Statistical Inference for Persistent Homology

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

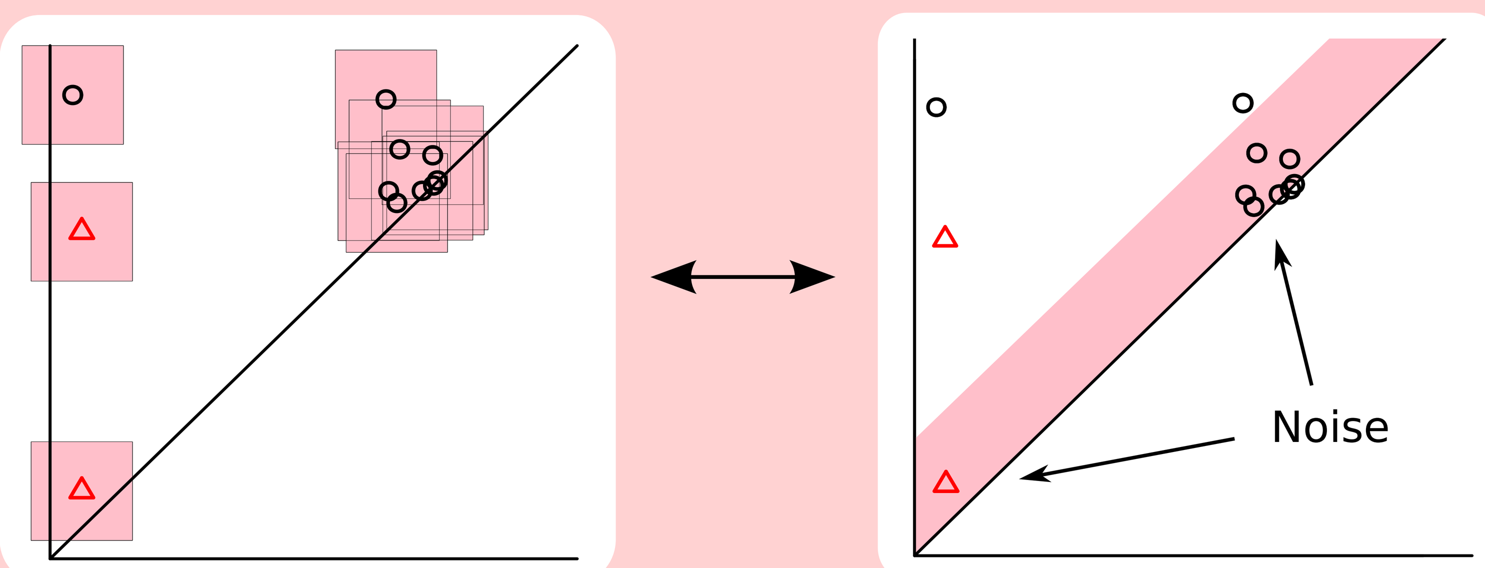
CMU TopStat
stat.cmu.edu/topstat

Confidence Intervals for Persistence Diagrams

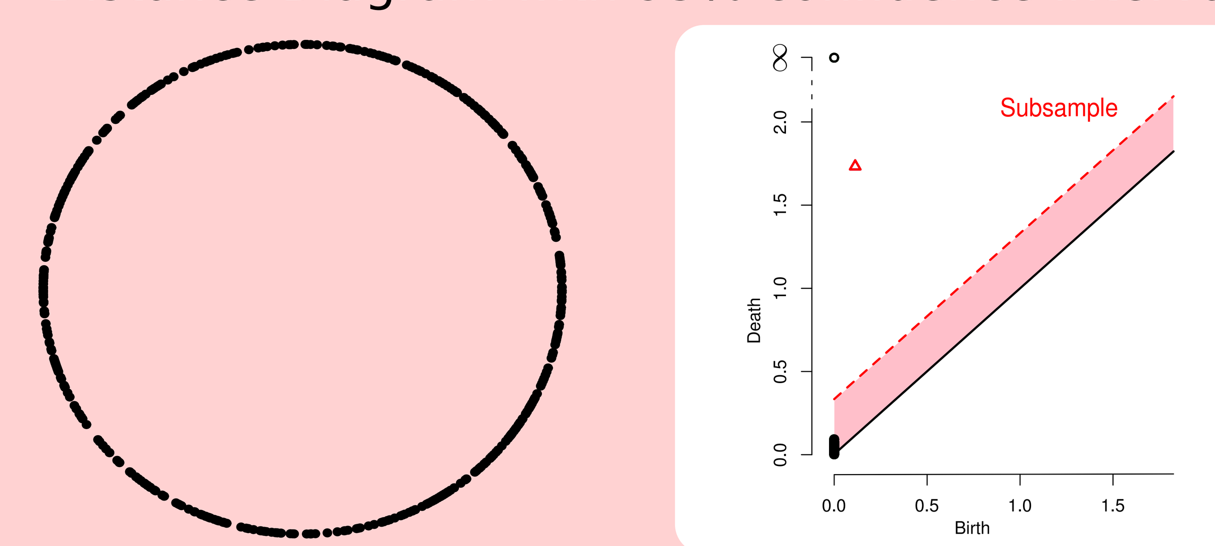
Goal

A 1-α confidence interval for the persistence diagram $\mathcal{P}$ consists of an estimate $\hat{\mathcal{P}}$ and $c > 0$ such that

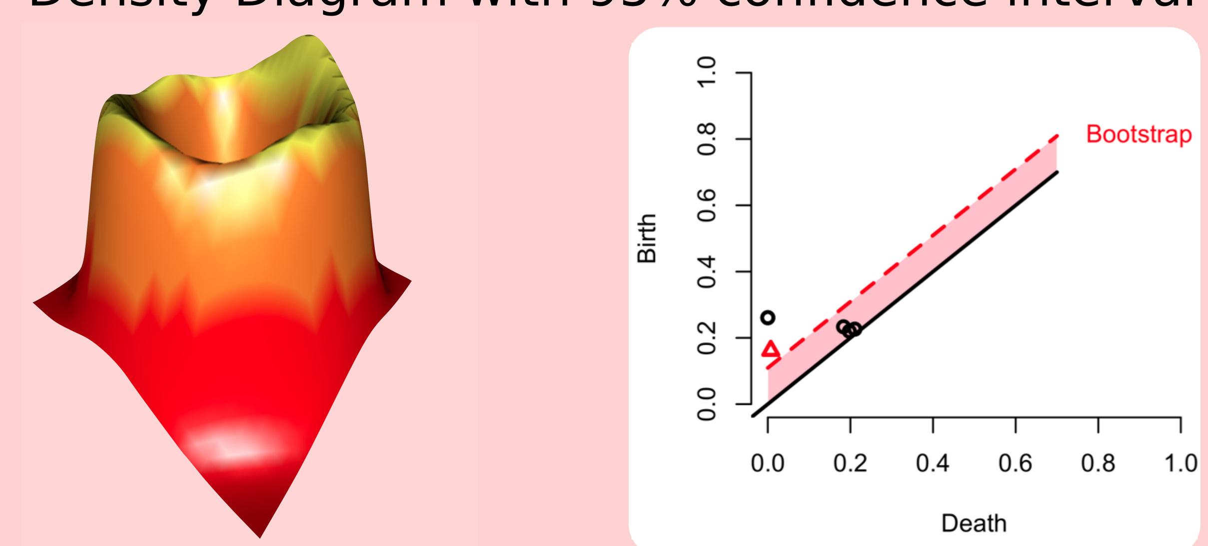
$$\mathbb{P}\left(W_\infty(\mathcal{P}, \hat{\mathcal{P}}) > c \right) \leq \alpha$$



Distance Diagram with 95% confidence interval



Density Diagram with 95% confidence interval



Distance Function

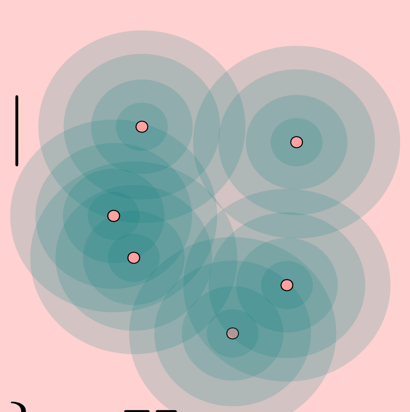
Notation

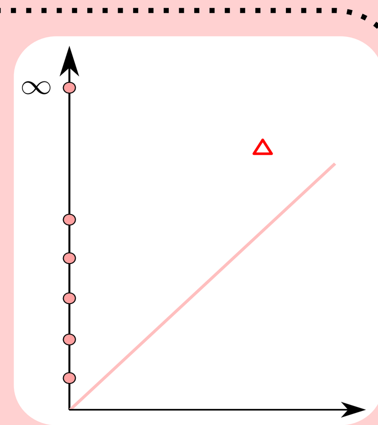
$$d_{\mathbb{X}}(a) = \inf_{x \in \mathbb{X}} ||x - a||$$

$$\mathcal{P}_1 = Dgm_p^-(d_{\mathbb{X}})$$

$$\mathcal{S}_n = \{X_1, \ldots, X_n\} \subset \mathbb{X}$$

$$\hat{\mathcal{P}}_1 = Dgm_p^-(d_{\mathcal{S}_n})$$





Stability Theorem.

$$W_\infty(Dgm_p(f), Dgm_p(g)) \leq ||f - g||_\infty$$

Subsampling Method

$\mathcal{S}_b^1, \ldots, \mathcal{S}_b^N$ are subsamples of size b.

$$L_b(t) = \frac{1}{N} \sum_{j=1}^{N} I\left(||d_{\mathcal{S}_b^j} - d_{\mathcal{S}_n}||_\infty > t \right)$$

Theorem. Almost surely, for all large n,

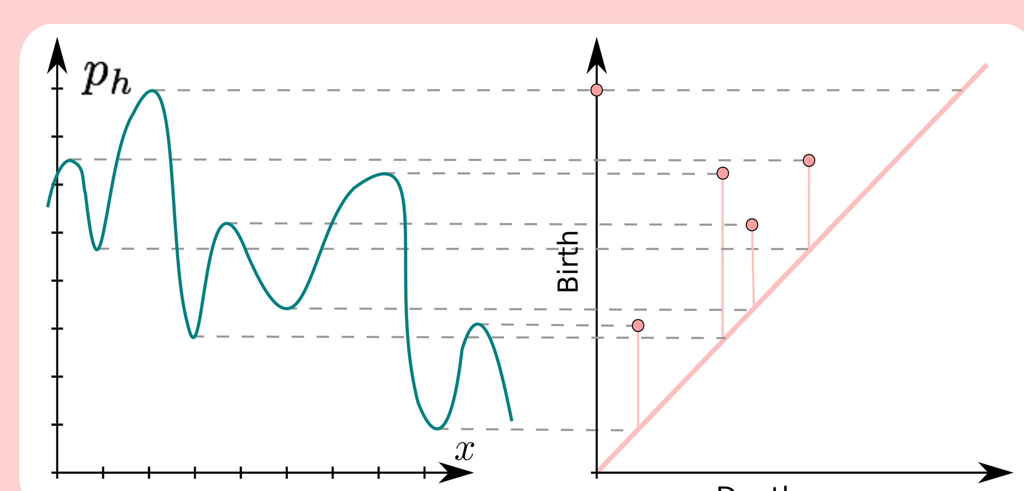
$$\mathbb{P}\left(W_\infty(\mathcal{P}_1, \hat{\mathcal{P}}_1) \geq 2L^{-1}(\alpha) \right) \leq \alpha$$

Density Function

Notation

$$p_h = X \star K \qquad \mathcal{P}_2 = Dgm_p^+(p_h)$$

$$\hat{p}_h = KDE(\mathcal{S}_n) \qquad \hat{\mathcal{P}}_2 = Dgm_p^+(\hat{p}_h)$$



Bootstrap Method

$\mathcal{S}_n^1, \ldots, \mathcal{S}_n^N$ are subsamples of size n.

$$\hat{p}_h^i = KDE(\mathcal{S}_n^i)$$

$$c = \inf \left\{ t : \frac{1}{N} \sum_{i=1}^{N} I(\sqrt{nh^D}\ ||\hat{p}_h - \hat{p}_h^i||_\infty > t) \leq \alpha \right\}$$

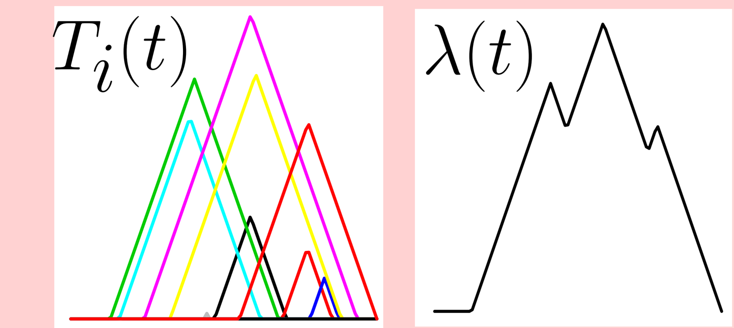
Theorem. As $n \to \infty$,

$$\mathbb{P}\left(W_\infty(\mathcal{P}_2, \hat{\mathcal{P}}_2) > \frac{c}{\sqrt{nh^D}} \right) \leq \alpha$$

Confidence Bands for Landscapes

Notation

Let $(a_i, b_i) \in \mathcal{P}$.
$T_i(t) = (t - a_i)_+ \wedge (b_i - t)_+$



The 1st Persistent Landscape (Bubenik, 2012) is the maximum contour of the triangles:

$$\lambda_j(t) = \max\{T_i(t) : (a_i, b_i) \in \mathcal{P}_j\}$$

Mean Landscape: $\mu(t) = \mathbb{E}\left[\lambda_{\mathcal{P}}(t) \right]$

Sample mean Landscape: $\bar{\mathcal{L}}_n(t) = \frac{1}{n} \sum_j \lambda_j(t)$

We want a confidence band for $\mu(t)$.

$$\mathcal{P}_1, \ldots, \mathcal{P}_n \sim P$$

Define the **empirical process** $\mathbb{G}_n = \sqrt{n}(\bar{\mathcal{L}}_n(t) - \mu(t))$.

Theorem. $\mathbb{G}_n$ converges to a Gaussian process:

$$\mathbb{G}_n \rightsquigarrow \mathbb{G}$$

Multiplier Bootstrap

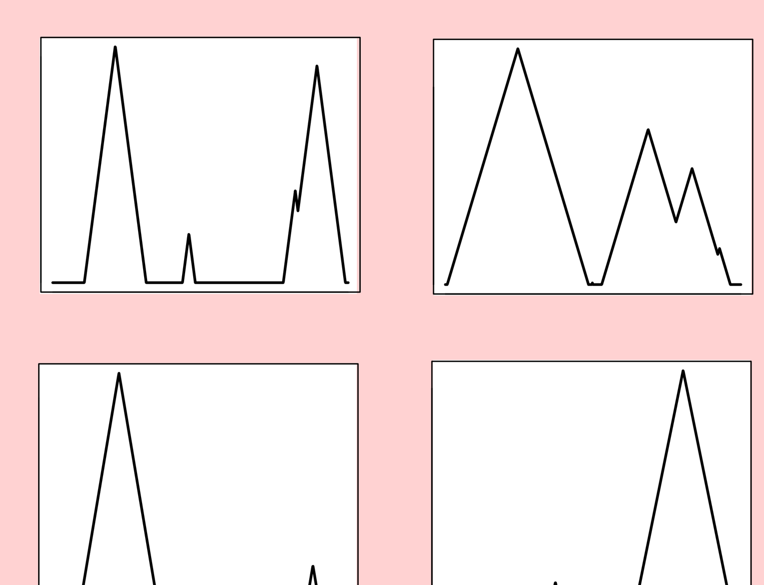
- Compute N copies of

$$\widetilde{\mathbb{G}}_n = \frac{1}{\sqrt{n}} \sum_j \xi_j(\lambda_j(t) - \bar{\mathcal{L}}_n(t))$$

- Compute c, the 1-α quantile of the bootstrapped $\sup_t |\widetilde{\mathbb{G}}_n(t)|$.

Theorem. For large n and N,

$$\mathbb{P}\left(\mu(t) \in \bar{\mathcal{L}}_n(t) \pm \frac{c}{\sqrt{n}} \quad \forall t \right) \geq 1 - \alpha$$



Mean Landscape
with 95% confidence band