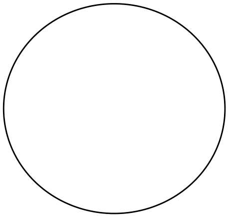
Statistical Inference on Topological Data Analysis

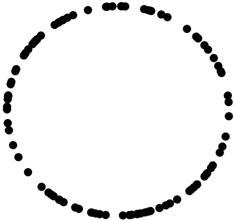
Jisu KIM

Carnege Mellon University

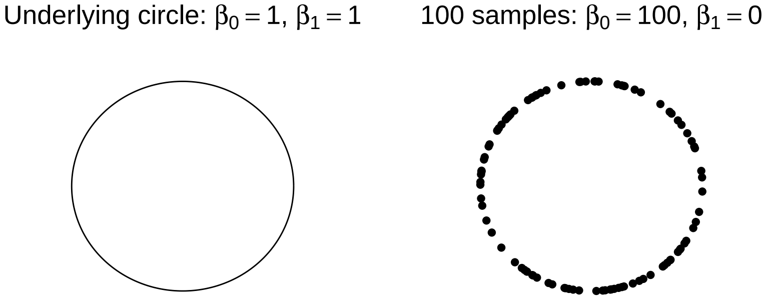
Jan 4, 2017

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

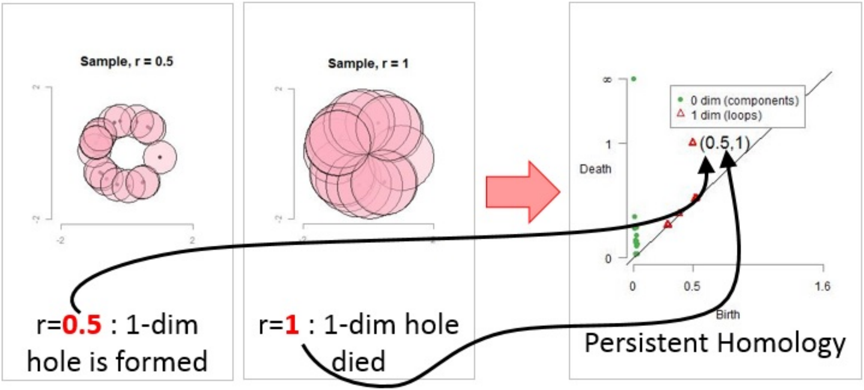




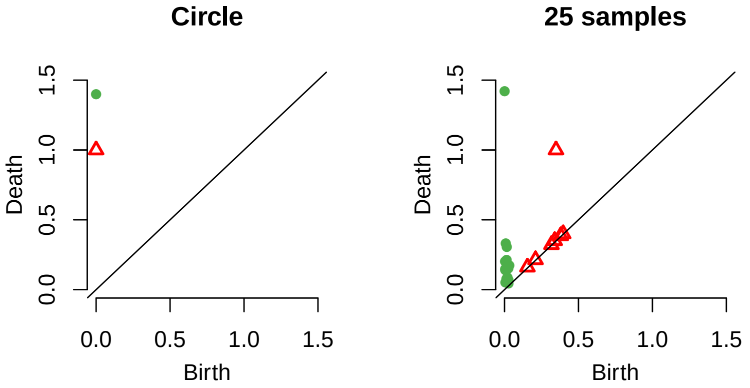
Homology of finite sample is different from homology of underlying manifold, hence it cannot be directly used for the inference.



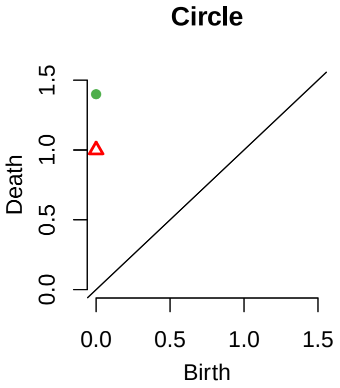
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



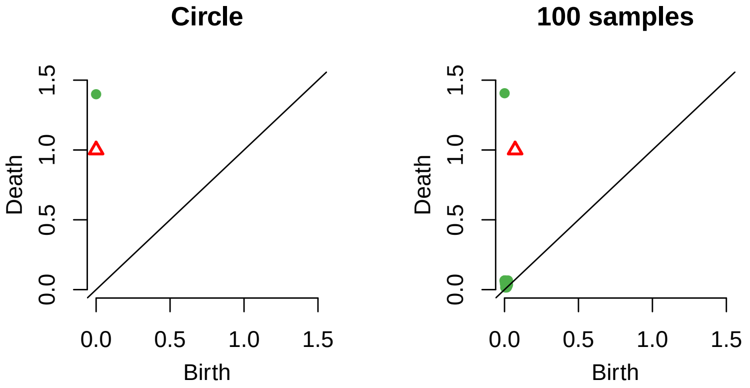
Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.



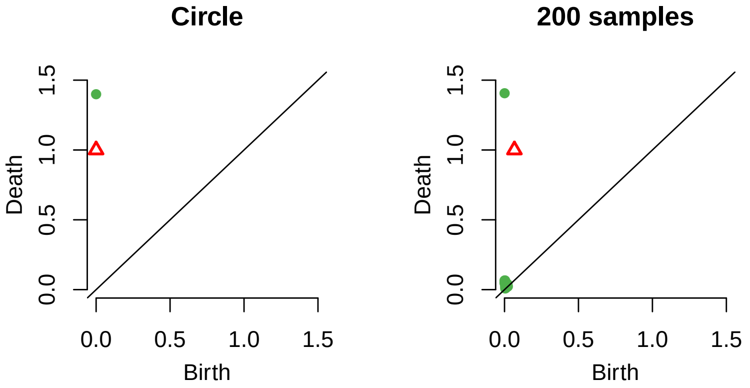
Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.



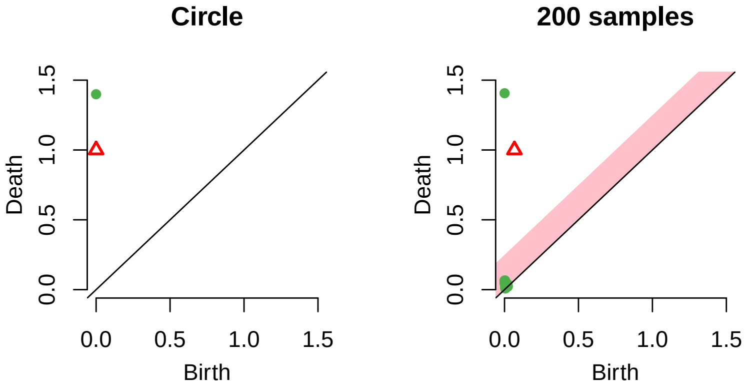
Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.



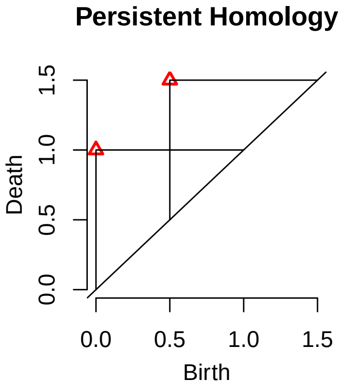
Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.

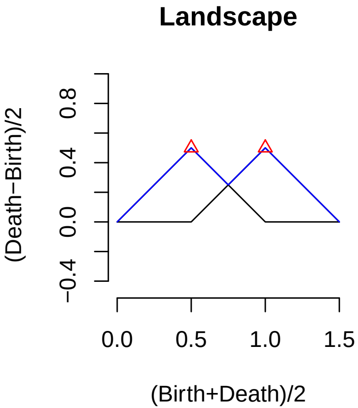


How can we distinguish statistically significant homological features from noisy homological features?

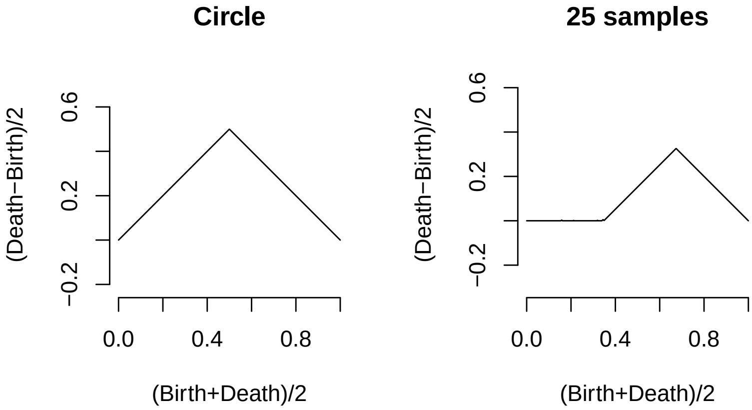


Landscape is a functional summary of the persistent homology.

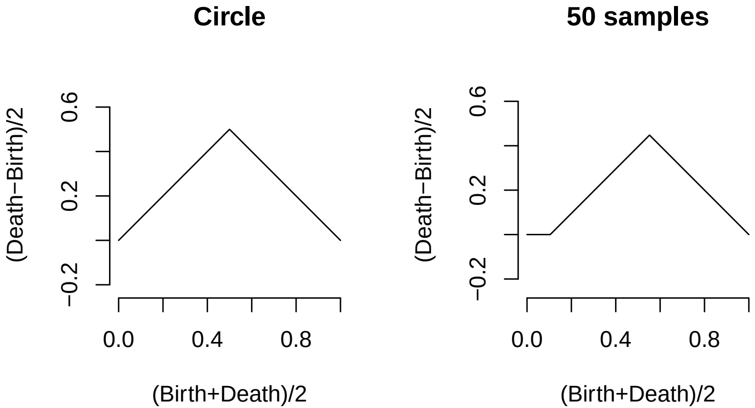




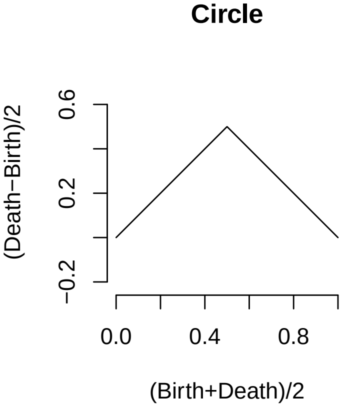
Landscape of the underlying manifold can be inferred from landscape of finite samples.

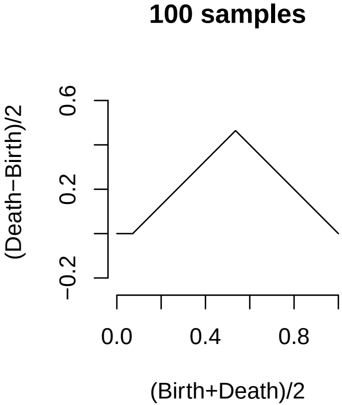


Landscape of the underlying manifold can be inferred from landscape of finite samples.

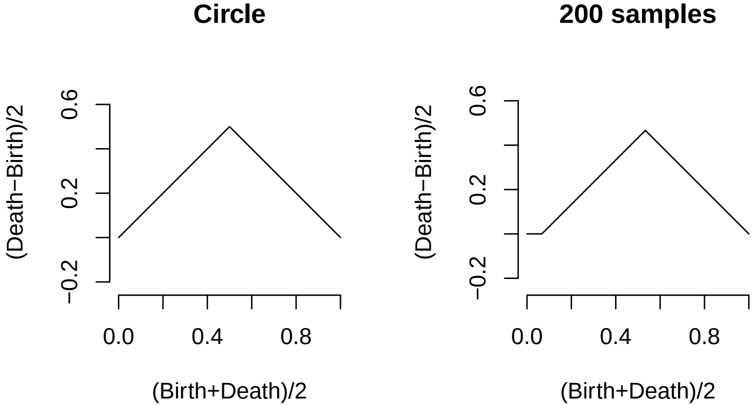


Landscape of the underlying manifold can be inferred from landscape of finite samples.

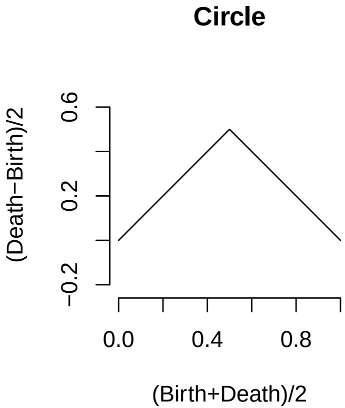


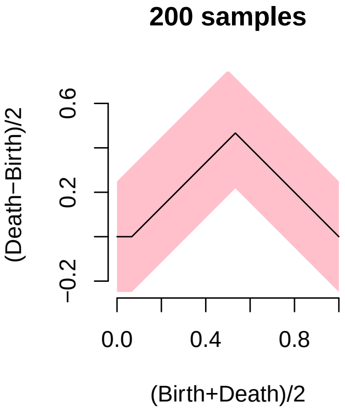


Landscape of the underlying manifold can be inferred from landscape of finite samples.



How can we statistically quantify the randomness of the landscape?





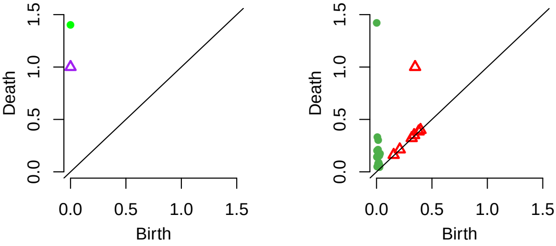
Bottleneck distance gives a metric on the space of the persistent homology.

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.



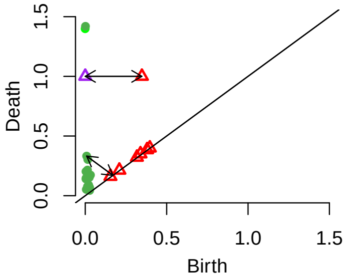
Bottleneck distance gives a metric on the space of persistent homology.

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.



Bottleneck distance can be controlled by the corresponding distance on functions: Stability Theorem.

Theorem

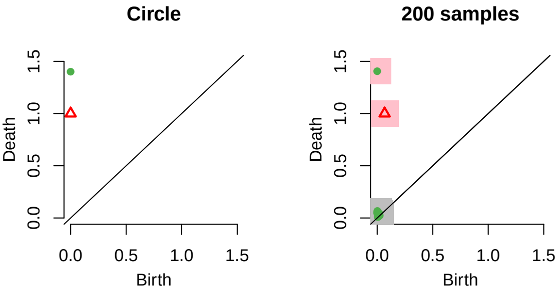
[Edelsbrunner and Harer, 2010][Chazal, de Silva, Glisse, and Oudot, 2012] Let $\mathbb{X}$ be finitely triangulable space and $f,\ g : \mathbb{X} \to \mathbb{R}$ be two continuous functions. Let $Dgm(f)$ and $Dgm(g)$ be corresponding persistent homologies. Then

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

Confidence band for the persistent homology is a random quantity containing the persistent homology with high probability.

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

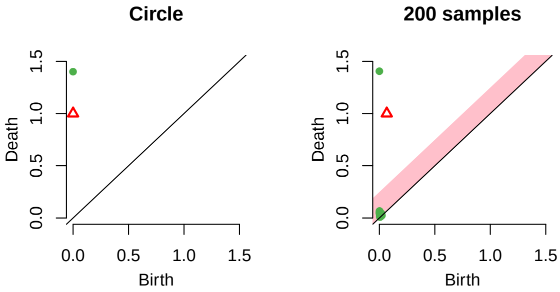
$$\mathbb{P}\left(W_\infty(Dgm(f_M),\, Dgm(f_X)) \leq c_n\right) \geq 1 - \alpha.$$



Confidence band for the persistent homology is a random quantity containing the persistent homology with high probability.

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 obtained by the corresponding confidence band for functions.

From Stability Theorem, $\mathbb{P}\left(||f_M - f_X|| \leq c_n\right) \geq 1 - \alpha$ implies

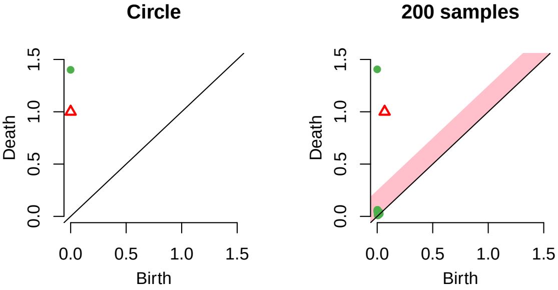
$$\mathbb{P}\left(W_\infty(Dgm(f_M), \, Dgm(f_X)) \leq c_n\right) \geq \mathbb{P}\left(||f_M - f_X||_\infty \leq c_n\right) \geq 1 - \alpha,$$

so the confidence band of corresponding functions f_M can be used for confidene band of persistent homologies $Dgm(f_M)$.

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

The validity of the bootstrap algorithm is proved and used in the framework of persistent homology.

- [Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, and Singh, 2014] proved for kernel density estimator,
- [Chazal, Fasy, Lecci, Michel, Rinaldo, and Wasserman, 2014a] proved for distance to measure and kernel distance.



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

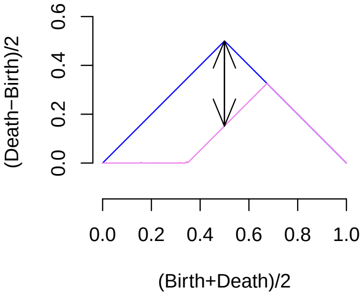
1. Given a sample $X = \{x_1, \ldots, x_n\}$, compute the kernel density estimator $\hat{p}_h$.
2. Draw $X^* = \{x_1^*, \ldots, x_n^*\}$ from $X = \{x_1, \ldots, x_n\}$ (with replacement), and compute $\theta^* = \sqrt{n}||\hat{p}_h^*(x) - \hat{p}_h(x)||_\infty$, where $\hat{p}_h^*$ is the density estimator 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 $\left[\hat{p}_h - \frac{q_\alpha}{\sqrt{n}}\,,\, \hat{p}_h + \frac{q_\alpha}{\sqrt{n}}\right]$.

∞-landscape distance gives a metric on the space of landscapes.

Definition

Let D_1, D_2 be multiset of points, and λ_1 , λ_2 be corresponding landscapes. ∞-landscape distance is defined as

$$\Lambda_\infty(D_1, D_2) = \|\lambda_1 - \lambda_2\|_\infty .$$



∞-landscape distance can be controlled by the corresponding distance on functions: Stability Theorem.

Theorem

[Bubenik, 2015] Let $f, g : \mathbb{X} \to \mathbb{R}$ be two functions, and let $Dgm(f)$ and $Dgm(g)$ be corresponding persistent homologies. Then

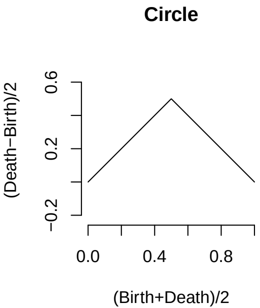
$$\Lambda_\infty(Dgm(f),\, Dgm(g)) \leq \|f - g\|_\infty.$$

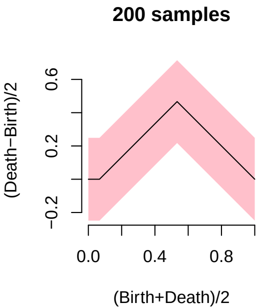
Confidence band for the landscape can be computed using the bootstrap algorithm.

- Let λ_M and λ_X be landscapes of the manifold M and samples X. From Stability Theorem, $\mathbb{P}\left(||f_M - f_X|| \leq c_n\right) \geq 1 - \alpha$ implies

  $$\mathbb{P}\left(\lambda_X(t) - c_n \leq \lambda_M(t) \leq \lambda_X(t) + c_n \, \forall t\right) \geq \mathbb{P}\left(||f_M - f_X|| \leq c_n\right) \geq 1-\alpha,$$

  so the confidence band of corresponding functions f_M can be used for confidene band of the landscape λ_M.





Confidence band for landscape can be computed using the bootstrap algorithm.

- Confidence band for landscape can be also computed using multiplier bootstrap; see [Chazal, Fasy, Lecci, Rinaldo, and Wasserman, 2014b].

Persistent Homology and Landscape as Robust Topological Features

Statistical Inference on Persistent Homology and Landscape

Reference

CMU TopStat

CMU TopStat ×

www.stat.cmu.edu/topstat/

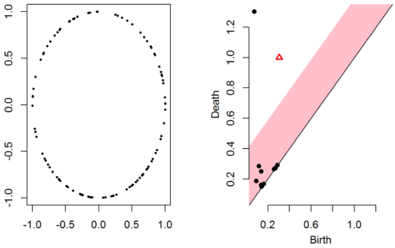
CMU TopStat | Home | People | Research | Papers | Presentations | Software | Private

CMU TopStat

The CMU Topological Statistics group is a research group at Carnegie Mellon University. The emphasis of our research is on statistical problems related to topological inference.

Visit the Projects page to see descriptions of our projects and relevant publications or preprints.

We meet every Friday 14:30.



You can send an email to the following address. The inbox is regularly checked.

topstat [at] stat [dot] cmu [dot] edu

Or you can contact us individually.

Reference

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1):77–102, January 2015. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2789272.2789275`.

Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance-to-a-measure and kernel distance. *Technical Report*, 2014a.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Annual Symposium on Computational Geometry*, pages 474–483. ACM, 2014b.

Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Society, 2010.

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 2014.

Thank you!