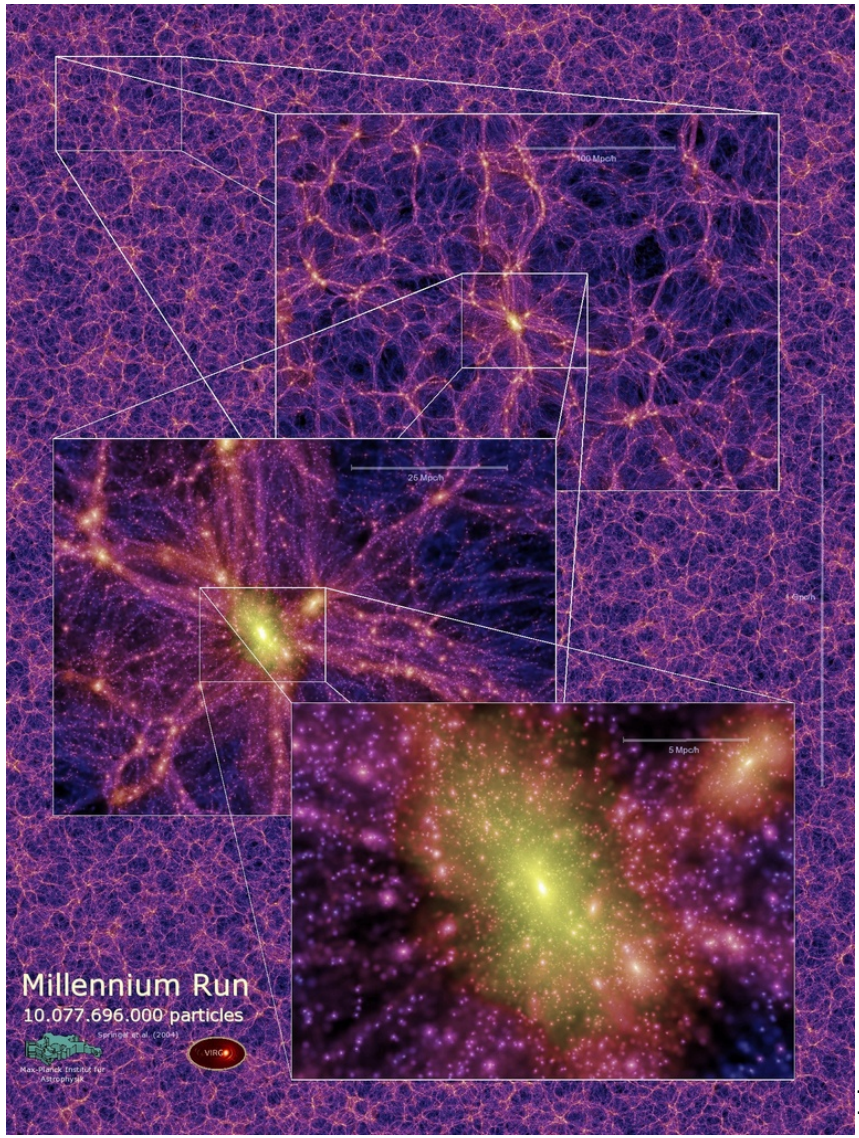
Statistical Inference for Topological Data Analysis

Jisu KIM

The 3rd Joint Conference on Statistics and Data Science in China
(2025 JCSDS)
2025-07-12

Topological structures in the data provide information.



[1]

[1] http://www.mpa-garching.mpg.de/galform/virgo/millennium/poster_half.jpg

Persistent Homology: observe topological structure with multi resolutions.

Persistent Homology: observe topological structure with multi resolutions.

Persistent Homology: observe topological structure with multi resolutions.

Persistent Homology: observe topological structure with multi resolutions.

- Georges Seurat, A Sunday afternoon on the island of La Grande Jatte (Un dimanche après-midi à l'Île de la Grande Jatte)

Statistic Inference for Topological Data Analysis is explored.

- Introduction to Topological Data Analysis
  - Computational Topology: An Introduction (Edelsbrunner, Harer, 2010)
  - Topological Data Analysis (Wasserman, 2016)
  - An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists (Chazal, Michel, 2021)
- Statistical Inference For Homological Features
  - Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)
- Statistical Inference for Persistent Homology
  - Confidence sets for persistence diagrams (Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2014)

The number of holes is used to summarize topological features.

- Geometrical objects:
  - 一, 二, 三, 四, 五, 六, 七, 八, 九, 十
  - 中,国,浙,江,杭,州
  - A, ㄱ, ㄴ, あ
- The number of holes of different dimensions is considered.
  1. β_0 =# of connected components 
  2. β_1 =# of loops (holes inside 1-dim sphere) 
  3. β_2 =# of voids (holes inside 2-dim sphere) 

Example : Objects are classified by homologies.

1. β_0 =# of connected components 
2. β_1 =# of loops (holes inside 1-dim sphere) 

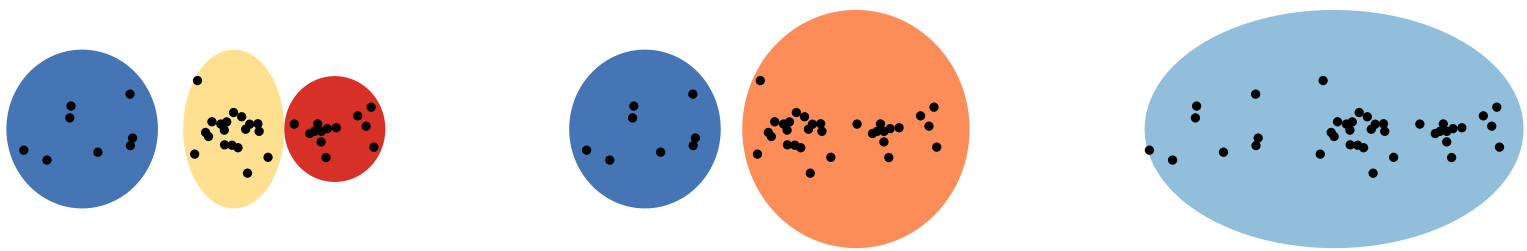
$\beta_0 \setminus \beta_1$	0	1	2
1	一, 七, 九, 十 ㄱ, ㄴ	五, A	四, 中, あ
2	二, 八		
3	三	国	
4	六, 江, 浙, 杭		
5			
6	州		

We want to cluster data.

- Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)

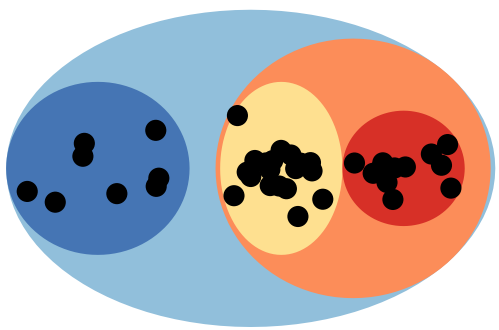
Different clusters can be formed by the desired level of resolution.

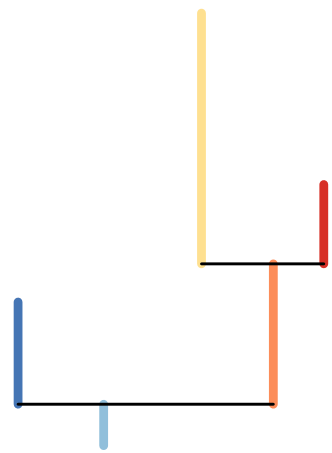
- If you want clusters to describe local and detailed information (high resolution), there will be more clusters with each of smaller sizes.
- If you want clusters to describe global and rough information (low resolution), there will be less clusters with each of larger sizes.



The network of clusters forms a tree: cluster tree

- Clusters from different levels of resolution have a natural network by inclusion relation.
- Inclusion network of clusters can be represented as a tree: cluster tree.

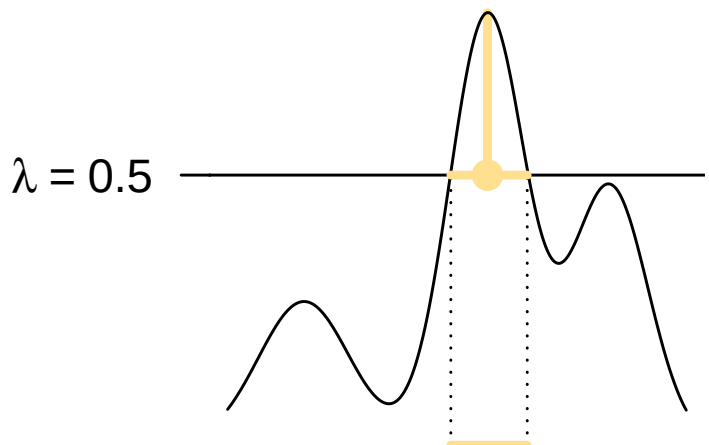




The cluster tree is the hierarchy of the high density clusters.

Definition

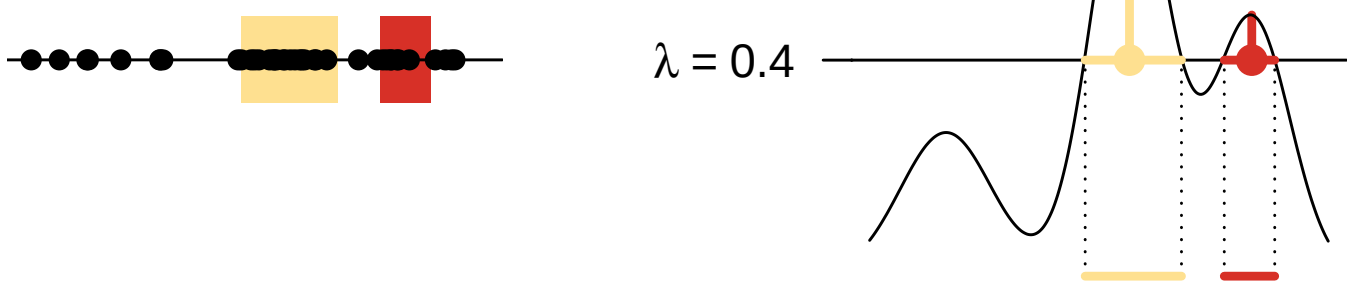
For a density function p, its cluster tree $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$ is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x \in \mathcal{X} : p(x) \geq \lambda\}$.



The cluster tree is the hierarchy of the high density clusters.

Definition

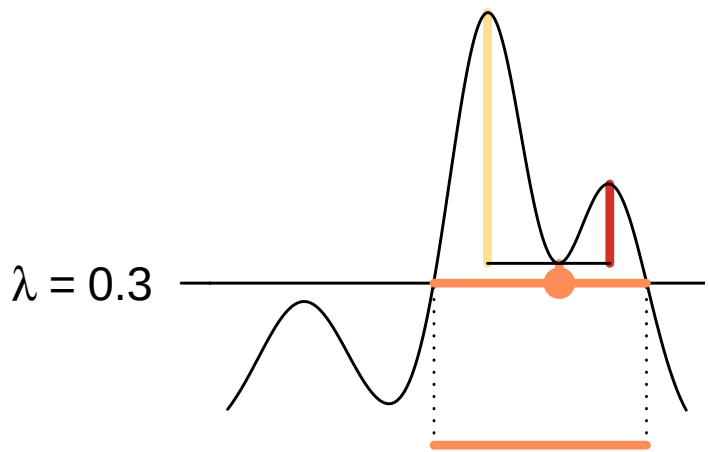
For a density function p, its cluster tree $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$ is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x \in \mathcal{X} : p(x) \geq \lambda\}$.



The cluster tree is the hierarchy of the high density clusters.

Definition

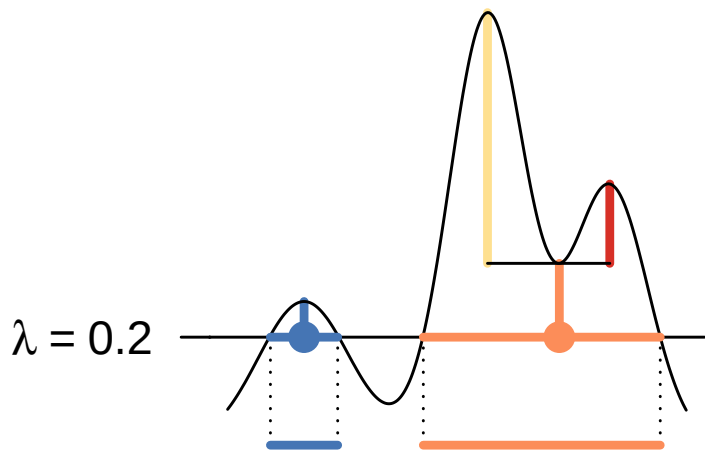
For a density function p, its cluster tree $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$ is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x \in \mathcal{X} : p(x) \geq \lambda\}$.



The cluster tree is the hierarchy of the high density clusters.

Definition

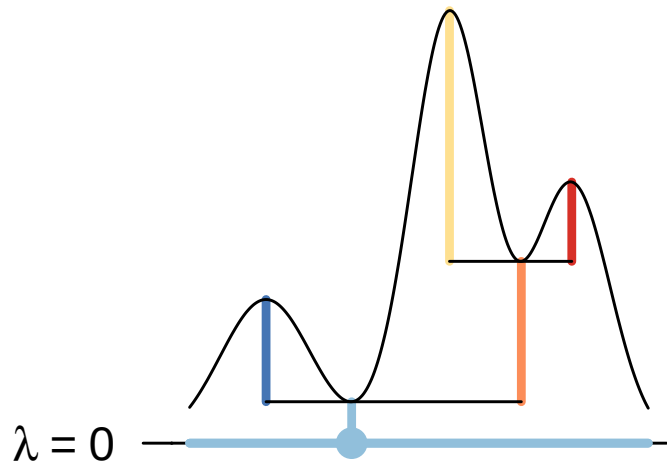
For a density function p, its cluster tree $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$ is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x \in \mathcal{X} : p(x) \geq \lambda\}$.



The cluster tree is the hierarchy of the high density clusters.

Definition

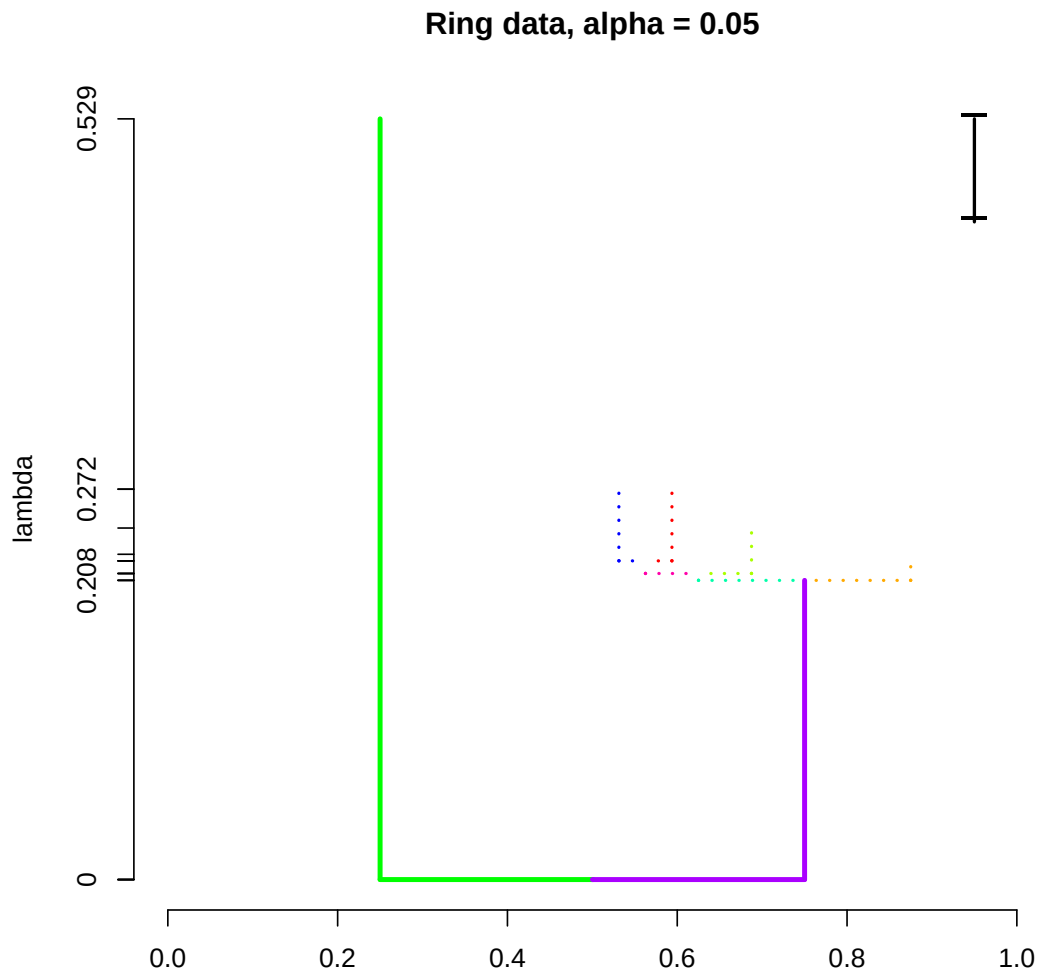
For a density function p, its cluster tree $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$ is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x \in \mathcal{X} : p(x) \geq \lambda\}$.



A confidence set helps denoising the empirical tree.

- An asymptotic $1-\alpha$ confidence set $\hat{C}_\alpha$ is a collection of trees with the property that

$$P(T_p \in \hat{C}_\alpha) = 1 - \alpha + o(1).$$



We use the bootstrap to compute $1-\alpha$ confidence set $\hat{C}_\alpha$.

- We let $T_{\hat{p}_h}$ be the cluster tree from the kernel density estimator $\hat{p}_h$, where

$$\hat{p}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

and the confidence set as the ball centered at $T_{\hat{p}_h}$ and radius t_α, i.e.

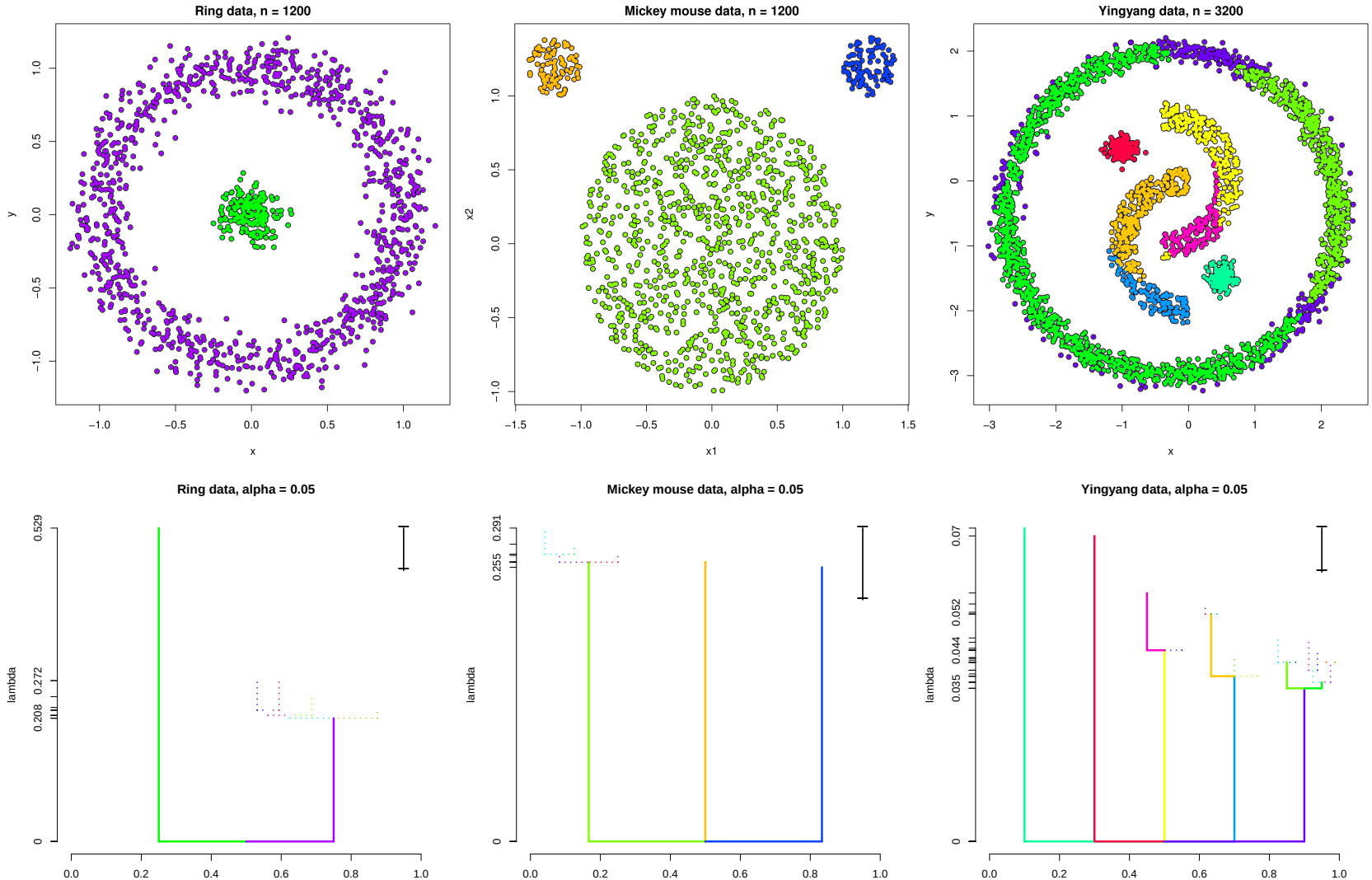
$$\hat{C}_\alpha = \{T : d_\infty(T, T_{\hat{p}_h}) \leq t_\alpha\}.$$

Theorem

(Theorem 3) Above confidence set $\hat{C}_\alpha$ satisfies

$$P\left(T_h \in \hat{C}_\alpha\right) = 1 - \alpha + O\left(\left(\frac{\log^7 n}{nh^d}\right)^{1/6}\right).$$

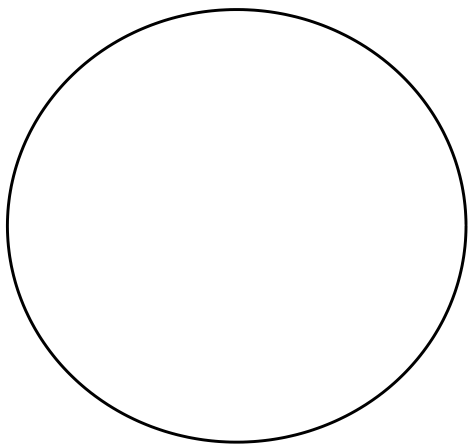
The pruned trees according to the confidence set recover the actual cluster trees.



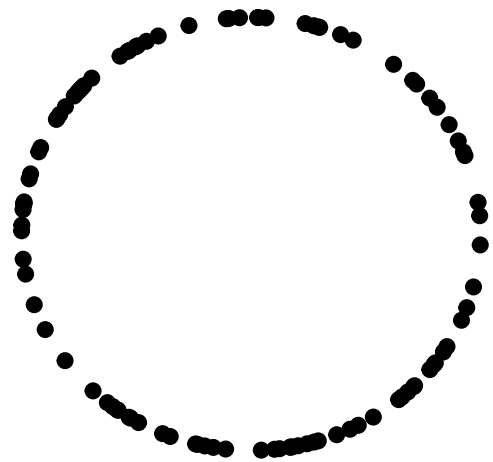
Homology of finite sample is different from homology of underlying manifold, hence it cannot be directly used for the inference.

- When analyzing data, we prefer robust features where features of the underlying manifold can be inferred from features of finite samples.
- Homology is not robust:

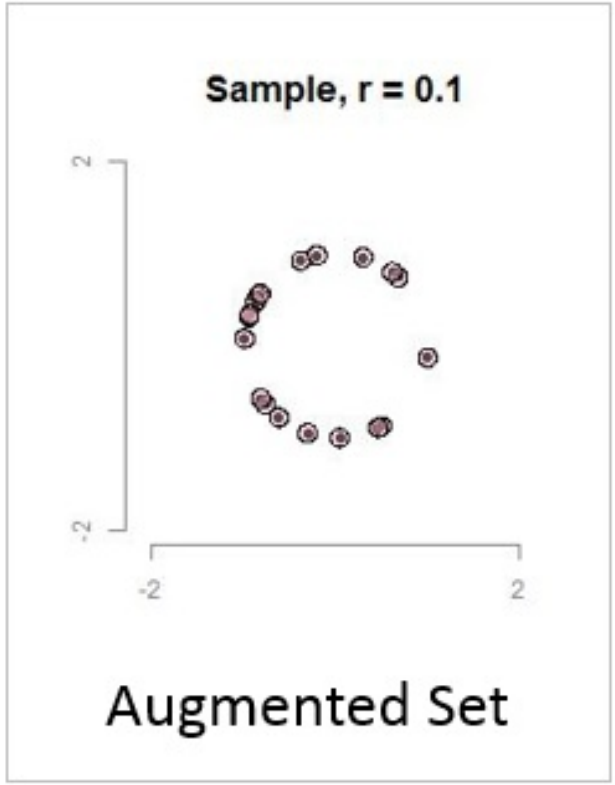
Underlying circle: $\beta_0 = 1, \beta_1 = 1$



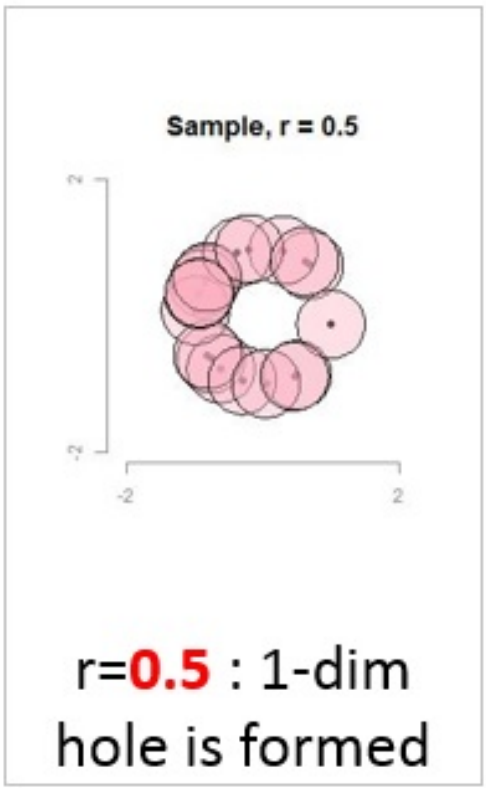
100 samples: $\beta_0 = 100, \beta_1 = 0$

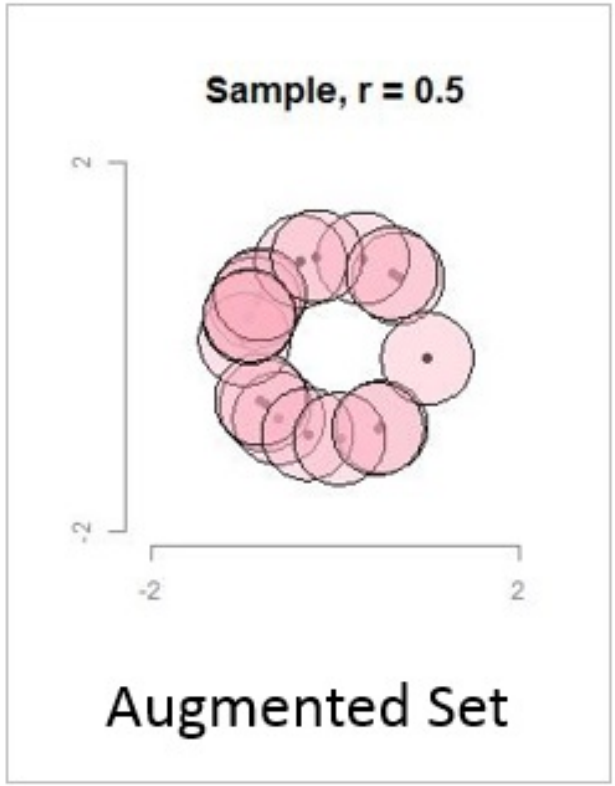


Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.

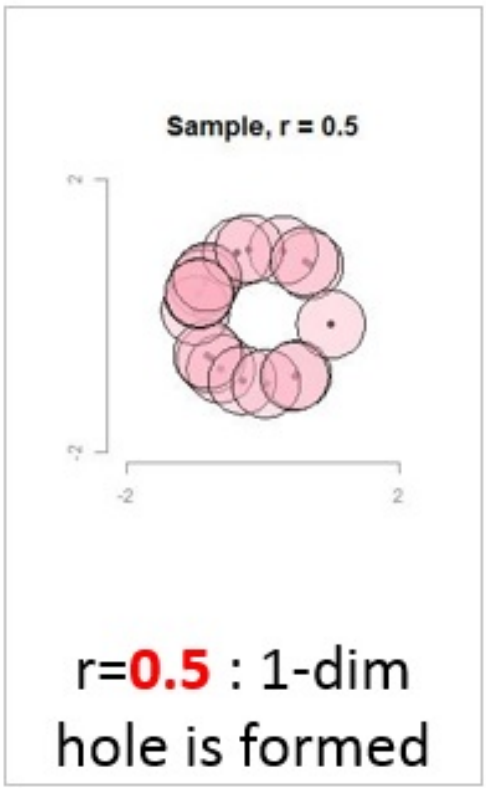


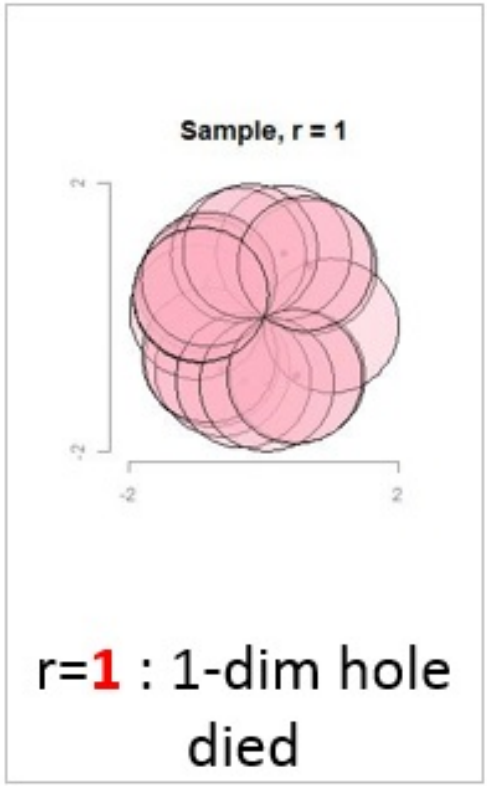
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.

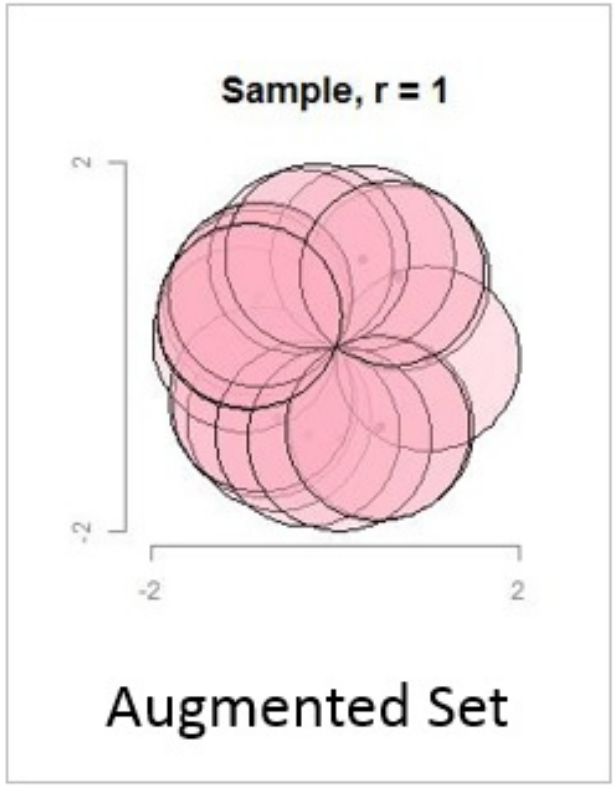




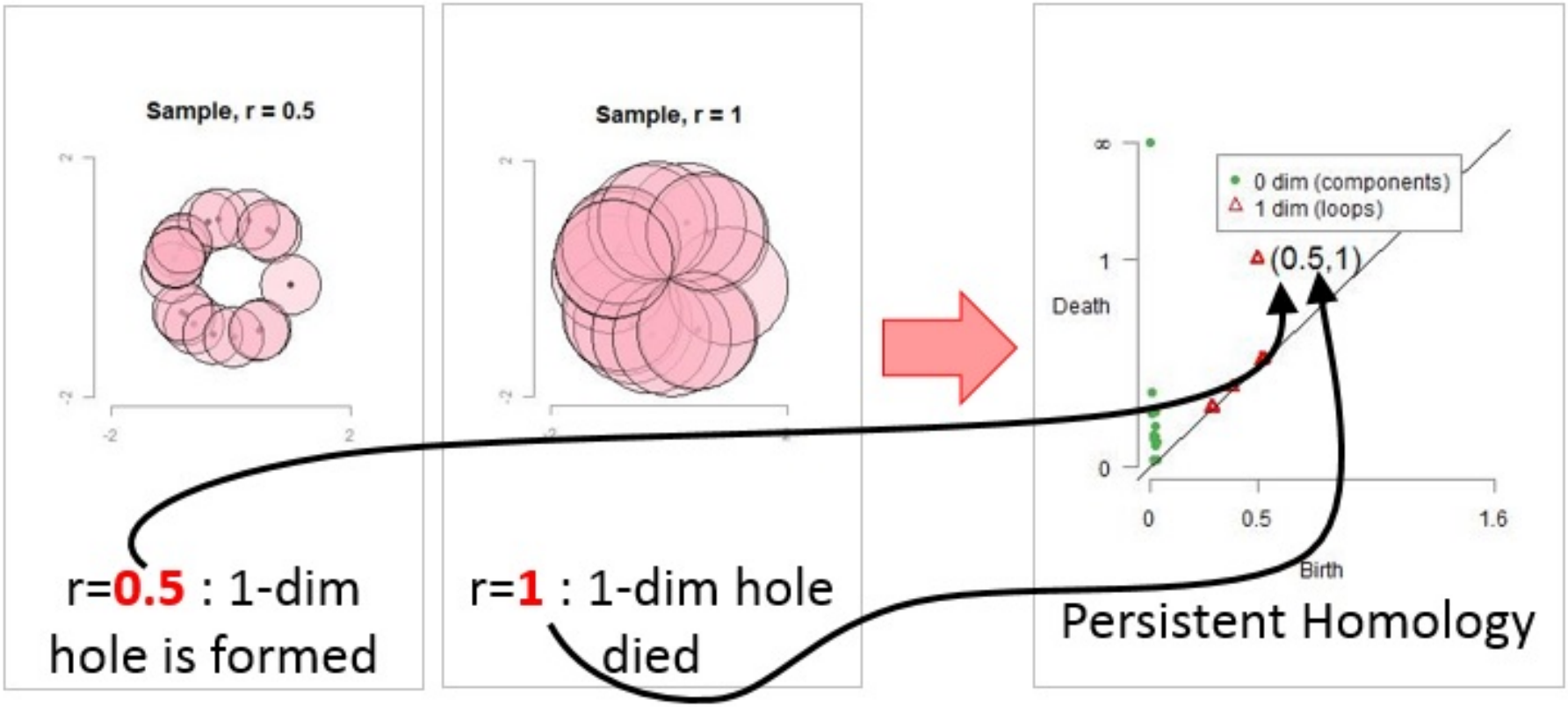
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.







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



We rely on the superlevel sets of the kernel density estimator to extract topological information of the underlying distribution.

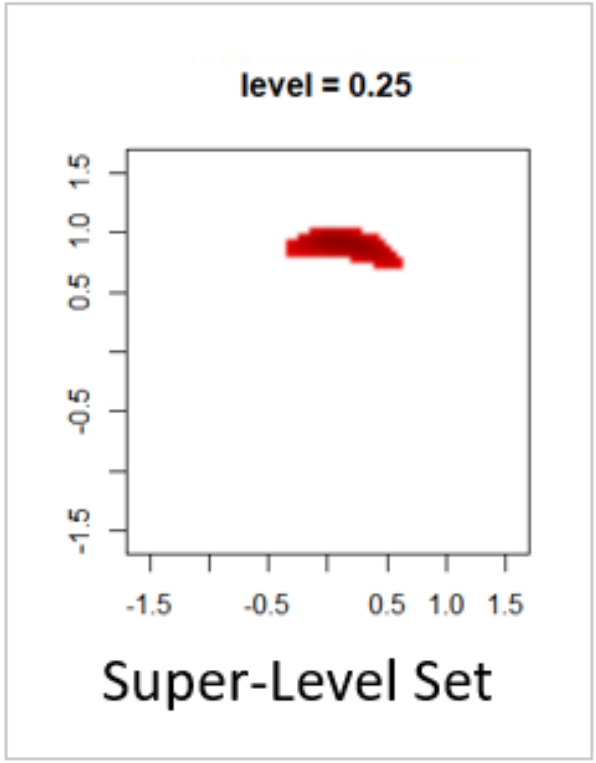
- The kernel density estimator is

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

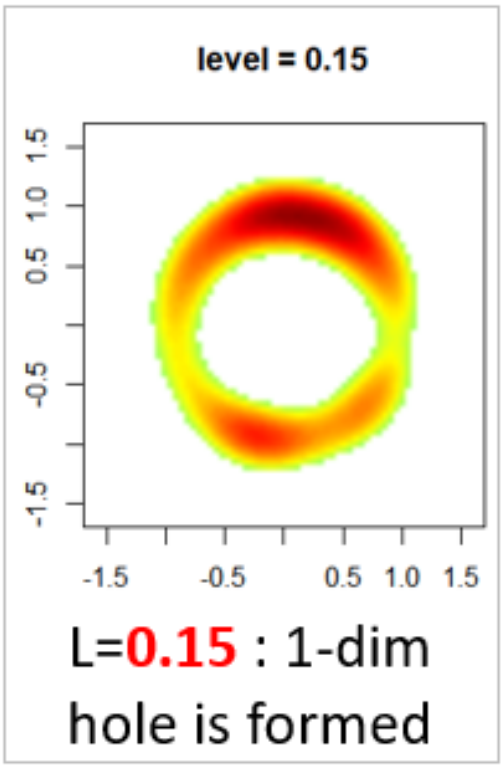
- We look at superlevel sets of the kernel density estimator as

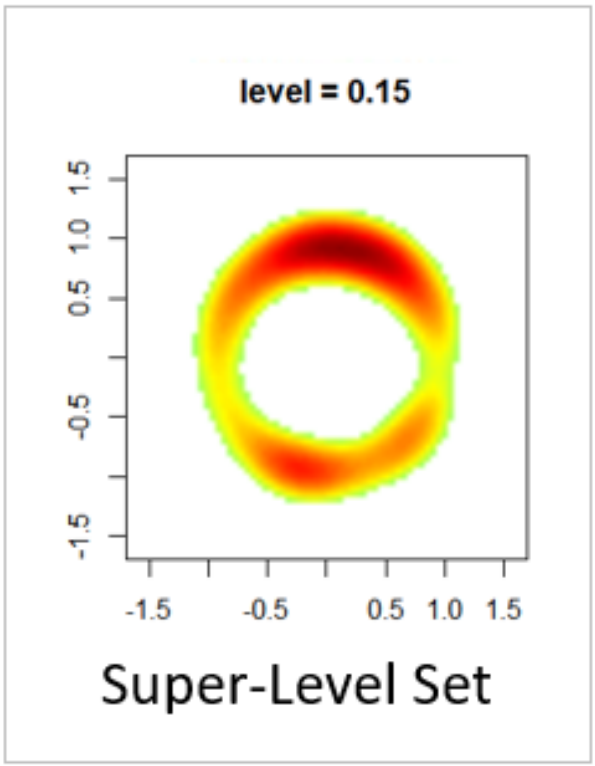
$$\left\{x \in \mathbb{R}^d : \hat{p}_h(x) \geq L\right\}_{L>0}.$$

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

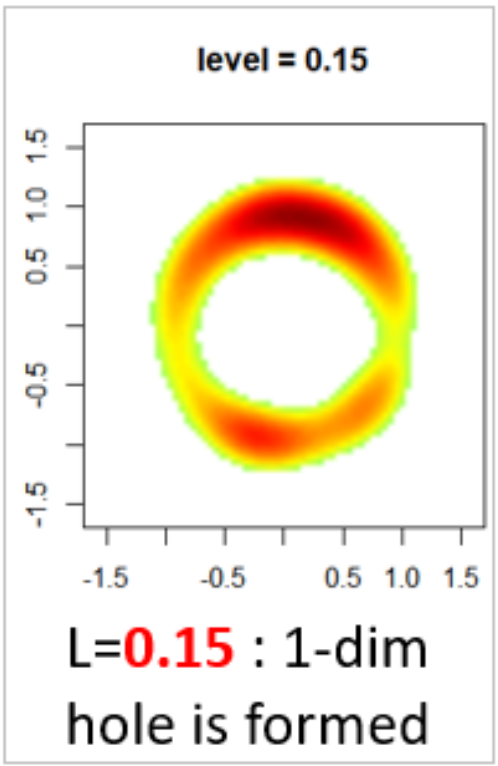


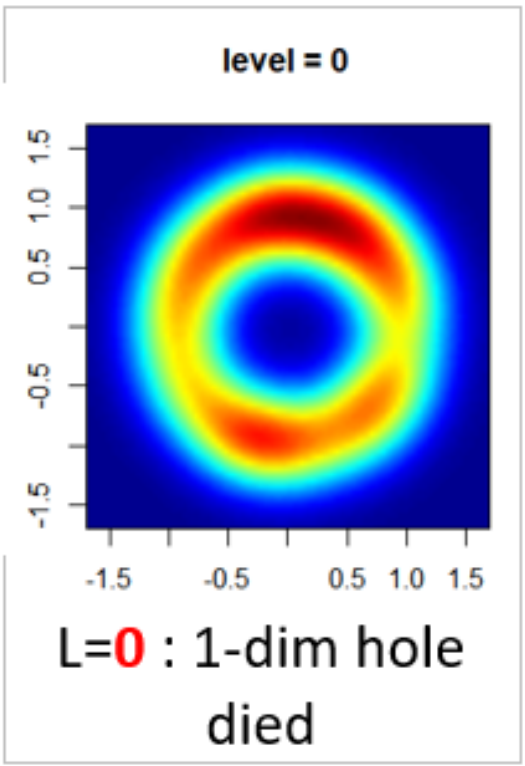
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.

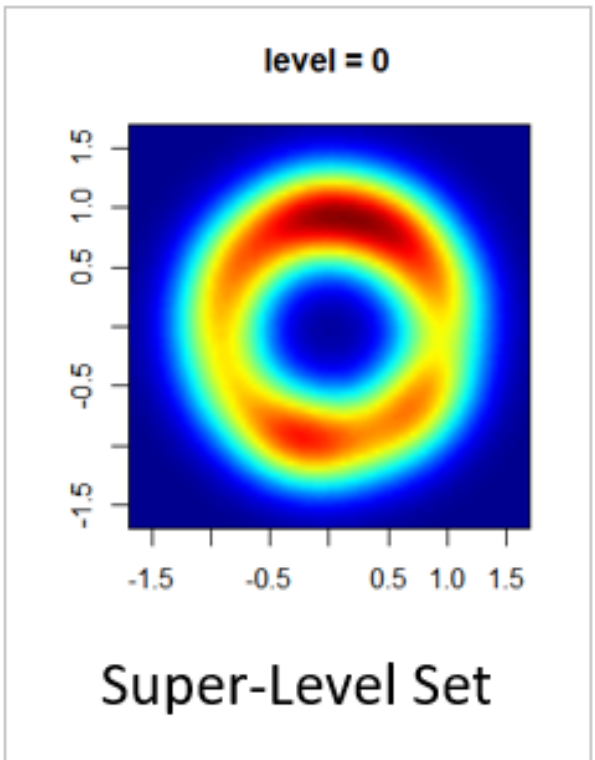




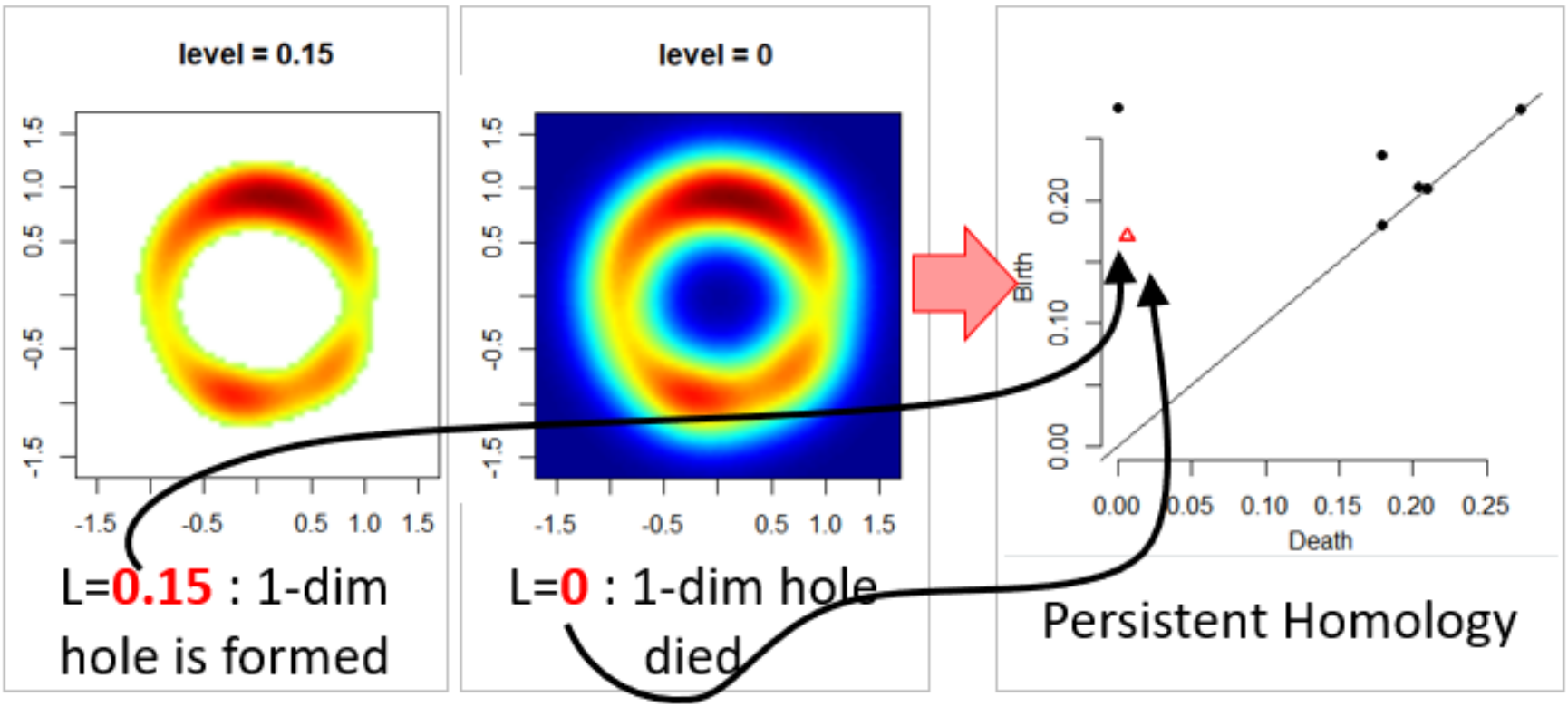
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



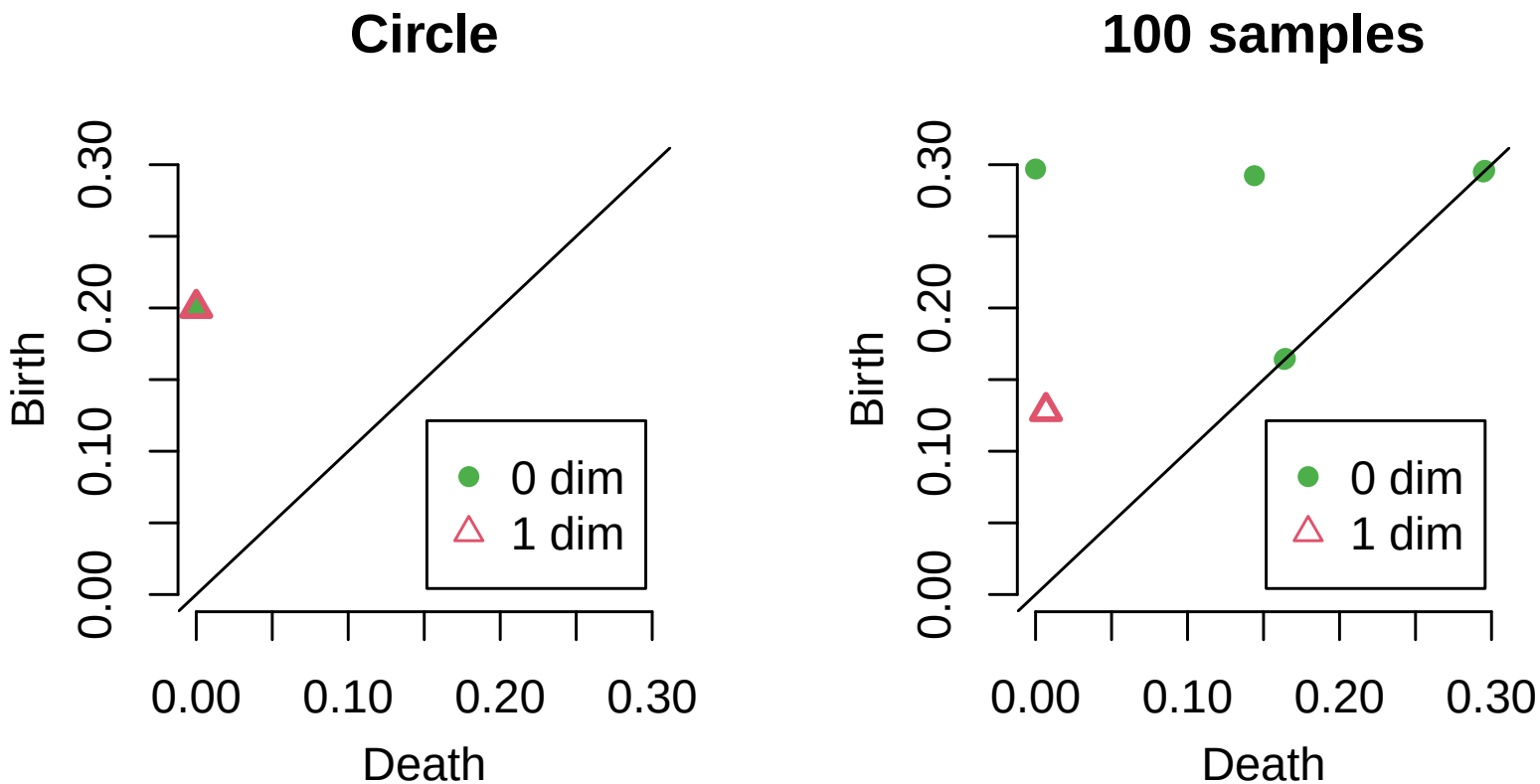




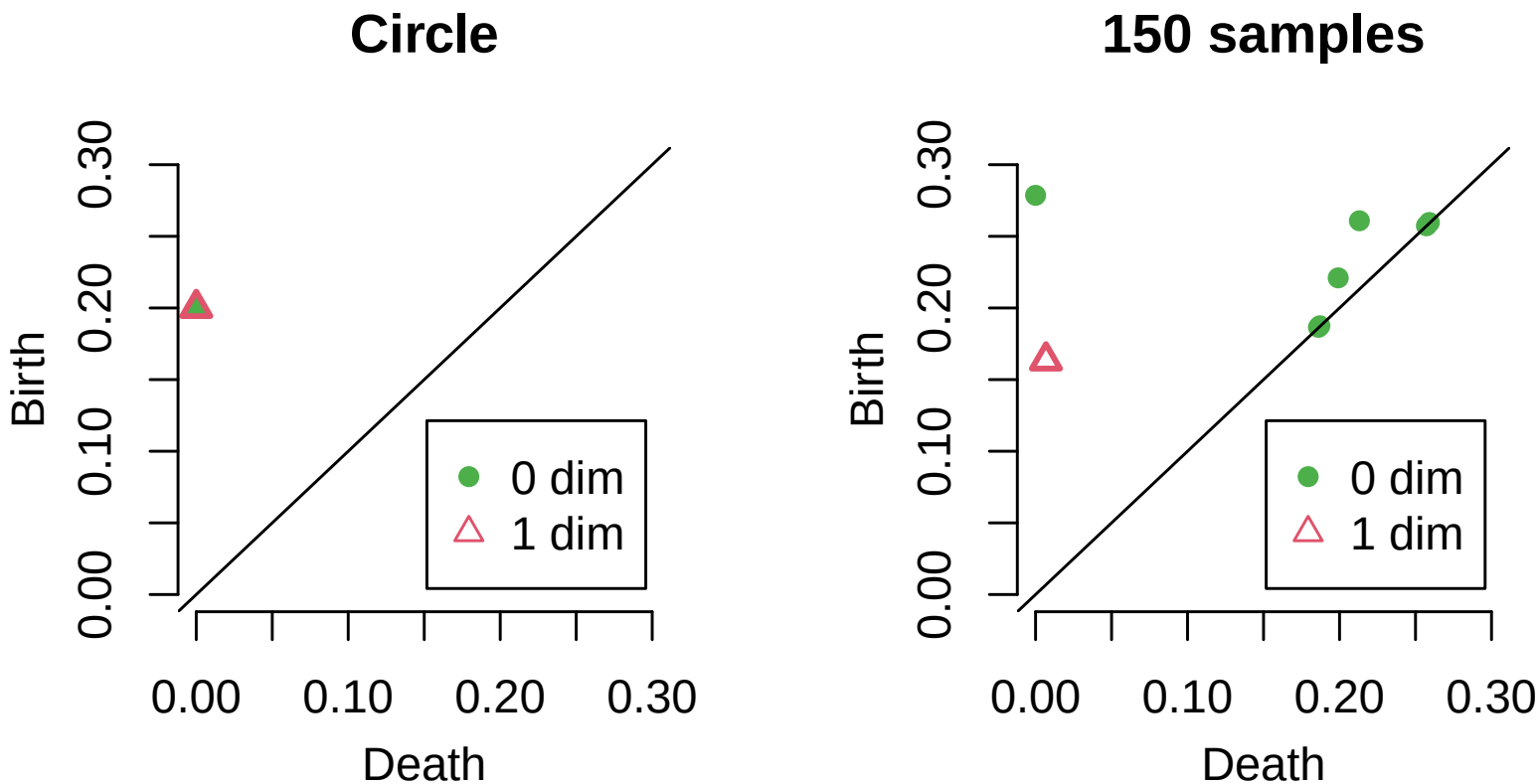
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



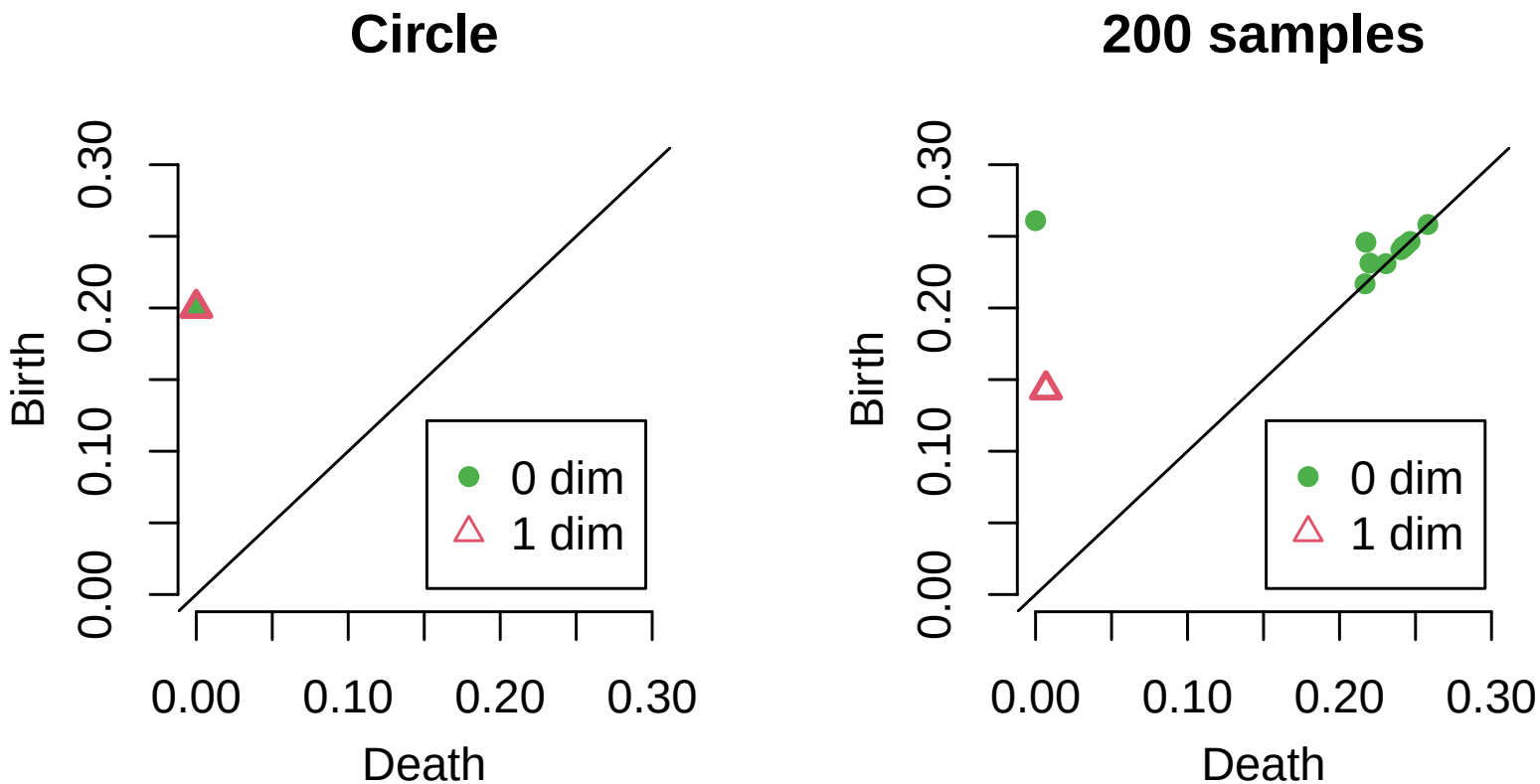
Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.



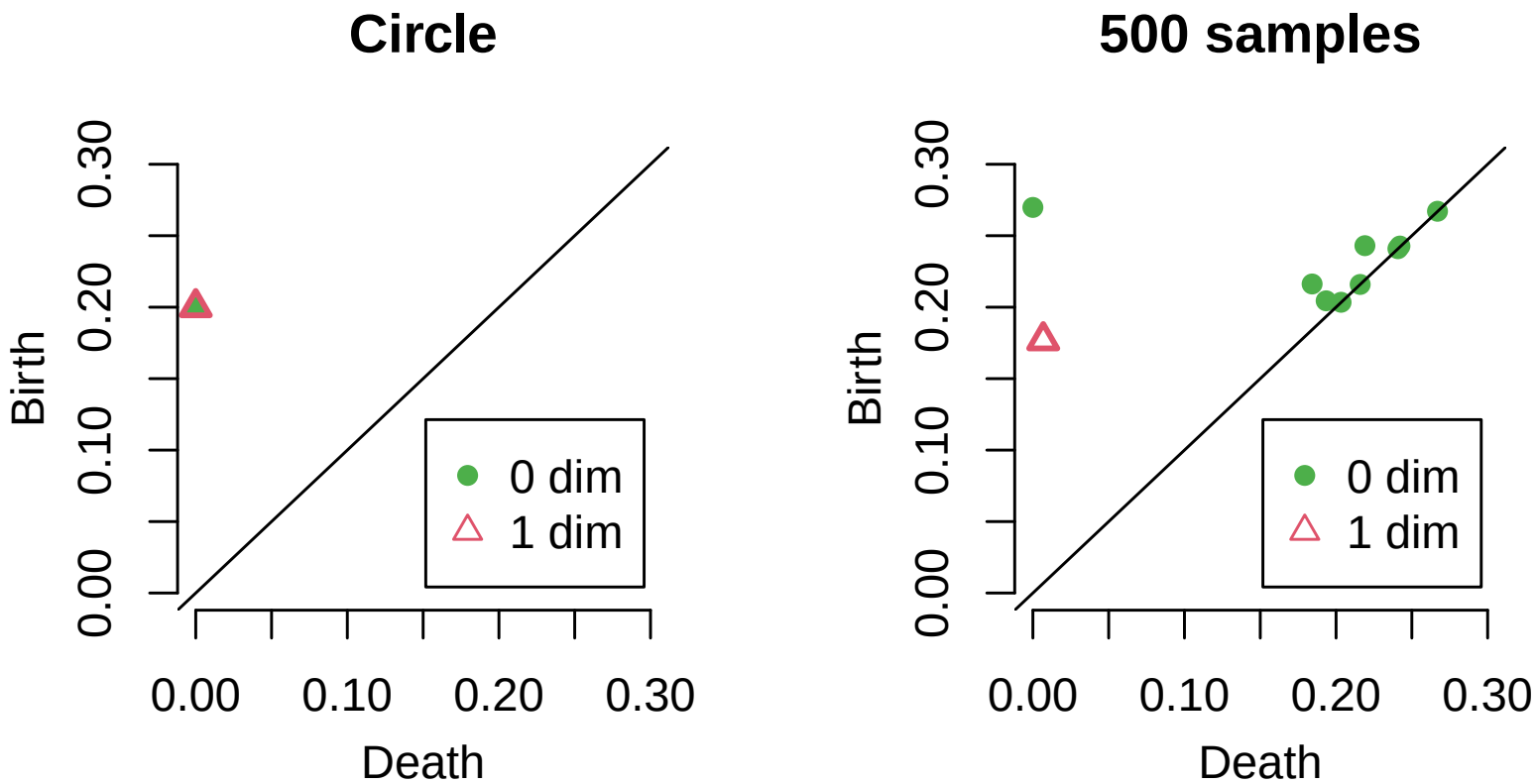
Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.



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

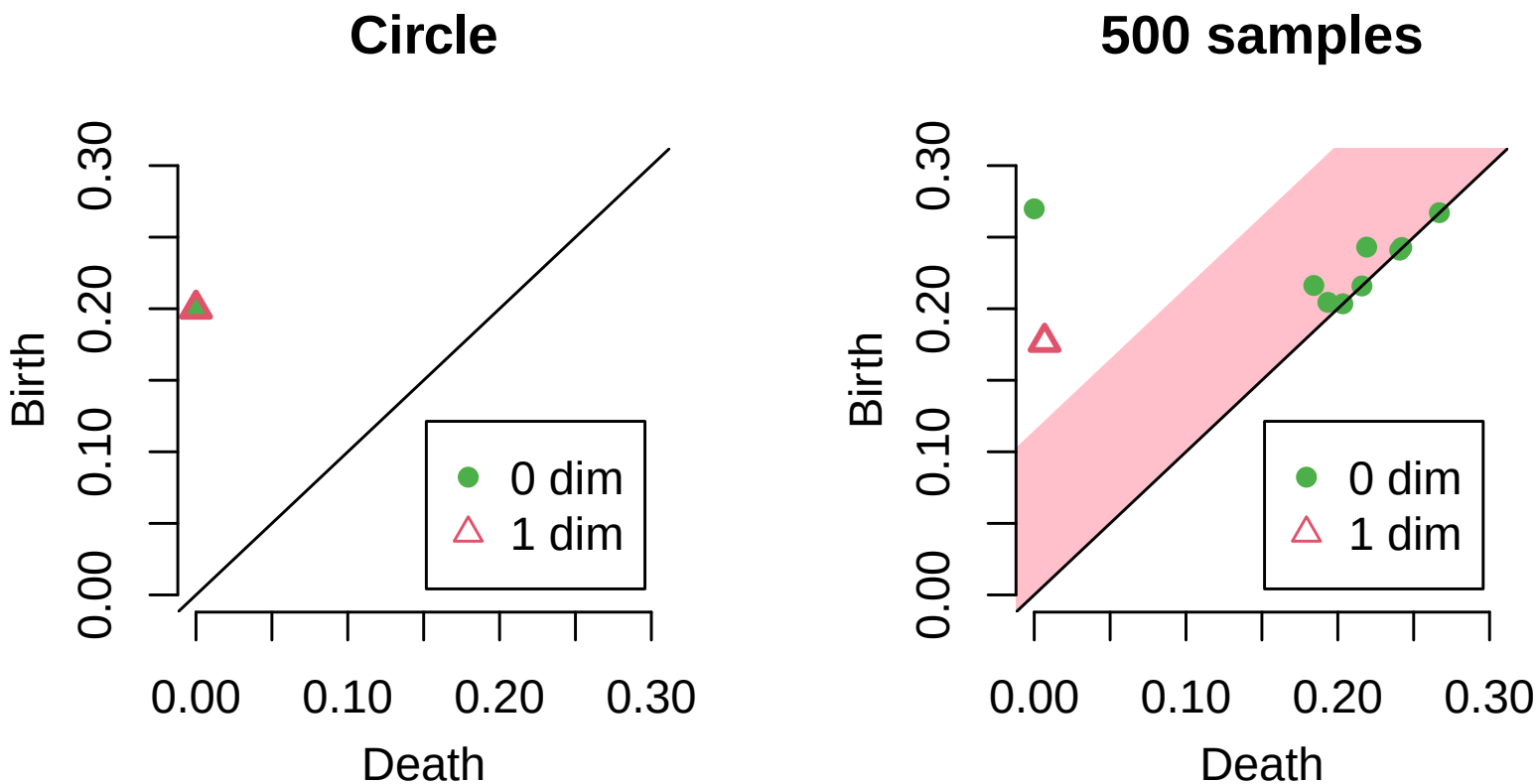


Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.



Statistically significant homological features can be distinguished from statistically insignificant ones.

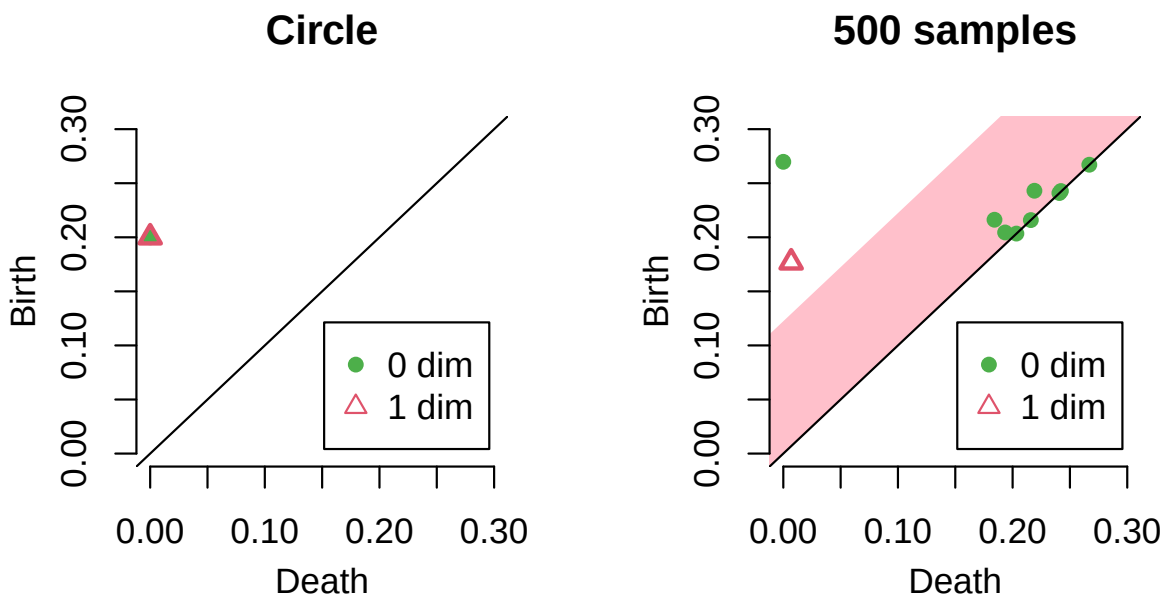
- Confidence sets for persistence diagrams (Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2014)



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

Let $Dgm(M)$ and $Dgm(X)$ be persistent homologies of the manifold M and the data X, respectively. Given the significance level $\alpha \in (0, 1)$, $(1 - \alpha)$ confidence band $c_n = c_n(X)$ is a random variable satisfying

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



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 $\theta^* = \sqrt{nh^d}||\hat{p}_h^*(x) - \hat{p}_h(x)||_\infty$, 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 $\hat{z}_\alpha = \inf\left\{q : \frac{1}{B}\sum_{j=1}^{B} I(\theta_j^* \geq q) \leq \alpha\right\}$
5. The $(1-\alpha)$ confidence band for $\mathbb{E}[\hat{p}_h]$ is $\left[\hat{p}_h - \frac{\hat{z}_\alpha}{\sqrt{nh^d}}\ ,\ \hat{p}_h + \frac{\hat{z}_\alpha}{\sqrt{nh^d}}\right]$.

Reference I

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers Artif. Intell.*, 4:667963, 2021. doi: $10.3389/\mathrm{frai}.2021.667963$. URL `https://doi.org/10.3389/frai.2021.667963`.

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*, 2014.

Herbert Edelsbrunner and John L. Harer. *Computational topology*. American Mathematical Society, Providence, RI, 2010. ISBN 978-0-8218-4925-5. doi: $10.1090/\mathrm{mbk}/069$. URL `https://doi.org/10.1090/mbk/069`. An introduction.

Reference II

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *Ann. Statist.*, 42(6):2301–2339, 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1252. URL https://doi.org/10.1214/14-AOS1252.

Jisu KIM, Yen-Chi Chen, Sivaraman Balakrishnan, Alessandro Rinaldo, and Larry Wasserman. Statistical inference for cluster trees. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1839–1847. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6508-statistical-inference-for-cluster-trees.pdf.

Larry Wasserman. Topological data analysis, 2016.

Thank you!

Statistical Inference for Cluster Trees

Persistent Homology

We can use ℓ_∞ metric to measure a distance between trees.

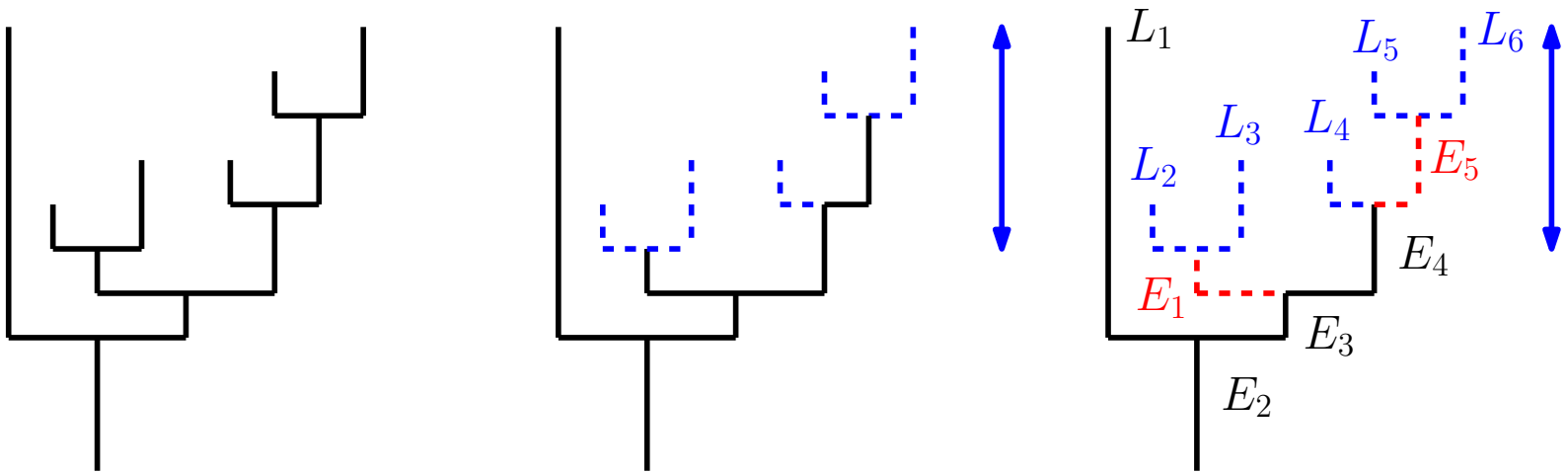
Definition

The l_∞ metric between trees are defined as

$$d_\infty(T_p, T_q) = \sup |p(x) - q(x)| .$$

Pruning finds the simpler trees that are in the confidence set.

- We propose two pruning schemes to find trees that are simpler the empirical tree $T_{\hat{p}_h}$ and are in the fconfidence set.
  - Pruning only leaves: remove all leaves of length less than $2t_\alpha$.
  - Pruning leaves and internal branches: iteratively remove all branches of cumulative length less than $2t_\alpha$.



Persistent Homology

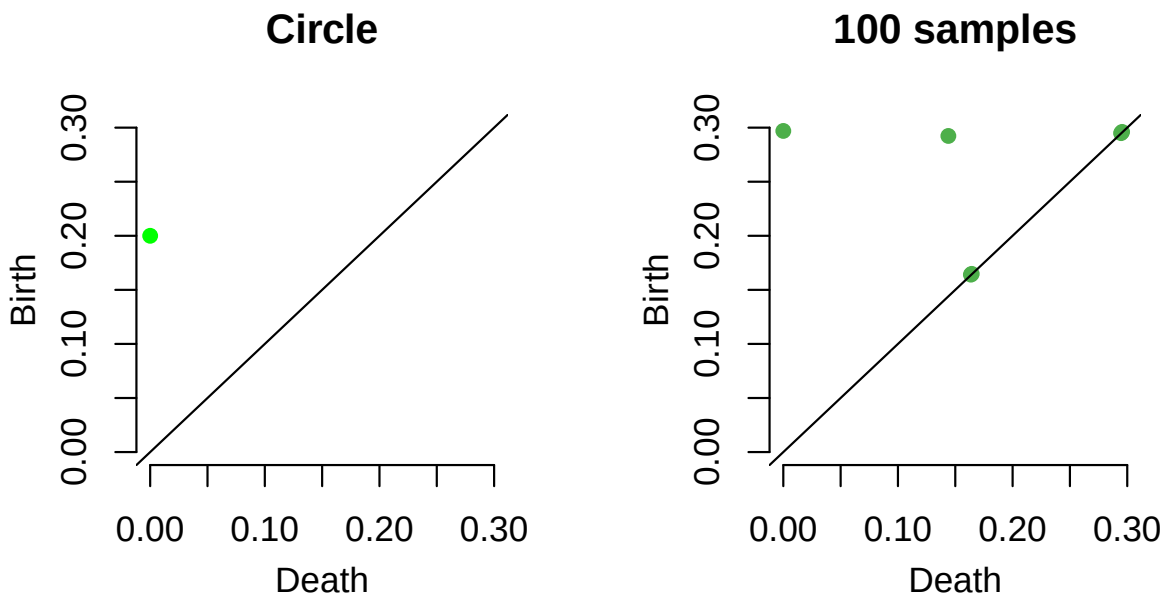
Bottleneck distance gives a metric on the space of persistent homology.

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.



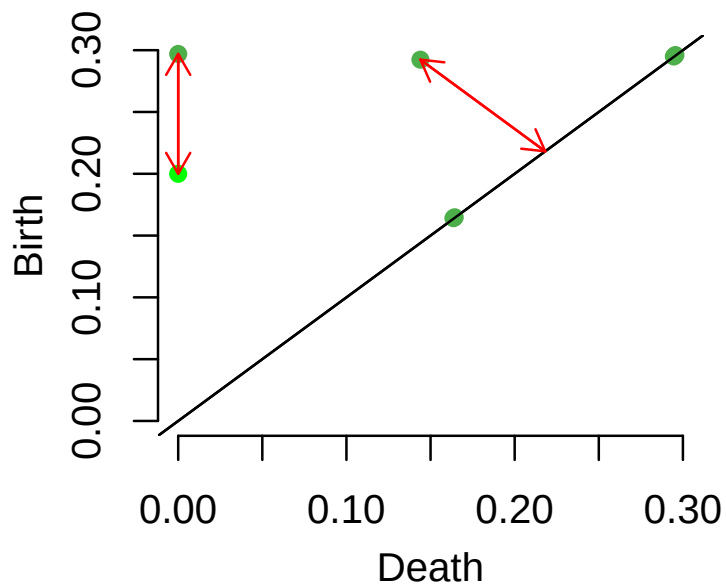
Bottleneck distance gives a metric on the space of persistent homology.

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.



$$\sup_{x \in D_1} \|x - \gamma_1(x)\|_\infty = 0.1$$

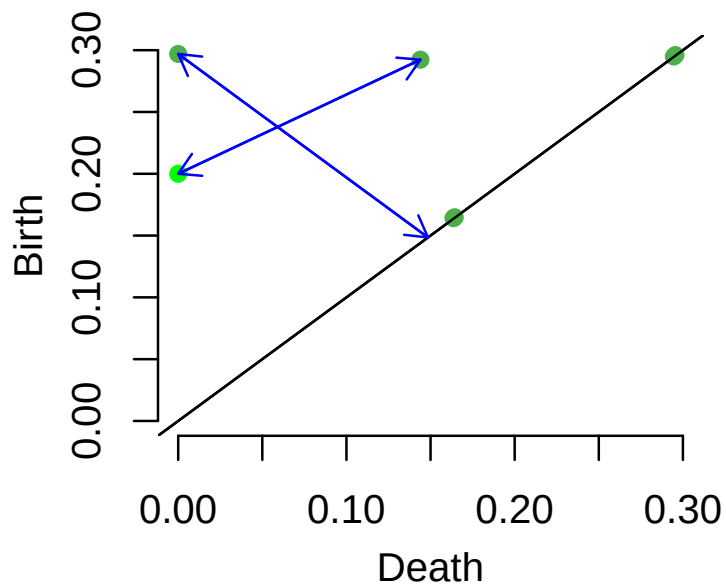
Bottleneck distance gives a metric on the space of persistent homology.

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.



$$\sup_{x \in D_1} \|x - \gamma_2(x)\|_\infty = 0.15$$

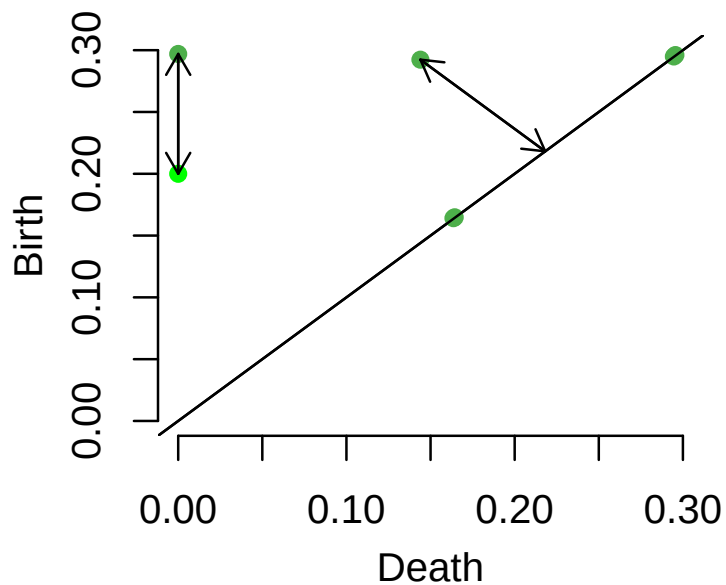
Bottleneck distance gives a metric on the space of persistent homology.

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.



$$\inf_\gamma \sup_{x \in D_1} \|x - \gamma(x)\|_\infty = 0.1$$

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

Theorem

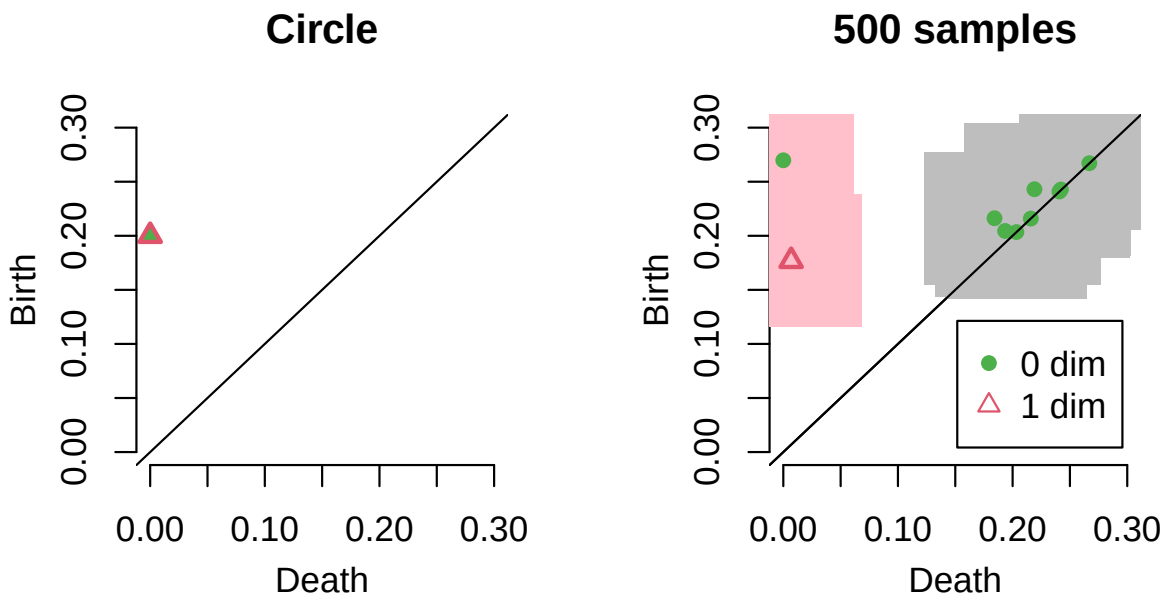
[Edelsbrunner and Harer, 2010][Chazal, de Silva, Glisse, and Oudot, 2012] Let $\mathbb{X}$ be finitely triangulable space and f, $g : \mathbb{X} \to \mathbb{R}$ be two continuous functions. Let $Dgm(f)$ and $Dgm(g)$ be corresponding persistence diagrams. Then

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

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

Let $Dgm(M)$ and $Dgm(X)$ be persistent homologies of the manifold M and the data X, respectively. Given the significance level $\alpha \in (0, 1)$, $(1-\alpha)$ confidence set $\{D \in Dgm : \ W_\infty(Dgm(X), D) \leq c_n\}$ is a random set satisfying

$$\mathbb{P}\left(Dgm(M) \in \{D \in Dgm : \ W_\infty(Dgm(X), D) \leq c_n\}\right) \geq 1 - \alpha.$$



Confidence band for the persistent homology can be obtained by the corresponding confidence band for functions.

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

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

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.

Bootstrap algorithm can be applied to peristent homology.

- for the case of kernel density estimator in Fasy et al. [2014],
- for the case of distance to measure and kernel distance in Chazal et al. [2014].