Statistical inference on persistent homology of density filtration on Rips complex

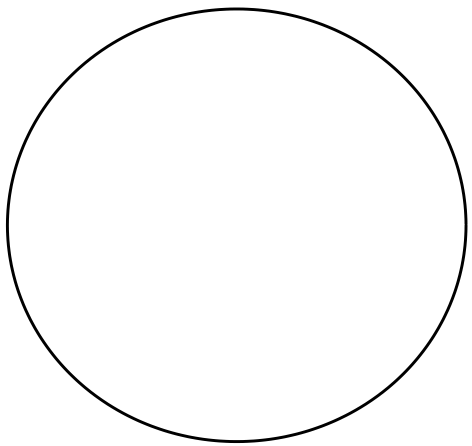
Jisu KIM

Carnege Mellon University

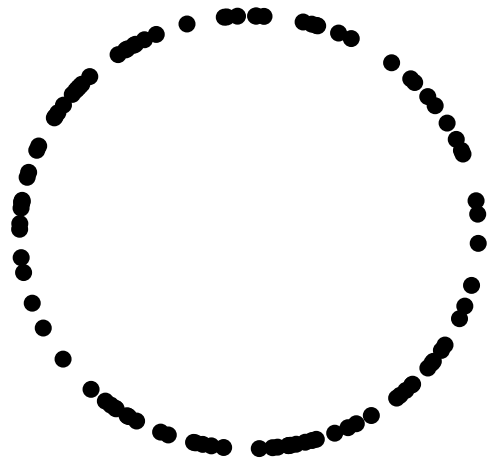
May 11, 2017

When analyzing data, we prefer robust features where features of the underlying manifold can be inferred from features of finite samples.

Underlying circle

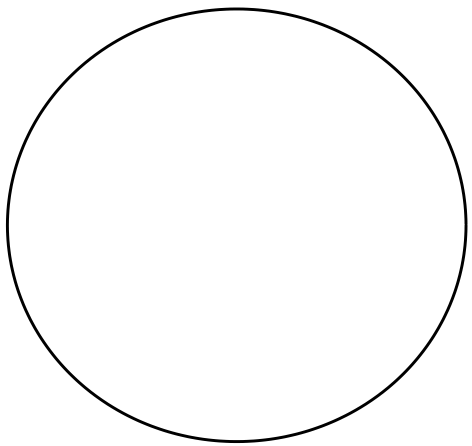


100 samples

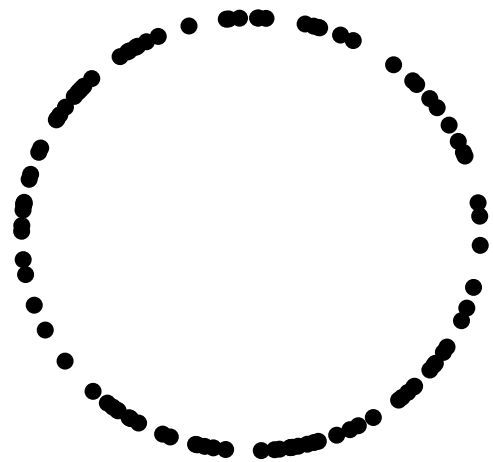


Homology of finite sample is different from homology of underlying manifold, hence it cannot be directly used for the inference.

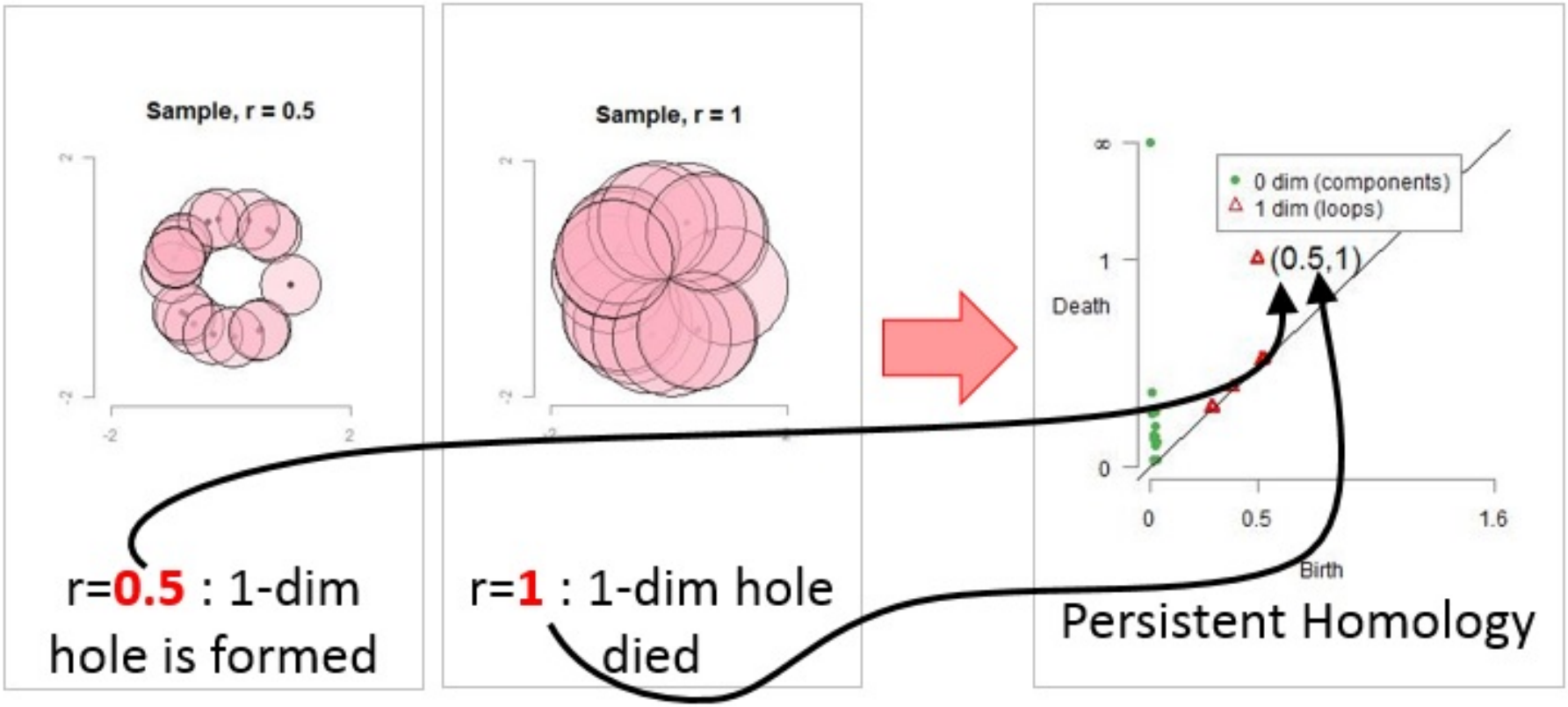
Underlying circle: $\beta_0 = 1$, $\beta_1 = 1$



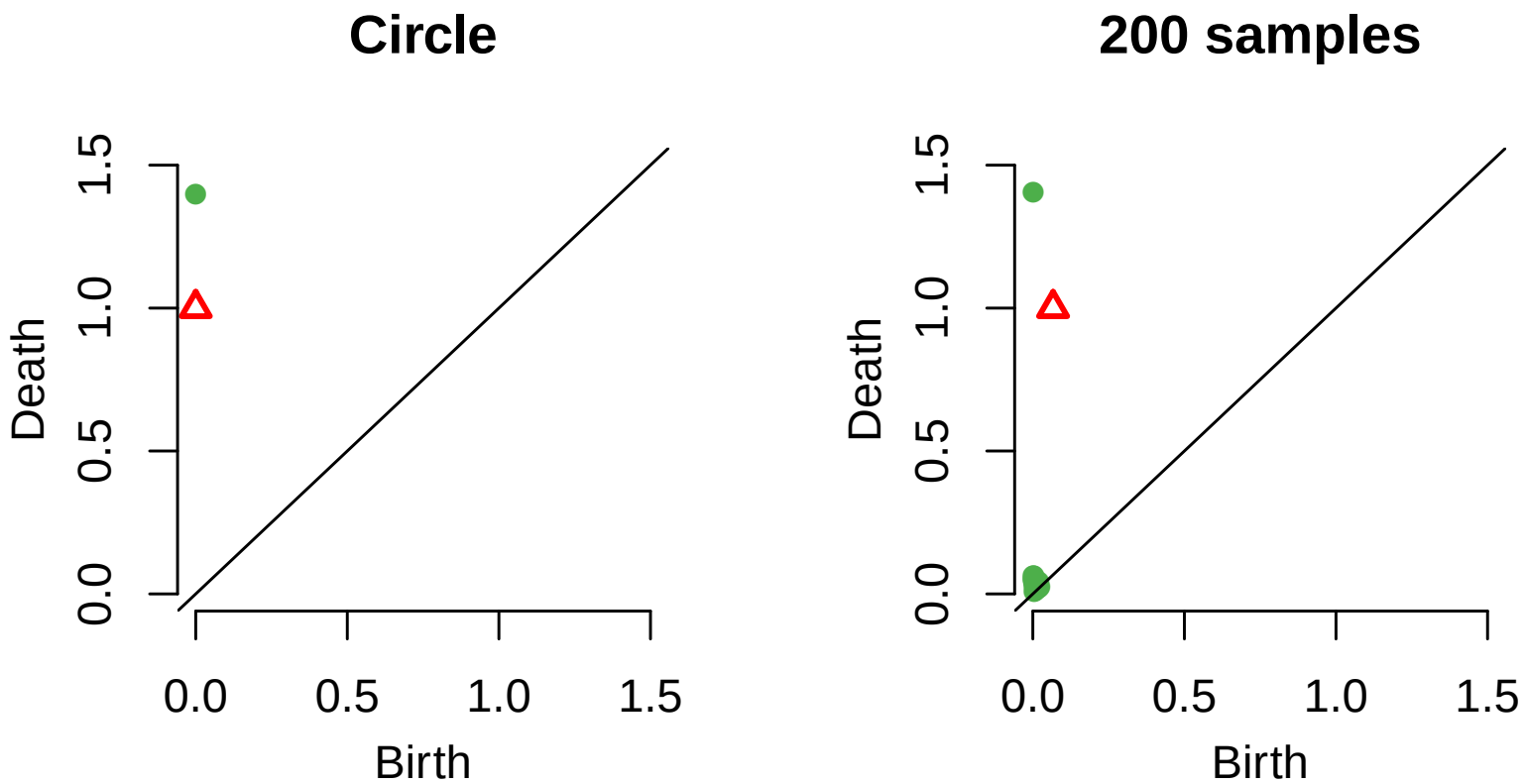
100 samples: $\beta_0 = 100$, $\beta_1 = 0$



Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.



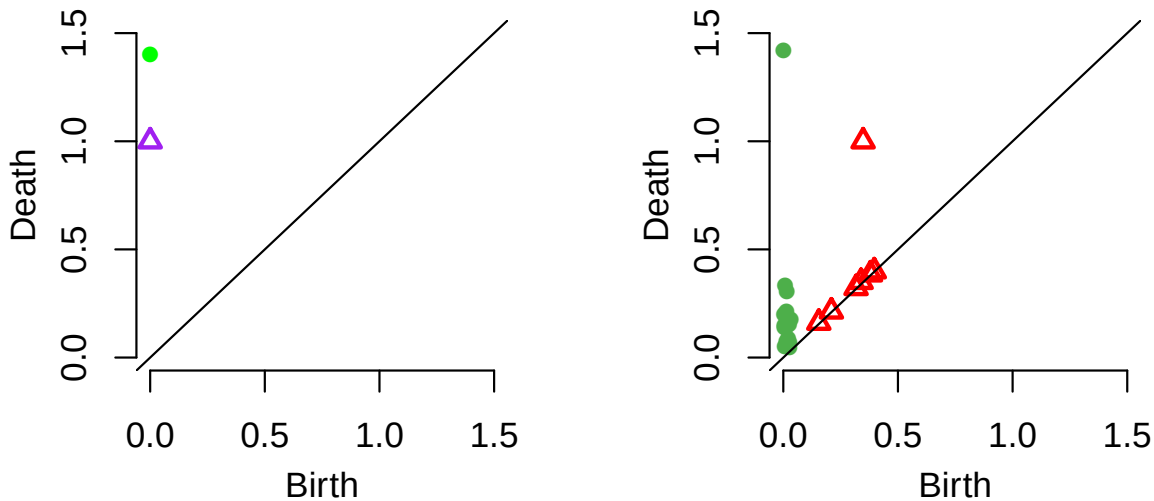
Bottleneck distance gives a metric on the space of persistent homology.

Definition

Let D_1, D_2 be multiset of points. Bottleneck distance is defined as

$$W_\infty(D_1, D_2) = \inf_\gamma \sup_{x \in D_1} \|x - \gamma(x)\|_\infty,$$

where γ ranges over all bijections from D_1 to D_2.



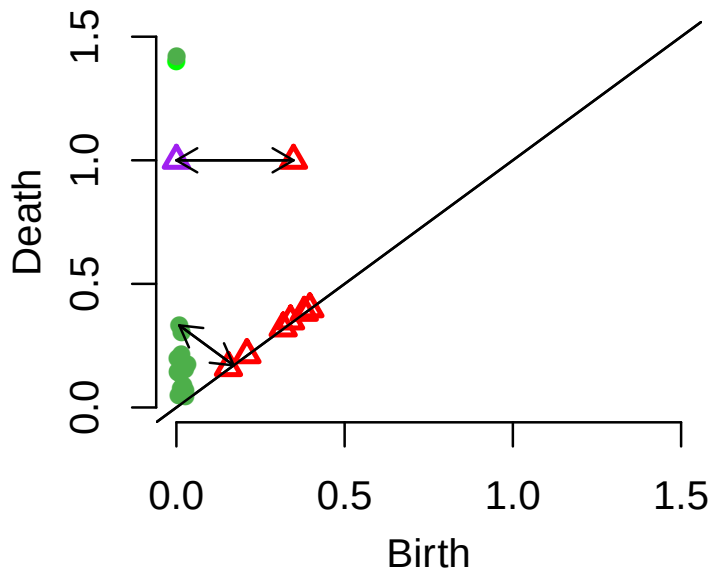
Bottleneck distance gives a metric on the space of persistent homology.

Definition

Let D_1, D_2 be multiset of points. Bottleneck distance is defined as

$$W_\infty(D_1, D_2) = \inf_\gamma \sup_{x \in D_1} \|x - \gamma(x)\|_\infty,$$

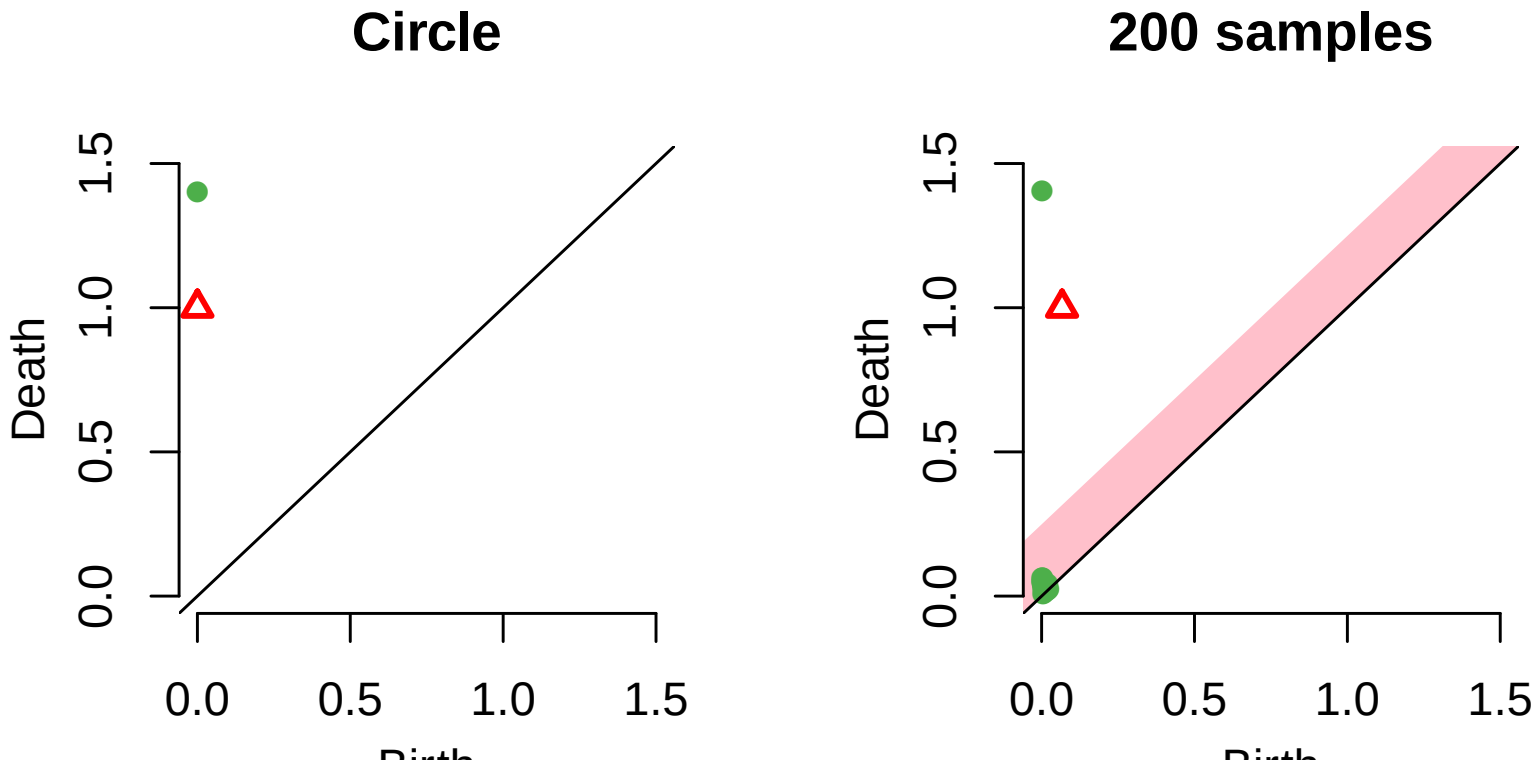
where γ ranges over all bijections from D_1 to D_2.



Confidence band for persistent homology separates homological signal from homological noise.

Let M be a compact manifold, and $X = \{X_1, \cdots, X_n\}$ be n samples. Let f_M and f_X be corresponding functions whose persistent homology is of interest. Given the significance level $\alpha \in (0, 1)$, $(1 - \alpha)$ confidence band $c_n = c_n(X)$ is a random variable satisfying

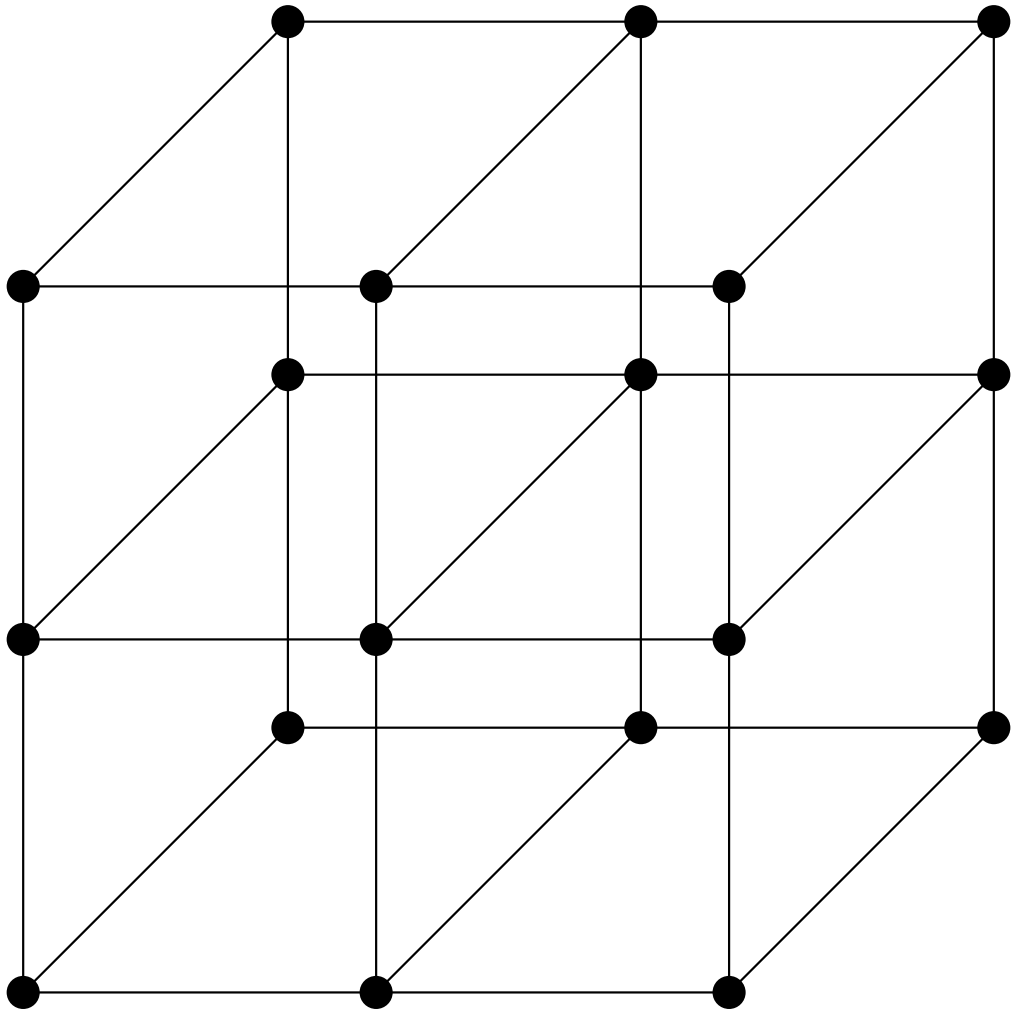
$$\mathbb{P}\left(W_\infty(Dgm(f_M),\, Dgm(f_X)) \leq c_n\right) \geq 1 - \alpha.$$



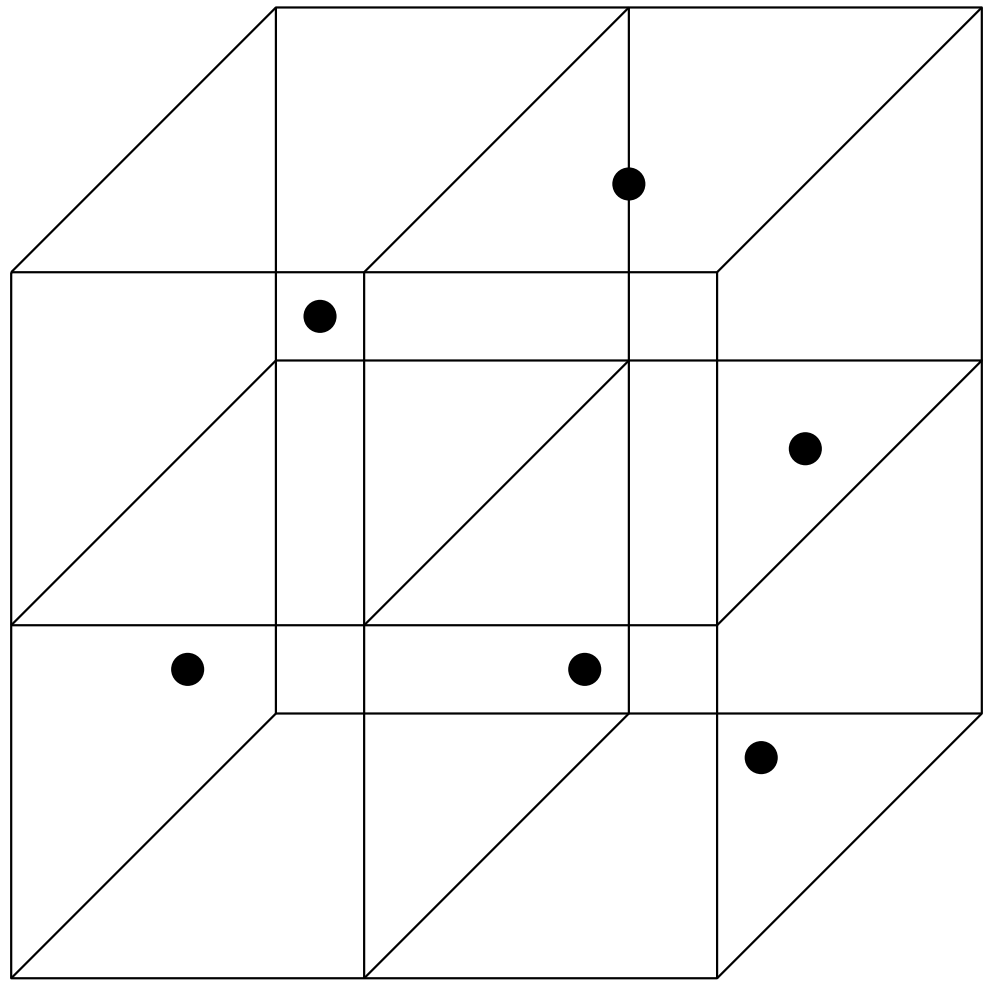
Confidence band for the persistent homology can be computed using the bootstrap algorithm.

1. Given a sample $X = \{x_1, \ldots, x_n\}$, compute the kernel density estimator $\hat{p}_h$.
2. Draw $X^* = \{x_1^*, \ldots, x_n^*\}$ from $X = \{x_1, \ldots, x_n\}$ (with replacement), and compute $\theta^* = \sqrt{n}||\hat{p}_h^*(x) - \hat{p}_h(x)||_\infty$, where $\hat{p}_h^*$ is the density estimator computed using X^*.
3. Repeat the previous step B times to obtain $\theta_1^*, \ldots, \theta_B^*$
4. Compute $q_\alpha = \inf\left\{q : \frac{1}{B}\sum_{j=1}^{B} I(\theta_j^* \geq q) \leq \alpha\right\}$
5. The $(1-\alpha)$ confidence band for $\mathbb{E}[\hat{p}_h]$ is $\left[\hat{p}_h - \frac{q_\alpha}{\sqrt{n}}\ ,\ \hat{p}_h + \frac{q_\alpha}{\sqrt{n}}\right]$, and we use $\frac{q_\alpha}{\sqrt{n}}$ as $\hat{c}_n$.

Computing a confidence band for the persistent homology incurs computing on a grid of points, which is infeasible in high dimensional space.



Computing the persistent homology of density function on data points reduces computational complexity.

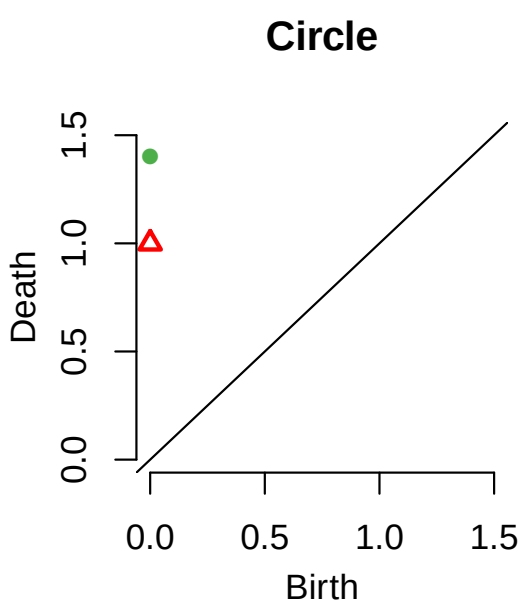


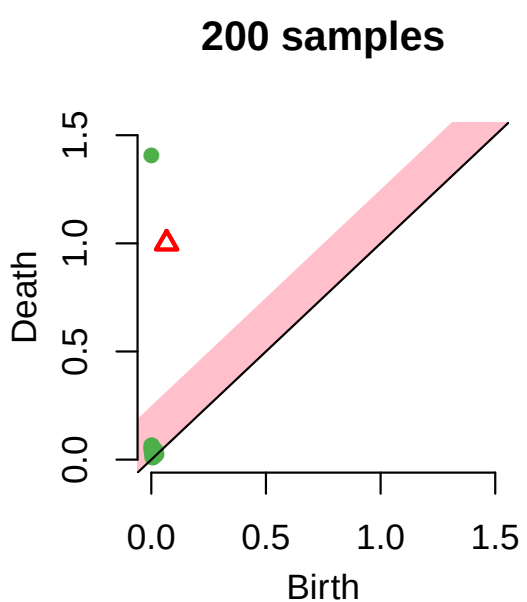
Confidence band for persistent homology separates homological signal from homological noise.

Theorem

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p$ *and let* $\hat{p}_h$ *be the kernel density estimator. Let* $\hat{\mathcal{X}}_n^L := \{X_i : \hat{p}_h(X_i) \geq L\}$, *and let* $\{\hat{R}_L(n, r, h)\}_{L \in \mathbb{R}}$ *be the Rips complex defined as* $\hat{R}_L(n, r, h) = \{\sigma \subset \hat{\mathcal{X}}_n^L : diam(\sigma) \leq 2r\}$. *Let* $\hat{c}_n$ *be from bootstrap algorithm and let* $\tilde{c}_r := \max_i \sup_{x \in \mathcal{B}(X_i, r)} |\hat{p}_h(x) - \hat{p}_h(X_i)|$. *Then given* $\alpha \in (0, 1)$,

$$\mathbb{P}(W_\infty(Dgm(p_h),\ Dgm(\hat{R}_L(n, r, h)) \leq \hat{c}_n + \tilde{c}_{2r}) \geq 1 - \alpha + o(1).$$





Thank you!