

# Statistical Inference for Geometric and Topological Data

Jisu KIM

INRIA

2020-09-30

## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

- Minimax Rates for Estimating the Dimension of a Manifold
- The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

### Statistical Inference For Homological Features

- Statistical Inference for Cluster Trees

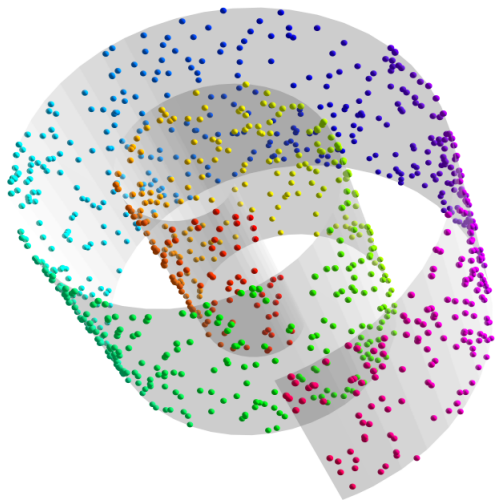
### Statistical Inference for Persistent Homology

- Confidence sets for persistence diagrams
- Persistent homology of KDE filtration on Vietoris-Rips complex

### Computation for Topological Data Analysis

- R Package TDA: Statistical Tools for Topological Data Analysis

The curse of dimensionality from the high dimensional data is mitigated when there is a low dimensional geometric and topological structure.

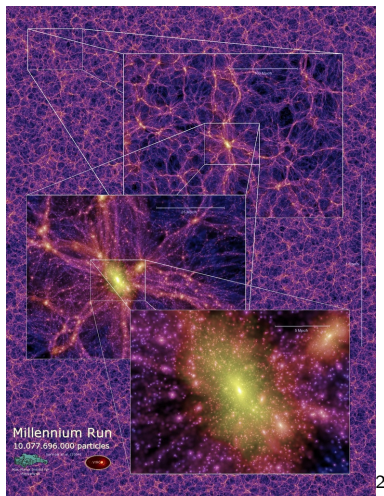


1

---

<sup>1</sup><http://www.skybluetrades.net/blog/posts/2011/10/30/machine-learning/>

Geometric and topological structures in the data provide information.



---

<sup>2</sup>[http://www.mpa-garching.mpg.de/galform/virgo/millennium/poster\\_half.jpg](http://www.mpa-garching.mpg.de/galform/virgo/millennium/poster_half.jpg)

# Statistic Inference for Geometric and Topological Data is explored.

- ▶ Minimax Rates for Geometric Parameters of a Manifold
  - ▶ Minimax Rates for Estimating the Dimension of a Manifold (Kim, Rinaldo, Wasserman, 2019)
  - ▶ The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory (Aamari, Kim, Chazal, Michel, Rinaldo, Wasserman, 2019)
- ▶ 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, 2014b)
  - ▶ Statistical inference on persistent homology of KDE filtration on Vietoris-Rips complex (Shin, Kim, Rinaldo, Wasserman, 2020?)
- ▶ Computation for Topological Data Analysis
  - ▶ R Package TDA: Statistical Tools for Topological Data Analysis (Fasy, Kim, Lecci, Maria, Milman, Rouvreau, 2014a)

## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

Minimax Rates for Estimating the Dimension of a Manifold

The Origin of the Reach: Better Understanding Regularity Through  
Minimax Estimation Theory

### Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

### Statistical Inference for Persistent Homology

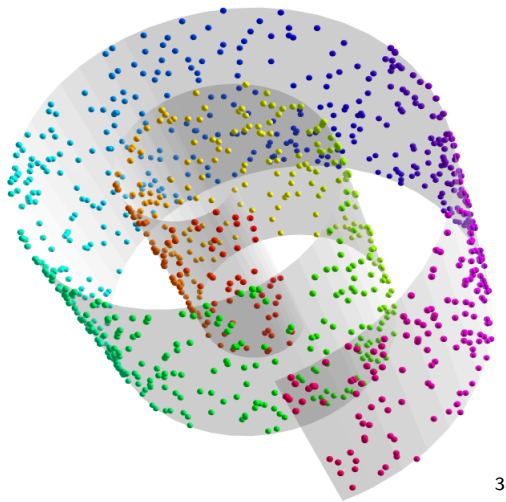
Confidence sets for persistence diagrams

Persistent homology of KDE filtration on Vietoris-Rips complex

### Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

A manifold is a low dimensional geometric structure that locally resembles Euclidean space.



---

<sup>3</sup><http://www.skybluetrades.net/blog/posts/2011/10/30/machine-learning/>

The maximum risk of an estimator is its worst expected error.

- ▶ the maximum risk of an estimator  $\hat{\theta}_n$  is the worst expected error that the estimator  $\hat{\theta}_n$  can make.



$$\sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ \ell \left( \hat{\theta}_n(X), \theta(P) \right) \right]$$

- ▶  $X = (X_1, \dots, X_n)$  is drawn from a fixed distribution  $P$ , where  $P$  is contained in set of distributions  $\mathcal{P}$ .
- ▶ estimator  $\hat{\theta}_n$  is any function of data  $X$ .
- ▶ The loss function  $\ell(\cdot, \cdot)$  measures the error of the estimator  $\hat{\theta}_n$ .



The minimax rate describes the statistical difficulty of estimating a parameter.

- ▶ The minimax rate  $R_n$  is the risk of an estimator that performs best in the worst case, as a function of sample size.



$$R_n = \inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ \ell \left( \hat{\theta}_n(X), \theta(P) \right) \right]$$

- ▶  $X = (X_1, \dots, X_n)$  is drawn from a fixed distribution  $P$ , where  $P$  is contained in set of distributions  $\mathcal{P}$ .
- ▶ estimator  $\hat{\theta}_n$  is any function of data  $X$ .
- ▶ The loss function  $\ell(\cdot, \cdot)$  measures the error of the estimator  $\hat{\theta}_n$ .

We measure the statistical difficulty of estimating geometric parameters of a manifold by their minimax rate.

- ▶ Minimax Rates for Estimating the Dimension of a Manifold (Kim, Rinaldo, Wasserman, 2019)
- ▶ The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory (Aamari, Kim, Chazal, Michel, Rinaldo, Wasserman, 2019)

## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

Minimax Rates for Estimating the Dimension of a Manifold

The Origin of the Reach: Better Understanding Regularity Through  
Minimax Estimation Theory

### Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

### Statistical Inference for Persistent Homology

Confidence sets for persistence diagrams

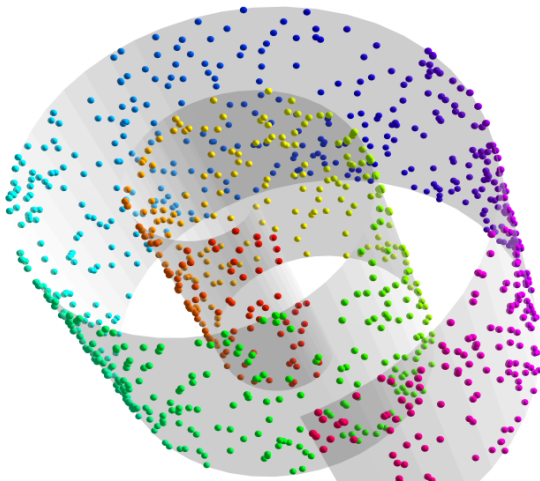
Persistent homology of KDE filtration on Vietoris-Rips complex

### Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

The intrinsic dimension of a manifold needs to be estimated a priori to the manifold learning.

- ▶ Most manifold learning algorithms require the intrinsic dimension of the manifold as input.
- ▶ The intrinsic dimension is rarely known in advance and therefore has to be estimated.



# Minimax rate for estimating the dimension



$$R_n = \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ 1 \left( \hat{\dim}_n(X) \neq \dim(P) \right) \right]$$

- ▶  $X = (X_1, \dots, X_n)$  is drawn from a fixed distribution  $P$ , where  $P$  is contained in set of distributions  $\mathcal{P}$ .
- ▶ estimator  $\hat{\dim}_n$  is any function of data  $X$ .
- ▶ 0 – 1 loss function is considered, so for all  $x, y \in \mathbb{R}$ ,  $\ell(x, y) = 1(x \neq y)$ .

Minimax rate for estimating the dimension: we first consider dimension  $d_1$  vs  $d_2$ .



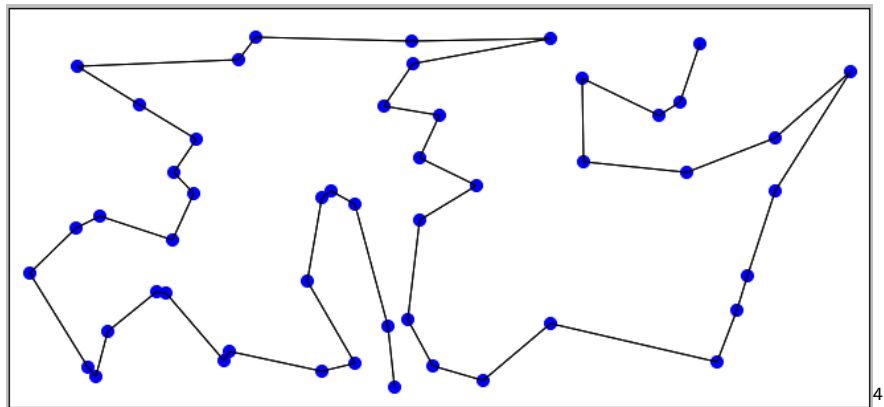
$$R_n = \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ 1 \left( \hat{\dim}_n(X) \neq \dim(P) \right) \right]$$

- ▶  $X = (X_1, \dots, X_n)$  is drawn from a fixed distribution  $P$ , where  $P$  is contained in set of distributions  $\mathcal{P} = \mathcal{P}^{d_1} \cup \mathcal{P}^{d_2}$ , where  $\mathcal{P}^d$  is a set of  $d$ -dimensional distributions..
- ▶ estimator  $\hat{\dim}_n$  is any function of data  $X$ .
- ▶ 0 – 1 loss function is considered, so for all  $x, y \in \mathbb{R}$ ,  
 $\ell(x, y) = 1(x \neq y)$ .

The Maximum Risk of any chosen Estimator Provides an Upper Bound on the Minimax Rate.

$$\begin{aligned} R_n &= \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ 1 \left( \hat{\dim}_n(X) \neq \dim(P) \right) \right] \\ &\leq \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ 1 \left( \hat{\dim}_n(X) \neq \dim(P) \right) \right]}_{\text{the maximum risk of any chosen estimator}} \end{aligned}$$

TSP(Travelling Salesman Problem) Path Finds Shortest Path that Visits Each Points exactly Once.



<sup>4</sup><http://www.heatonresearch.com/fun/tsp/anneal>



Our Estimator estimates Dimension to be  $d_2$  if  $d_1$ -squared Length of TSP Generated by the Data is Long.

- ▶ When intrinsic dimension is higher, length of TSP path is likely to be longer.
- ▶

$$\hat{\dim}_n(X) = d_1 \iff \min_{\sigma \in S_n} \sum_{i=1}^{n-1} \|X_{\sigma(i+1)} - X_{\sigma(i)}\|_{\mathbb{R}^m}^{d_1} \leq C,$$

where  $C$  is some constant.

Mimimax rate is upper bounded by  $O\left(n^{-\left(\frac{d_2}{d_1}-1\right)n}\right)$ .

### Proposition

(Proposition 9) Let  $1 \leq d_1 < d_2 \leq m$ . Then

$$\inf_{\hat{\dim}_n P \in \mathcal{P}^{d_1} \cup \mathcal{P}^{d_2}} \sup_{P \in \mathcal{P}^{(n)}} \mathbb{E}_{P^{(n)}} \left[ \mathbf{1} \left( \hat{\dim}_n(X) \neq \dim(P) \right) \right] \lesssim n^{-\left(\frac{d_2}{d_1}-1\right)n}.$$

Le Cam's Lemma provides lower bounds for estimating the dimension.

### Lemma

(Lemma 10, Le Cam's Lemma) Let  $\mathcal{P}$  be a set of probability measures, and  $\mathcal{P}^{d_1}, \mathcal{P}^{d_2} \subset \mathcal{P}$  be such that for all  $P \in \mathcal{P}^{d_i}$ ,  $\theta(P) = \theta_i$  for  $i = 1, 2$ . For any  $Q_i \in \text{co}(\mathcal{P}_i)$ , let  $q_i$  be density of  $Q_i$  with respect to measure  $\nu$ . Then

$$\begin{aligned} & \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ 1 \left( \hat{\dim}_n(X) \neq \dim(P) \right) \right] \\ & \geq \frac{1(\theta_1 \neq \theta_2)}{4} \sup_{Q_i \in \text{co}(\mathcal{P}^{d_i})} \int [q_1(x) \wedge q_2(x)] d\nu(x). \end{aligned}$$

Mimimax rate is lower bounded by  $\Omega \left( n^{-2(d_2-d_1)n} \right)$ .

- ▶ Lower bound below is now combination of Le Cam's lemma, constructions of  $T$ ,  $\mathcal{P}_1^{d_1}$ ,  $\mathcal{P}_2^{d_2}$ , and claim.

## Proposition

*(Proposition 14)*

$$\inf_{\hat{\dim} P \in \mathcal{P}^{d_1} \cup \mathcal{P}^{d_2}} \sup_{P^{(n)}} \mathbb{E}_{P^{(n)}} \left[ \mathbb{1} \left( \hat{\dim}_n(X) \neq \dim(P) \right) \right] \gtrsim n^{-2(d_2-d_1)n}.$$

# Minimax rate for estimating the dimension

## Theorem

(Proposition 16 and 17)

$$n^{-2n} \lesssim \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ \mathbf{1} \left( \hat{\dim}_n(X) \neq \dim(P) \right) \right] \lesssim n^{-\frac{1}{m-1}n}.$$

## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

Minimax Rates for Estimating the Dimension of a Manifold

The Origin of the Reach: Better Understanding Regularity Through  
Minimax Estimation Theory

### Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

### Statistical Inference for Persistent Homology

Confidence sets for persistence diagrams

Persistent homology of KDE filtration on Vietoris-Rips complex

### Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

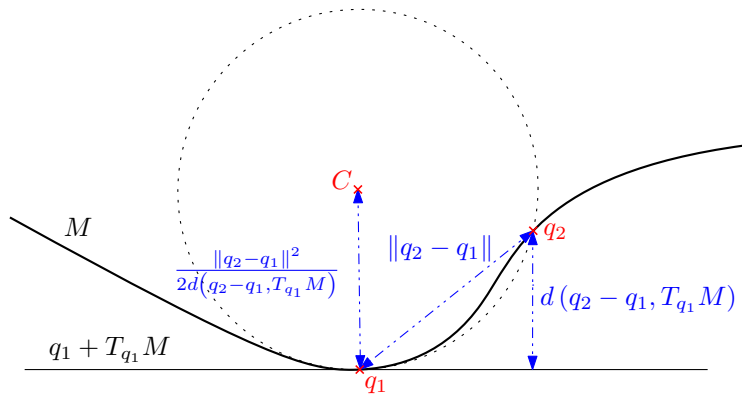
The reach is the maximum radius of a ball that can roll over the manifold.

### Definition

When  $M \subset \mathbb{R}^m$  is a manifold, the reach of  $M$ , denoted by  $\tau(M)$ , can be defined as

$$\tau(M) = \inf_{q_2 \neq q_1 \in M} \frac{\|q_2 - q_1\|_2^2}{2d(q_2 - q_1, T_{q_1} M)},$$

where  $T_a M$  is the tangent space of  $M$  at  $a$ .



The reach is a regularity parameter in many geometrical inference problem.

- ▶ The reach is a key parameter in:
  - ▶ Dimension estimation
  - ▶ Homology inference
  - ▶ Volume estimation
  - ▶ Manifold clustering
  - ▶ Diffusion maps



# Minimax rate for estimating the reach



$$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ \left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}_n(X)} \right|^q \right]$$

- ▶  $X = (X_1, \dots, X_n)$  is drawn from a fixed distribution  $P$ , where  $P$  is contained in set of distributions  $\mathcal{P}$ .
- ▶ estimator  $\hat{\tau}_n$  is any function of data  $X$ .
- ▶ inverse  $l_q$  loss function is considered, so for all  $x, y \in \mathbb{R}$ ,  
 $\ell(x, y) = \left| \frac{1}{x} - \frac{1}{y} \right|^q$ .

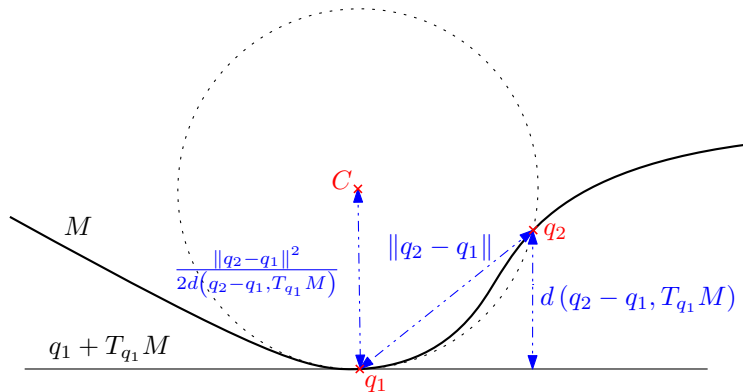
The maximum risk of our estimator provides an upper bound on the minimax rate.

$$\begin{aligned} R_n &= \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[ \left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}_n(X)} \right|^q \right] \\ &\leq \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[ \left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}(X)} \right|^q \right]}_{\text{the maximum risk of our estimator}} \end{aligned}$$

We define the reach estimator  $\hat{\tau}_n$  as the maximum radius of a ball that you can roll over the point cloud.

- Given observation  $X = (X_1, \dots, X_n)$ , then the reach estimator  $\hat{\tau}_n$  is a plugin estimator as

$$\hat{\tau}_n(X) = \inf_{1 \leq i \neq j \leq n} \frac{\|X_j - X_i\|_2^2}{2d(X_j - X_i, T_{X_i}M)}$$



Minimax rate is upper bounded by  $O\left(n^{-\frac{2p}{3d-1}}\right)$ .

**Theorem**  
(Theorem 5.1)

$$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[ \left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}_n(X)} \right|^q \right] \lesssim n^{-\frac{2q}{3d-1}}.$$

Le Cam's lemma provides a lower bound based on the reach difference and the statistical difference of two distributions.

- ▶ Total variance distance between two distributions is defined as

$$TV(P, P') = \sup_{A \in \mathcal{B}(\mathbb{R}^D)} |P(A) - P'(A)|.$$

### Lemma

(Lemma 5.2) Let  $P, P' \in \mathcal{P}$  with respective supports  $M$  and  $M'$ . Then

$$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[ \left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}_n(X)} \right|^q \right] \gtrsim \left| \frac{1}{\tau(M)} - \frac{1}{\tau(M')} \right|^q (1 - TV(P, P'))^{2n}.$$

Mimimax rate is lower bounded by  $\Omega\left(n^{-\frac{p}{d}}\right)$ .

## Proposition

*(Proposition 5.6)*

$$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[ \left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}_n(X)} \right|^q \right] \gtrsim n^{-\frac{q}{d}}.$$

# Minimax rate for estimating the reach

## Theorem

(Theorem 5.1 and Proposition 5.6)

$$n^{-\frac{q}{d}} \lesssim \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ \left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}_n(X)} \right|^q \right] \lesssim n^{-\frac{2q}{3d-1}}.$$

## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

Minimax Rates for Estimating the Dimension of a Manifold

The Origin of the Reach: Better Understanding Regularity Through  
Minimax Estimation Theory

### Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

### Statistical Inference for Persistent Homology

Confidence sets for persistence diagrams

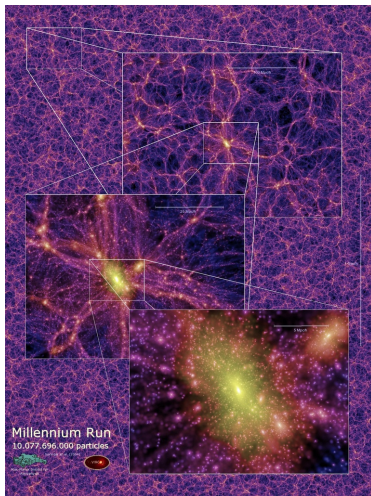
Persistent homology of KDE filtration on Vietoris-Rips complex

### Computation for Topological Data Analysis




R Package TDA: Statistical Tools for Topological Data Analysis



Topological holes in the data provide information.



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.  $\beta_0$  = # of connected components 
  2.  $\beta_1$  = # of loops (holes inside 1-dim sphere) 
  3.  $\beta_2$  = # 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		

# Statistical inference for homological features.

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

## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

Minimax Rates for Estimating the Dimension of a Manifold

The Origin of the Reach: Better Understanding Regularity Through  
Minimax Estimation Theory

### Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

### Statistical Inference for Persistent Homology

Confidence sets for persistence diagrams

Persistent homology of KDE filtration on Vietoris-Rips complex

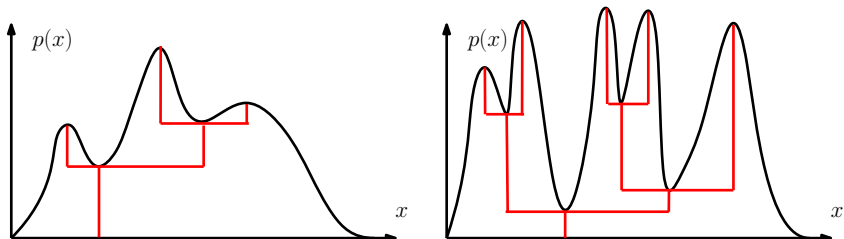
### Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

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

### Definition

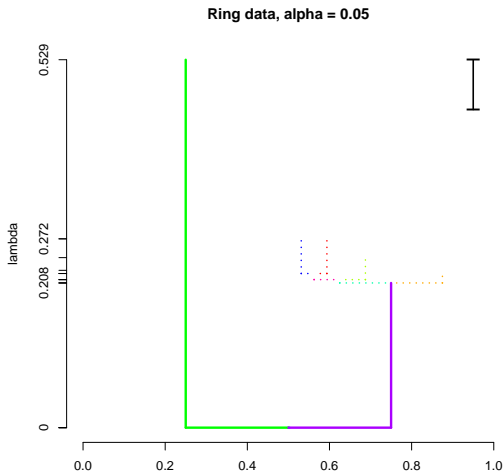
For a density function  $p$ , its cluster tree  $T_p$  is a function where  $T_p(\lambda)$  is the set of connected components of the upper level set  $\{x : p(x) \geq \lambda\}$ .



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  $T_{\hat{\rho}_h}$  be the cluster tree from the kernel density estimator  $\hat{\rho}_h$ , where

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

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

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

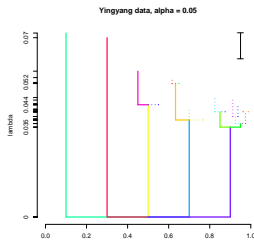
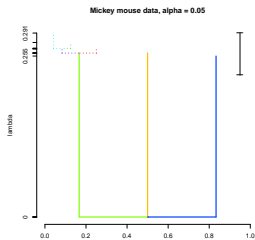
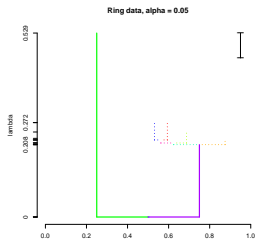
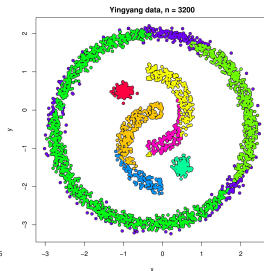
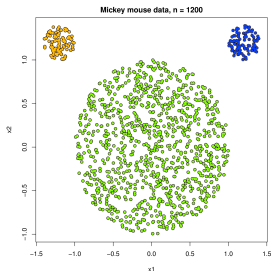
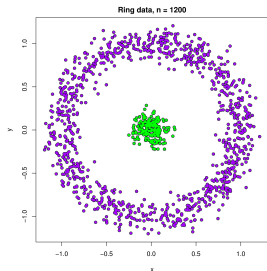
## Theorem

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

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



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



## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

Minimax Rates for Estimating the Dimension of a Manifold  
The Origin of the Reach: Better Understanding Regularity Through  
Minimax Estimation Theory

### Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

### Statistical Inference for Persistent Homology

Confidence sets for persistence diagrams  
Persistent homology of KDE filtration on Vietoris-Rips complex

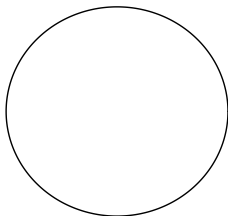
### Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

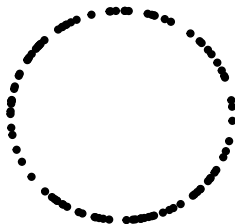
Homology of finite sample is different from homology of underlying manifold, hence it cannot be directly used for the inference.

- ▶ 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:

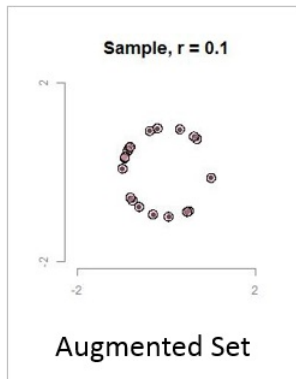
Underlying circle:  $\beta_0 = 1, \beta_1 = 1$



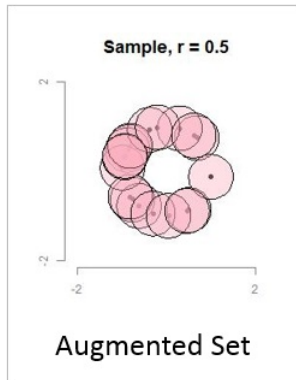
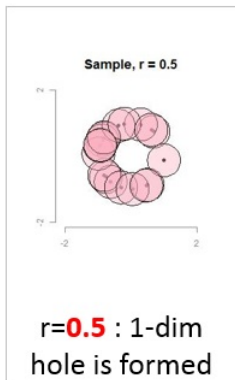
100 samples:  $\beta_0 = 100, \beta_1 = 0$



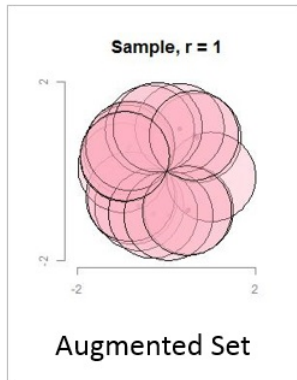
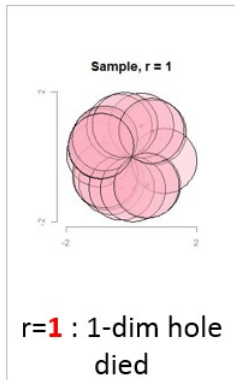
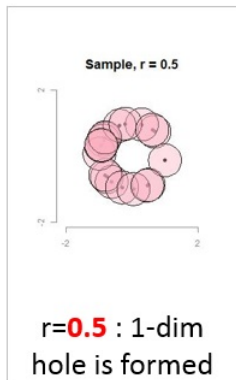
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



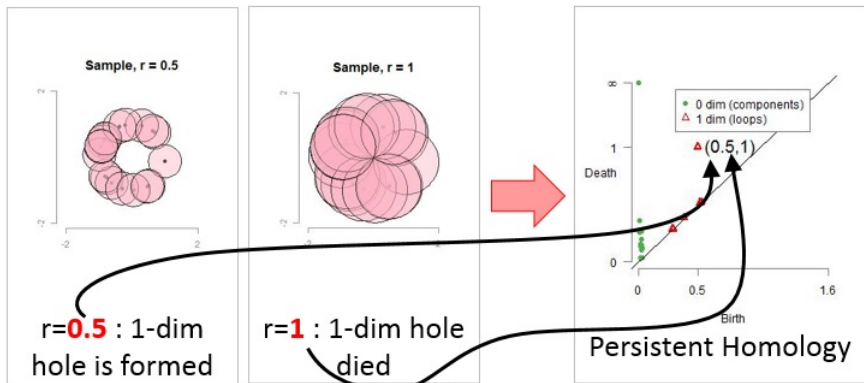
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



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.



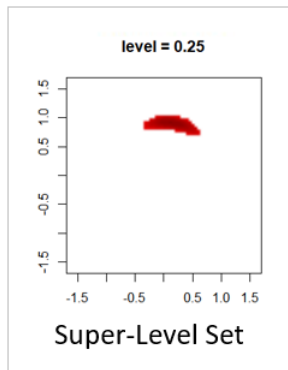
We rely on the kernel density estimator to extract topological information of the underlying distribution.

- ▶ The kernel density estimator is

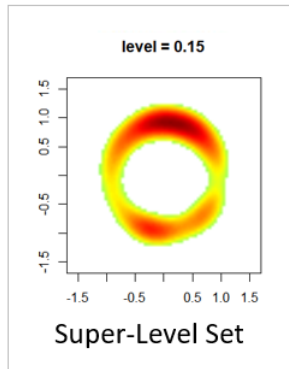
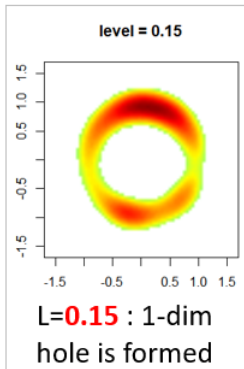
$$\hat{p}_h(x) = \frac{1}{nh^m} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$



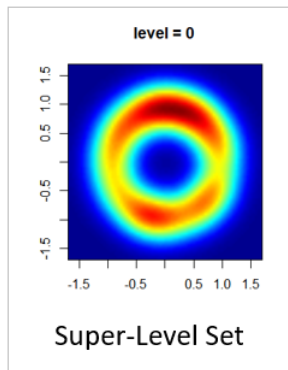
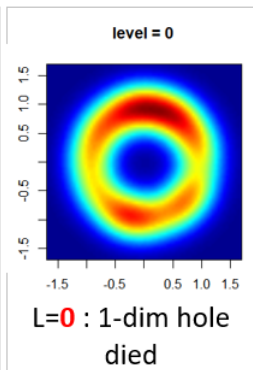
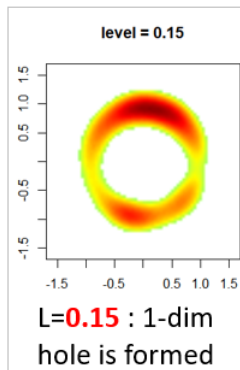
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



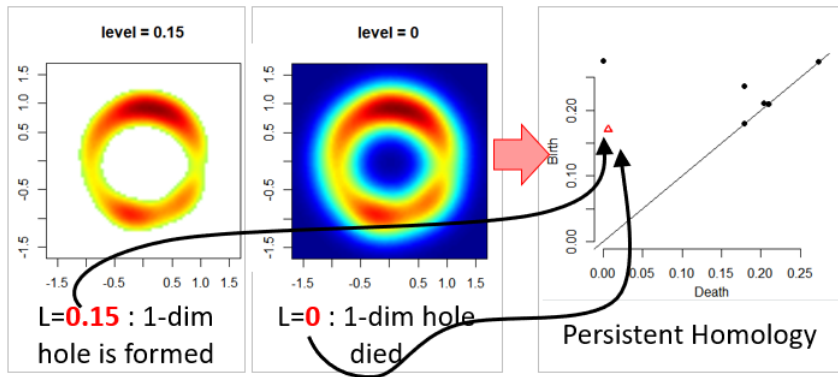
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



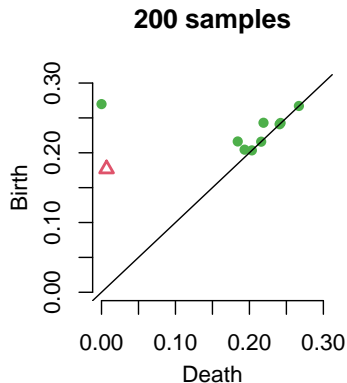
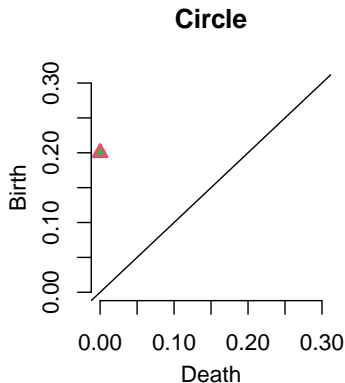
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



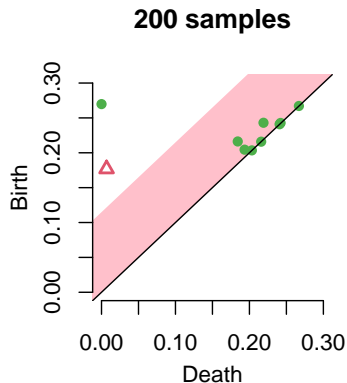
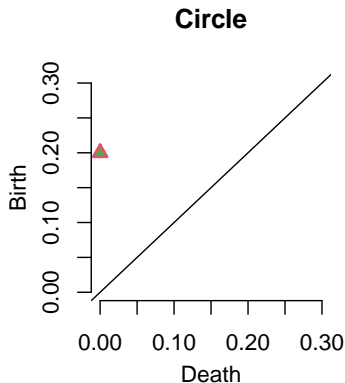
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



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.



# Statistical inference for persistent homology.

- ▶ Confidence sets for persistence diagrams (Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2014b)
- ▶ Persistent homology of KDE filtration on Vietoris-Rips complex (Shin, Kim, Rinaldo, Wasserman, 2020?)

## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

Minimax Rates for Estimating the Dimension of a Manifold

The Origin of the Reach: Better Understanding Regularity Through  
Minimax Estimation Theory

### Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

### Statistical Inference for Persistent Homology

**Confidence sets for persistence diagrams**

Persistent homology of KDE filtration on Vietoris-Rips complex

### Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis



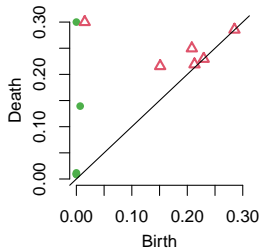
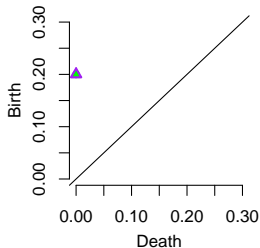
Bottleneck distance gives a metric on the space of persistent homology.

### 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  $\gamma$  ranges over all bijections from  $D_1$  to  $D_2$ .



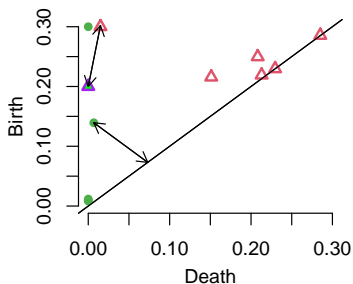
Bottleneck distance gives a metric on the space of persistent homology.

### Definition

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

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

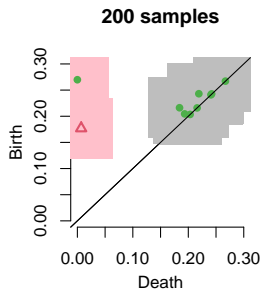
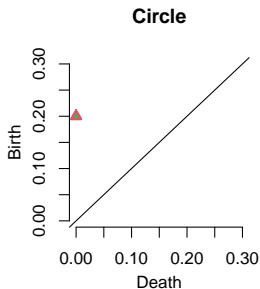
where  $\gamma$  ranges over all bijections from  $D_1$  to  $D_2$ .



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

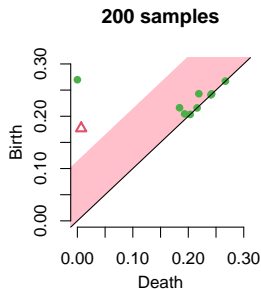
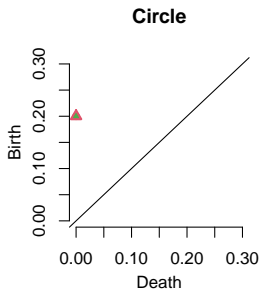
$$\mathbb{P}(Dgm(f_M) \in \{\mathcal{P} : d_B(\mathcal{P}, Dgm(f_X)) \leq c_n\}) \geq 1 - \alpha.$$



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  $\theta^* = \sqrt{nh^m} \|\hat{p}_h^*(x) - \hat{p}_h(x)\|_\infty$ , 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  $\left[ \hat{p}_h - \frac{\hat{z}_\alpha}{\sqrt{nh^m}}, \hat{p}_h + \frac{\hat{z}_\alpha}{\sqrt{nh^m}} \right]$ .

## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

Minimax Rates for Estimating the Dimension of a Manifold  
The Origin of the Reach: Better Understanding Regularity Through  
Minimax Estimation Theory

### Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

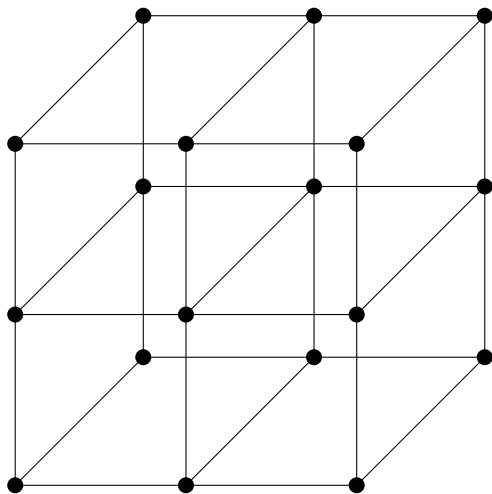
### Statistical Inference for Persistent Homology

Confidence sets for persistence diagrams  
Persistent homology of KDE filtration on Vietoris-Rips complex

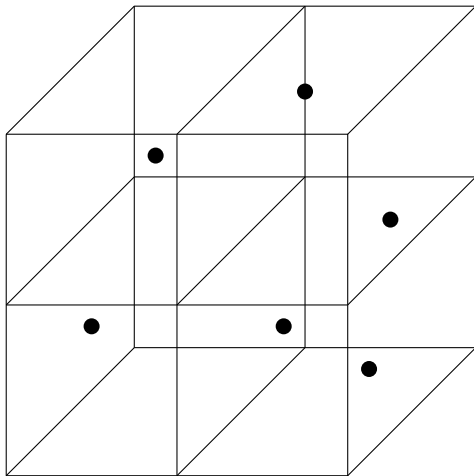
### Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

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



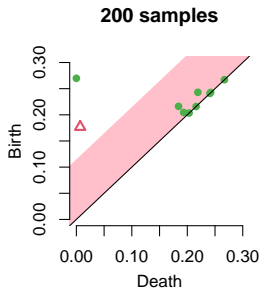
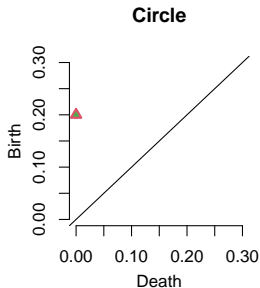
Computing the persistent homology of density function on data points reduces computational complexity.





# 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. [2014b], Bobrowski et al. [2014], Chazal et al. [2011].

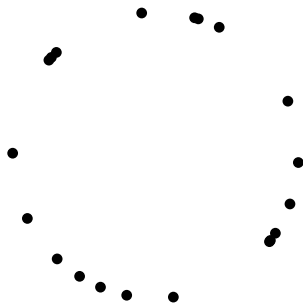


We use the Vietoris-Rips complex to estimate the target persistent homology.

- ▶ For  $\mathcal{X} \subset \mathbb{R}^m$  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

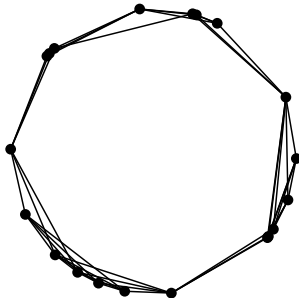


We use the Vietoris-Rips complex to estimate the target persistent homology.

- ▶ For  $\mathcal{X} \subset \mathbb{R}^m$  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

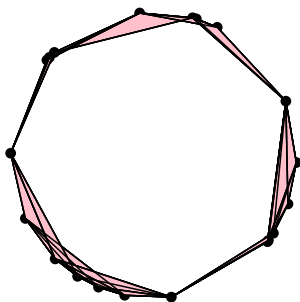


We use the Vietoris-Rips complex to estimate the target persistent homology.

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

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

- ▶ The KDE (kernel density estimator) is

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

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

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

We estimate the target 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,

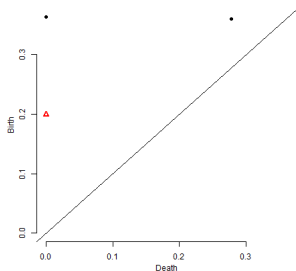
$$\left\{ \text{Rips} \left( \mathcal{X}_{n,L}^{\hat{p}_h}, r \right) \right\}_{L>0}, \text{ where } \mathcal{X}_{n,L}^{\hat{p}_h} = \{X_i \in \mathcal{X}_n : \hat{p}_h(X_i) \geq L\}.$$

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



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

### Theorem

(Theorem 16, Corollary 17) Let  $\{r_n\}_{n \in \mathbb{N}}$  and  $\{h_n\}_{n \in \mathbb{N}}$  be satisfying  $r_n = \Omega\left(\left(\frac{\log n}{n}\right)^{1/m}\right)$ ,  $r_n = o(1)$ , and  $\frac{\log(1/h_n)}{nh_n^m} = O(1)$ . Then

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



# Confidence set

- ▶ An asymptotic  $1 - \alpha$  confidence set  $\hat{C}_\alpha$  is a random set of persistent homologies satisfying

$$\mathbb{P}(PH_*(p_{h_n}) \in \hat{C}_\alpha) \geq 1 - \alpha + o(1).$$

# 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  $\hat{b}_\alpha$ , i.e.

$$\hat{C}_\alpha = \left\{ \mathcal{P} \text{ : } d_B(\mathcal{P}, PH_*^R(\hat{p}_{h_n}, r_n)) \leq \hat{b}_\alpha \right\}.$$

This is a valid confidence set by the following theorem.

## Theorem

*(Theorem 20)*

$$\mathbb{P} \left( PH_*(p_{h_n}) \in \hat{C}_\alpha \right) \geq 1 - \alpha + o(1).$$

## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

Minimax Rates for Estimating the Dimension of a Manifold  
The Origin of the Reach: Better Understanding Regularity Through  
Minimax Estimation Theory

### Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

### Statistical Inference for Persistent Homology

Confidence sets for persistence diagrams  
Persistent homology of KDE filtration on Vietoris-Rips complex

### Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

## Introduction

### Minimax Rates for Geometric Parameters of a Manifold

Minimax Rates for Estimating the Dimension of a Manifold

The Origin of the Reach: Better Understanding Regularity Through  
Minimax Estimation Theory

### Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

### Statistical Inference for Persistent Homology

Confidence sets for persistence diagrams

Persistent homology of KDE filtration on Vietoris-Rips complex

### Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

# R Package TDA provides an R interface for C++ libraries for Topological Data Analysis.

- ▶ website:  
<https://cran.r-project.org/web/packages/TDA/index.html>
- ▶ Author: Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, Clément Maria, David Milman, and Vincent Rouvreau.
- ▶ R is a programming language for statistical computing and graphics.
- ▶ R has short development time, while C/C++ has short execution time.
- ▶ R package TDA provides an R interface for C++ library GUDHI/Dionysus/PHAT, which are for Topological Data Analysis.

Thank you!

## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

- Statistical Inference on Persistence Homology

## References

## Minimax Rates for Estimating the Dimension of a Manifold

### Regularity conditions

Upper Bound

Lower Bound

Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

Reach and its Geometry

Reach estimator and its analysis

Minimax Estimates

## Statistical Inference for Cluster Trees

Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology

Statistical Inference on Persistence Homology

## References

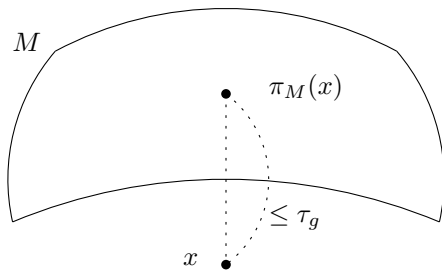


The supporting manifold  $M$  is assumed to be bounded.

$$M \subset I := [-K_I, K_I]^m \subset \mathbb{R}^m \text{ with } K_I \in (0, \infty)$$

The reach is assumed to be lower bounded to avoid an arbitrarily complicated manifold.

- ▶  $\mathcal{P}$  is a set of distributions  $P$  that is supported on a bounded manifold  $M$ , with its reach  $\tau(M) \geq \tau_g$ , and with other regularity assumptions.



The reach is assumed to be lower bounded to avoid an arbitrarily complicated manifold.

- ▶  $M$  is of local reach  $\geq \tau_\ell$ , if for all points  $p \in M$ , there exists a neighborhood  $U_p \subset M$  such that  $U_p$  is of reach  $\geq \tau_\ell$ .

Density is bounded away from  $\infty$  with respect to the uniform measure.

- ▶ Distribution  $P$  is absolutely continuous to induced Lebesgue measure  $\text{vol}_M$ , and  $\frac{dP}{d\text{vol}_M} \leq K_p$  for fixed  $K_p$ .
- ▶ This implies that the distribution on the manifold is of essential dimension  $d$ .
- ▶  $\mathcal{P}_{\kappa_l, \kappa_g, K_p}^d$  denotes set of distributions  $P$  that is supported on  $d$ -dimensional manifold of (global) reach  $\geq \tau_g$ , local reach  $\geq \tau_\ell$ , and density is bounded by  $K_p$ .

## Minimax Rates for Estimating the Dimension of a Manifold

Regularity conditions

Upper Bound

Lower Bound

Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

Reach and its Geometry

Reach estimator and its analysis

Minimax Estimates

## Statistical Inference for Cluster Trees

Confidence band for Persistent Homology of KDEs computed on  
Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology

Statistical Inference on Persistence Homology

## References

Our Estimator has Maximum Risk of  $O\left(n^{-\left(\frac{d_2}{d_1}-1\right)n}\right)$ .

- ▶ Our estimator makes error with probability at most  $O\left(n^{-\left(\frac{d_2}{d_1}-1\right)n}\right)$  if intrinsic dimension is  $d_2$ .
- ▶ Our estimator is always correct when the intrinsic dimension is  $d_1$ .

Our Estimator makes Error with Probability at most  $O\left(n^{-\left(\frac{d_2}{d_1}-1\right)n}\right)$  if Intrinsic Dimension is  $d_2$ .

- ▶ Based on the following lemma:

### Lemma

(Lemma 6) Let  $X_1, \dots, X_n \sim P \in \mathcal{P}_{\kappa_l, \kappa_g, K_p}^{d_2}$ , then

$$P^{(n)} \left[ \sum_{i=1}^{n-1} \|X_{i+1} - X_i\|^{d_1} \leq L \right] \lesssim n^{-\frac{d_2}{d_1} n}.$$

Our Estimator is always Correct when the Intrinsic Dimension is  $d_1$ .

► Based on following lemma:

### Lemma

(Lemma 7) Let  $M$  be a  $d_1$ -dimensional manifold with global reach  $\geq \tau_g$  and local reach  $\geq \tau_\ell$ , and  $X_1, \dots, X_n \in M$ . Then there exists  $C$  which depends only on  $m$ ,  $d_1$  and  $K_I$ , and there exists  $\sigma \in S_n$  such that

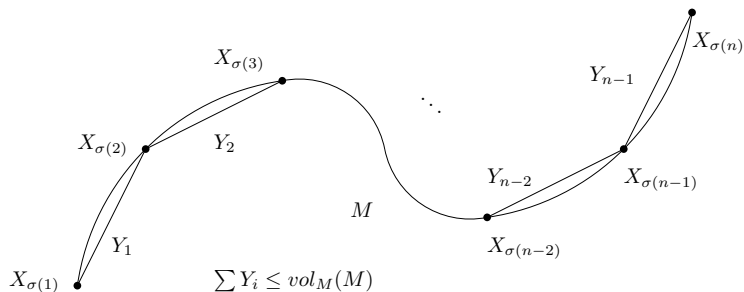
$$\sum_{i=1}^{n-1} \|X_{\sigma(i+1)} - X_{\sigma(i)}\|_{\mathbb{R}^m}^{d_1} \leq C.$$



Our estimator is always correct when the intrinsic dimension is  $d_1$ .

$$\sum_{i=1}^{n-1} \|X_{\sigma(i+1)} - X_{\sigma(i)}\|_{\mathbb{R}^m}^{d_1} \leq C.$$

- ▶ When  $d_1 = 1$  so that the manifold is a curve, length of TSP path is bounded by length of curve  $vol_M(M)$ .

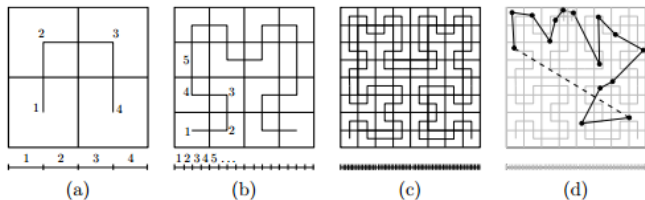


- ▶ Global reach  $\geq \tau_g$  implies  $vol_M(M)$  is bounded.

Our estimator is always correct when the intrinsic dimension is  $d_1$ .

$$\sum_{i=1}^{n-1} \|X_{\sigma(i+1)} - X_{\sigma(i)}\|_{\mathbb{R}^m}^{d_1} \leq C.$$

- When  $d_1 > 1$ , Several conditions implied by regularity conditions combined with Hölder continuity of  $d_1$ -dimensional space-filling curve is used.



Our estimator is always correct when the intrinsic dimension is  $d_1$ .

$$\sum_{i=1}^{n-1} \|X_{\sigma(i+1)} - X_{\sigma(i)}\|_{\mathbb{R}^m}^{d_1} \leq C.$$

- ▶ When  $d_1 > 1$ , Several conditions implied by regularity conditions combined with Hölder continuity of  $d_1$ -dimensional space-filling curve is used.

### Lemma

(Lemma 22, Space-filling curve) There exists surjective map  $\psi_d : \mathbb{R} \rightarrow \mathbb{R}^d$  which is Hölder continuous of order  $1/d$ , i.e.

$$0 \leq \forall s, t \leq 1, \|\psi_d(s) - \psi_d(t)\|_{\mathbb{R}^d} \leq 2\sqrt{d+3}|s - t|^{1/d}.$$

## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound**

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

- Statistical Inference on Persistence Homology

## References

A subset  $T \subset [-K_I, K_I]^n$  and set of distributions  $\mathcal{P}_1^{d_1}, \mathcal{P}_2^{d_2}$  are found so that, whenever  $X = (X_1, \dots, X_n) \in T$ , we cannot distinguish two models.

- ▶ The lower bound measures how hard it is to tell whether the data come from a  $d_1$  or  $d_2$ -dimensional manifold.
- ▶  $T, \mathcal{P}_1^{d_1}$  and  $\mathcal{P}_2^{d_2}$  are linked to the lower bound by using Le Cam's lemma.

Le Cam's Lemma provides lower bounds based on the minimum of two densities  $q_1 \wedge q_2$ , where  $q_1, q_2$  are in convex hull of  $\mathcal{P}_1^{d_1}$  and convex hull of  $\mathcal{P}_2^{d_2}$ , respectively.

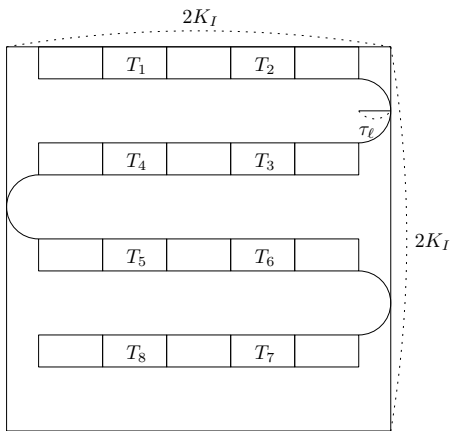
### Lemma

(Lemma 10, Le Cam's Lemma) Let  $\mathcal{P}$  be a set of probability measures, and  $\mathcal{P}^{d_1}, \mathcal{P}^{d_2} \subset \mathcal{P}$  be such that for all  $P \in \mathcal{P}^{d_i}$ ,  $\theta(P) = \theta_i$  for  $i = 1, 2$ . For any  $Q_i \in \text{co}(\mathcal{P}_i)$ , let  $q_i$  be density of  $Q_i$  with respect to measure  $\nu$ . Then

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \mathbf{1} \left( \hat{\text{dim}}_n(X) \neq \text{dim}(P) \right) \right] \geq \frac{\mathbf{1}(\theta_1 \neq \theta_2)}{4} \sup_{Q_i \in \text{co}(\mathcal{P}^{d_i})} \int [q_1(x) \wedge q_2(x)] d\nu(x)$$

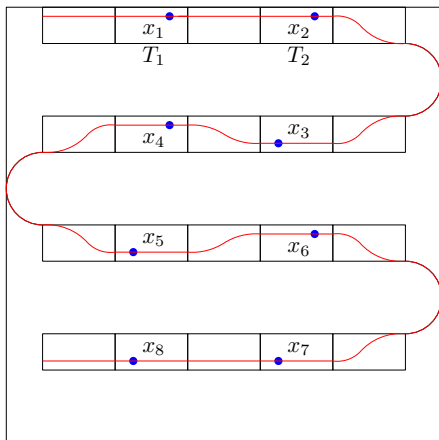
$T$  is constructed so that for any  $x = (x_1, \dots, x_n) \in T$ , there exists a  $d_1$ -dimensional manifold that satisfies regularity conditions and passes through  $x_1, \dots, x_n$ .

- ▶  $T_i$ 's are cylinder sets in  $[-K_I, K_I]^{d_2}$ , and then  $T$  is constructed as  $T = S_n \prod_{i=1}^n T_i$ , where the permutation group  $S_n$  acts on  $\prod_{i=1}^n T_i$  as a coordinate change.



$T$  is constructed so that for any  $x = (x_1, \dots, x_n) \in T$ , there exists a  $d_1$ -dimensional manifold that satisfies regularity conditions and passes through  $x_1, \dots, x_n$ .

- ▶ Given  $x_1, \dots, x_n \in T$  (blue points), manifold of global reach  $\geq \tau_g$  and local reach  $\geq \tau_\ell$  (red line) passes through  $x_1, \dots, x_n$ .





$\mathcal{P}_1^{d_1}$  is constructed as set of distributions that are supported on manifolds that passes through  $x_1, \dots, x_n$  for  $x = (x_1, \dots, x_n) \in T$ , and  $\mathcal{P}_2^{d_2}$  is a singleton set consisting of the uniform distribution on  $[-K_l, K_l]^{d_2}$ .

If  $X \in T$ , it is hard to determine whether  $X$  is sampled from distribution  $P$  in either  $\mathcal{P}_1^{d_1}$  or  $\mathcal{P}_2^{d_2}$ .

- ▶ There exists  $Q_1 \in \text{co}(\mathcal{P}_1^{d_1})$  and  $Q_2 \in \text{co}(\mathcal{P}_2^{d_2})$  such that  $q_1(x) \geq Cq_2(x)$  for every  $x \in T$  with  $C < 1$ .
- ▶ Then  $q_1(x) \wedge q_2(x) \geq Cq_2(x)$  if  $x \in T$ , so  $C \int_T q_2(x) dx$  can serve as lower bound of minimax rate.
- ▶ Based on following claim:

### Claim

(Claim 25) Let  $T = S_n \prod_{i=1}^n T_i$ . Then for all  $x \in \text{int } T$ , there exists  $C > 0$  that depends only on  $\kappa_I$ ,  $K_I$ , and  $r_x > 0$  such that for all  $r < r_x$ ,

$$Q_1(B(x_i, r)) \geq CQ_2(B(x_i, r)).$$

## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

- Statistical Inference on Persistence Homology

## References

## Multinary Classification and 0 – 1 Loss are Considered.



$$R_n = \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ \mathbf{1} \left( \hat{\dim}_n(X) \neq \dim(P) \right) \right]$$

- ▶ Now the manifolds are of any dimensions between 1 and  $m$ , so considered distribution set is  $\mathcal{P} = \bigcup_{d=1}^m \mathcal{P}^d$ .
- ▶ 0 – 1 loss function is considered, so for all  $x, y \in \mathbb{R}$ ,  $\ell(x, y) = I(x \neq y)$ .

Mimimax Rate is Upper Bounded by  $O\left(n^{-\frac{1}{m-1}n}\right)$ , and Lower Bounded by  $\Omega\left(n^{-2n}\right)$ .

## Proposition

*(Proposition 16 and 17)*

$$n^{-2n} \lesssim \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[ \mathbf{1} \left( \hat{\dim}_n \neq \dim(P) \right) \right] \lesssim n^{-\frac{1}{m-1}n}.$$

## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

- Statistical Inference on Persistence Homology

## References

## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

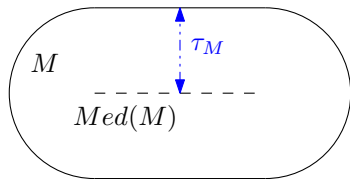
- Statistical Inference on Persistence Homology

## References

The medial axis of a set  $M$  is the set of points that have at least two nearest neighbors on the set  $M$ .



$$\text{Med}(M) = \{z \in \mathbb{R}^m : \text{there exists } p \neq q \in M \text{ with} \\ \|p - z\| = \|q - z\| = d(z, M)\}.$$

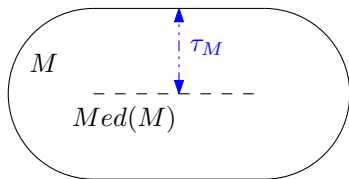




The reach of  $M$ , denoted by  $\tau_M$ , is the minimum distance from  $Med(M)$  to  $M$ .



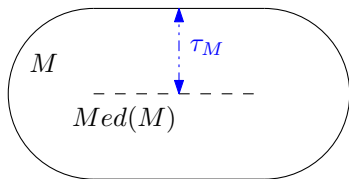
$$\tau_M = \inf_{x \in Med(M), y \in M} \|x - y\|.$$



The reach  $\tau_M$  gives the maximum offset size of  $M$  on which the projection is well defined.



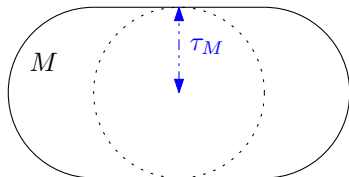
$$\tau_M = \inf_{x \in \text{Med}(M), y \in M} \|x - y\|.$$



The reach  $\tau_M$  gives the maximum radius of a ball that you can roll over  $M$ .

- ▶ When  $M \subset \mathbb{R}^m$  is a manifold,

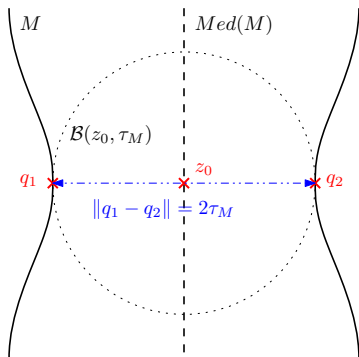
$$\tau_M = \inf_{q_2 \neq q_1 \in M} \frac{\|q_2 - q_1\|^2}{2d(q_2 - q_1, T_{q_1}M)}.$$



The bottleneck is a geometric structure where the manifold is nearly self-intersecting.

### Definition

(Definition 3.1) A pair of points  $(q_1, q_2)$  in  $M$  is said to be a bottleneck of  $M$  if there exists  $z_0 \in \text{Med}(M)$  such that  $q_1, q_2 \in \mathcal{B}(z_0, \tau_M)$  and  $\|q_1 - q_2\| = 2\tau_M$ .

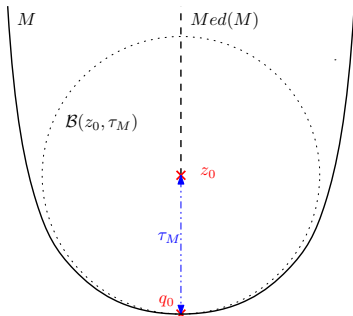
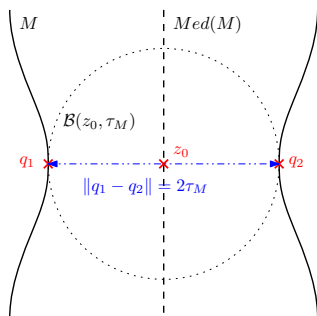


The reach is attained either from the bottleneck (global case) or the area of high curvature (local case).

## Theorem

(Theorem 3.4) At least one of the following two assertions holds:

- ▶ (Global Case)  $M$  has a bottleneck  $(q_1, q_2) \in M^2$ .
- ▶ (Local case) There exists  $q_0 \in M$  and an arc-length parametrized  $\gamma_0$  such that  $\gamma_0(0) = q_0$  and  $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$ .



## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

- Statistical Inference on Persistence Homology

## References

The statistical efficiency of the reach estimator  $\hat{\tau}$  is analyzed through its risk.

- ▶ The risk of the estimator  $\hat{\tau}$  is the expected loss the estimator.

$$\mathbb{E}_{P^{(n)}} [\ell(\hat{\tau}(\mathcal{X}), \tau_M)].$$

- ▶  $\mathcal{X} = \{X_1, \dots, X_n\}$  is drawn from a fixed distribution  $P$  with its support  $M$ .
- ▶ The loss function used is  $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^p$ ,  $p \geq 1$ .

The risk of the reach estimator  $\hat{\tau}$  is analyzed.

- ▶ The risk of the estimator  $\hat{\tau}$  is the expected loss the estimator

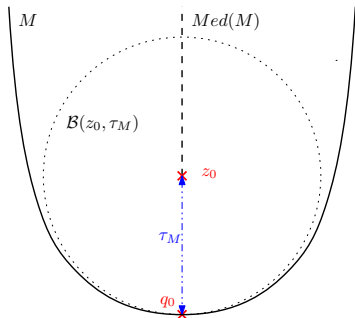
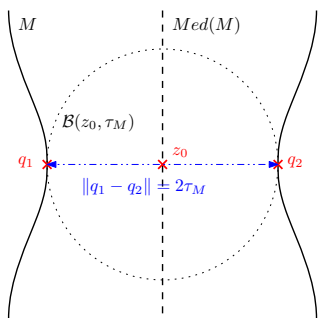
$$\mathbb{E}_{P^{(n)}} \left[ \left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathcal{X})} \right|^q \right].$$

- ▶  $\mathcal{X} = \{X_1, \dots, X_n\}$  is drawn from a fixed distribution  $P$  with its support  $M$ .
- ▶ The loss function used is  $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^q$ ,  $q \geq 1$ .



The reach estimator has the risk of  $O\left(n^{-\frac{2q}{3d-1}}\right)$ .

- ▶ The reach estimator has the risk of  $O\left(n^{-\frac{q}{d}}\right)$  for the global case.
- ▶ The reach estimator has the risk of  $O\left(n^{-\frac{2q}{3d-1}}\right)$  for the local case.

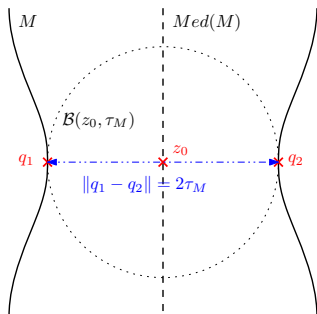


The reach estimator has the maximum risk of  $O\left(n^{-\frac{q}{d}}\right)$  for the global case.

### Proposition

(Proposition 4.3) Assume that the support  $M$  has a bottleneck. Then,

$$\mathbb{E}_{P^n} \left[ \left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathcal{X})} \right|^q \right] \lesssim n^{-\frac{q}{d}}.$$

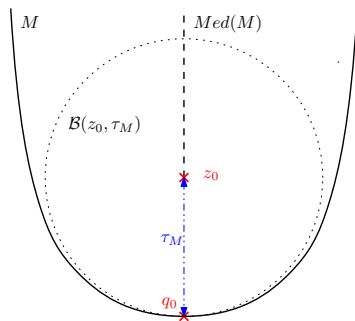


The reach estimator has the maximum risk of  $O\left(n^{-\frac{2q}{3d-1}}\right)$  for the local case.

### Proposition

(Proposition 4.7) Suppose there exists  $q_0 \in M$  and a geodesic  $\gamma_0$  with  $\gamma_0(0) = q_0$  and  $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$ . Then,

$$\mathbb{E}_{P^n} \left[ \left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathcal{X})} \right|^q \right] \lesssim n^{-\frac{2q}{3d-1}}.$$



## Minimax Rates for Estimating the Dimension of a Manifold

Regularity conditions

Upper Bound

Lower Bound

Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

Reach and its Geometry

Reach estimator and its analysis

**Minimax Estimates**

## Statistical Inference for Cluster Trees

Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology

Statistical Inference on Persistence Homology

## References

The statistical difficulty of the reach estimation problem is analyzed by the minimax rate.

- ▶ Minimax rate is the risk of an estimator that performs best in the worst case, as a function of sample size.



$$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} [\ell(\hat{\tau}_n(\mathcal{X}), \tau_M)].$$

- ▶  $\mathcal{X} = \{X_1, \dots, X_n\}$  is drawn from a fixed distribution  $P$  with its support  $M$ , where  $P$  is contained in set of distributions  $\mathcal{P}$ .
- ▶ An estimator  $\hat{\tau}_n$  is any function of data  $\mathcal{X}$ .
- ▶ The loss function used is  $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^q$ ,  $q \geq 1$ .

The statistical difficulty of the reach estimation problem is analyzed by the minimax rate.

- ▶ Minimax rate is the risk of an estimator that performs best in the worst case, as a function of sample size.



$$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[ \left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n(\mathcal{X})} \right|^q \right].$$

- ▶  $\mathcal{X} = \{X_1, \dots, X_n\}$  is drawn from a fixed distribution  $P$  with its support  $M$ , where  $P$  is contained in set of distributions  $\mathcal{P}$ .
- ▶ An estimator  $\hat{\tau}_n$  is any function of data  $\mathcal{X}$ .
- ▶ The loss function used is  $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^q$ ,  $q \geq 1$ .

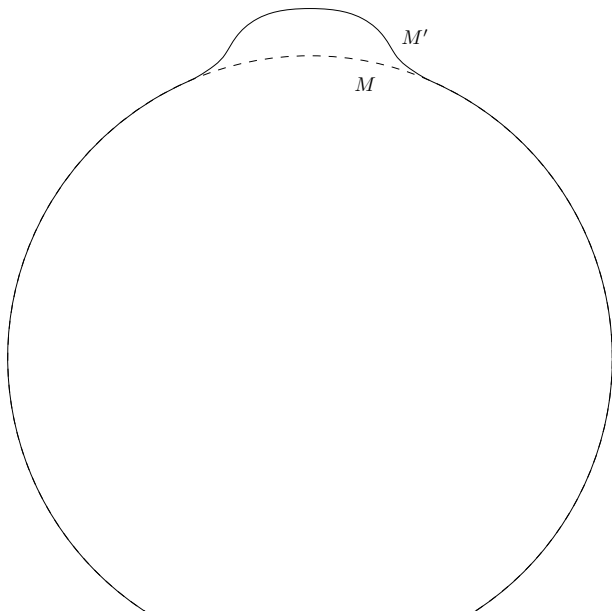
Two distributions  $P, P'$  are found so that their reaches differ but they are statistically difficult to distinguish.



$$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[ \left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n} \right|^q \right] \gtrsim \left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^q (1 - TV(P, P'))^{2n}.$$

- ▶ The lower bound measures how hard it is to tell whether the data is from distributions with different reaches.
- ▶  $P$  and  $P'$  are found so that  $\left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^q$  is large while  $(1 - TV(P, P'))^{2n}$  is small.

$P$  is a distribution supported on a sphere while  $P'$  is a distribution supported on a bumped sphere.





## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

- Statistical Inference on Persistence Homology

## References

We can use  $l_\infty$  metric to measure a distance between trees.

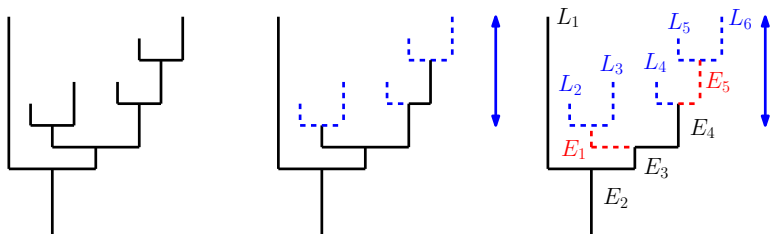
### Definition

The  $l_\infty$  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  $T_{\hat{\rho}_h}$  and are in the confidence 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  $2t_\alpha$ .



## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

- Statistical Inference on Persistence Homology

## References

We are considering the upper level set of the average kernel density estimator on the support.

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

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

- ▶ We are considering the upper level sets of the average kernel density estimator

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

We are considering the upper level set of the average kernel density estimator on the support.

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

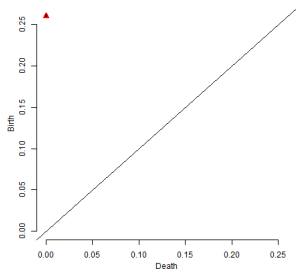
$$\{D_L\}_{L>0}, \text{ where } D_L := \{x \in \text{supp}(P) : p_h(x) \geq L\}.$$

We are targeting the persistent homology of the upper level set of the average kernel density estimator on the support.

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

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

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



We 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,

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



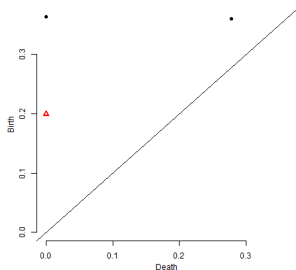
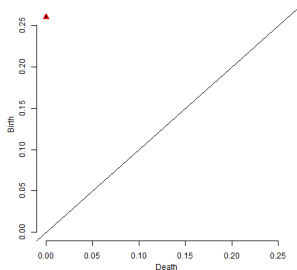
We estimate the target persistent homology by the persistent homology of the KDE filtration on Vietoris-Rips complexes.

- ▶ We estimate the target persistent homology

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

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

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



## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

- Statistical Inference on Persistence Homology

## References

## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

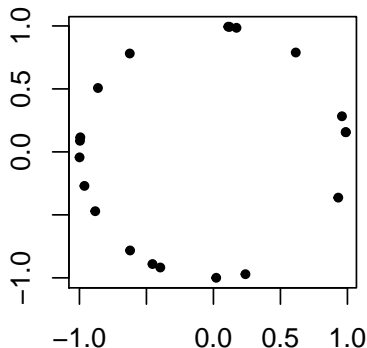
- Statistical Inference on Persistence Homology

## References

R Package TDA provides a function to sample on a circle.

The function `circleUnif()` generates  $n$  sample from the uniform distribution on the circle in  $\mathbb{R}^2$  with radius  $r$ .

```
circleSample <- circleUnif(n = 20, r = 1)
plot(circleSample, xlab = "", ylab = "", pch = 20)
```



R Package TDA provides distance functions and density functions over a grid.

Suppose  $n = 400$  points are generated from the unit circle, and grid of points are generated.

```
X <- circleUnif(n = 400, r = 1)

lim <- c(-1.7, 1.7)
by <- 0.05
margin <- seq(from = lim[1], to = lim[2], by = by)
Grid <- expand.grid(margin, margin)
```

## R Package TDA provides KDE function over a grid.

The Gaussian Kernel Density Estimator (KDE)  $\hat{p}_h : \mathbb{R}^d \rightarrow [0, \infty)$  is defined as

$$\hat{p}_h(y) = \frac{1}{n(\sqrt{2\pi}h)^d} \sum_{i=1}^n \exp\left(\frac{-\|y - x_i\|_2^2}{2h^2}\right),$$

where  $h$  is a smoothing parameter.

The function `kde()` computes the KDE function  $\hat{p}_h$  on a grid of points.

```
h <- 0.3
KDE <- kde(X = X, Grid = Grid, h = h)

par(mfrow = c(1,2))
plot(X, xlab = "", ylab = "", main = "Sample X", pch = 20)
persp(x = margin, y = margin,
      z = matrix(KDE, nrow = length(margin), ncol = length(margin)),
      xlab = "", ylab = "", zlab = "", theta = -20, phi = 35, scale = FALSE,
      expand = 3, col = "red", border = NA, ltheta = 50, shade = 0.5,
      main = "KDE")
```

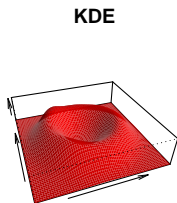
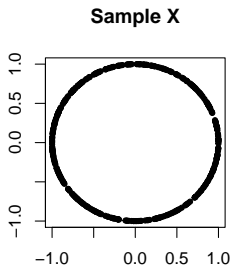
## R Package TDA provides KDE function over a grid.

The Gaussian Kernel Density Estimator (KDE)  $\hat{\rho}_h : \mathbb{R}^d \rightarrow [0, \infty)$  is defined as

$$\hat{\rho}_h(y) = \frac{1}{n(\sqrt{2\pi}h)^d} \sum_{i=1}^n \exp\left(\frac{-\|y - x_i\|_2^2}{2h^2}\right),$$

where  $h$  is a smoothing parameter.

The function `kde()` computes the KDE function  $\hat{\rho}_h$  on a grid of points.



## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology**

- Statistical Inference on Persistence Homology

## References



## R Package TDA computes Persistent Homology over a grid.

- ▶ The function `gridDiag()` computes the persistence diagram of sublevel (and superlevel) sets of the input function.
  - ▶ `gridDiag()` evaluates the real valued input function over a grid.
  - ▶ `gridDiag()` constructs a filtration of simplices using the values of the input function.
  - ▶ `gridDiag()` computes the persistent homology of the filtration.
- ▶ The user can choose to compute persistent homology using either C++ library GUDHI, Dionysus, or PHAT.

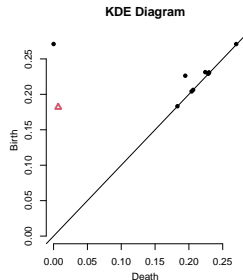
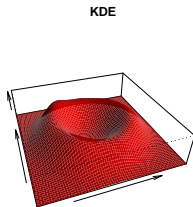
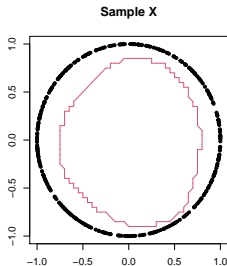
# R Package TDA computes Persistent Homology over a grid.

```
DiagGrid <- gridDiag(X = X, FUN = kde, lim = c(lim, lim), by = by,
  sublevel = FALSE, library = "Dionysus", location = TRUE,
  printProgress = FALSE, h = h)

par(mfrow = c(1,3))
plot(X, xlab = "", ylab = "", main = "Sample X", pch = 20)
one <- which(DiagGrid[["diagram"]][, 1] == 1)
for (i in seq(along = one)) {
  for (j in seq_len(dim(DiagGrid[["cycleLocation"]][[one[i]]])[1])) {
    lines(DiagGrid[["cycleLocation"]][[one[i]]][j, , ], pch = 19, cex = 1,
      col = i + 1)
  }
}
persp(x = margin, y = margin,
  z = matrix(KDE, nrow = length(margin), ncol = length(margin)),
  xlab = "", ylab = "", zlab = "", theta = -20, phi = 35, scale = FALSE,
  expand = 3, col = "red", border = NA, ltheta = 50, shade = 0.9,
  main = "KDE")
plot(x = DiagGrid[["diagram"]], main = "KDE Diagram")
```

# R Package TDA computes Persistent Homology over a grid.

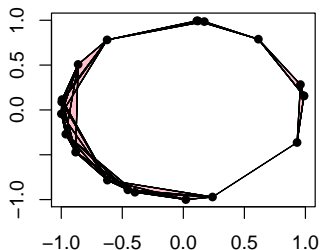
- ▶ The function `gridDiag()` computes the persistent homology of sublevel (and superlevel) sets of the input function.
  - ▶ `gridDiag()` evaluates the real valued input function over a grid.
  - ▶ `gridDiag()` constructs a filtration of simplices using the values of the input function.
  - ▶ `gridDiag()` computes the persistent homology of the filtration.
- ▶ The user can choose to compute persistent homology using either GUDHI, Dionysus, or PHAT.



# R Package TDA computes Vietoris-Rips Persistent Homology.

- ▶ Vietoris-Rips complex consists of simplices whose pairwise distances of vertices are at most  $\epsilon$  apart, i.e.

$$R(X, \epsilon) = \{[X_{n_1}, \dots, X_{n_r}] : d(X_{n_i}, X_{n_j}) \leq \epsilon\}.$$



- ▶ Rips filtration is formed by Rips complexes with gradually increasing  $\epsilon$ .

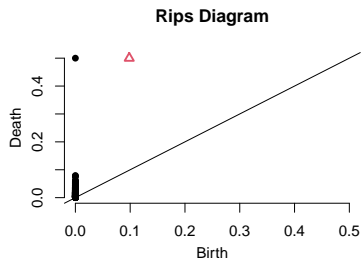
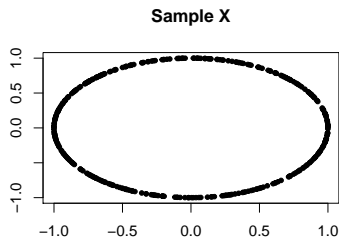
# R Package TDA computes Vietoris-Rips Persistent Homology.

- ▶ The function `ripsDiag()` computes the persistence diagram of the Rips filtration built on top of a point cloud.
  - ▶ `ripsDiag()` constructs the Vietoris-Rips filtration using the data points.
  - ▶ `ripsDiag()` computes the persistent homology of the Vietoris-Rips filtration.
- ▶ The user can choose to compute persistent homology using either C++ library GUDHI, Dionysus, or PHAT.

```
DiagRips <- ripsDiag(X = X, maxdimension = 1, maxscale = 0.5,  
  library = c("GUDHI", "Dionysus"), location = TRUE)  
  
par(mfrow = c(1,2))  
plot(X, xlab = "", ylab = "", main = "Sample X", pch = 20)  
plot(x = DiagRips[["diagram"]], main = "Rips Diagram")
```

# R Package TDA computes Vietoris-Rips Persistent Homology.

- ▶ The function `ripsDiag()` computes the persistence diagram of the Rips filtration built on top of a point cloud.
  - ▶ `ripsDiag()` constructs the Vietoris-Rips filtration using the data points.
  - ▶ `ripsDiag()` computes the persistent homology of the Vietoris-Rips filtration.
- ▶ The user can choose to compute persistent homology using either C++ library GUDHI, Dionysus, or PHAT.



## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

- Statistical Inference on Persistence Homology

## References

R Package TDA computes the bootstrap confidence band for a function.

The function `bootstrapBand()` computes  $(1 - \alpha)$  bootstrap confidence band for  $\mathbb{E}[\hat{\rho}_h]$ .

```
bandKDE <- bootstrapBand(X = X, FUN = kde, Grid = Grid, B = 20,  
  parallel = FALSE, alpha = 0.1, h = h)  
print(bandKDE[["width"]])
```

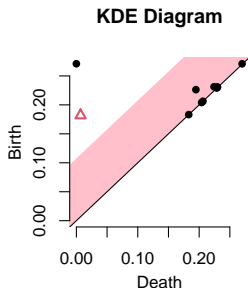
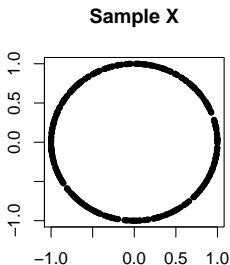
```
##          90%  
## 0.0537502
```



# R Package TDA computes the bootstrap confidence band for the persistent homology.

The  $(1 - \alpha)$  bootstrap confidence band for  $\mathbb{E}[\hat{\rho}_h]$  is used as the confidence band for the persistent homology.

```
par(mfrow = c(1,2))
plot(X, xlab = "", ylab = "", main = "Sample X", pch = 20)
plot(x = DiagGrid[["diagram"]], band = 2 * bandKDE[["width"]],
     main = "KDE Diagram")
```



## Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions

- Upper Bound

- Lower Bound

- Upper Bound and Lower Bound for General Case

## The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry

- Reach estimator and its analysis

- Minimax Estimates

## Statistical Inference for Cluster Trees

- Confidence band for Persistent Homology of KDEs computed on Vietoris-Rips complexes

## R Package TDA: Statistical Tools for Topological Data Analysis

- Sample on manifolds, Distance Functions, and Density Estimators

- Persistent Homology

- Statistical Inference on Persistence Homology

## References

## References

- Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the Reach of a Manifold. *ArXiv e-prints*, May 2019.
- 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 T. Fasy, Jisu Kim, Fabrizio Lecci, Clément Maria, David L. Millman, and Vincent Rouvreau. Introduction to the R package TDA. *CoRR*, abs/1411.1830, 2014a. URL <http://arxiv.org/abs/1411.1830>.
- 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 2014b. doi: 10.1214/14-AOS1252. URL <http://dx.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