

Confidence Band for Persistent Homology

Jisu KIM

Inria

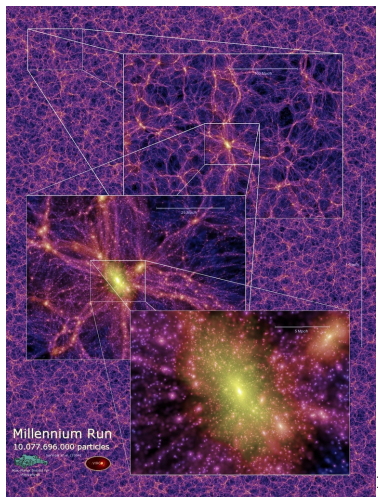
2020-08-06

Background

Confidence band for Persistent Homology of KDEs computed on a grid




Confidence band for Persistent Homology of KDEs computed on Rips complexes

Geometric structures in the data provide information.



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

The number of holes is used to summarize geometrical features.

- ▶ Geometrical objects :
 - ▶ A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z,
 - ▶ 가, 字, 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) : if $dim \geq 3$ 

Example : Objects are classified by homologies.

1. $\beta_0 = \#$ of connected components ●

2. $\beta_1 = \#$ of loops ○

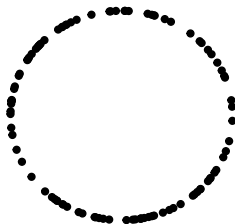
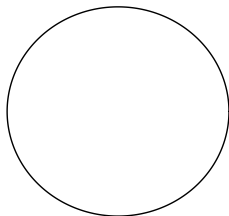
| $\beta_0 \setminus \beta_1$ | 0 | 1 | 2 |
|-----------------------------|---|------------------|------|
| 1 | C, G, I, J, L, M, N, S, U, V, W, Z, E, F, T, Y, H, K, X | A, R, D, O, P, Q | B, あ |
| 2 | 가, 字 | | |

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

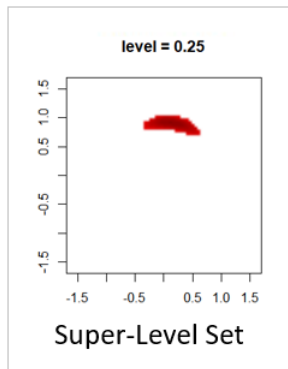


We rely on the kernel density estimator to extract topological information of the underlying distribution.

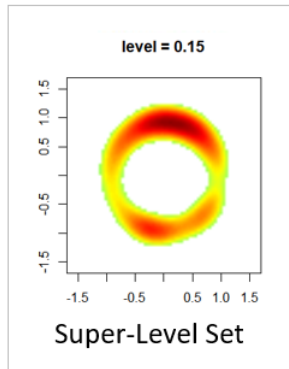
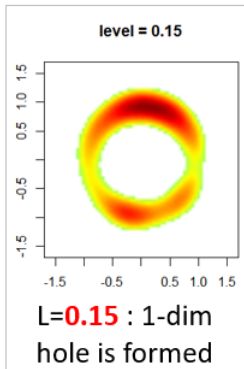
- ▶ The kernel density estimator is

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

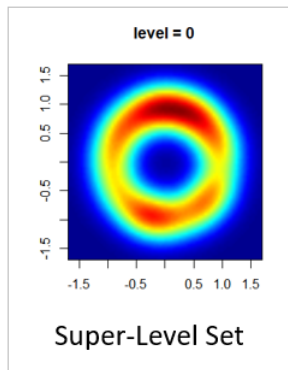
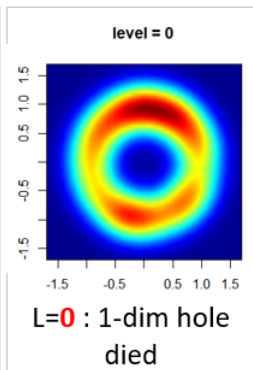
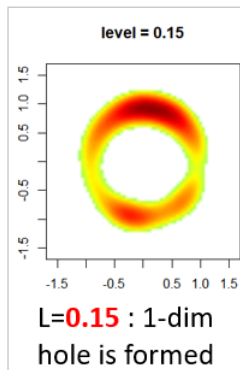
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



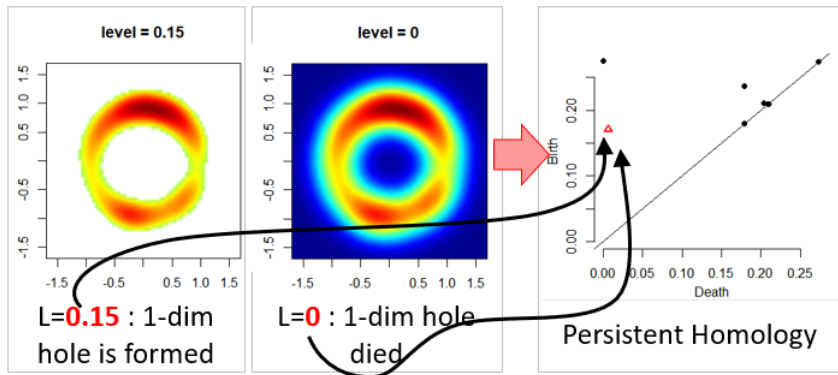
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



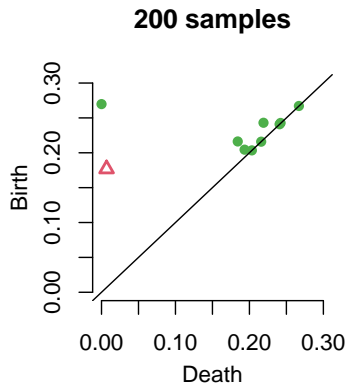
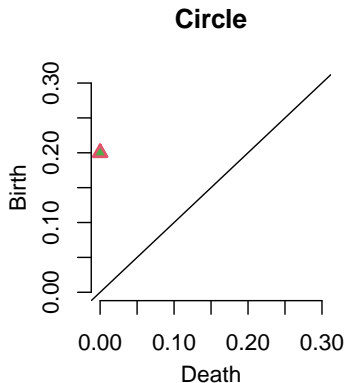
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



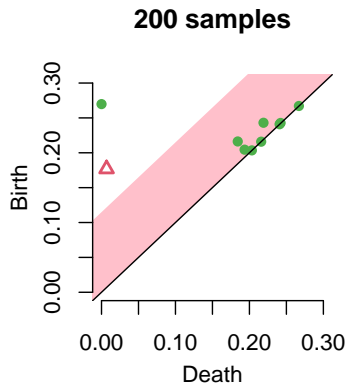
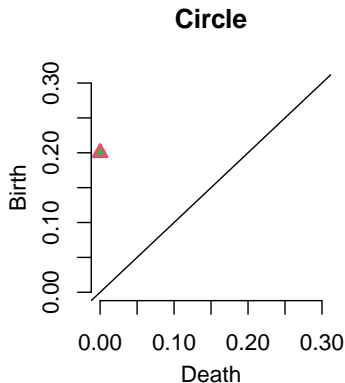
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



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



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



Background

Confidence band for Persistent Homology of KDEs computed on a grid

Confidence band for Persistent Homology of KDEs computed on Rips complexes

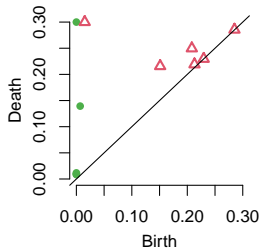
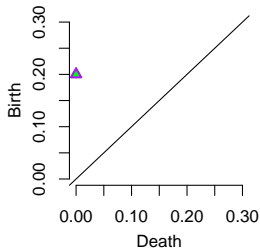
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

where γ ranges over all bijections from D_1 to D_2 .



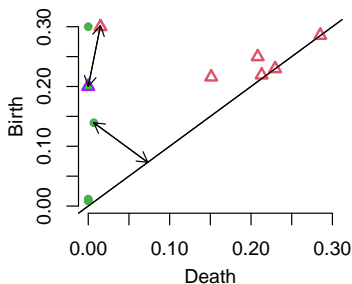
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

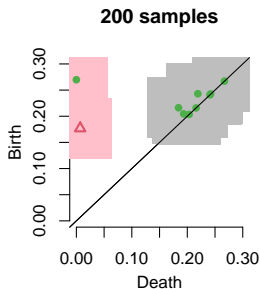
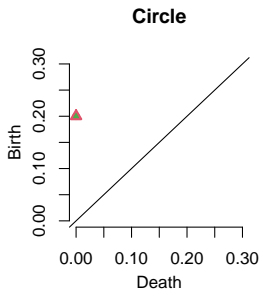
where γ ranges over all bijections from D_1 to D_2 .



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

Let M be a compact manifold, and $X = \{X_1, \dots, X_n\}$ 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

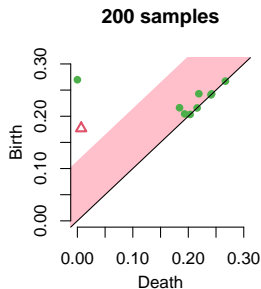
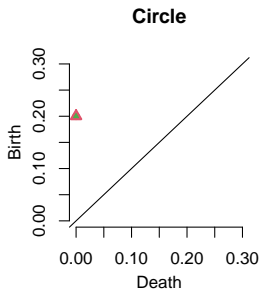
$$\mathbb{P}(Dgm(f_M) \in \{\mathcal{P} : d_B(\mathcal{P}, Dgm(f_X)) \leq c_n\}) \geq 1 - \alpha.$$



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

Let M be a compact manifold, and $X = \{X_1, \dots, X_n\}$ 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

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



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

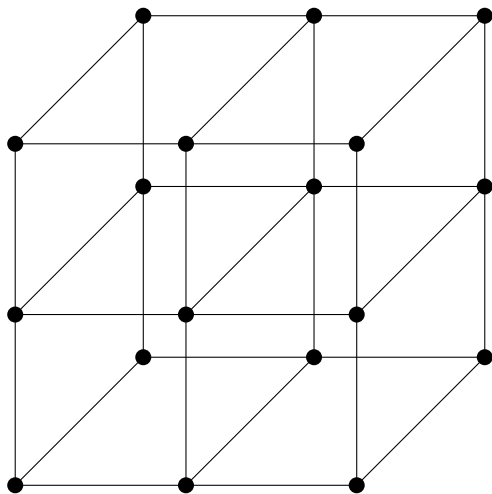
1. Given a sample $X = \{x_1, \dots, x_n\}$, compute the kernel density estimator \hat{p}_h .
2. Draw $X^* = \{x_1^*, \dots, x_n^*\}$ from $X = \{x_1, \dots, x_n\}$ (with replacement), and compute $\theta^* = \sqrt{nh^d} \|\hat{p}_h^*(x) - \hat{p}_h(x)\|_\infty$, where \hat{p}_h^* is the density estimator computed using X^* .
3. Repeat the previous step B times to obtain $\theta_1^*, \dots, \theta_B^*$
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 $\mathbb{E}[p_h]$ is $\left[\hat{p}_h - \frac{\hat{z}_\alpha}{\sqrt{nh^d}}, \hat{p}_h + \frac{\hat{z}_\alpha}{\sqrt{nh^d}} \right]$.

Background

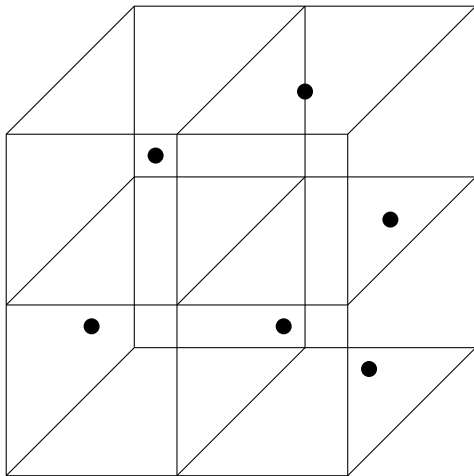
Confidence band for Persistent Homology of KDEs computed on a grid

Confidence band for Persistent Homology of KDEs computed on Rips complexes

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

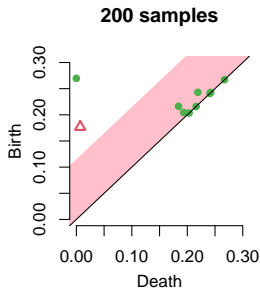
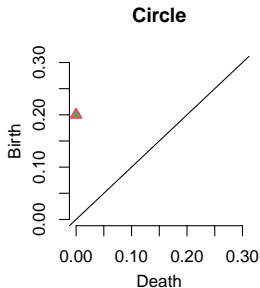


Computing the persistent homology of density function on data points reduces computational complexity.



How can we compute a confidence band for the persistent homology with computation on data points?

- ▶ (Shin, Kim, Rinaldo, Wasserman, 2020?) : extending work from Fasy et al. [2014], Bobrowski et al. [2014], Chazal et al. [2011].



We are considering the upper level set of the average KDE on the support.

- ▶ Let $X_1, \dots, X_n \sim P$, then the average kernel density estimator (KDE) is

$$p_h(x) = \mathbb{E}_P[\hat{p}_h(x)] = \frac{1}{h^d} \mathbb{E}_P \left[K \left(\frac{x - X}{h} \right) \right].$$

- ▶ We are considering the upper level sets of the average KDE as

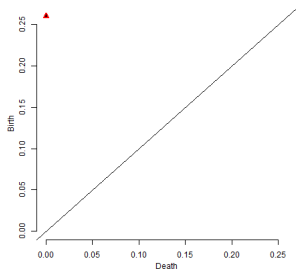
$$\{D_L\}_{L>0}, \text{ where } D_L := \{x \in \text{supp}(P) : p_h(x) \geq L\}.$$

We are targeting the persistent homology of the upper level set of the average KDE on the support.

- ▶ We are considering the upper level sets of the KDE as

$$\{D_L\}_{L>0}, \text{ where } D_L := \{x \in \text{supp}(P) : p_h(x) \geq L\}.$$

and targeting its persistent homology $PH_*^{\text{supp}(P)}(p_h)$.

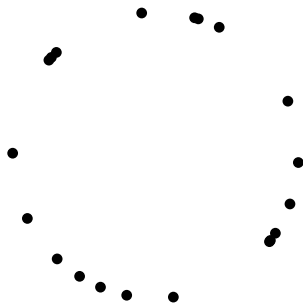


We use the Vietoris-Rips complex to estimate the target persistent homology.

- ▶ For $\mathcal{X} \subset \mathbb{R}^d$ and $r > 0$, the Vietoris-Rips complex $\text{Rips}(\mathcal{X}, r)$ is defined as

$$\text{Rips}(\mathcal{X}, r) = \{ \{x_1, \dots, x_k\} \subset \mathcal{X} : d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k \}.$$

Vietoris-Rips complex

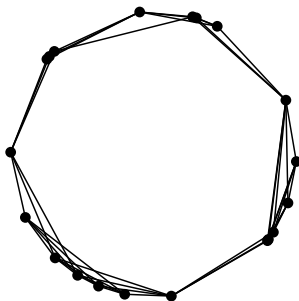


We use the Vietoris-Rips complex to estimate the target persistent homology.

- ▶ For $\mathcal{X} \subset \mathbb{R}^d$ and $r > 0$, the Vietoris-Rips complex $\text{Rips}(\mathcal{X}, r)$ is defined as

$$\text{Rips}(\mathcal{X}, r) = \{ \{x_1, \dots, x_k\} \subset \mathcal{X} : d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k \}.$$

Vietoris-Rips complex

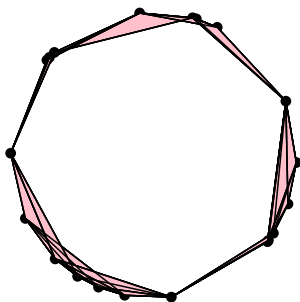


We use the Vietoris-Rips complex to estimate the target persistent homology.

- ▶ For $\mathcal{X} \subset \mathbb{R}^d$ and $r > 0$, the Vietoris-Rips complex $\text{Rips}(\mathcal{X}, r)$ is defined as

$$\text{Rips}(\mathcal{X}, r) = \{ \{x_1, \dots, x_k\} \subset \mathcal{X} : d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k \}.$$

Vietoris-Rips complex



We estimate the target level set by considering the Vietoris-Rips complex generated from the level set of the KDE.

- ▶ For $\mathcal{X} \subset \mathbb{R}^d$ and $r > 0$, the Vietoris-Rips complex $\text{Rips}(\mathcal{X}, r)$ is defined as

$$\text{Rips}(\mathcal{X}, r) = \{ \{x_1, \dots, x_k\} \subset \mathcal{X} : d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k \}.$$

- ▶ Given the KDE $\hat{\rho}_h$ and for $\mathcal{X}_n = \{X_1, \dots, X_n\}$, we consider the Vietoris-Rips complex generated from the level set of the $\hat{\rho}_h$ as

$$\left\{ \text{Rips} \left(\mathcal{X}_{n,L}^{\hat{\rho}_h}, r \right) \right\}_{L>0}, \text{ where } \mathcal{X}_{n,L}^{\hat{\rho}_h} = \{X_i \in \mathcal{X}_n : \hat{\rho}_h(X_i) \geq L\}.$$

We estimate the target level set by considering the Vietoris-Rips complex generated from the level set of the KDE.

- ▶ For $\mathcal{X}_n = \{X_1, \dots, X_n\}$, we estimate the target level set by the level sets of the KDE \hat{p}_h on Vietoris-Rips complexes,

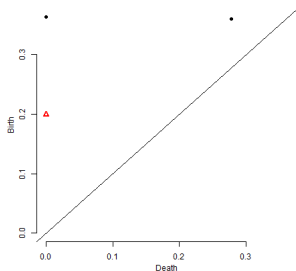
$$\left\{ \text{Rips} \left(\mathcal{X}_{n,L}^{\hat{p}_h}, r \right) \right\}_{L>0}, \text{ where } \mathcal{X}_{n,L}^{\hat{p}_h} = \{X_i \in \mathcal{X}_n : \hat{p}_h(X_i) \geq L\}.$$

We estimate the target persistent homology by the persistent homology of the KDE filtration on Vietoris-Rips complexes.

- ▶ We estimate the target persistent homology by the persistent homology of the level sets of the KDE $\hat{\rho}_h$ on Vietoris-Rips complexes,

$$\left\{ \text{Rips} \left(\mathcal{X}_{n,L}^{\hat{\rho}_h}, r \right) \right\}_{L>0}, \text{ where } \mathcal{X}_{n,L}^{\hat{\rho}_h} = \{X_i \in \mathcal{X}_n : \hat{\rho}_h(X_i) \geq L\}.$$

and denote the persistent homology as $PH_*^R(\hat{\rho}_h, r)$.



We estimate the target level set by Vietoris-Rips complexes from the KDE level sets.

- ▶ We approximate the target level set

$$\{D_L\}_{L>0}, \text{ where } D_L := \{x \in \mathbb{X} : p_h(x) \geq L\},$$

by the level sets of the KDE on Vietoris-Rips complexes,

$$\left\{ \text{Rips} \left(\mathcal{X}_{n,L}^{\hat{p}_h}, r \right) \right\}_{L>0}, \text{ where } \mathcal{X}_{n,L}^{\hat{p}_h} = \{X_i \in \mathcal{X}_n : \hat{p}_h(X_i) \geq L\}.$$

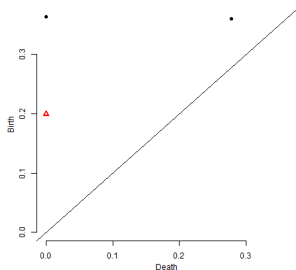
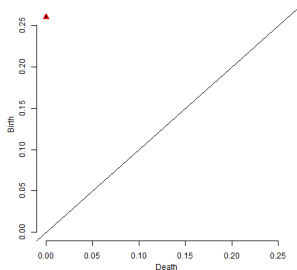
We estimate the target persistent homology by the persistent homology of the KDE filtration on Vietoris-Rips complexes.

- ▶ We estimate the target persistent homology

$$PH_*^{\text{supp}(P)}(\rho_h),$$

by the persistent homology of the KDE filtration on Vietoris-Rips complexes,

$$PH_*^R(\hat{\rho}_h, r).$$



The persistent homology of the KDE filtration on Vietoris-Rips complexes is consistent.

Theorem

(Theorem 14, Corollary 15) Let $\{r_n\}_{n \in \mathbb{N}}$ and $\{h_n\}_{n \in \mathbb{N}}$ be satisfying $r_n = \Omega\left(\left(\frac{\log n}{n}\right)^{1/d}\right)$, $r_n = o(1)$, and $\frac{\log(1/h_n)}{nh_n^d} = O(1)$. Then

$$d_B\left(PH_*^R(\hat{p}_{h_n}, r_n), PH_*^{\text{supp}(P)}(p_{h_n})\right) = O_P\left(\sqrt{\frac{\log(1/h_n)}{nh_n^d}} + \|r_n\|_\infty\right).$$

Confidence band

- ▶ An asymptotic $1 - \alpha$ confidence band c_n is a random variable satisfying

$$\mathbb{P}(d_B \left(PH_*^{\text{supp}(P)}(p_{h_n}), PH_*^R(\hat{p}_{h_n}, r_n) \right) \leq c_n) \geq 1 - \alpha + o(1).$$

Confidence band for the persistent homology of the KDE filtration.

Theorem

(Theorem 18)

$$\mathbb{P} \left(d_B \left(PH_*^{supp(P)}(p_{h_n}), PH_*^R(\hat{p}_{h_n}, r_n) \right) \leq \frac{\hat{z}_\alpha}{\sqrt{nh^d}} + \hat{c} \right) \geq 1 - \alpha + o(1),$$

where \hat{z}_α is from the bootstrap algorithm and \hat{c} is computed from data points.

Thank you!

Reference

- O. Bobrowski, S. Mukherjee, and J. E. Taylor. Topological consistency via kernel estimation. *ArXiv e-prints*, July 2014.
- Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Scalar field analysis over point cloud data. *Discrete & Computational Geometry*, 46(4):743–775, 2011.
- 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, 12 2014. doi: 10.1214/14-AOS1252. URL <http://dx.doi.org/10.1214/14-AOS1252>.

Confidence set for the persistent homology of the KDE filtration.

- ▶ Since covering condition is not verifiable, we instead consider the threshold where the covering is satisfied, i.e. consider $\epsilon > 0$ satisfying

$$\{x : \hat{\rho}_h(x) \geq \epsilon\} \subset \bigcup_i \mathbb{B}_{\mathbb{R}^d}(X_i, r).$$

- ▶ Set

$$\mathcal{E}_r = \left\{ \epsilon \in \mathbb{R}_+ : \{x : \hat{\rho}_h(x) \geq \epsilon\} \subset \bigcup_i \mathbb{B}_{\mathbb{R}^d}(X_i, r) \right\},$$

and let

$$\hat{c}_r = \inf\{\epsilon \in \mathcal{E}_r\} \vee \max_i \sup_{x \in \mathbb{B}_{\mathbb{R}^d}(X_i, r)} |\hat{\rho}_h(X_i) - \hat{\rho}_h(x)|.$$

Confidence set for the persistent homology of the KDE filtration.

- ▶ We let the confidence set as the ball centered at $PH_*^R(\hat{p}_{h_n}, r_n)$ and radius $\frac{\hat{z}_\alpha}{\sqrt{nh^d}} + \hat{c}_r \vee \hat{c}_{2r}$, i.e.

$$\hat{C}_\alpha = \left\{ \mathcal{P} \text{ : } d_B(\mathcal{P}, PH_*^R(\hat{p}_{h_n}, r_n)) \leq \frac{\hat{z}_\alpha}{\sqrt{nh^d}} + \hat{c}_r \vee \hat{c}_{2r} \right\}.$$

This is a valid confidence set by the following theorem.

Theorem

(Theorem 18)

$$\mathbb{P} \left(d_B \left(PH_*^{\text{supp}(P)}(p_{h_n}), PH_*^R(\hat{p}_{h_n}, r_n) \right) \leq \frac{\hat{z}_\alpha}{\sqrt{nh^d}} + \hat{c}_r \vee \hat{c}_{2r} \right) \geq 1 - \alpha + o(1).$$