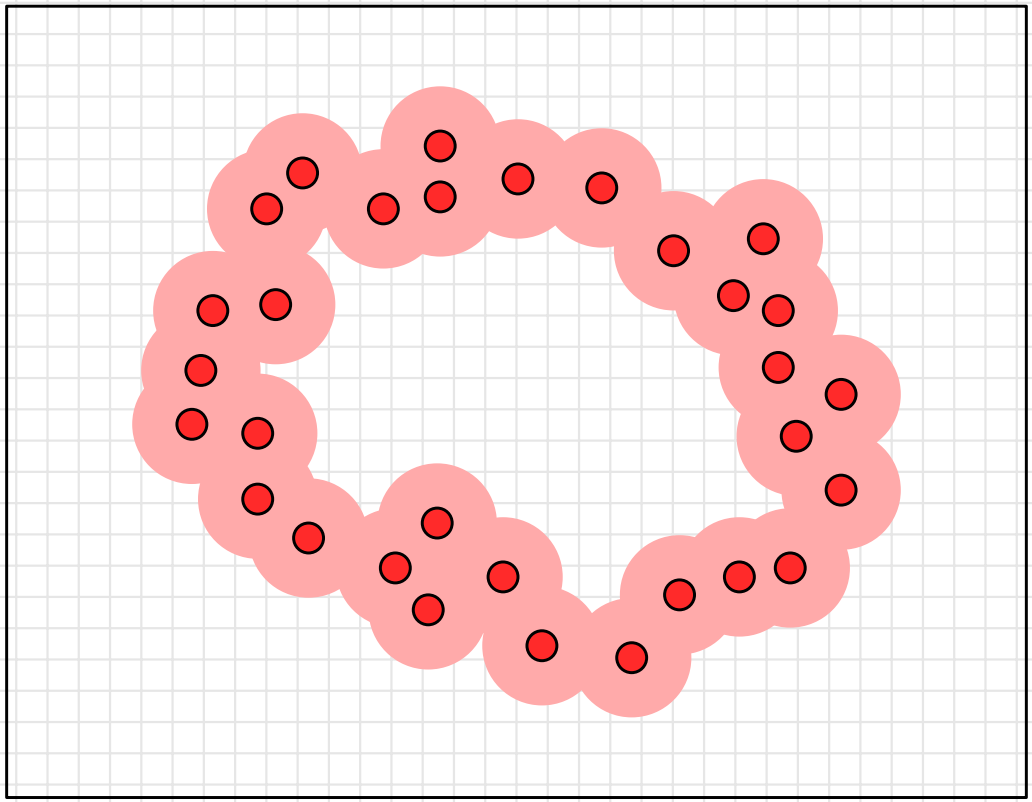
Motivation

Extracting Structure from Point Cloud Data



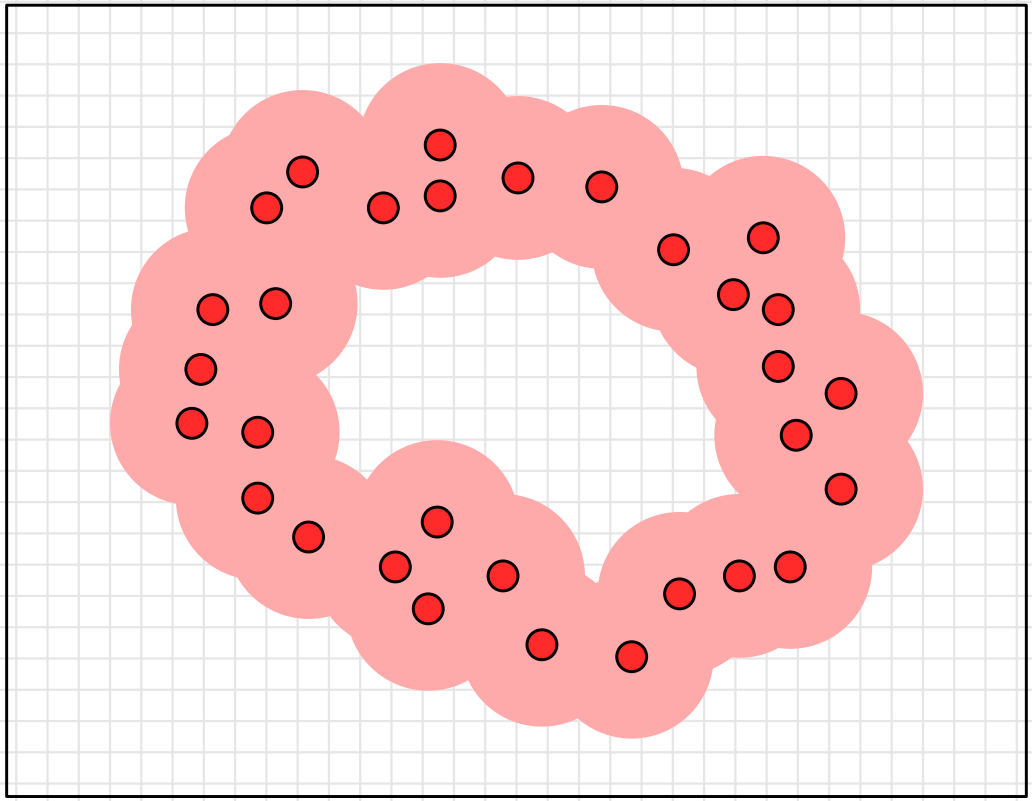
Motivation

Extracting Structure from Point Cloud Data



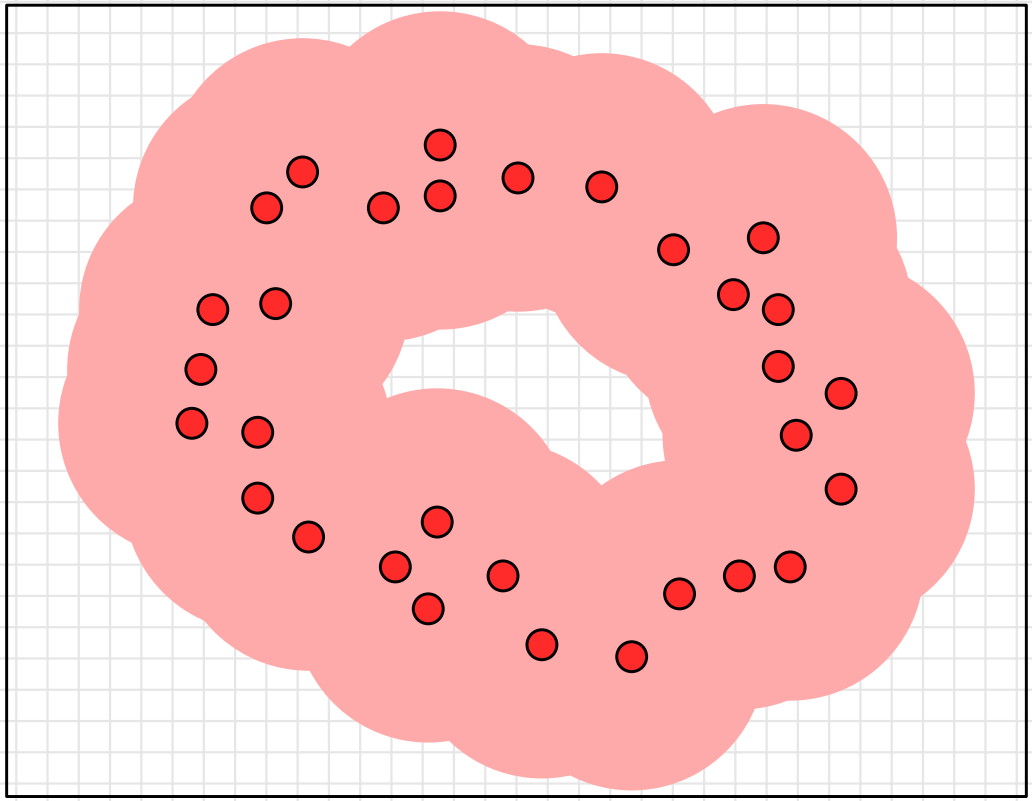
Motivation

Extracting Structure from Point Cloud Data



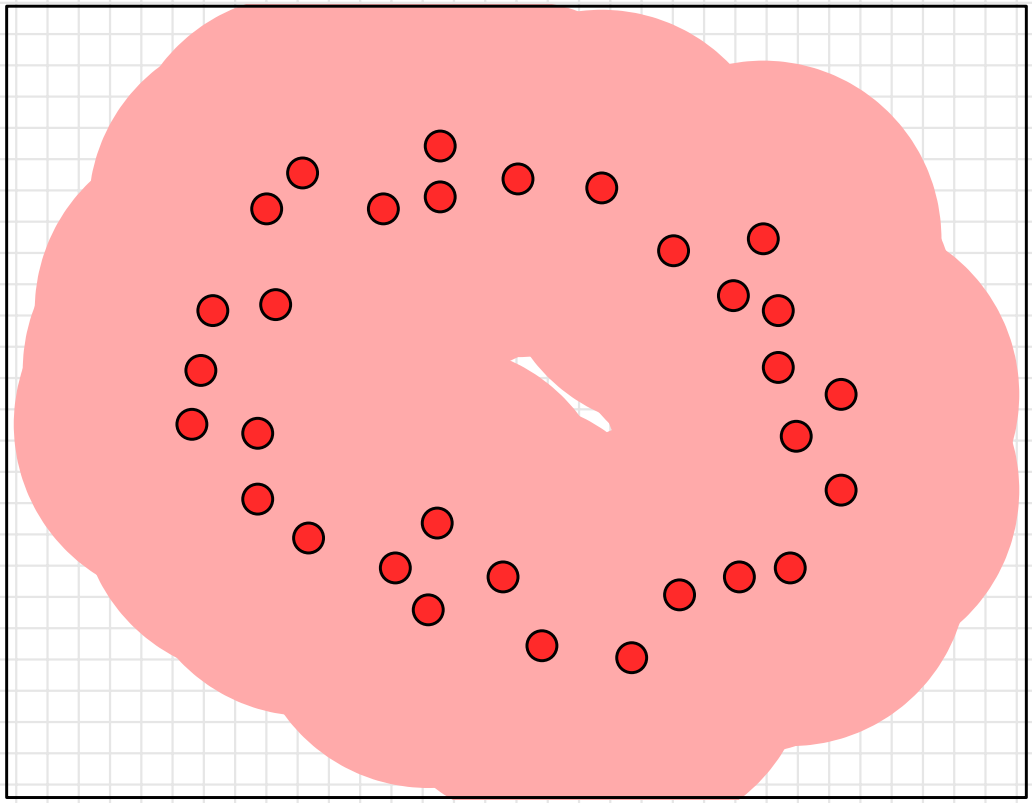
Motivation

Extracting Structure from Point Cloud Data



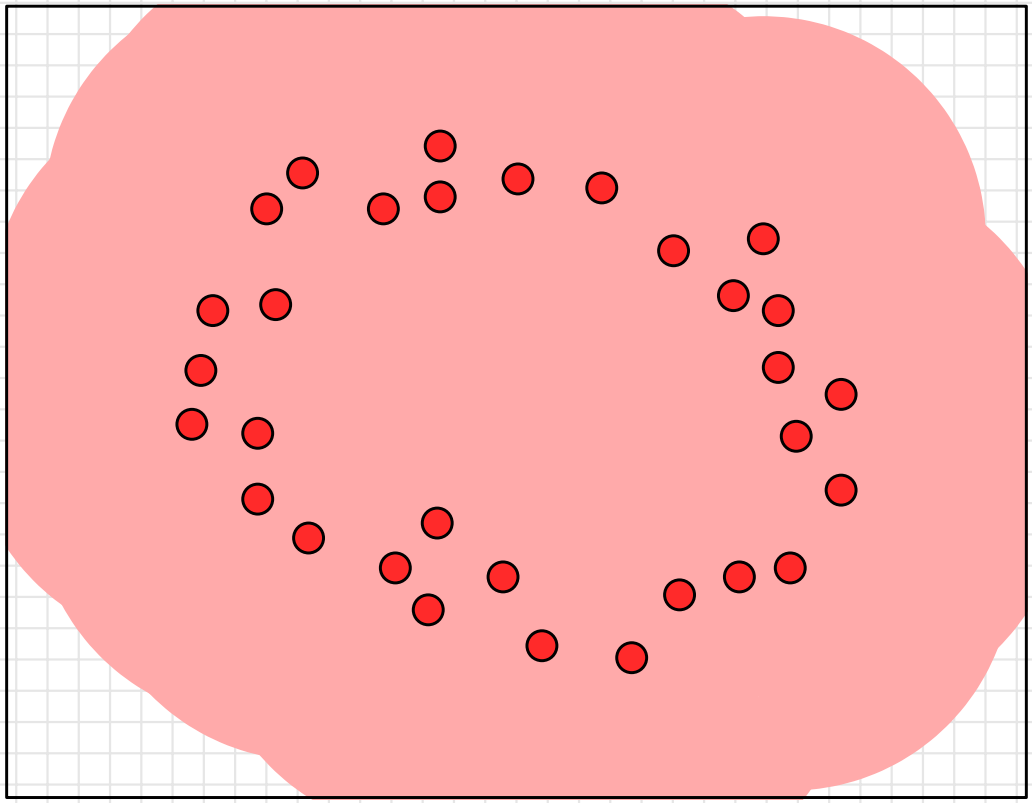
Motivation

Extracting Structure from Point Cloud Data



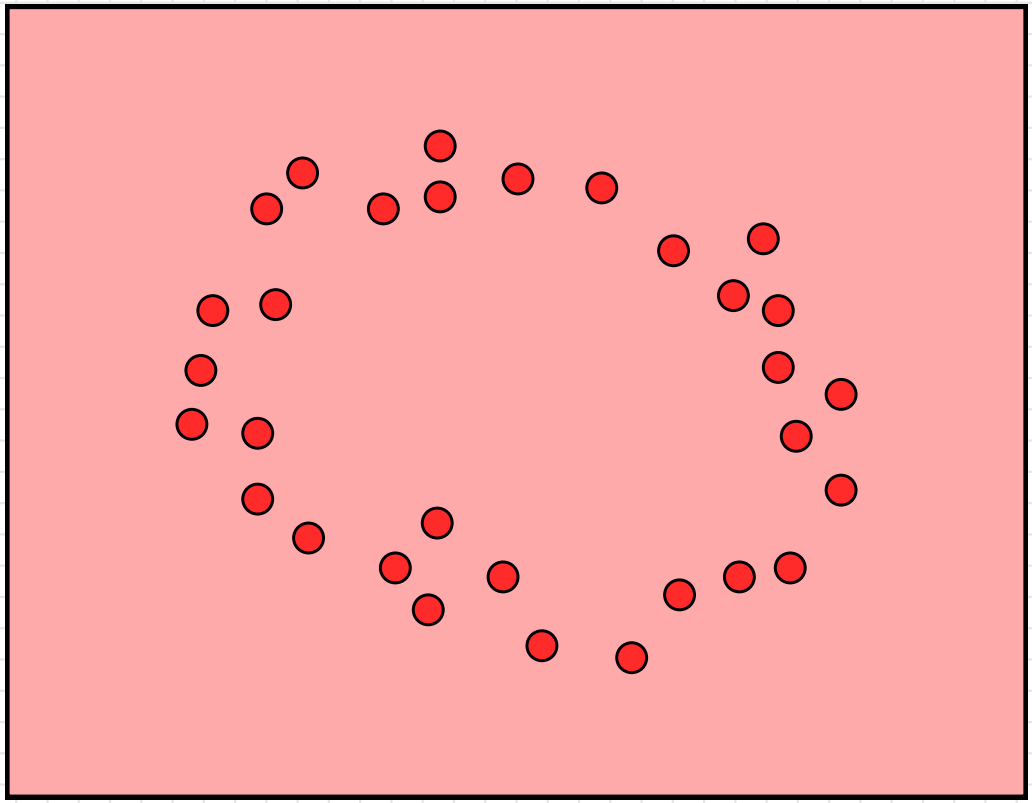
Motivation

Extracting Structure from Point Cloud Data



Motivation

Extracting Structure from Point Cloud Data



Homeomorphisms

An Informal Discussion

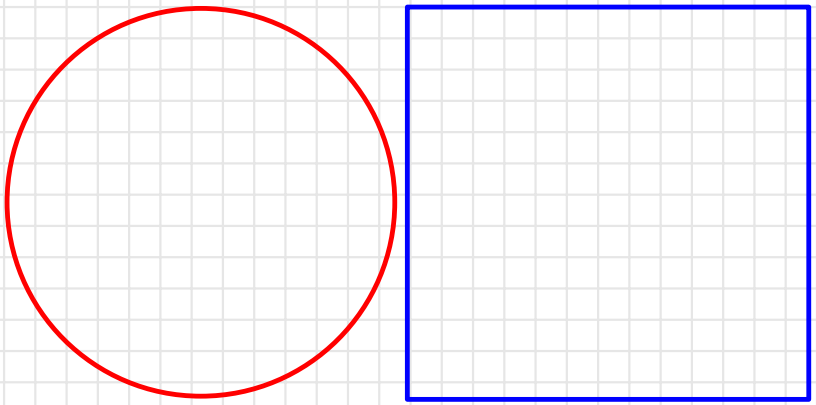
A *homeomorphism* is a continuous bending, stretching, or shrinking (but not tearing or glueing) of one object into another object.

Homeomorphisms

An Informal Discussion

A *homeomorphism* is a continuous bending, stretching, or shrinking (but not tearing or glueing) of one object into another object.

Circle = Square

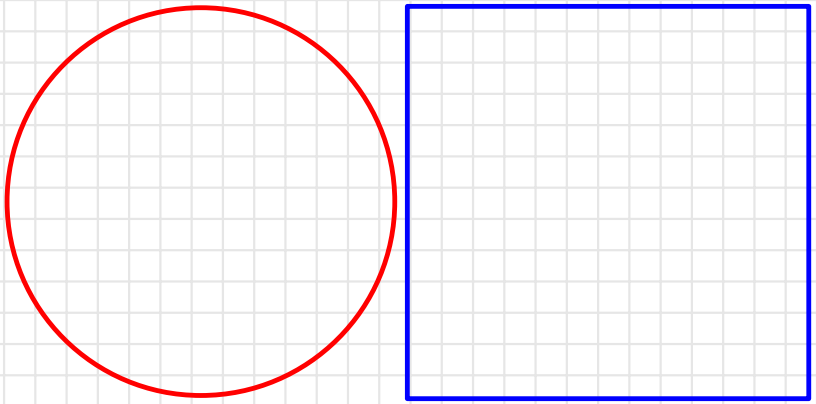


Homeomorphisms

An Informal Discussion

A *homeomorphism* is a continuous bending, stretching, or shrinking (but not tearing or glueing) of one object into another object.

Circle = Square



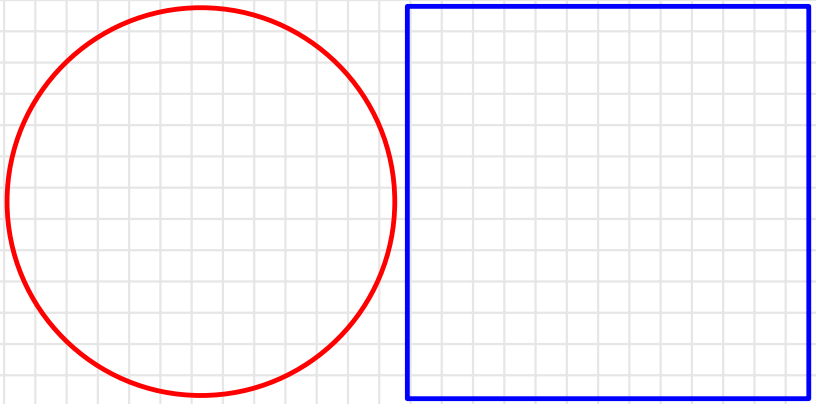
Donut = Coffee Cup

Homeomorphisms

An Informal Discussion

A *homeomorphism* is a continuous bending, stretching, or shrinking (but not tearing or glueing) of one object into another object.

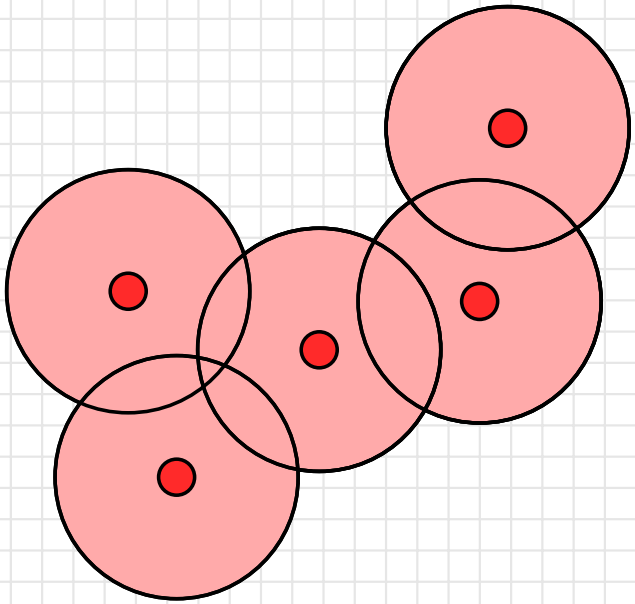
Circle = Square



Donut ≠ Sugar Bowl

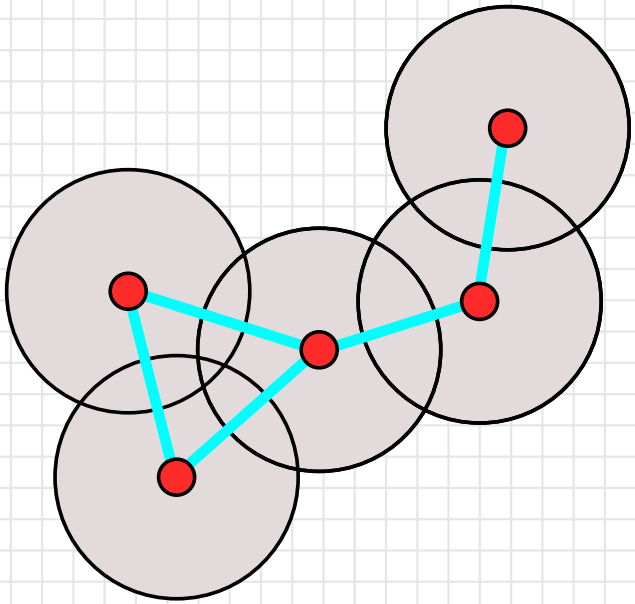
Topological Spaces

Nerve Complexes



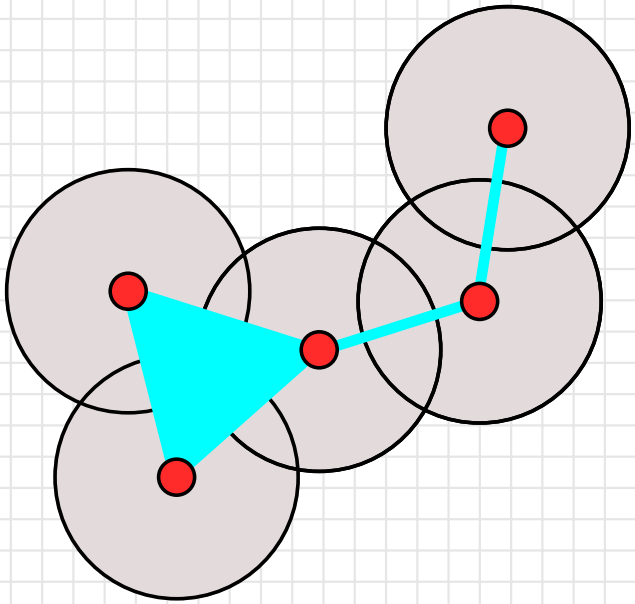
Topological Spaces

Nerve Complexes



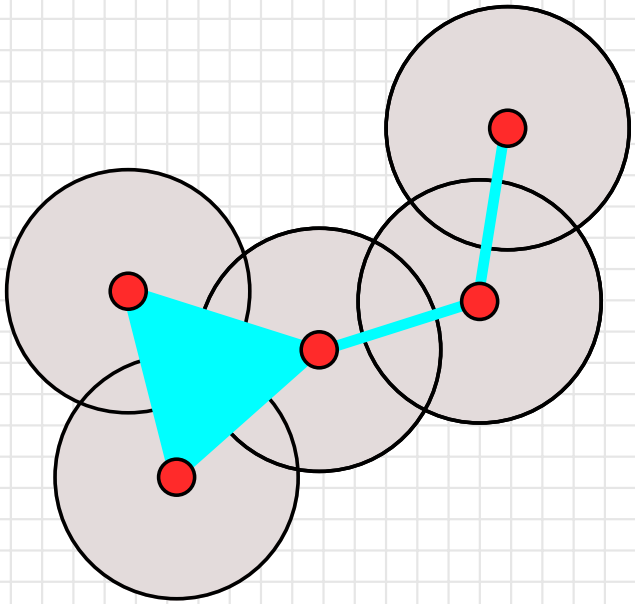
Topological Spaces

Nerve Complexes



Topological Spaces

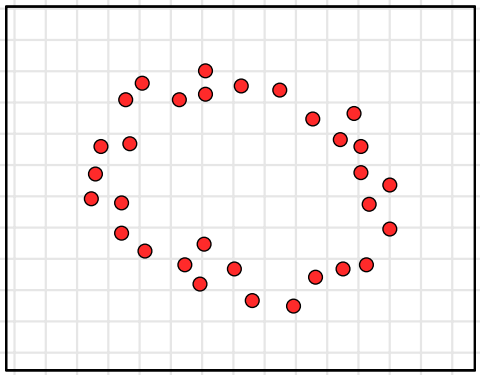
Nerve Complexes



Nerve Lemma

If X is a set of closed convex sets, then $\mathrm{Nrv}(X)$ is topologically equivalent to $\bigcup X$ (up to homotopy type).

Homology



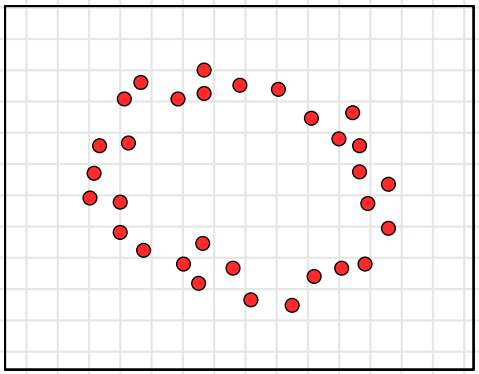
$\beta_0 = n$

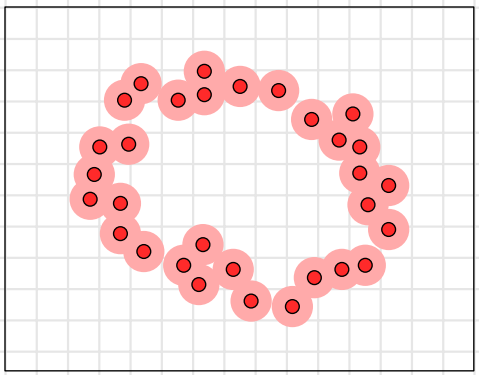
$\beta_1 = 0$

Simplicial Homology

- X_p is the power set of all p-simplices.
- $C_p = (X_p, +_2)$.
- $\partial_p : C_p \to C_{p-1}$ is the boundary map.
- $H_p = \mathrm{Ker}(\partial_p)/\mathrm{Im}(\partial_{p+1})$
- $\beta_p = \mathrm{rank}(H_p)$.

Homology





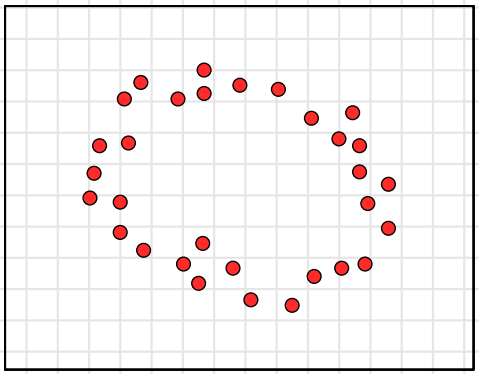
$\beta_0 = n$

$\beta_1 = 0$

$\beta_0 = 5$

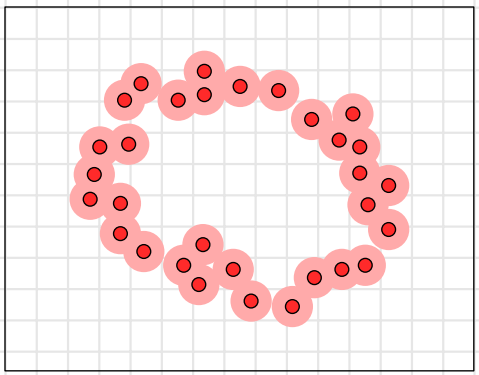
$\beta_1 = 1$

Homology



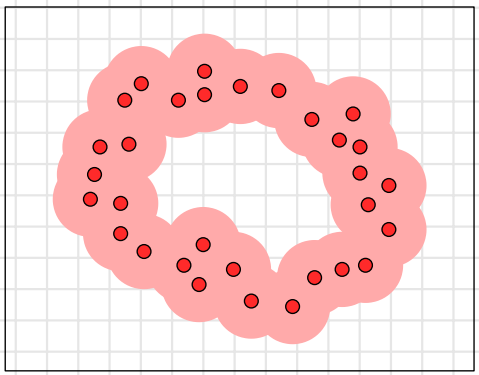
$\beta_0 = n$

$\beta_1 = 0$



$\beta_0 = 5$

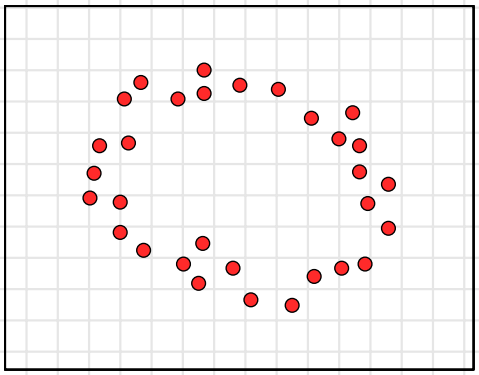
$\beta_1 = 1$



$\beta_0 = 1$

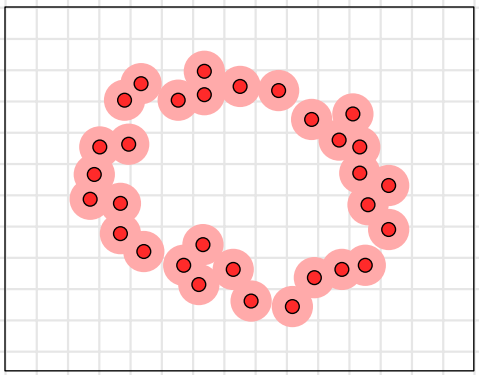
$\beta_1 = 1$

Homology



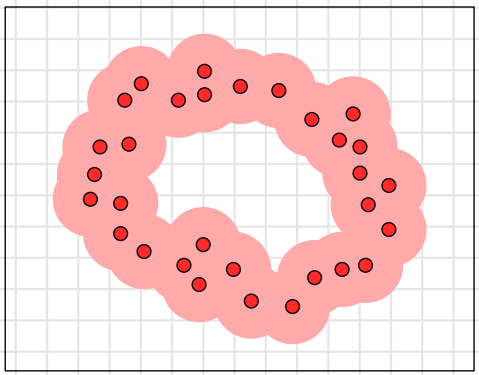
$\beta_0 = n$

$\beta_1 = 0$



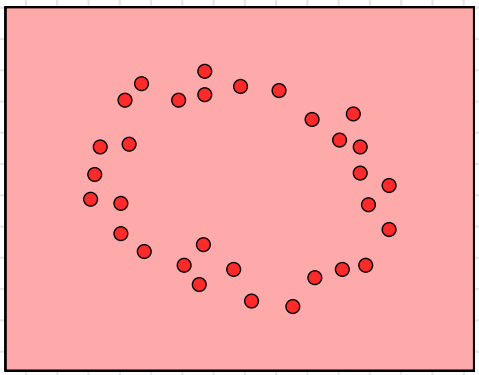
$\beta_0 = 5$

$\beta_1 = 1$



$\beta_0 = 1$

$\beta_1 = 1$

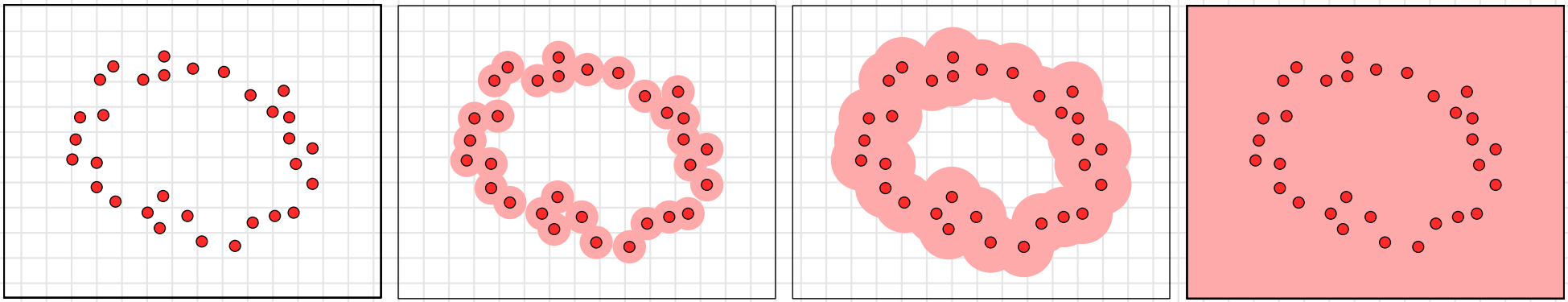


$\beta_0 = 1$

$\beta_1 = 0$

Homology

$d^{-1}(0)$ $d^{-1}([0, r_1])$ $d^{-1}([0, r_2])$ $d^{-1}([0, r_4])$



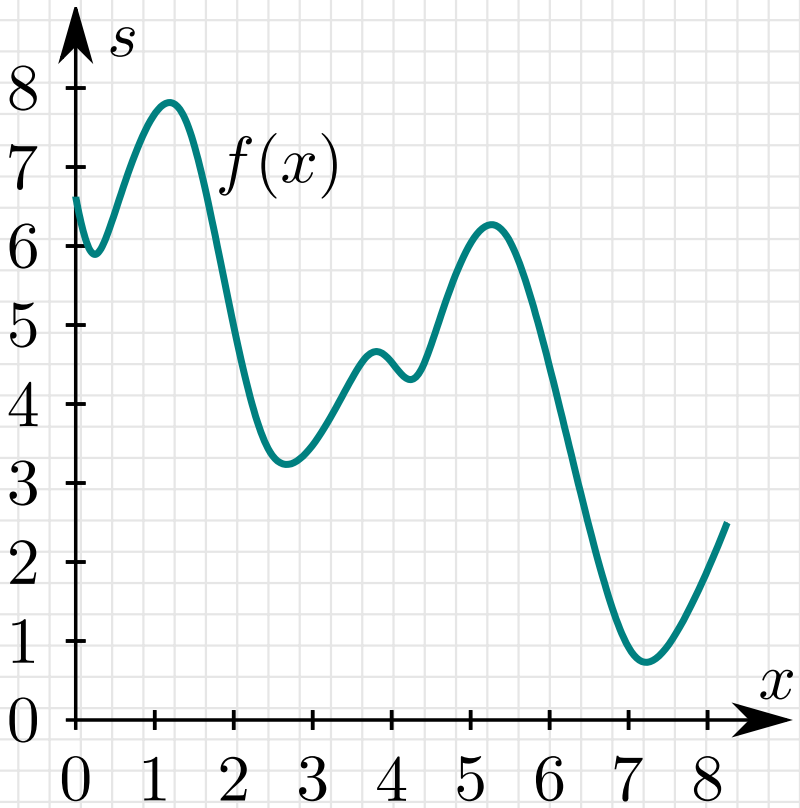
$\beta_0 = n$ $\beta_0 = 5$ $\beta_0 = 1$ $\beta_0 = 1$

$\beta_1 = 0$ $\beta_1 = 1$ $\beta_1 = 1$ $\beta_1 = 0$

Persistent Homology

Lower-Level Set Filtration

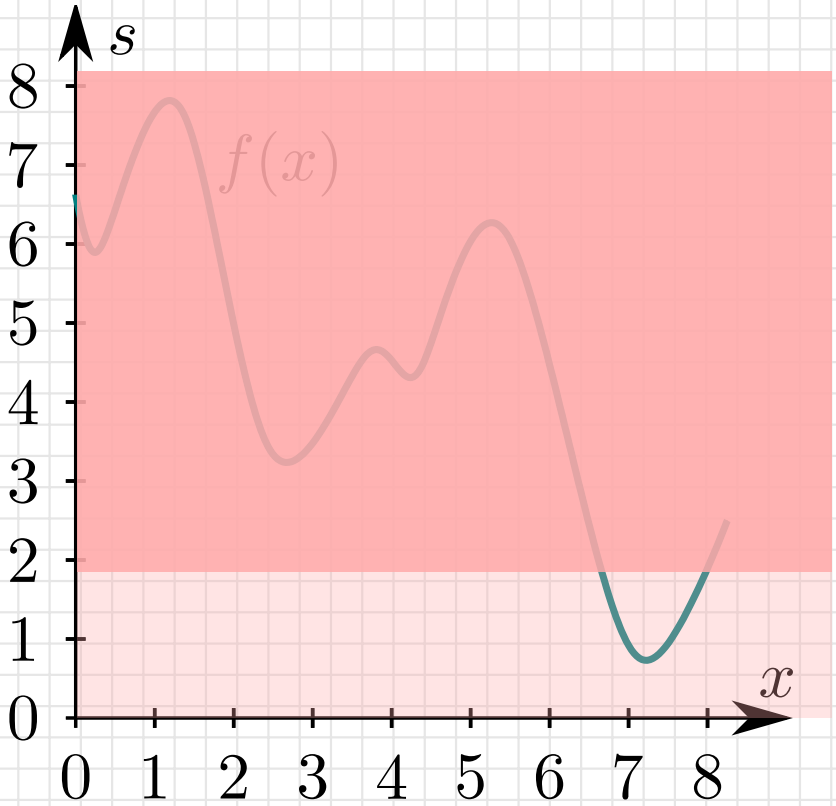
Let $\mathbb{M}$ be a compact topological space.
Let $f : \mathbb{M} \to \mathbb{R}$ be continuous.
We consider all sets of the form $\mathbb{M}_t := f^{-1}((-\infty, t])$.



Persistent Homology

Lower-Level Set Filtration

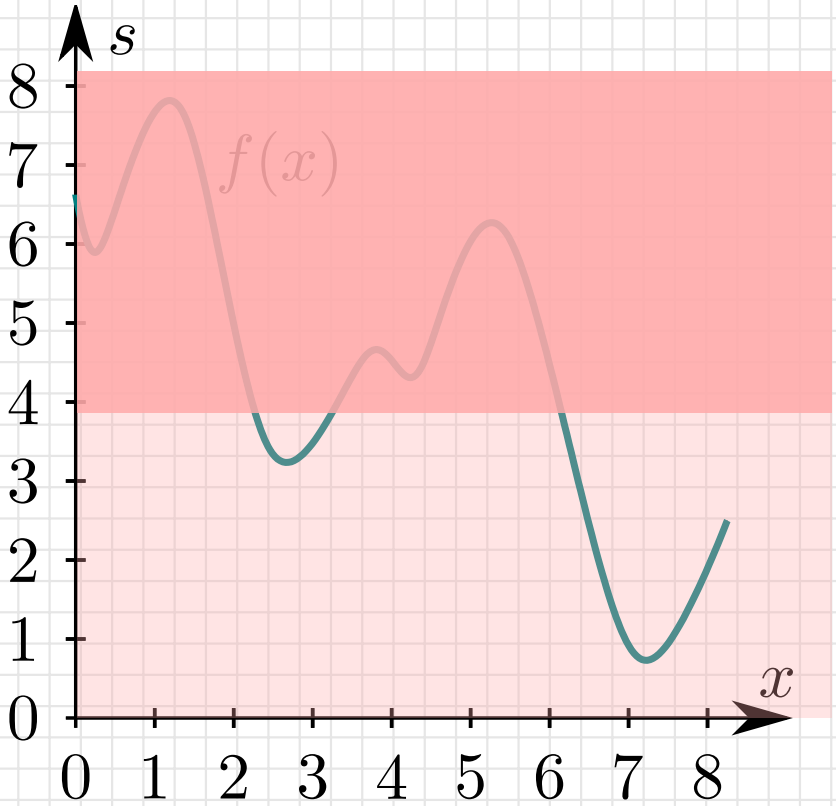
Let $\mathbb{M}$ be a compact topological space.
Let $f : \mathbb{M} \to \mathbb{R}$ be continuous.
We consider all sets of the form $\mathbb{M}_t := f^{-1}((-\infty, t])$.



Persistent Homology

Lower-Level Set Filtration

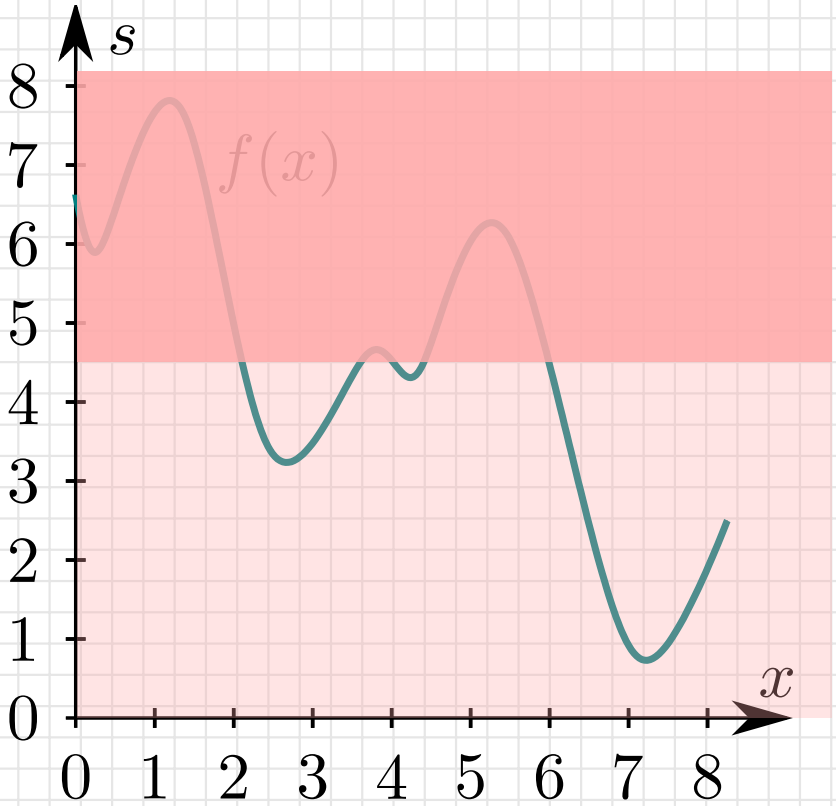
Let $\mathbb{M}$ be a compact topological space.
Let $f : \mathbb{M} \to \mathbb{R}$ be continuous.
We consider all sets of the form $\mathbb{M}_t := f^{-1}((-\infty, t])$.



Persistent Homology

Lower-Level Set Filtration

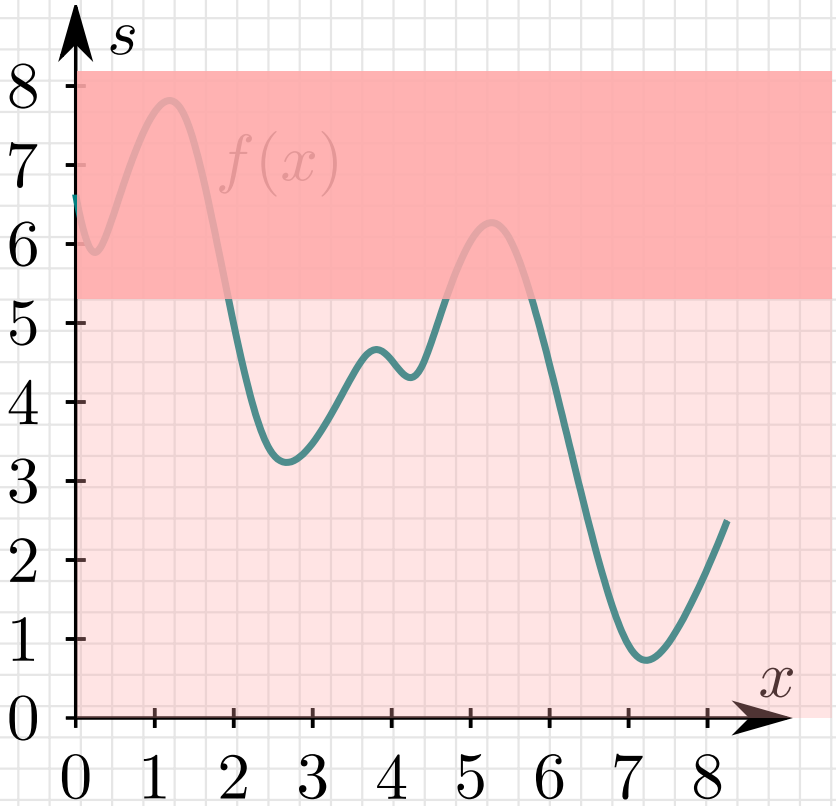
Let $\mathbb{M}$ be a compact topological space.
Let $f : \mathbb{M} \to \mathbb{R}$ be continuous.
We consider all sets of the form $\mathbb{M}_t := f^{-1}((-\infty, t])$.



Persistent Homology

Lower-Level Set Filtration

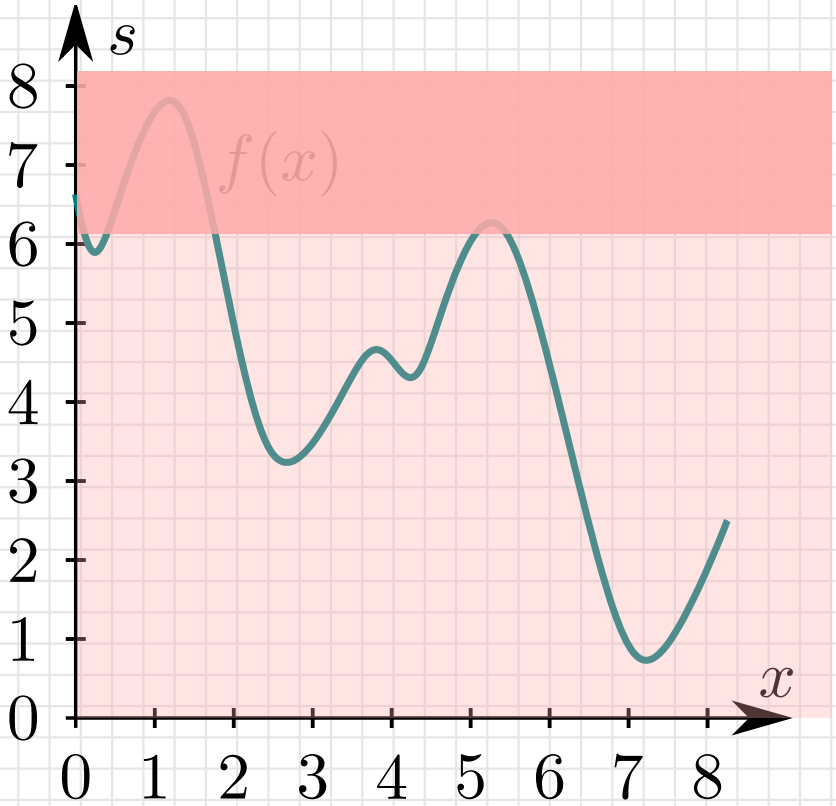
Let $\mathbb{M}$ be a compact topological space.
Let $f : \mathbb{M} \to \mathbb{R}$ be continuous.
We consider all sets of the form $\mathbb{M}_t := f^{-1}((-\infty, t])$.



Persistent Homology

Lower-Level Set Filtration

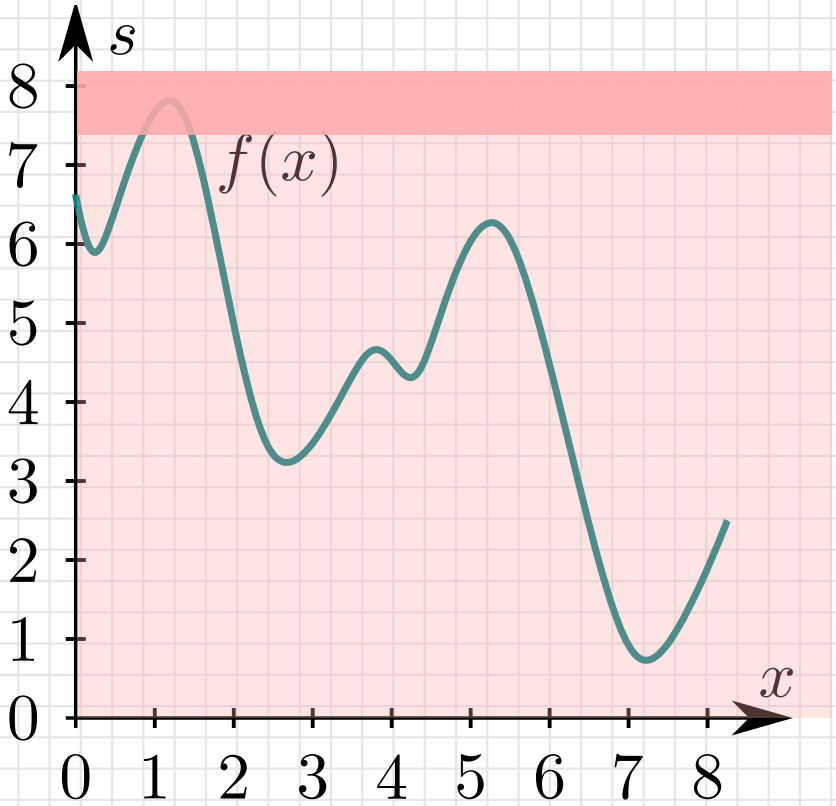
Let $\mathbb{M}$ be a compact topological space.
Let $f : \mathbb{M} \to \mathbb{R}$ be continuous.
We consider all sets of the form $\mathbb{M}_t := f^{-1}((-\infty, t])$.



Persistent Homology

Lower-Level Set Filtration

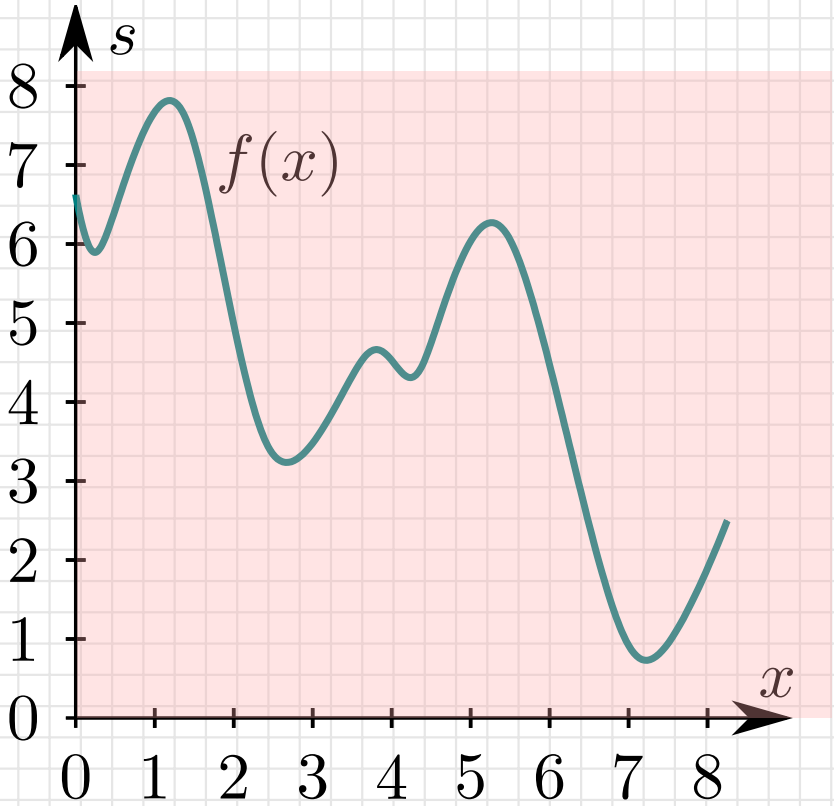
Let $\mathbb{M}$ be a compact topological space.
Let $f : \mathbb{M} \to \mathbb{R}$ be continuous.
We consider all sets of the form $\mathbb{M}_t := f^{-1}((-\infty, t])$.



Persistent Homology

Lower-Level Set Filtration

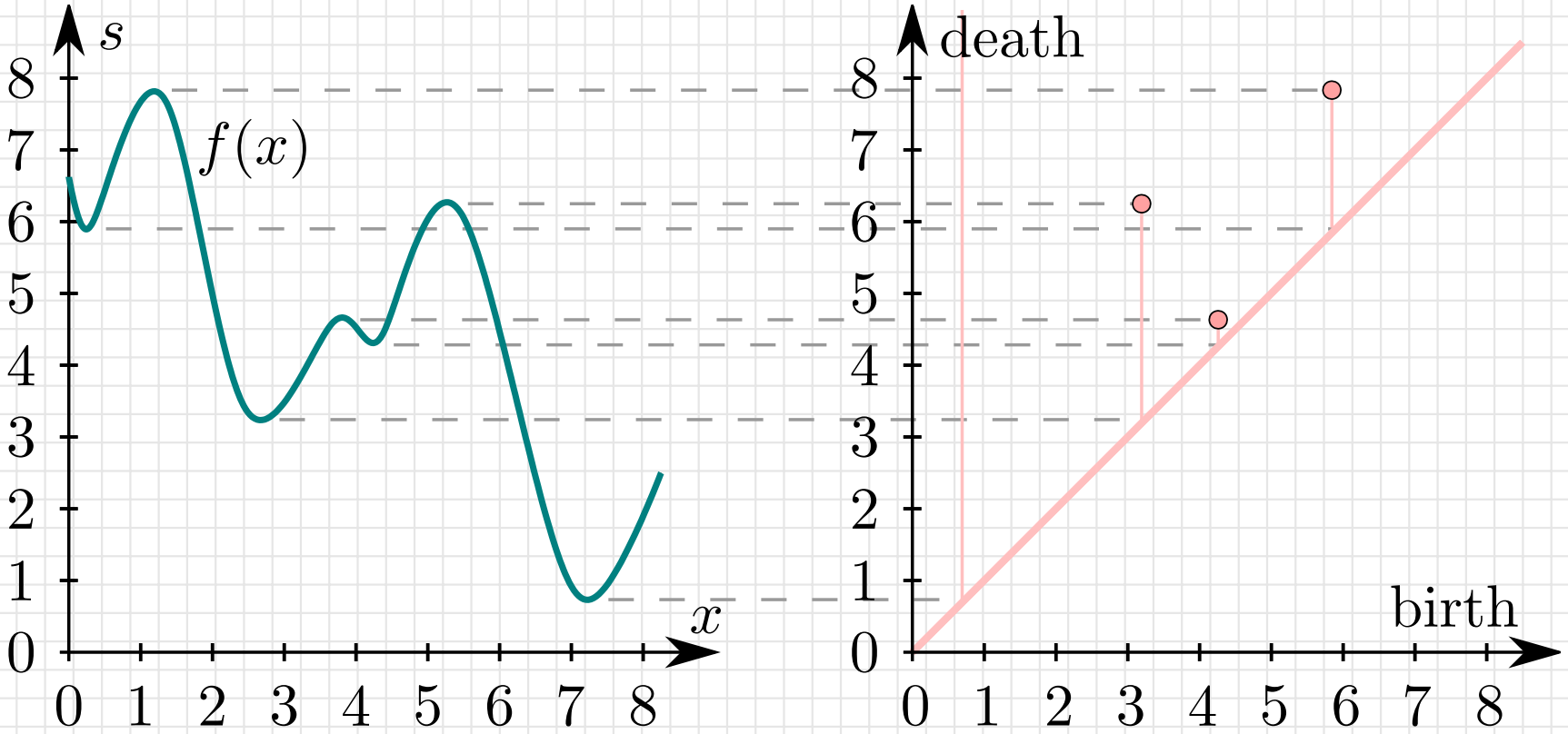
Let $\mathbb{M}$ be a compact topological space.
Let $f : \mathbb{M} \to \mathbb{R}$ be continuous.
We consider all sets of the form $\mathbb{M}_t := f^{-1}((-\infty, t])$.



Persistent Homology

Lower-Level Set Filtration

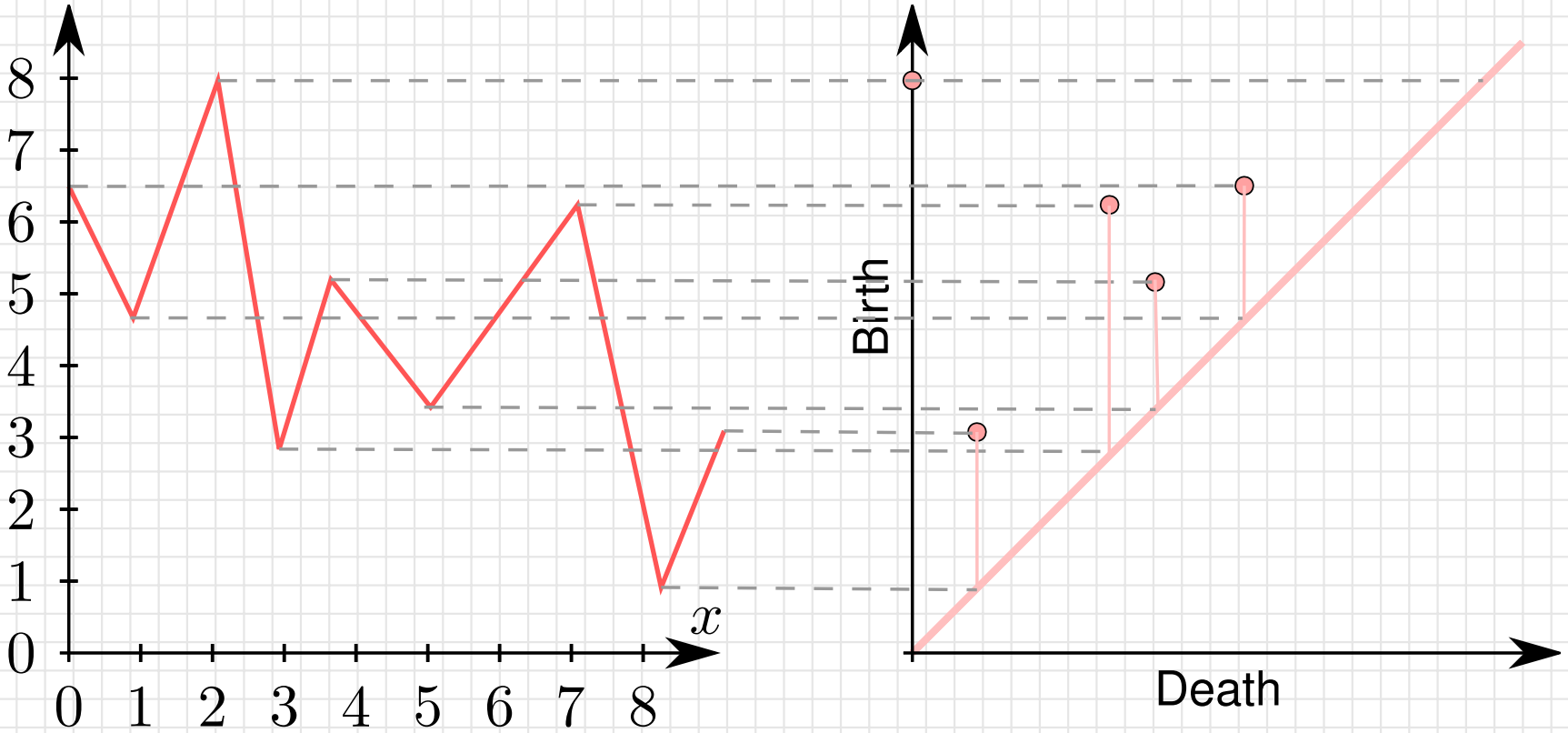
Let $\mathbb{M}$ be a compact topological space.
Let $f : \mathbb{M} \to \mathbb{R}$ be continuous.
We consider all sets of the form $\mathbb{M}_t := f^{-1}((-\infty, t])$.



Persistent Homology

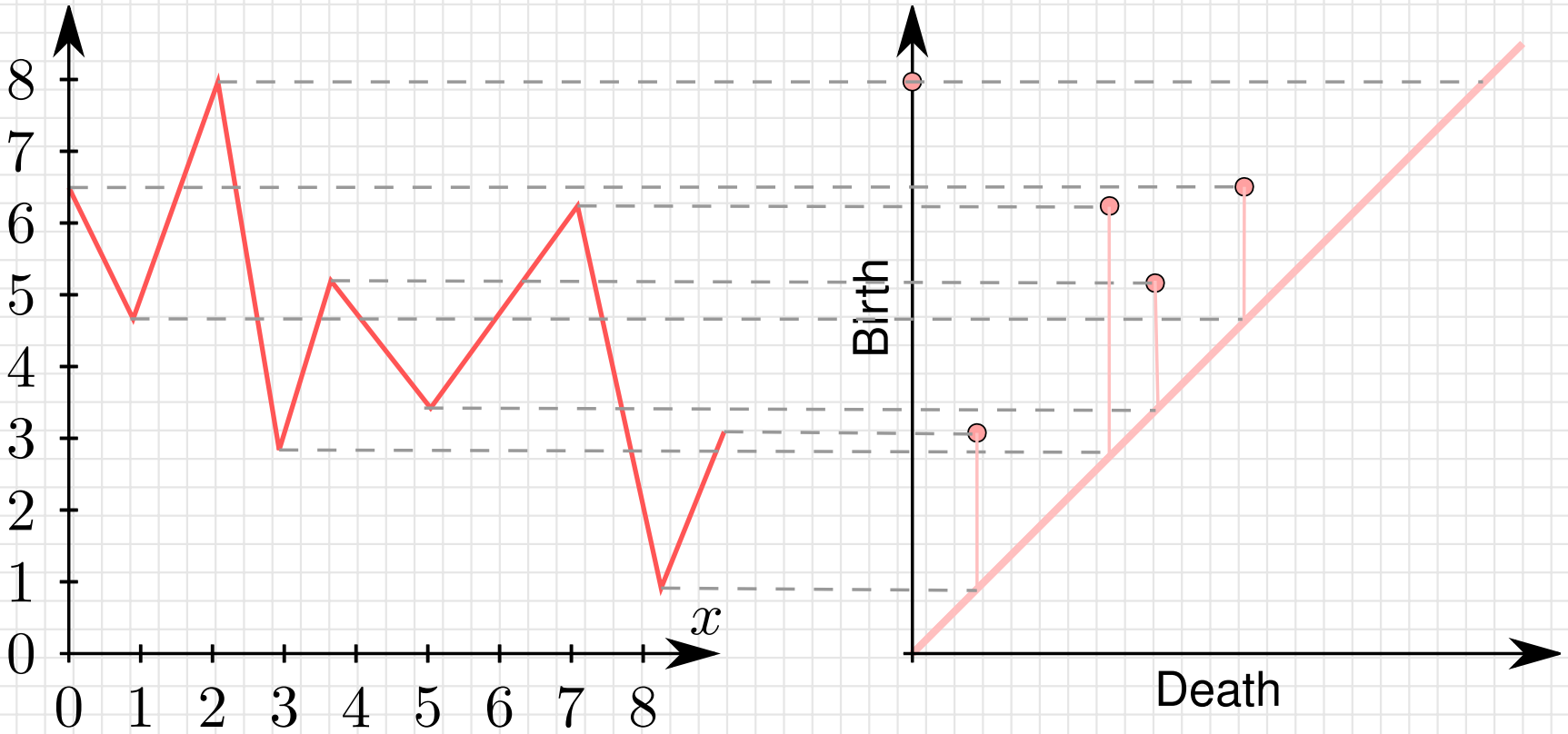
Upper-Level Set Filtration

Let $\mathbb{M}$ be a compact topological space.
Let $f : \mathbb{M} \to \mathbb{R}$ be continuous.
We consider all sets of the form $\mathbb{M}^t := f^{-1}([t, \infty))$.



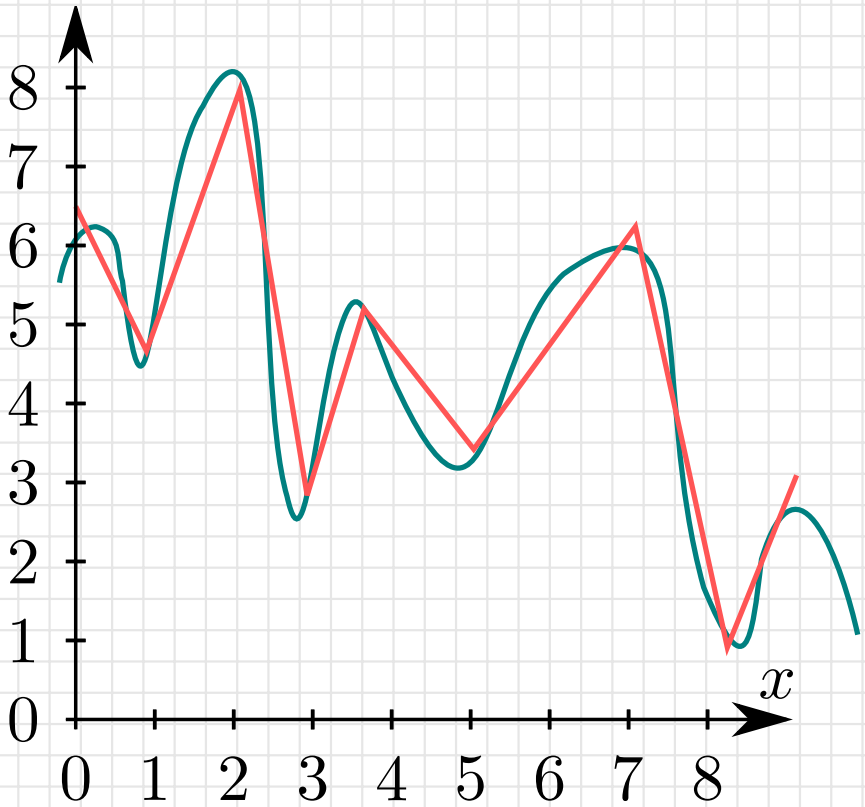
Distance and Stability

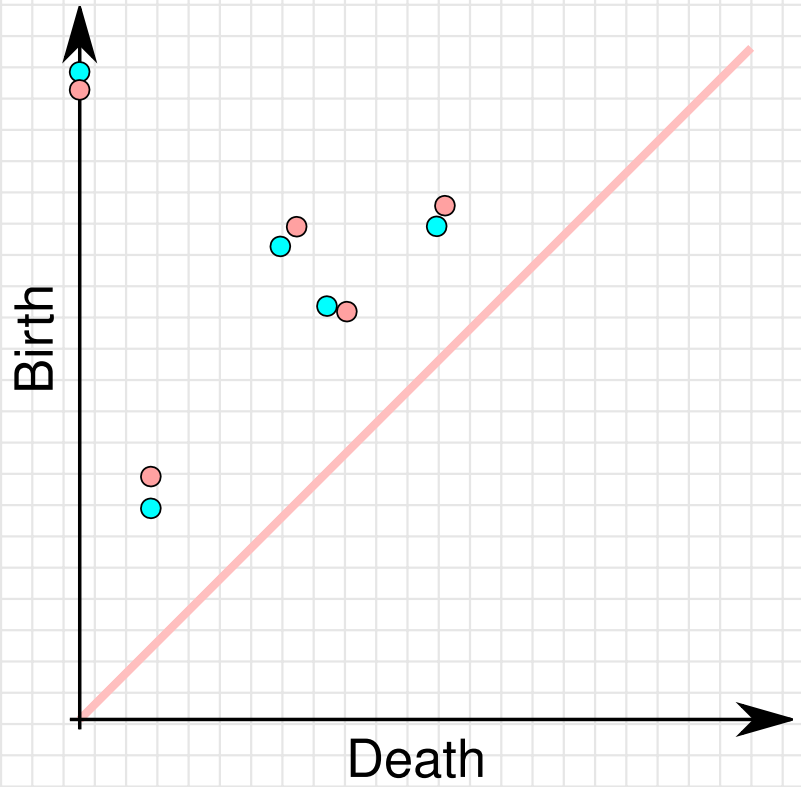
A Metric on Persistence Diagrams



Distance and Stability

A Metric on Persistence Diagrams





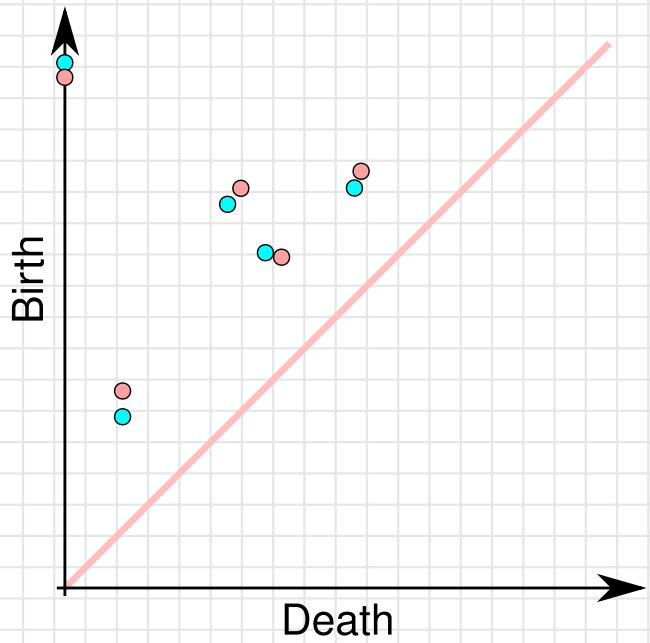
Bottleneck Distance

Given two persistence diagrams, find the best *perfect matching* M between the point sets.

Minimize Cost

We wish to find

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



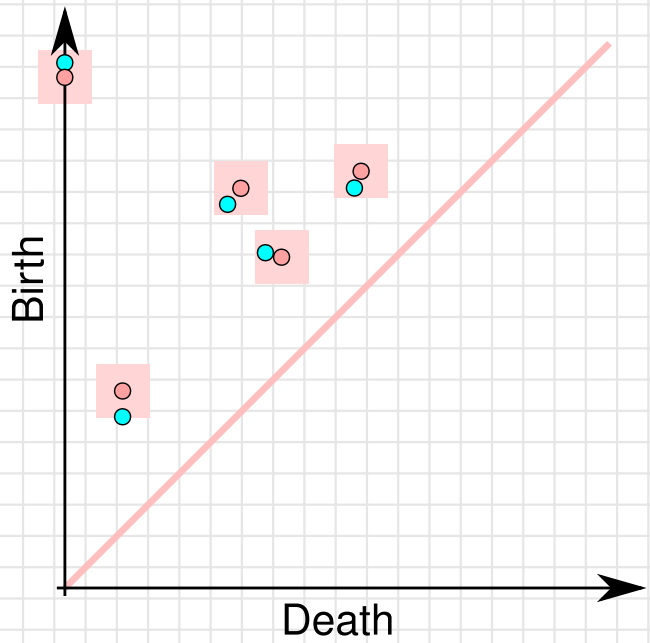
Bottleneck Distance

Given two persistence diagrams, find the best *perfect matching* M between the point sets.

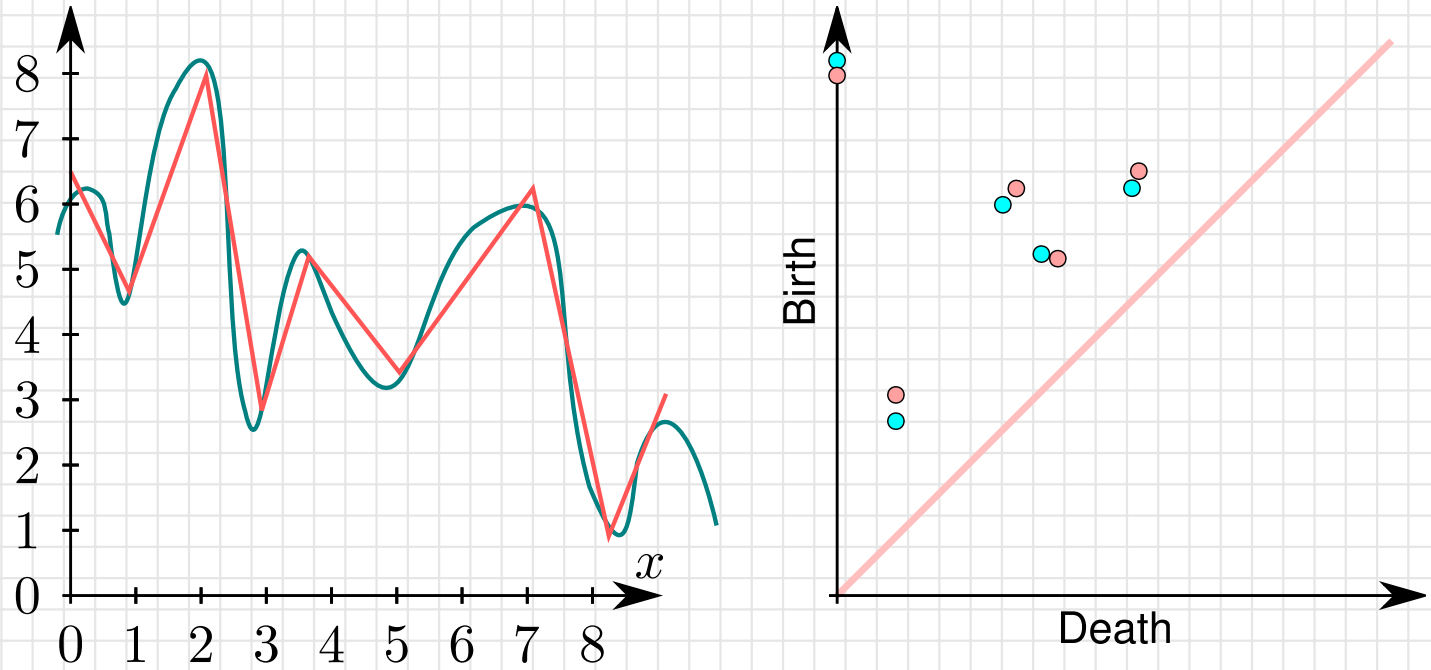
Minimize Cost

We wish to find

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



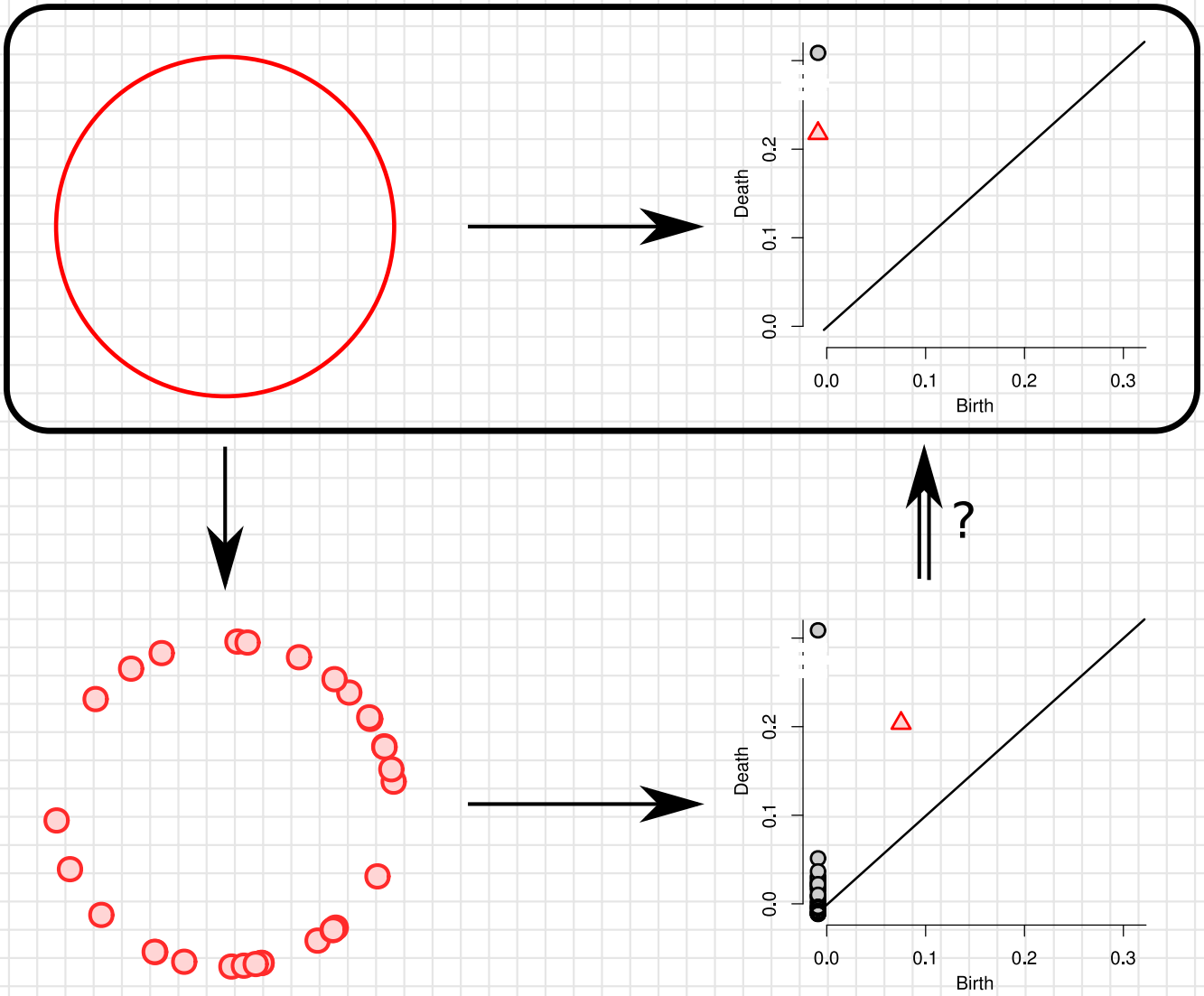
Stability of Matchings



Bottleneck Stability Theorem [CDGO]

$$||f_1 - f_2||_\infty \quad \geq \quad W_\infty(D_1, D_2)$$

Topological Inference



Objective

Let $\mathcal{D}_T$ denote the set of all T-bounded persistence diagrams.

Confidence Sets

Given $\alpha \in (0,1)$ and unknown diagram D, we want $\mathcal{C}_\alpha \subset \mathcal{D}_T$ such that

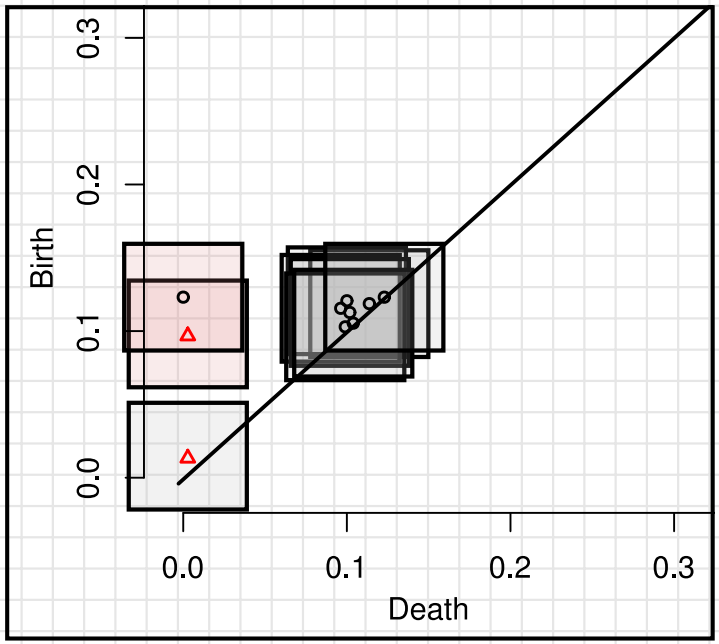
$$\mathbb{P}(D \in \mathcal{C}_\alpha) \geq 1 - \alpha.$$

Question

If $\widehat{D}$ is an estimate of D, how close is $\widehat{D}$ to D ?

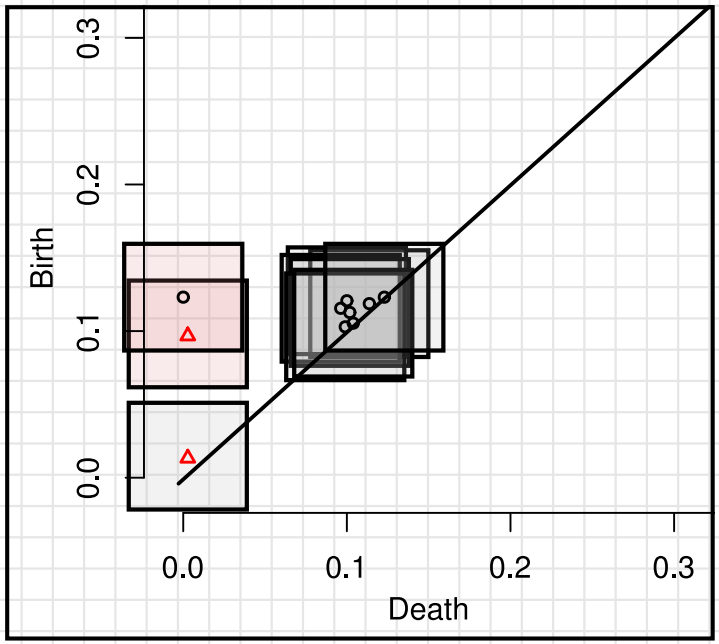
Confidence Sets for Persistent Diagrams

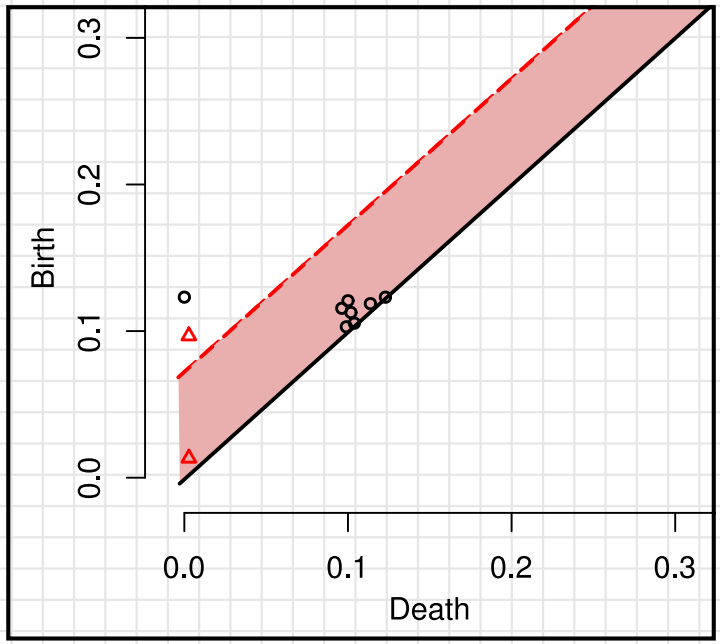
$$C_\alpha = \{D \in \mathcal{D}_T \ : \ W_\infty(D, \hat{D}) \leq t(\alpha)\}$$



Confidence Sets for Persistent Diagrams

$$C_\alpha = \{D \in \mathcal{D}_T \ : \ W_\infty(D, \hat{D}) \leq t(\alpha)\}$$





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 $\widehat{D}_n = \mathrm{Dgm}^-(X_1, \ldots, X_n)$

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 $\widehat{D}_n = \text{Dgm}^-(X_1, \ldots, X_n)$

Question

How does $\widehat{D}_n$ compare to $\mathbb{E}(\widehat{D}_n) \approx \text{Dgm}^-(\mathbb{M})$?

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 $\widehat{D}_n = \text{Dgm}^-(X_1, \ldots, X_n)$

Question

How does $\widehat{D}_n$ compare to $\mathbb{E}(\widehat{D}_n) \approx \text{Dgm}^-(\mathbb{M})$?

Answer

Find $\mathcal{C}_\alpha$ such that $\mathbb{P}(\mathbb{E}(\widehat{D}_n) \in \mathcal{C}_\alpha) \geq 1 - \alpha$.

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 $\widehat{D}_n = \text{Dgm}^-(X_1, \ldots, X_n)$

Question

How does $\widehat{D}_n$ compare to $\mathbb{E}(\widehat{D}_n) \approx \text{Dgm}^-(\mathbb{M})$?

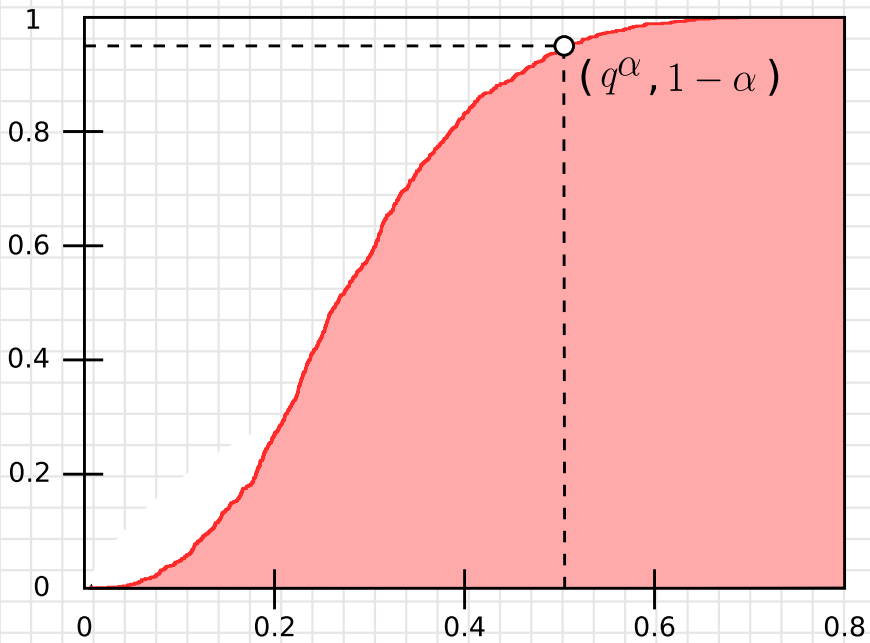
Answer

Find $\mathcal{C}_\alpha$ such that $\mathbb{P}(\mathbb{E}(\widehat{D}_n) \in \mathcal{C}_\alpha) \geq 1 - \alpha$. How?

Computing a Confidence Interval

With Infinite Resources

Repeatedly sample n data points, obtaining:



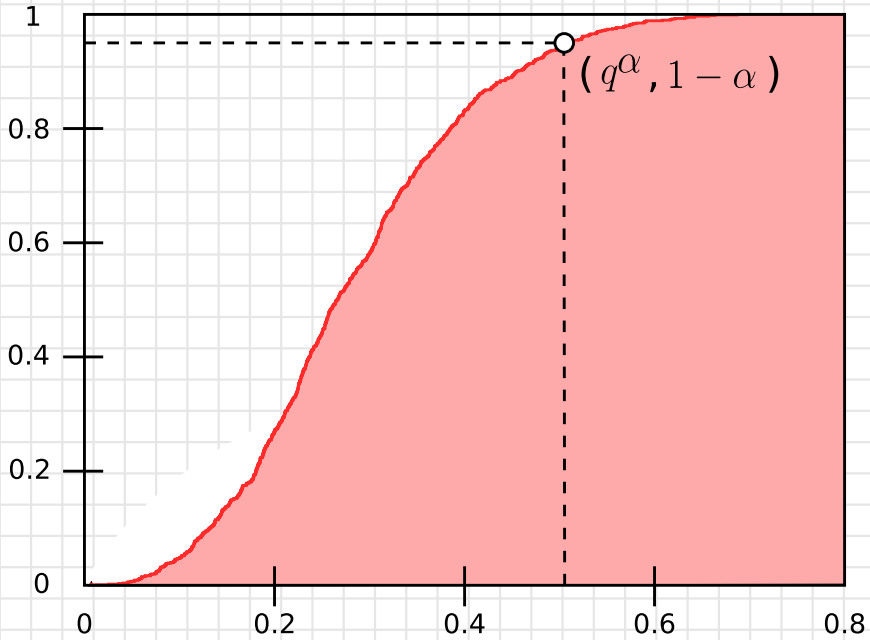
Confidence Intervals

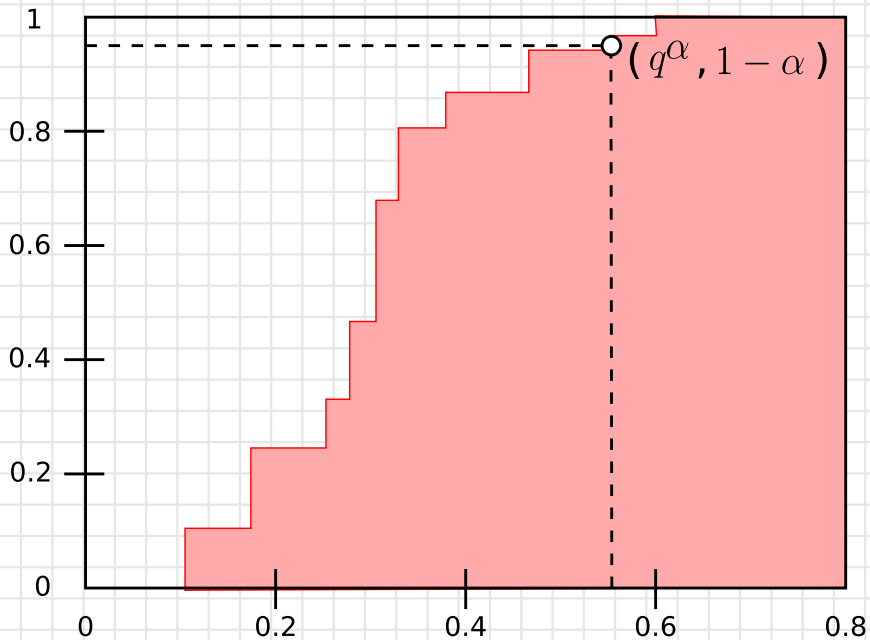
$$\mathbb{P}(\Theta_n(\mathbb{M}) \in [0, q^\alpha]) \geq 1 - \alpha.$$

Computing a Confidence Interval

With Infinite Resources

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





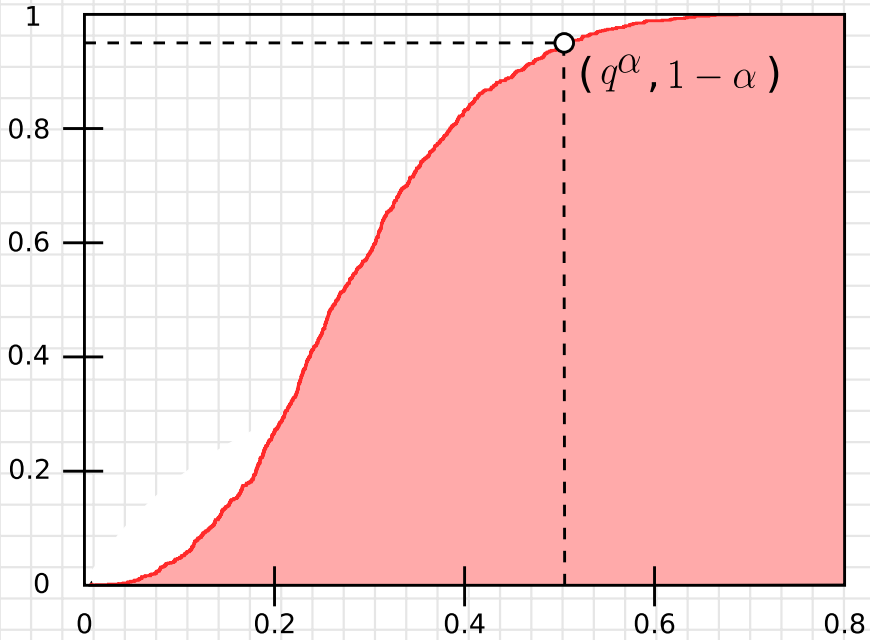
Confidence Intervals

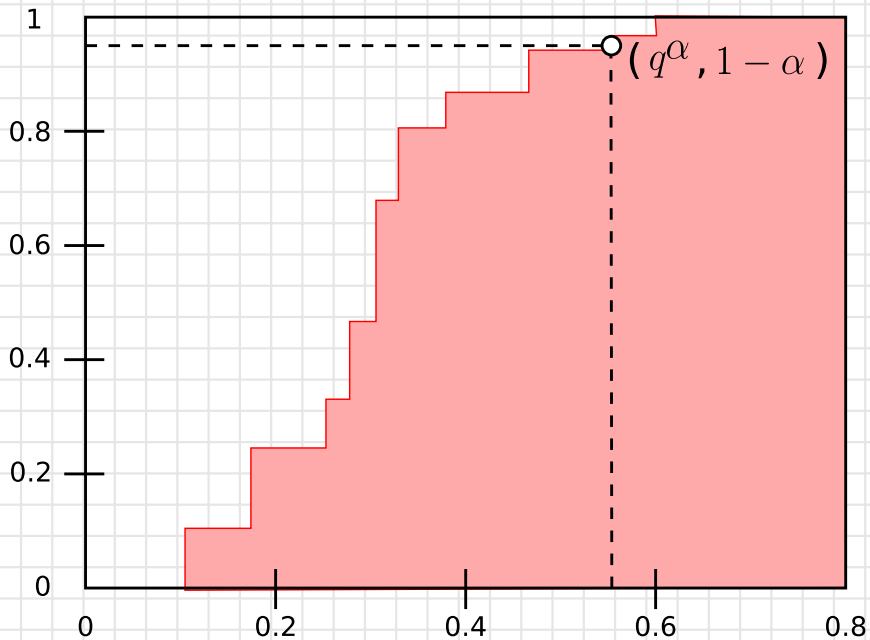
$$\mathbb{P}(\Theta_n(\mathbb{M}) \in [0, q^\alpha]) \geq 1 - \alpha.$$

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

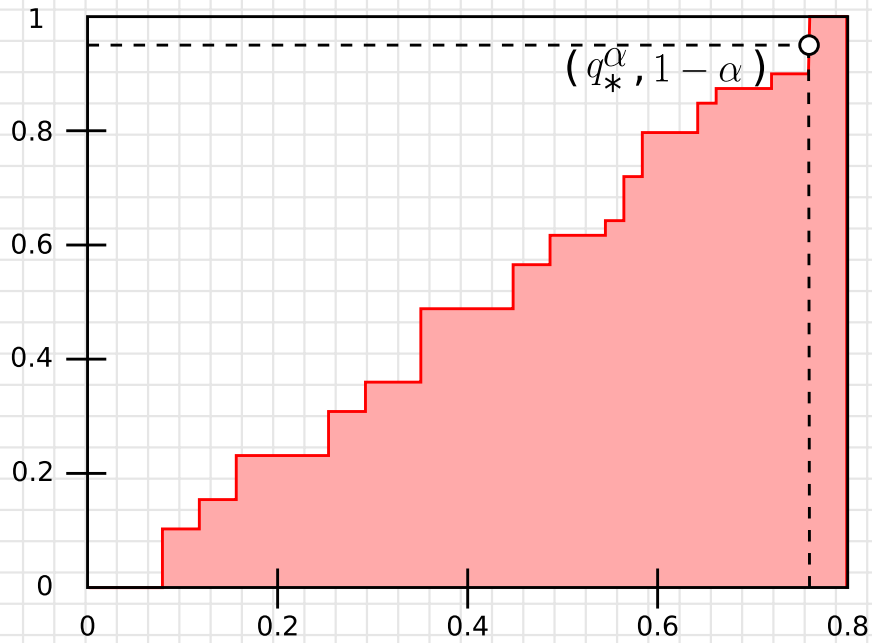
When We Can Only Take One Sample

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

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

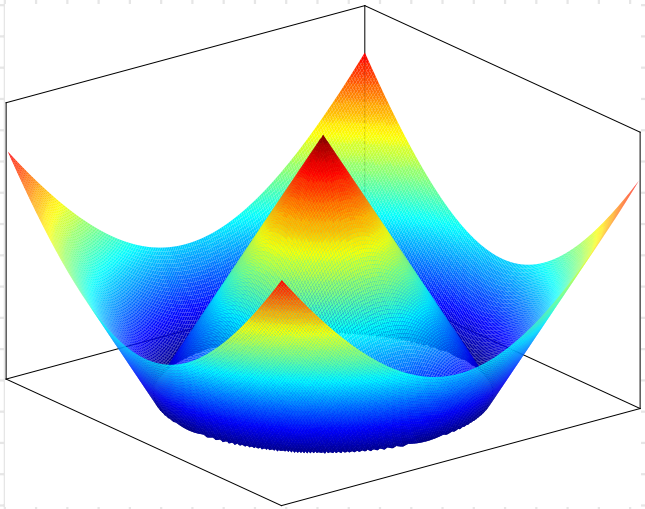
Compute $\hat{\Theta}_n^* = \Theta(X_1^*, \ldots, X_n^*)$.

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



Distance to a Subset

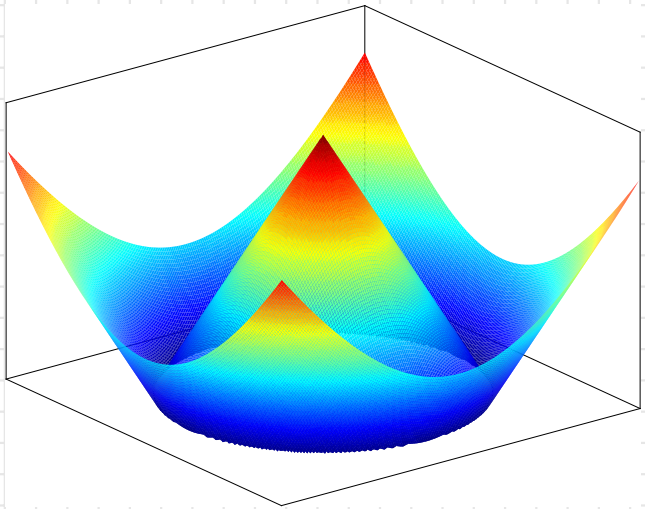
$d_{\mathbb{M}}(a) = \inf_{x \in \mathbb{M}} ||x - a||$
$D = \mathrm{Dgm}_p^-(d_{\mathbb{M}})$



Distance to a Subset

$d_{\mathbb{M}}(a) = \inf_{x \in \mathbb{M}} ||x - a||$
$D = \text{Dgm}_p^-(d_{\mathbb{M}})$

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



Subsampling

A Variant of the Bootstrap

Let $\mathcal{S}_b^i$ be a subsample of size $b < n$, for $i = 1, \ldots, B$.
L_b is the CDF of $H(\mathcal{S}_b^i, \mathcal{S}_n)$.

Subsampling

A Variant of the Bootstrap

Let $\mathcal{S}_b^i$ be a subsample of size $b < n$, for $i = 1, \ldots, B$.
L_b is the CDF of $H(\mathcal{S}_b^i, \mathcal{S}_n)$.

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

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

$$\mathbb{P}\left(||d_{\mathbb{M}} - d_{\mathcal{S}_n}||_\infty > 2L_b^{-1}(\alpha)\right) \leq \alpha + O\left(\sqrt{1/n}\right)$$

Subsampling

A Variant of the Bootstrap

Let $\mathcal{S}_b^i$ be a subsample of size $b < n$, for $i = 1, \ldots, B$.
L_b is the CDF of $H(\mathcal{S}_b^i, \mathcal{S}_n)$.

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

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

$$\mathbb{P}\left(||d_{\mathbb{M}} - d_{\mathcal{S}_n}||_\infty > 2L_b^{-1}(\alpha)\right) \leq \alpha + O\left(\sqrt{1/n}\right)$$

[RS-2002] On the uniform asymptotic validity of subsampling and the bootstrap. Annals of Statistics.

Putting It All Together

Subsampling Theorem

$$\mathbb{P}\left(||d_{\mathbb{M}} - d_{\mathcal{S}_n}||_\infty > 2L_b^{-1}(\alpha)\right) \leq \alpha + O\left(\sqrt{1/n}\right)$$

Putting It All Together

Subsampling Theorem

$$\mathbb{P}\left(||d_{\mathbb{M}} - d_{\mathcal{S}_n}||_\infty > 2L_b^{-1}(\alpha)\right) \leq \alpha + O\left(\sqrt{1/n}\right)$$

Bottleneck Stability Theorem

$$||d_{\mathbb{M}} - d_{\mathcal{S}_n}||_\infty \quad \geq \quad W_\infty(D, \widehat{D}_n)$$

Putting It All Together

Subsampling Theorem

$$\mathbb{P}\left(||d_{\mathbb{M}} - d_{\mathcal{S}_n}||_\infty > 2L_b^{-1}(\alpha)\right) \leq \alpha + O\left(\sqrt{1/n}\right)$$

Bottleneck Stability Theorem

$$||d_{\mathbb{M}} - d_{\mathcal{S}_n}||_\infty \quad \geq \quad W_\infty(D, \widehat{D}_n)$$

Confidence Sets for Persistence Diagrams

$$\mathbb{P}\left(W_\infty(D, \widehat{D}_n) > 2L_b^{-1}(\alpha)\right) \leq \alpha + O\left(\sqrt{1/n}\right)$$

Putting It All Together

Subsampling Theorem

$$\mathbb{P}\left(||d_{\mathbb{M}} - d_{\mathcal{S}_n}||_\infty > 2L_b^{-1}(\alpha)\right) \leq \alpha + O\left(\sqrt{1/n}\right)$$

Bottleneck Stability Theorem

$$||d_{\mathbb{M}} - d_{\mathcal{S}_n}||_\infty \quad \geq \quad W_\infty(D, \widehat{D}_n)$$

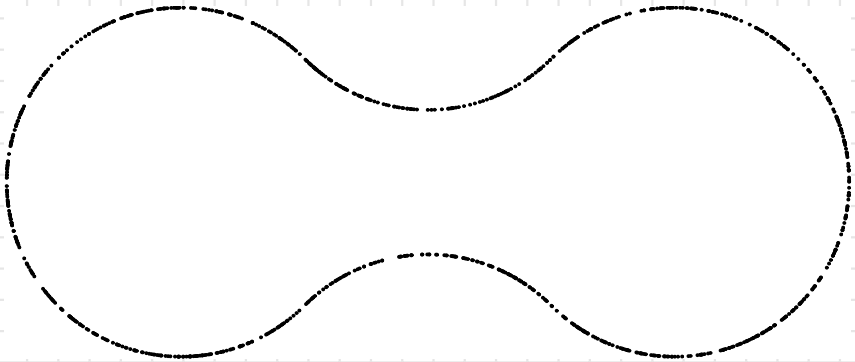
Confidence Sets for Persistence Diagrams

$$\mathbb{P}\left(W_\infty(D, \widehat{D}_n) > 2L_b^{-1}(\alpha)\right) \leq \alpha + O\left(\sqrt{1/n}\right)$$

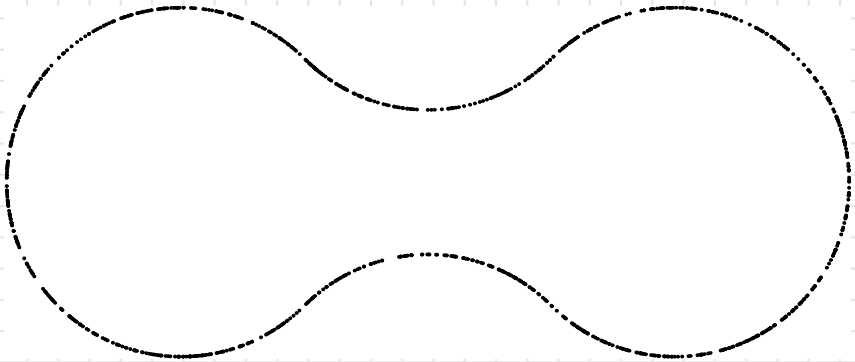
Asymptotic Confidence Sets for Persistence Diagrams

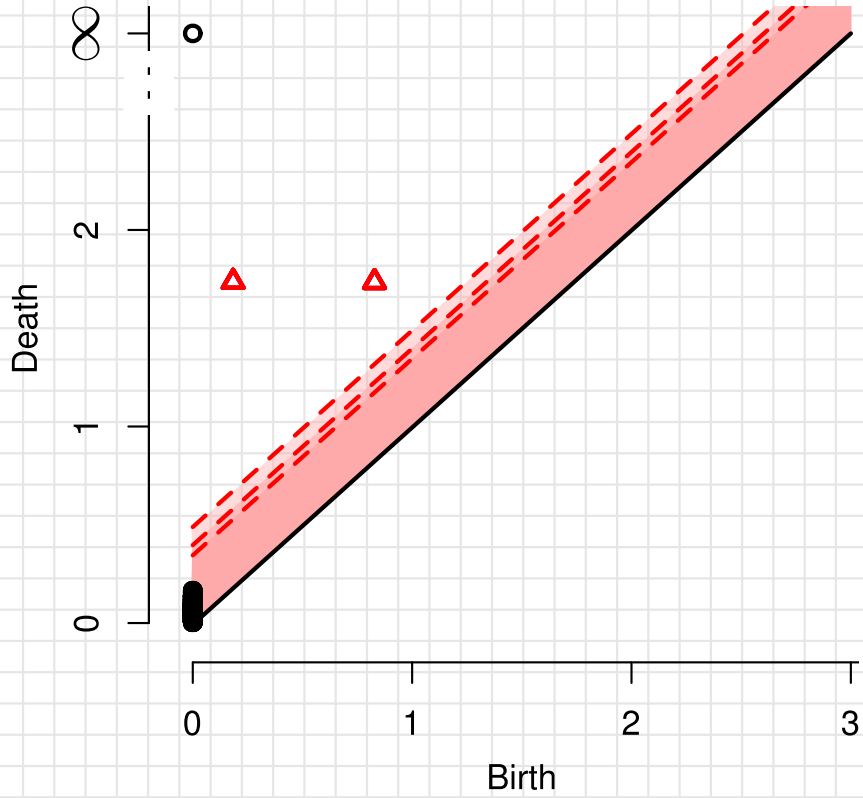
$$\lim_{n\to\infty} \mathbb{P}\left(W_\infty(D, \widehat{D}_n) \leq 2L_b^{-1}(\alpha)\right) \quad \geq \quad 1 - \alpha$$

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*

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

Theorem (Method of Shells)

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

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

These Methods are Different

Concentration of Measure

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

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

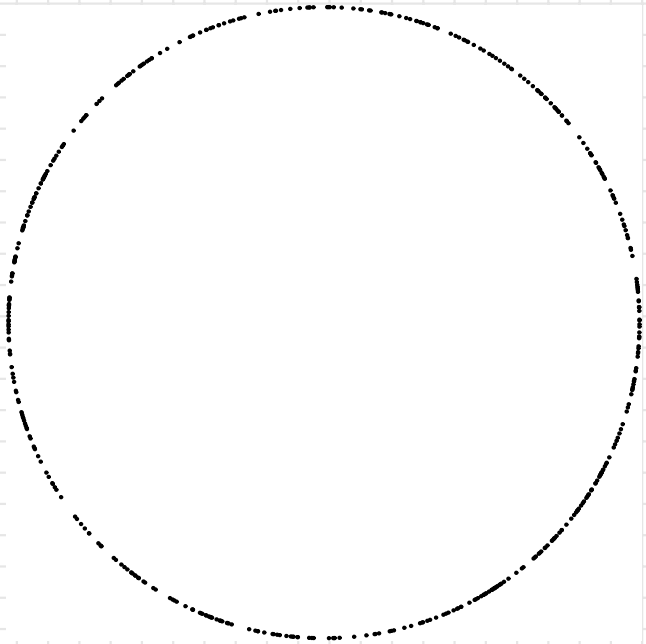
Shells

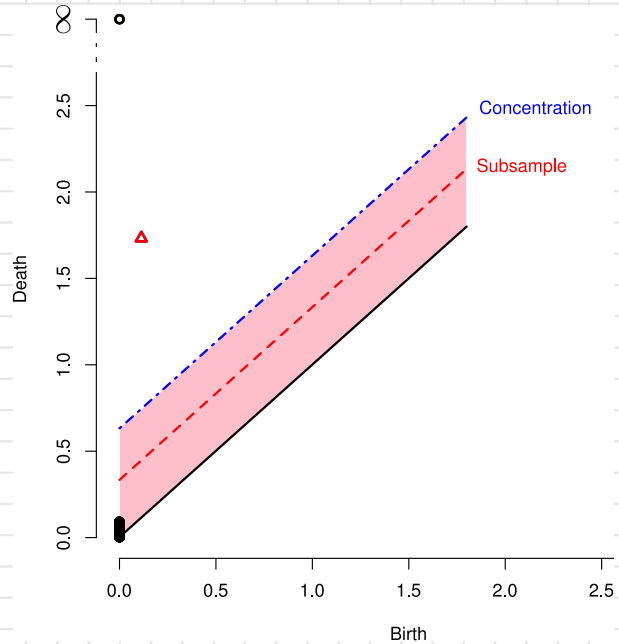
$\hat{t}_s$ is found by solving the following for t:

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

Distance Function Examples

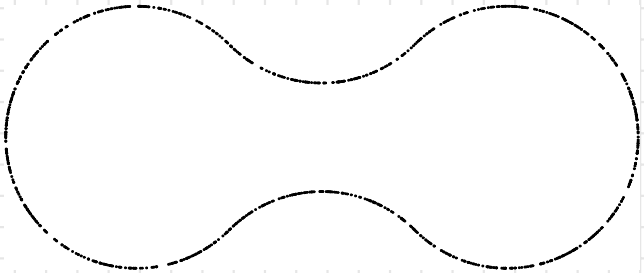
Uniform Distribution on Unit Circle

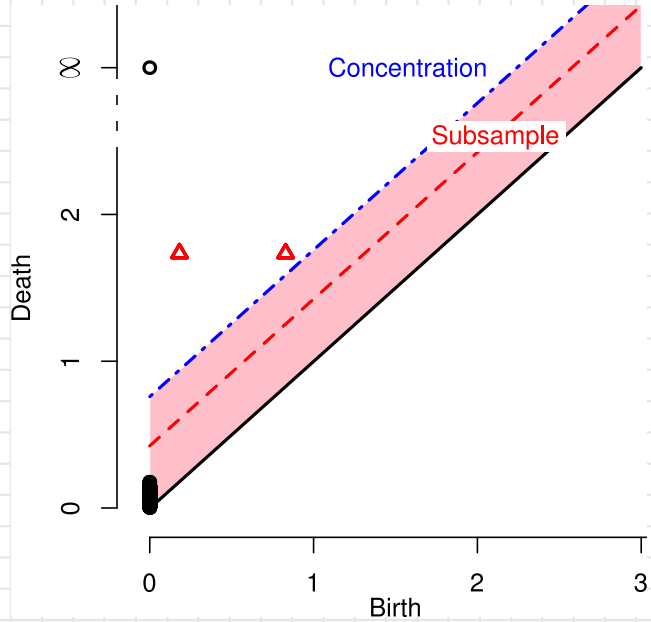




Distance Function Examples

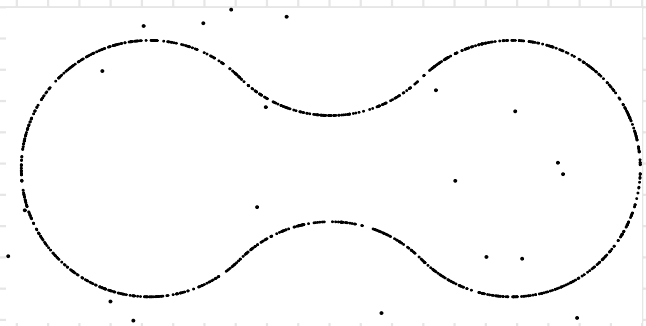
Uniform Distribution on Cassini Curve

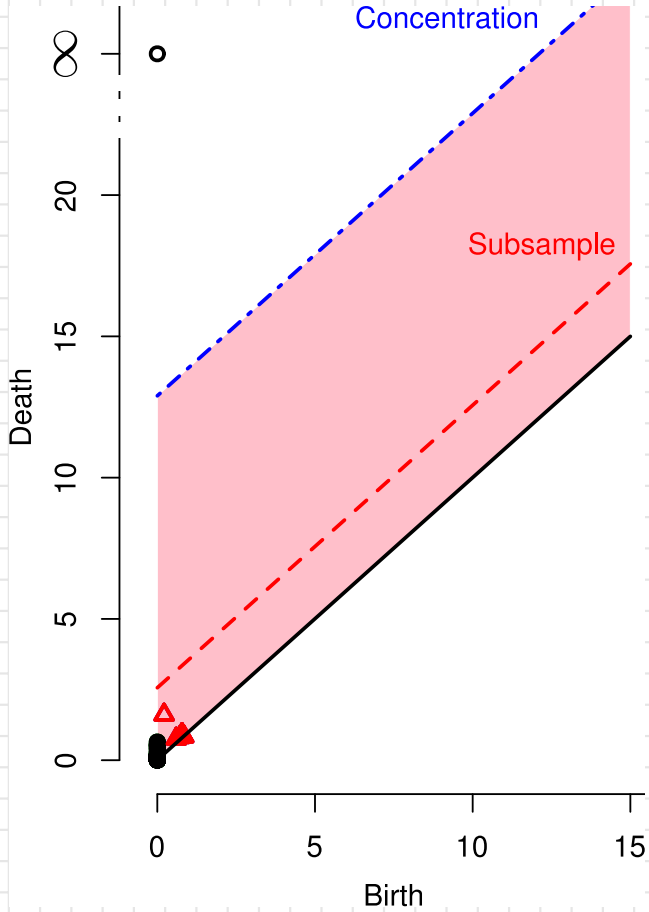




Distance Function Examples

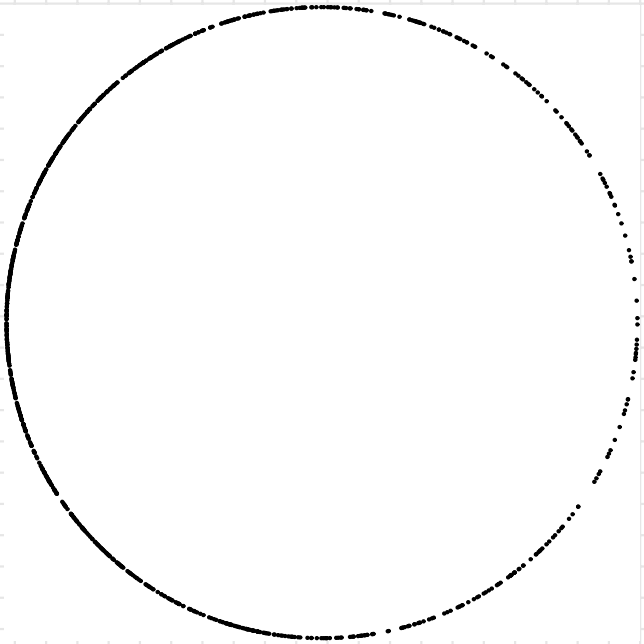
Cassini Curve with Outliers

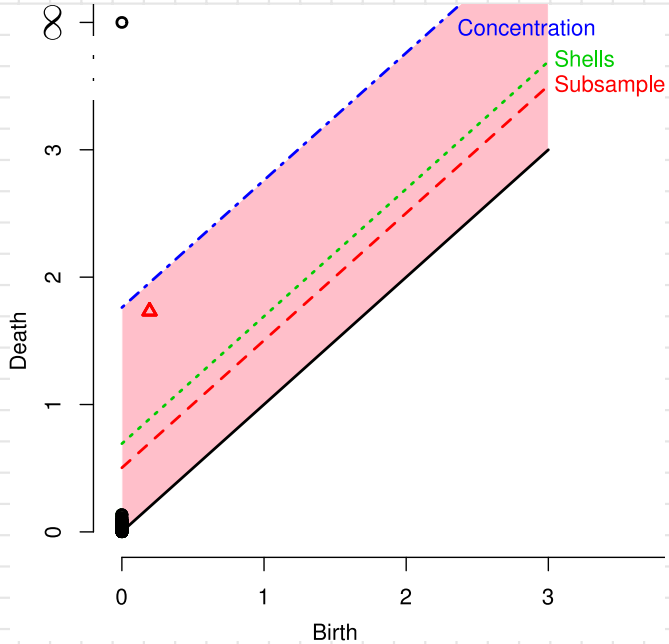




Distance Function Examples

Normal Distribution on Unit Circle





The Intersection of Statistics and Topology:

Stochastic Convergence of Persistence Landscapes and Silhouettes

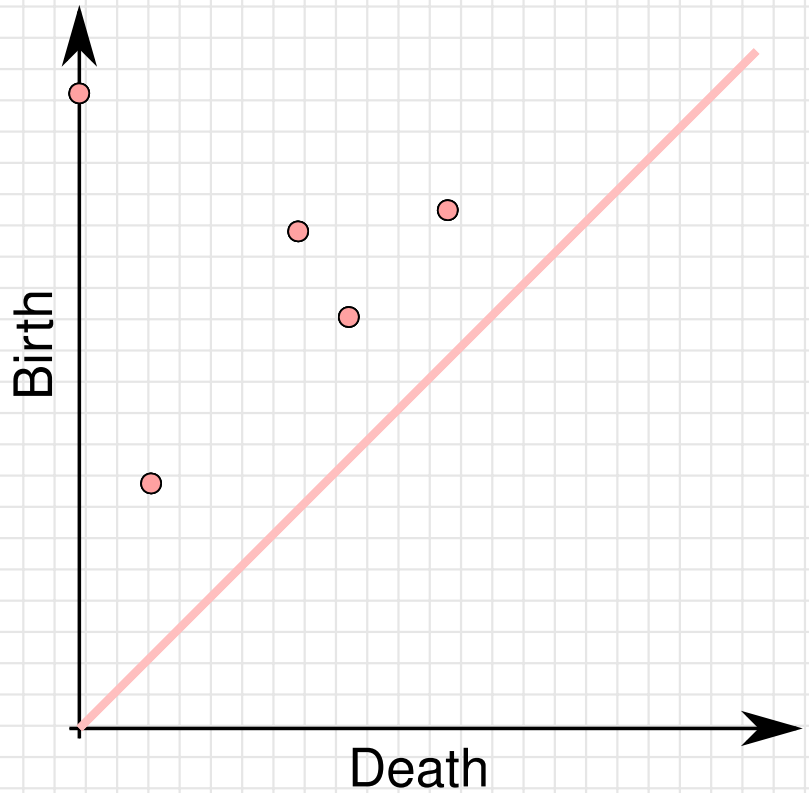
Brittany Terese Fasy

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

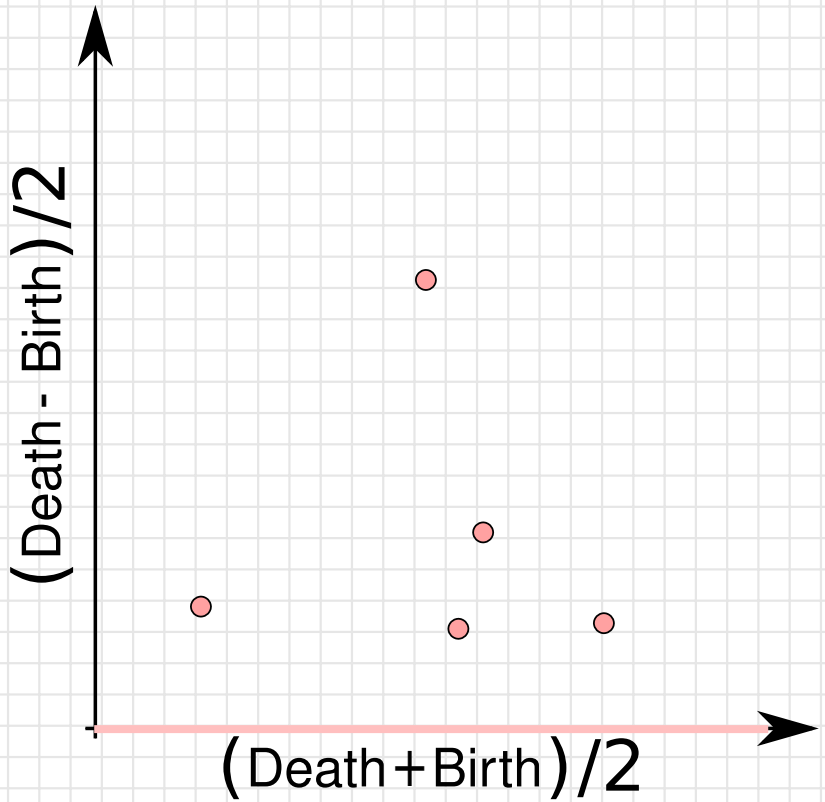
3 June 2014

[C**F**LRW] Stochastic Convergence of Persistence Landscapes and Silhouettes. ArXiv. SoCG 2014 Proceedings, Kyoto. Journal version pending.

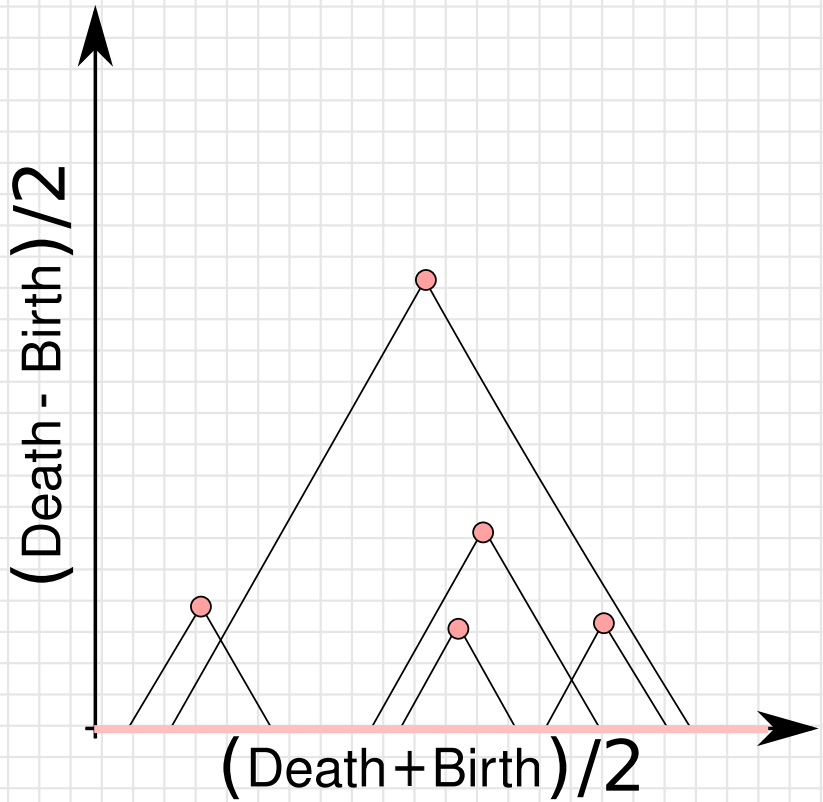
Persistence Landscapes



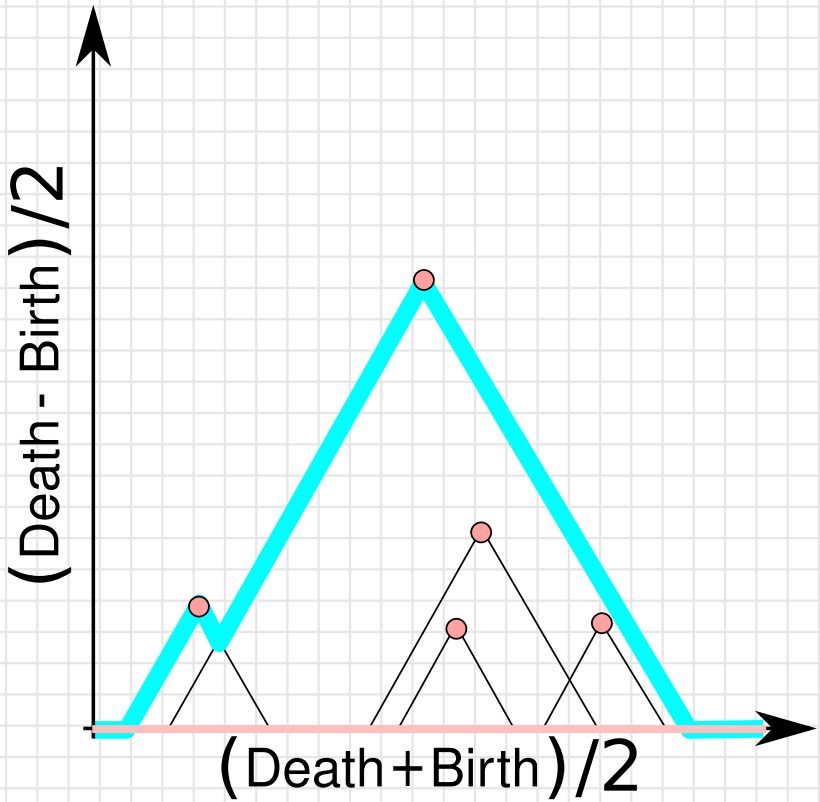
Persistence Landscapes



Persistence Landscapes

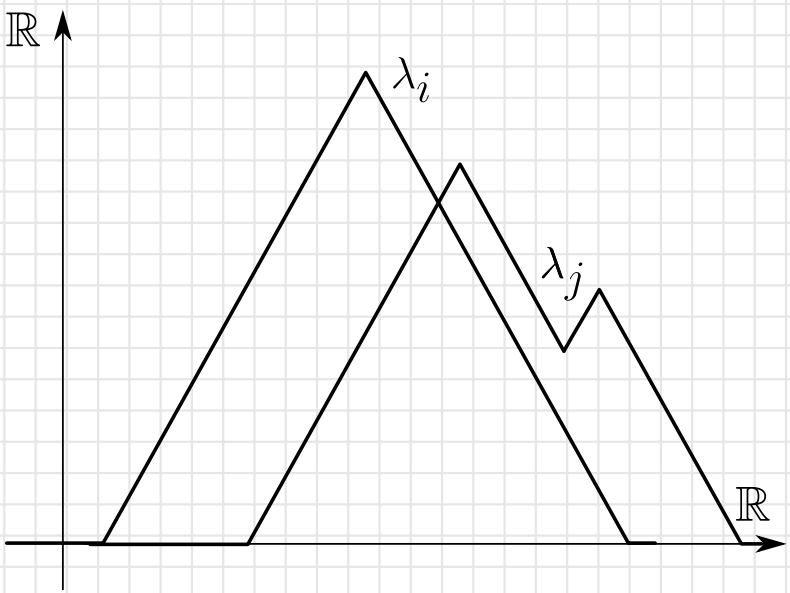


Persistence Landscapes



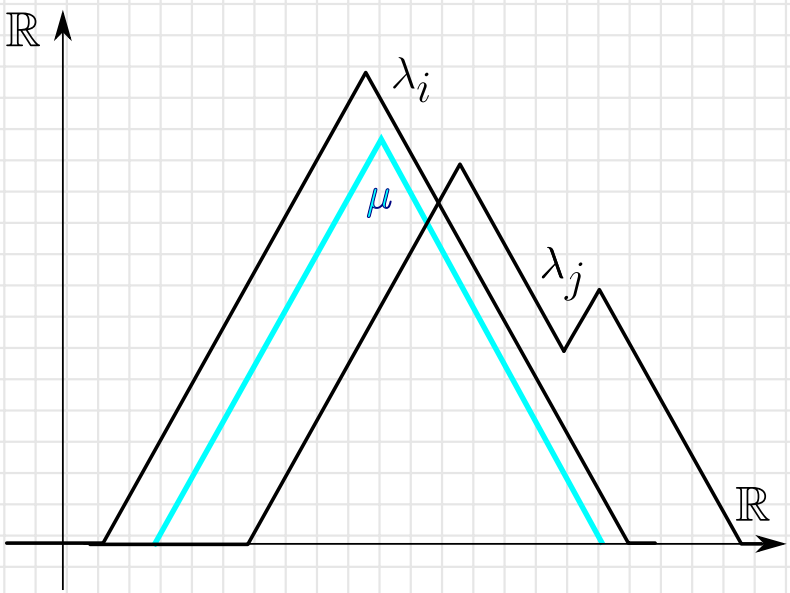
[B-2012] Statistical Topology Using Persistent Homology. ArXiv 1207.6437

Weak Convergence of Landscapes



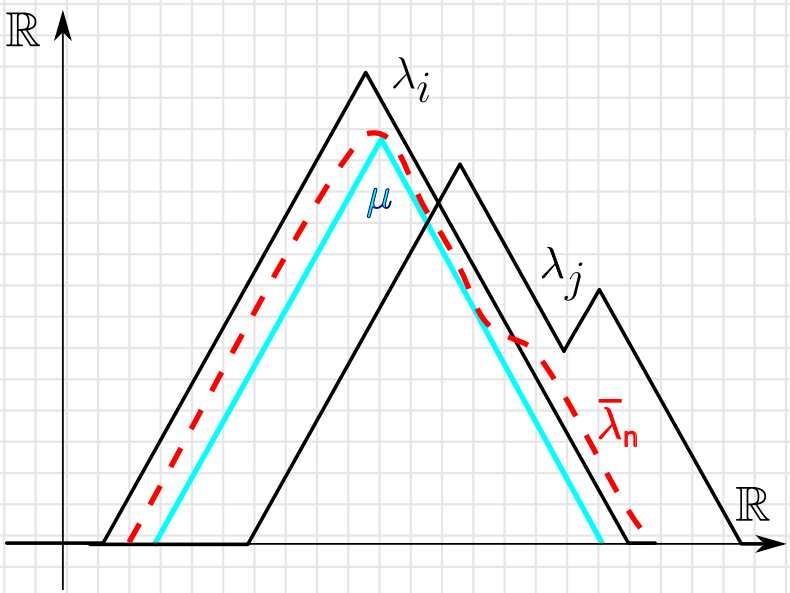
Let $\lambda_1, \ldots, \lambda_n \sim \mathcal{L}_T$.

Weak Convergence of Landscapes



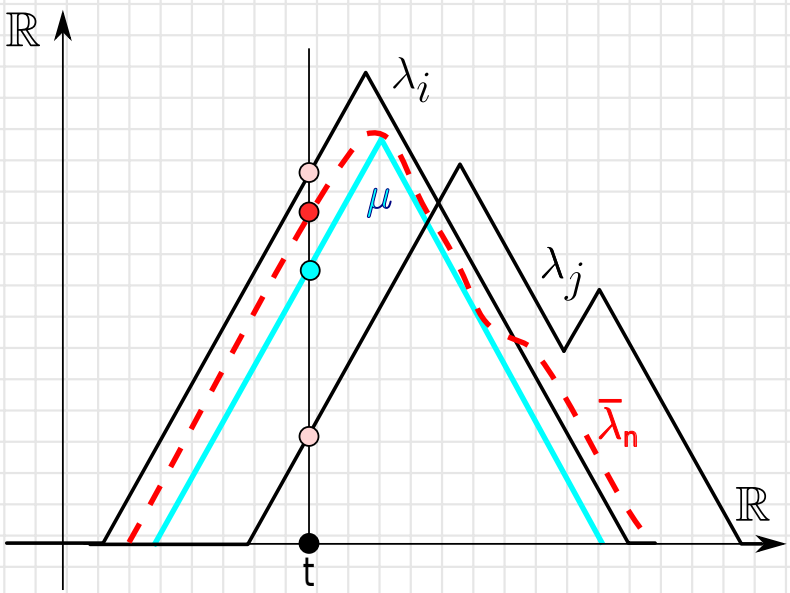
Let $\lambda_1, \ldots, \lambda_n \sim \mathcal{L}_T$.
λ : true (unknown) landscape
$\mu = \mathbb{E}(\lambda_i)$

Weak Convergence of Landscapes



Let $\lambda_1, \ldots, \lambda_n \sim \mathcal{L}_T$.
λ : true (unknown) landscape
$\mu = \mathbb{E}(\lambda_i)$
$\bar{\lambda}_n$: average landscape

Weak Convergence of Landscapes

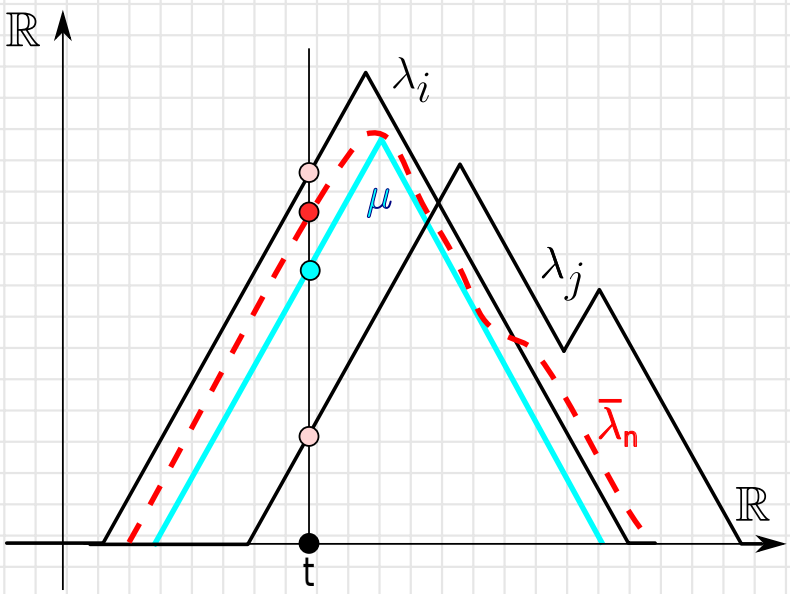


Let $\lambda_1, \ldots, \lambda_n \sim \mathcal{L}_T$.
λ : true (unknown) landscape
$\mu = \mathbb{E}(\lambda_i)$
$\bar{\lambda}_n$: average landscape

Pointwise Convergence [B-2012].

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

Weak Convergence of Landscapes



Let $\lambda_1, \ldots, \lambda_n \sim \mathcal{L}_T$.
λ : true (unknown) landscape
$\mu = \mathbb{E}(\lambda_i)$
$\bar{\lambda}_n$: average landscape

Pointwise Convergence [B-2012].

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

Gaussian Process on $[0, T]$

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

Uniform Convergence

Weak Convergence

$\mathbb{G}_n$ converges weakly to the Brownian bridge $\mathbb{G}$ with covariance function

$$\kappa(f,g) = \int f(u)g(y)dP(u) - (\int f(u)dP(u))(\int g(u)dP(u)).$$

Uniform Convergence

Weak Convergence

$\mathbb{G}_n$ converges weakly to the Brownian bridge $\mathbb{G}$ with covariance function

$$\kappa(f,g) = \int f(u)g(y)dP(u) - (\int f(u)dP(u))(\int g(u)dP(u)).$$

Uniform CLT

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

$$\sup_{z\in\mathbb{R}} \left| \mathbb{P}\Big(\sup_{t\in[t_*,t^*]} |\mathbb{G}_n(t)| \le z \Big) - \mathbb{P}(W \le z) \right| = O\Big(\frac{(\log n)^{\frac{7}{8}}}{n^{\frac{1}{8}}} \Big).$$

Uniform Convergence

Weak Convergence

$\mathbb{G}_n$ converges weakly to the Brownian bridge $\mathbb{G}$ with covariance function

$$\kappa(f,g) = \int f(u)g(y)dP(u) - (\int f(u)dP(u))(\int g(u)dP(u)).$$

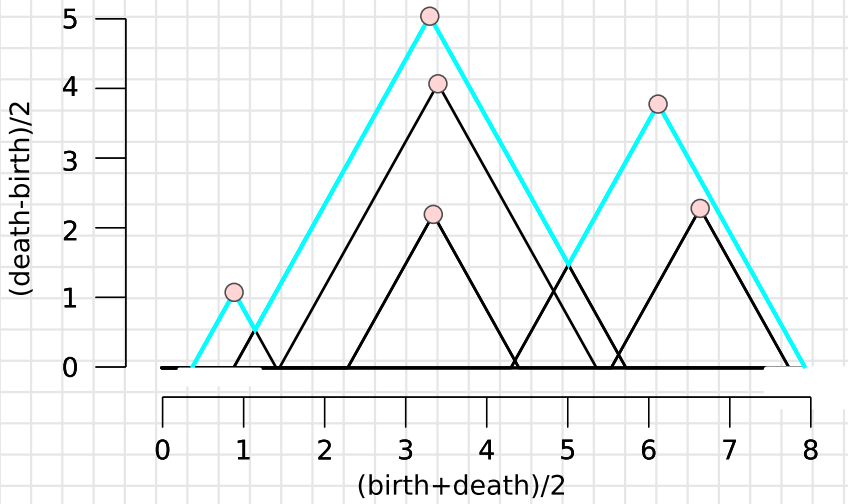
Uniform CLT

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

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

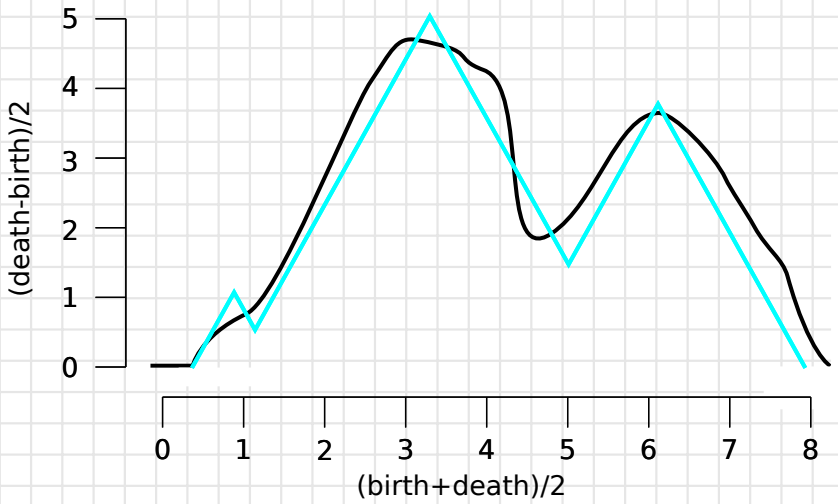
Proofs rely on [CCK-2013]: Anti-concentration and honest adaptive confidence bands. ArXiv 1303.7152.

Confidence Bands



$$\mathbb{E}||\lambda - \bar{\lambda}_n||_\infty \leq \overbrace{\mathbb{E}||\mu - \bar{\lambda}_n||_\infty}^{\text{variance}} + \underbrace{\mathbb{E}||\lambda - \mu||_\infty}_{\text{bais}}$$

Confidence Bands



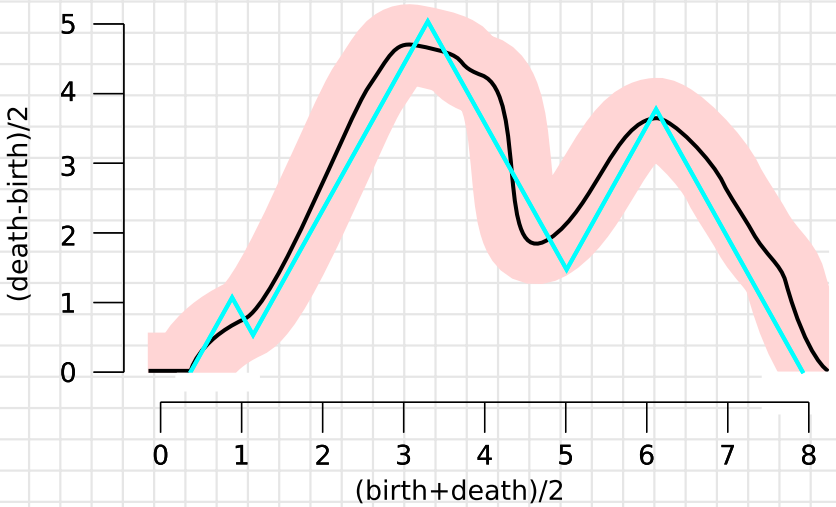
$$\mathbb{E}||\lambda - \bar{\lambda}_n||_\infty \leq \overbrace{\mathbb{E}||\mu - \bar{\lambda}_n||_\infty}^{\text{variance}}$$

Confidence Band

Find c such that $\ell_n = \bar{\lambda}_n - c$ and $u_n = \bar{\lambda}_n + c$ such that

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

Asymptotic Confidence Bands



$$\mathbb{E}||\lambda - \bar{\lambda}_n||_\infty \leq \overbrace{\mathbb{E}||\mu - \bar{\lambda}_n||_\infty}^{\text{variance}}$$

Asymptotic Confidence Band

Find c such that $\ell_n = \bar{\lambda}_n - c$ and $u_n = \bar{\lambda}_n + c$ such that

$$\lim_{n\to\infty} \mathbb{P}(\ell_n(t) \leq \mu(t) \leq u_n(t) \text{ for all } t) \geq 1 - \alpha.$$

Confidence Bands for Landscapes

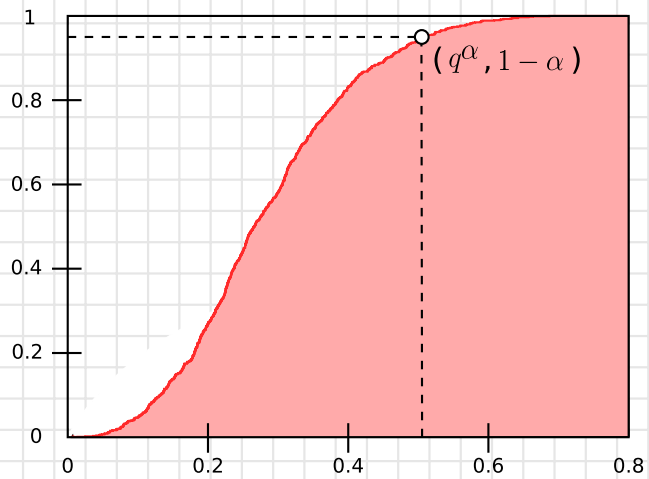
$\xi_1, \ldots, \xi_n \sim N(0,1)$

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

α-Quantile

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

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



Confidence Bands for Landscapes

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

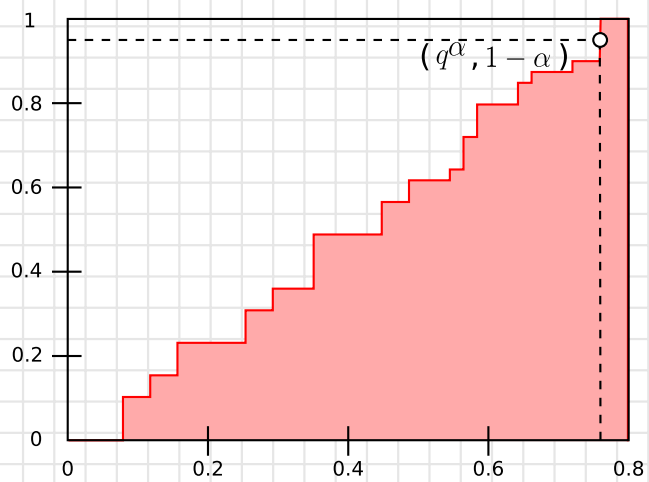
$$\widetilde{\mathbb{G}}_n(f_t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i (\lambda_i(t) - \bar{\lambda}_n(t))$$

α-Quantile

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

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

*Approx. $\widetilde{Z}_\alpha$ by MC simulation

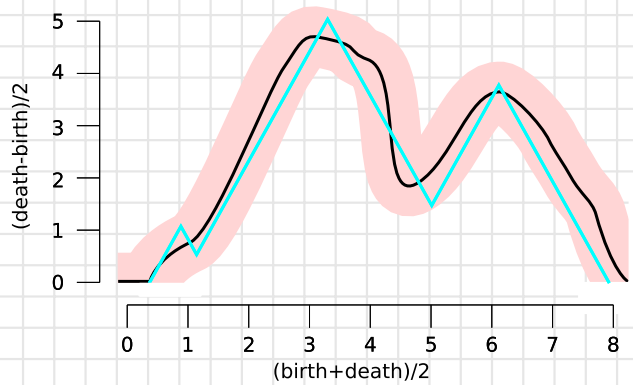


Confidence Bands for Landscapes

The Multiplier Bootstrap

$$\ell_n = \bar{\lambda}_n(t) - \frac{\tilde{Z}(\alpha)}{\sqrt{n}}$$

$$u_n = \bar{\lambda}_n(t) + \frac{\tilde{Z}(\alpha)}{\sqrt{n}}$$



Uniform Band

$$\mathbb{P}\left(\ell_n(t) \leq \mu(t) \leq u_n(t) \text{ for all } t\right) \geq 1 - \alpha - \left(\frac{(\log n)^{\frac{7}{8}}}{n^{\frac{1}{8}}}\right).$$

Variable Width Confidence Bands

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

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

$\widehat{Q}_\alpha$ is the unique value such that

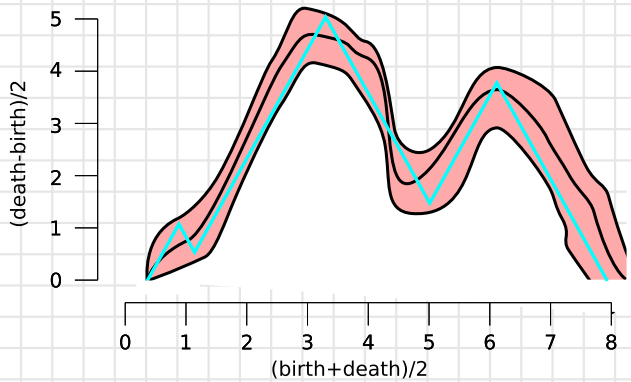
$$\mathbb{P}\left(\sup_t |\widetilde{\mathbb{H}}_n(f_t)| > \widetilde{Q}_\alpha \;\middle|\; \{\lambda_i\} \right) = \alpha$$

Confidence Bands for Landscapes

The Multiplier Bootstrap

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

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



Variable Band

$$\mathbb{P}\left(\ell_n(t) \leq \mu(t) \leq u_n(t) \text{ for all } t\right) \geq 1 - \alpha - \left(\frac{(\log n)^{\frac{7}{8}}}{n^{\frac{1}{8}}}\right).$$

The Intersection of Statistics and Topology:

... And More!

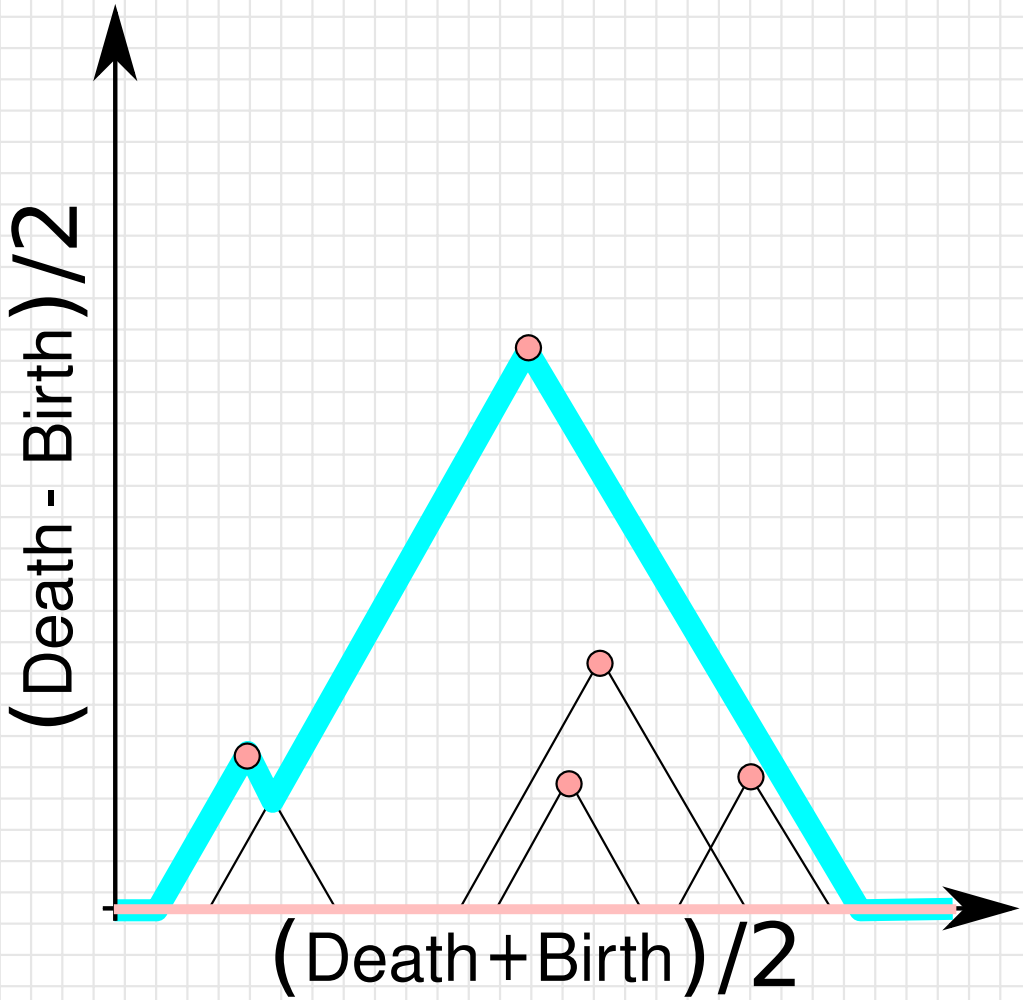
Brittany Terese Fasy

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

3 June 2014

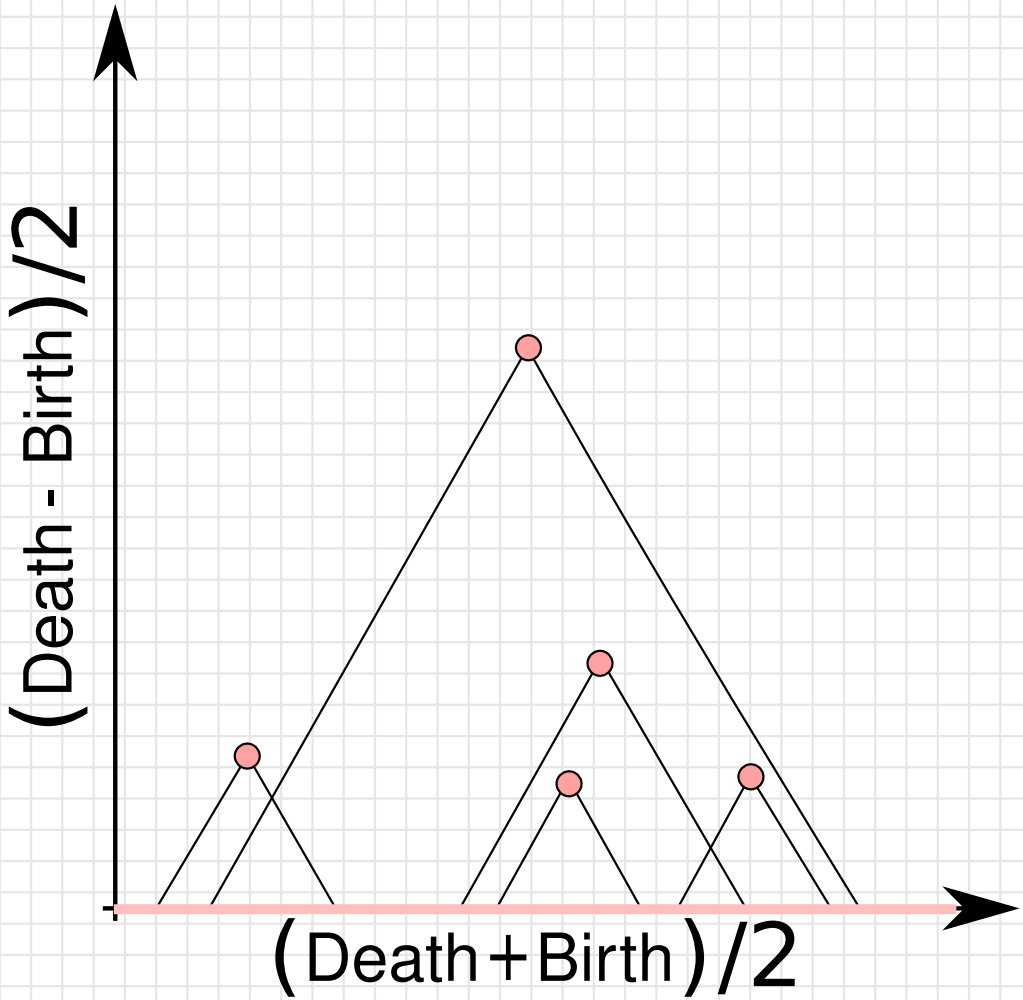
Persistence Silhouettes

Definitions



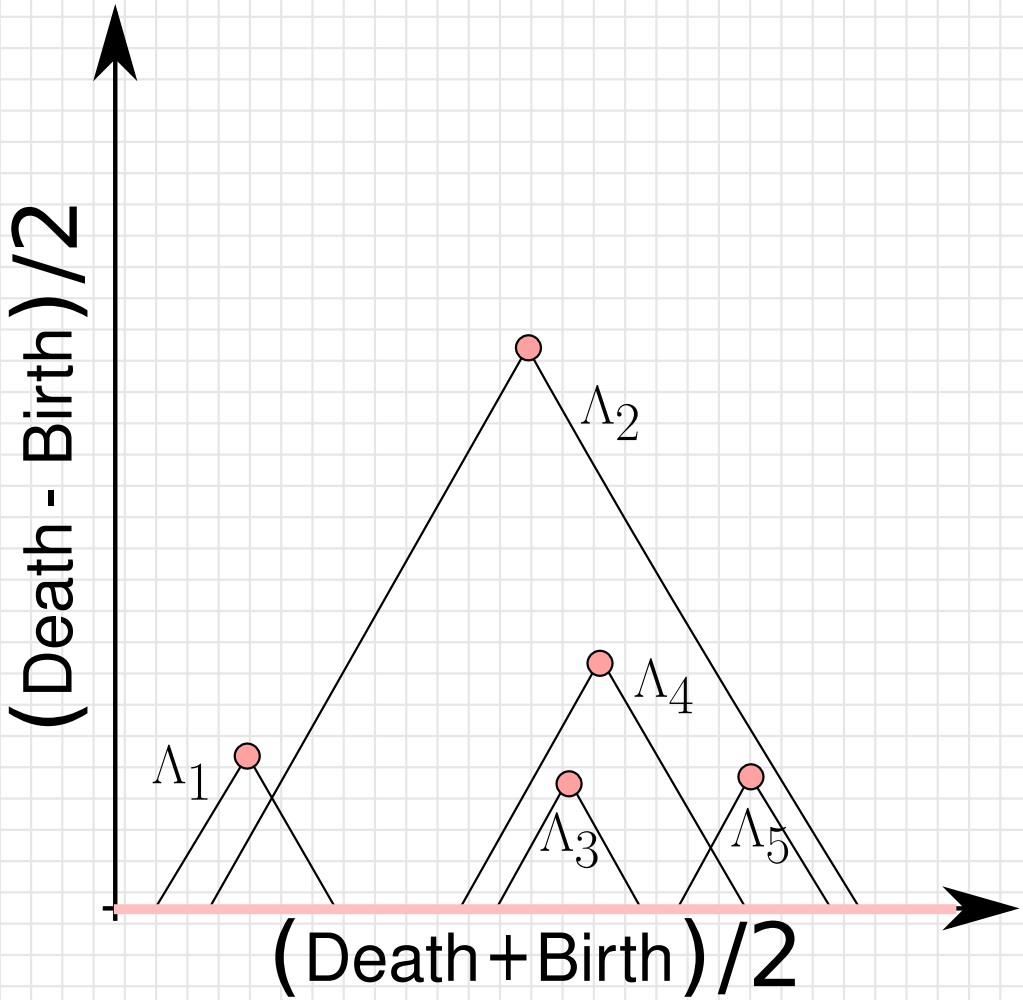
Persistence Silhouettes

Definitions



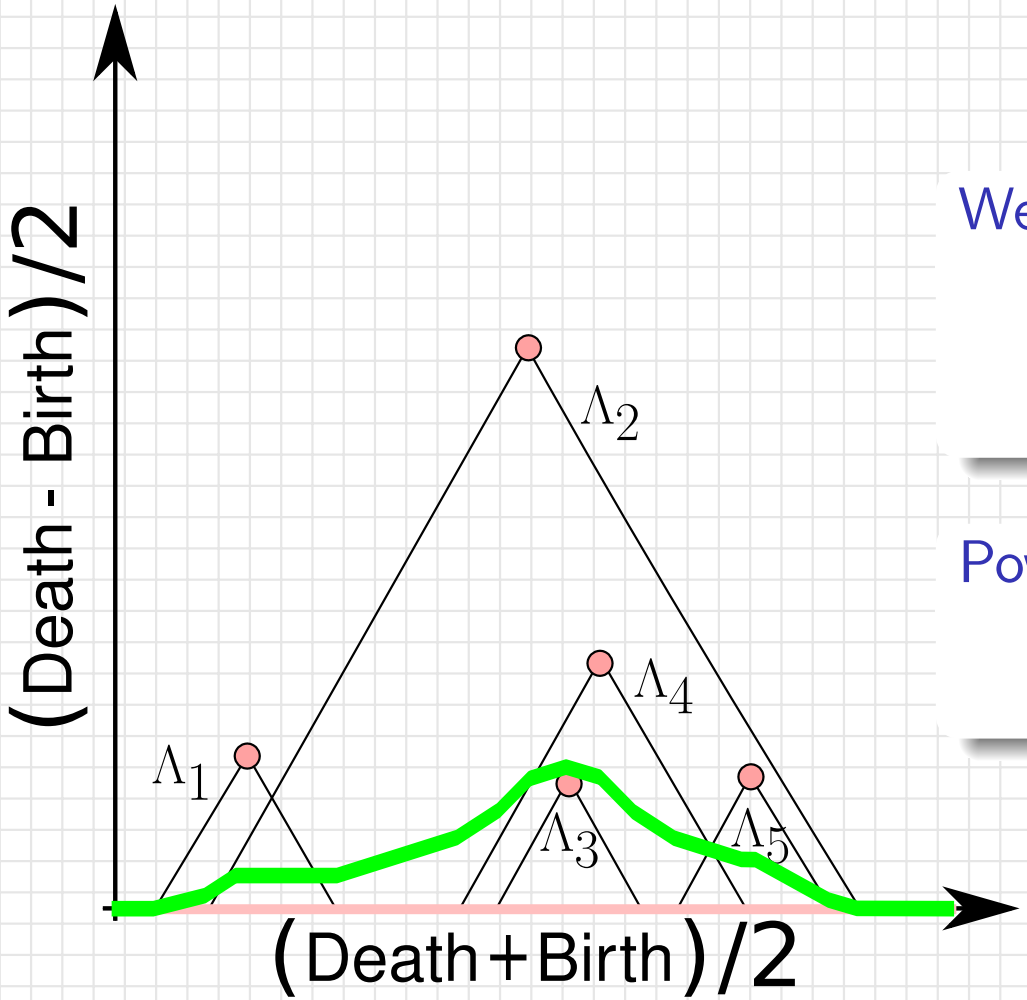
Persistence Silhouettes

Definitions



Persistence Silhouettes

Definitions



Weighted Silhouette

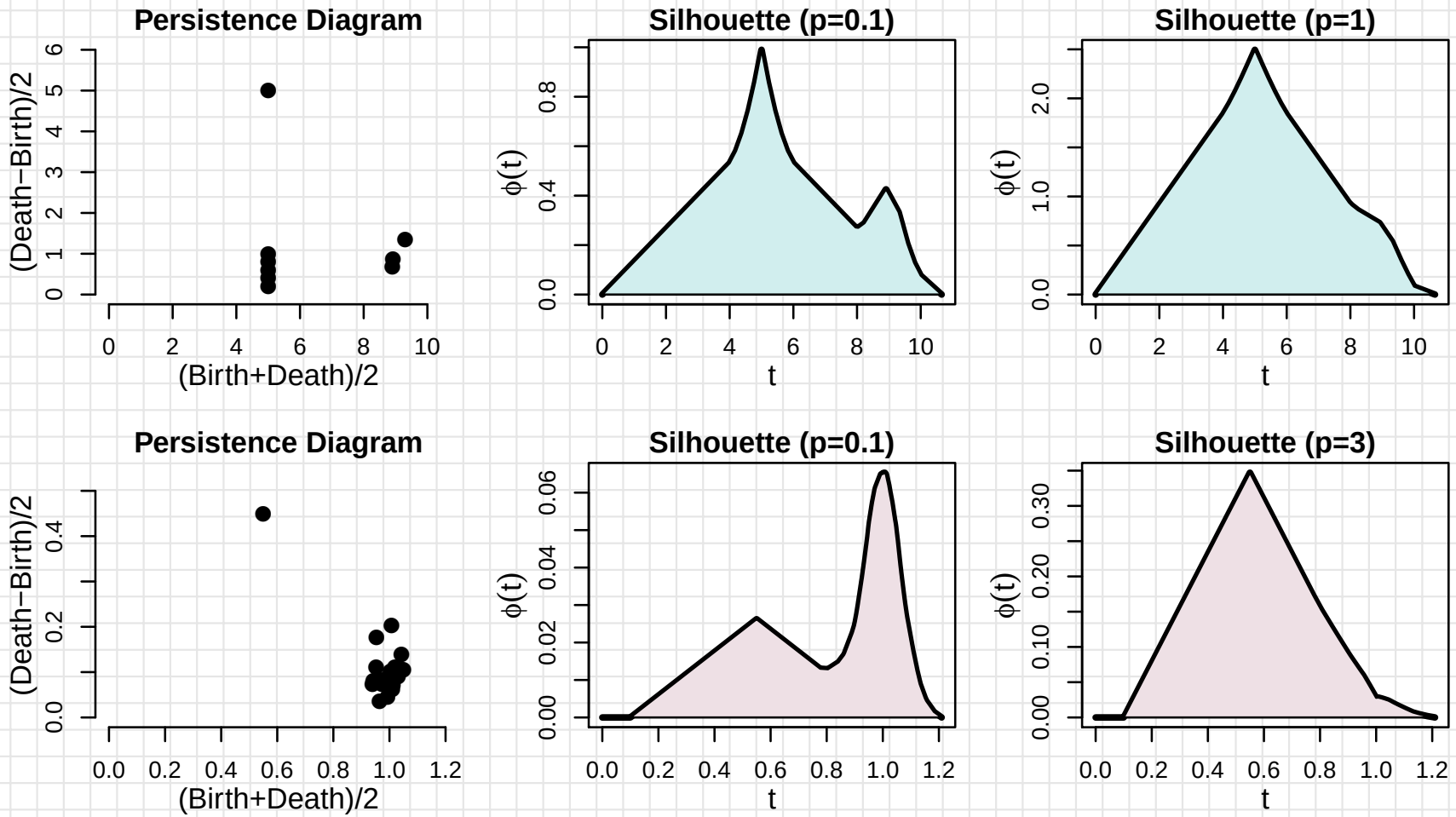
$$\phi(t) = \frac{\sum_{i=1}^{n} w_i \Lambda_j(t)}{\sum_{j=1}^{m} \sum w_j}$$

Power-Weighted Silhouette

$$w_i = |d_i - b_i|^p$$

Power-Weighted Silhouettes

Two Examples



Persistence Silhouettes

Results

Since ϕ 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.

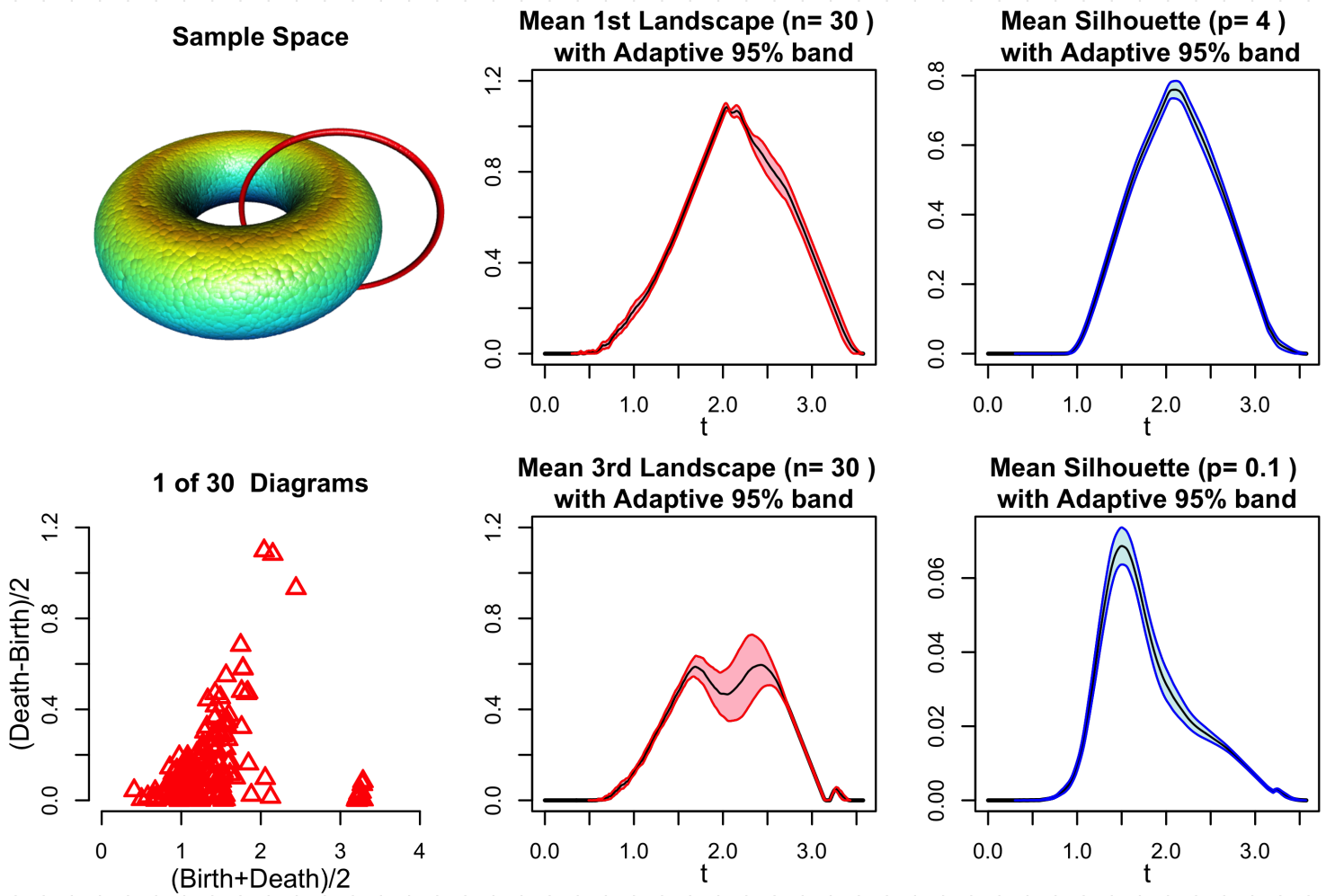
Confidence Bands

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

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

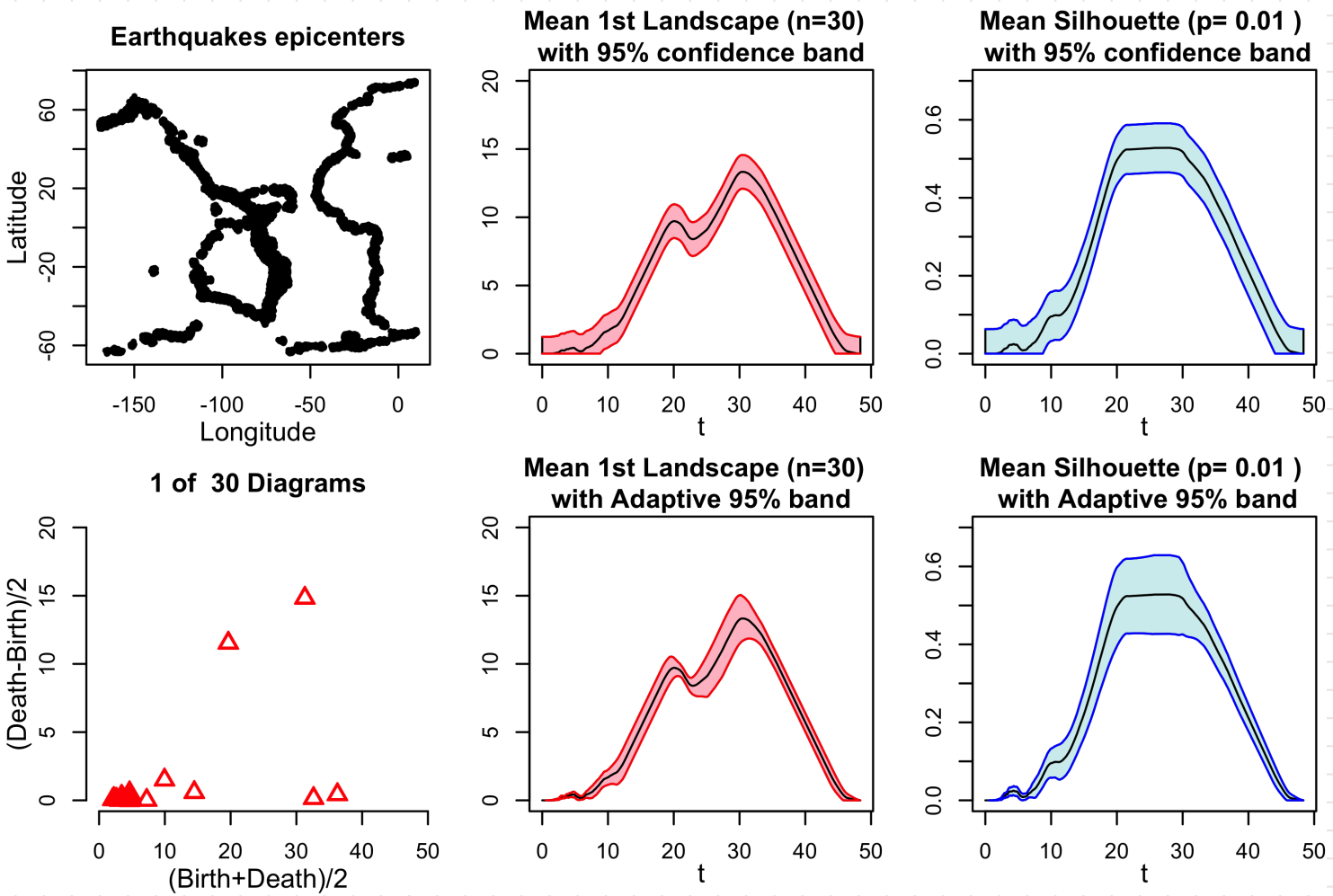
Example I

A Toy Example

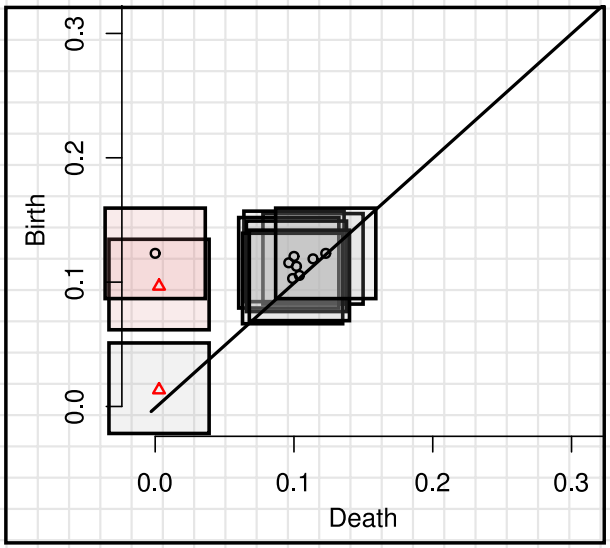


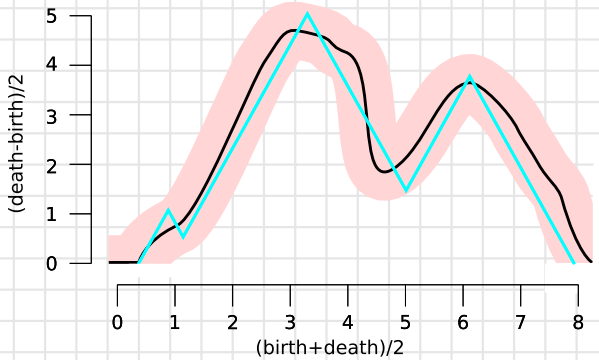
Example II

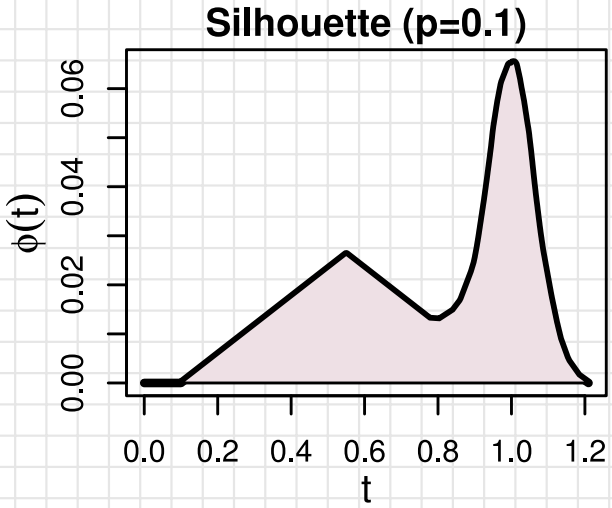
Earthquake Epicenters



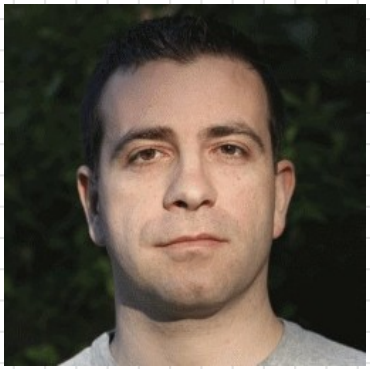
Summary


Birth
Death
0.0
0.1
0.2
0.3


(death-birth)/2
(birth+death)/2
0
1
2
3
4
5
6
7
8


Silhouette (p=0.1)
ϕ(t)
t
0.00
0.02
0.04
0.06
0.0
0.2
0.4
0.6
0.8
1.0
1.2

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.