

Statistical Inference and Featurization on Geometric and Topological Data

Jisu KIM



대학원생 세미나
2023-09-08

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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

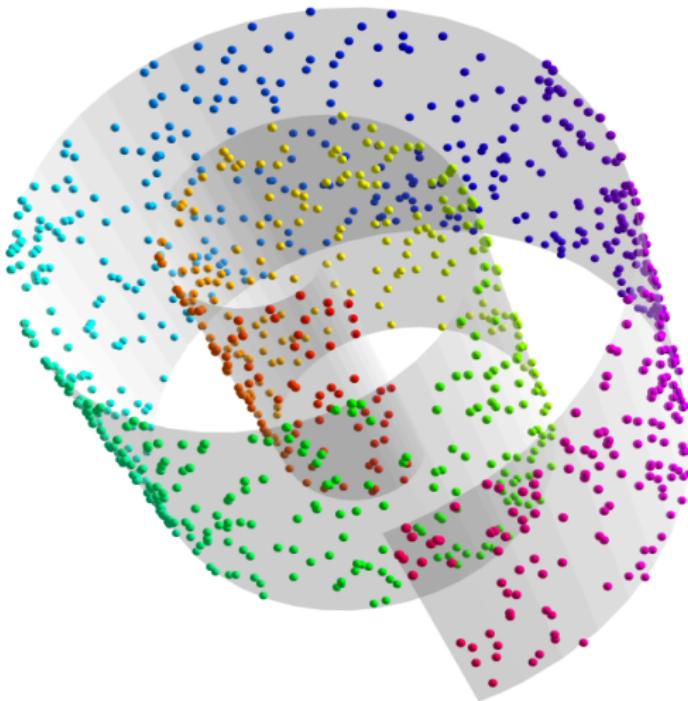
Circular Coordinates and its Applications

Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

References

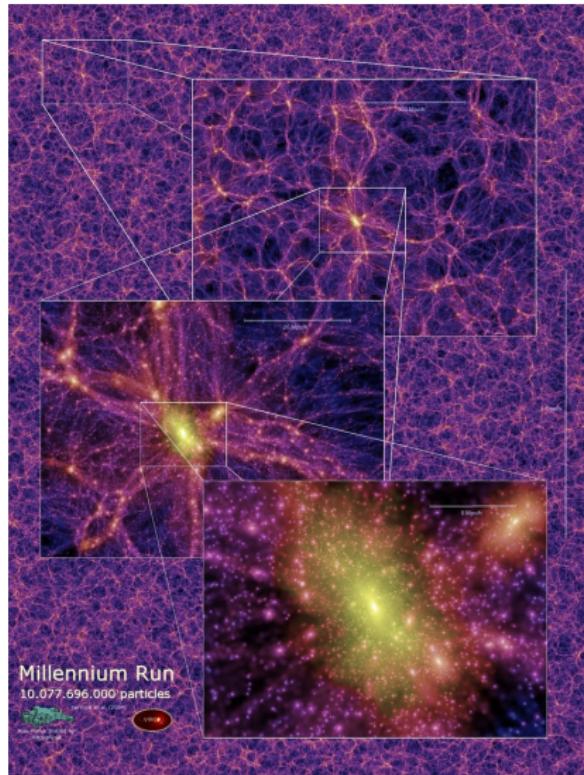
The curse of dimensionality from the high dimensional data is mitigated when there is a low dimensional geometric and topological structure.



1

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

Geometric and topological structures in the data provide information.



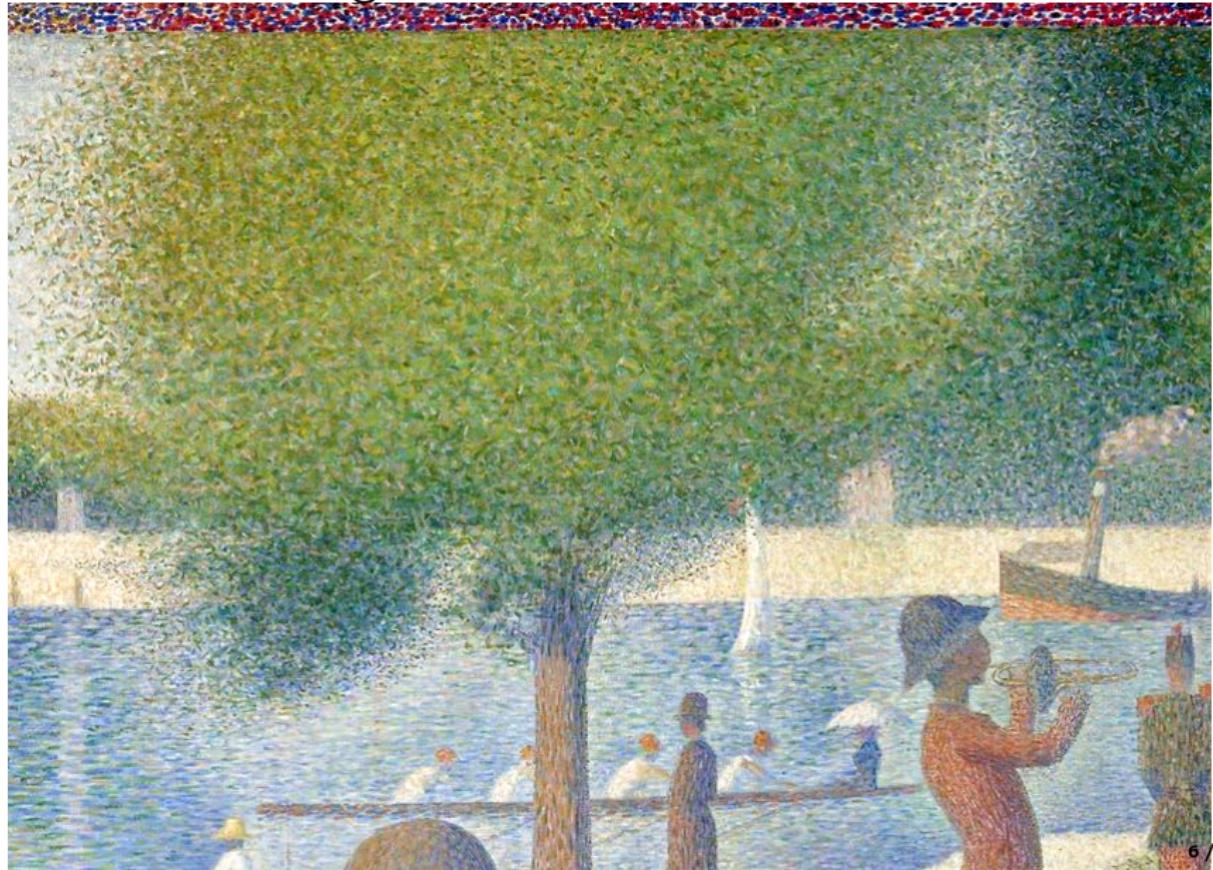
2

²http://www.mpa-garching.mpg.de/galform/virgo/millennium/poster_half.jpg

Inferencing topology from data: observe topological structure with a single resolution, or with multi resolutions.



Inferencing topology from data: observe topological structure with a single resolution, or with multi resolutions.



Inferencing topology from data: observe topological structure with a single resolution, or with multi resolutions.



Inferencing topology from data: observe topological structure with a single resolution, or with multi resolutions.

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



Statistic Inference for 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)
- ▶ Inference for Homotopy
 - ▶ Homotopy Reconstruction via the Cech Complex and the Vietoris-Rips Complex (Shin, Kim, Chazal, Rinaldo, Wasserman, 2020a)
- ▶ Statistical Inference for Homological Features
 - ▶ Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)
- ▶ Statistical Inference for Persistent Homology
 - ▶ Statistical inference on persistent homology of KDE filtration on Vietoris-Rips complex (Shin, Kim, Rinaldo, Wasserman, 2023?)

Featurization using Topological Data Analysis is explored.

- ▶ Featurization using Persistent Homology
 - ▶ Time Series Featurization via Topological Data Analysis (Kim, Kim, Rinaldo, Chazal, 2020b)
 - ▶ Efficient Topological Layer based on Persistent Landscapes (Kim, Kim, Zaheer, Kim, Chazal, Wasserman, 2020)
- ▶ Featurization using Circular Coordinates
 - ▶ Generalized penalty for circular coordinate representation (Luo, Patania, Kim, Vejdemo-Johansson, 2021)
- ▶ 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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

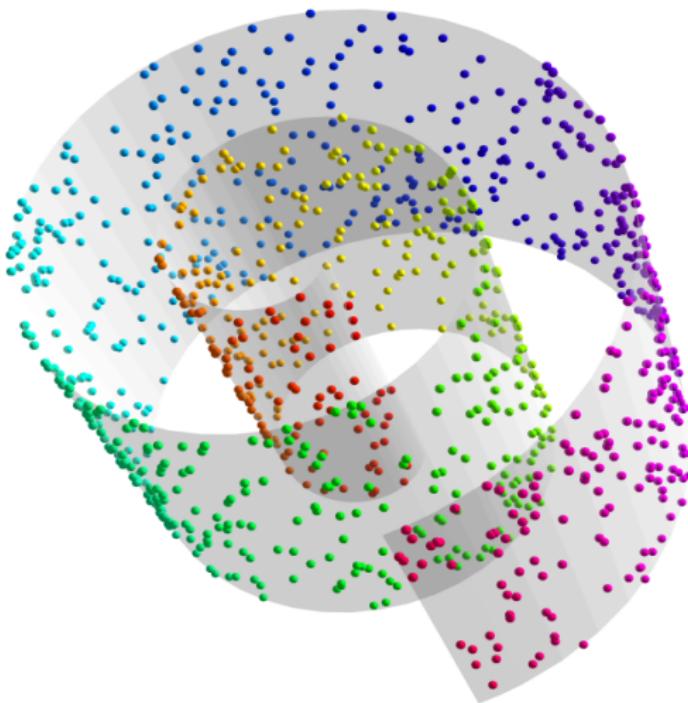
Circular Coordinates and its Applications

Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

References

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



3

³<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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

Circular Coordinates and its Applications

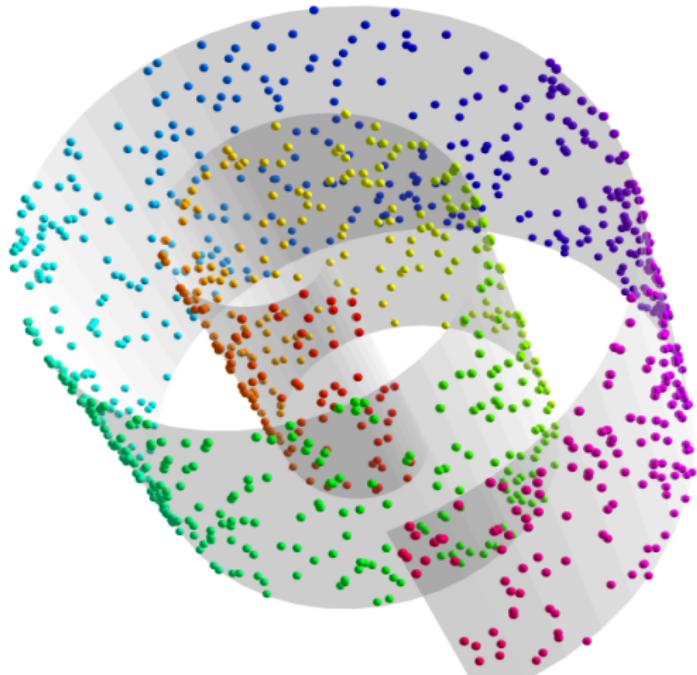
Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

References

The intrinsic dimension of a manifold needs to be estimated a prior 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 $x, y \in \mathbb{R}$,
 $\ell(x, y) = 1(x \neq y)$.

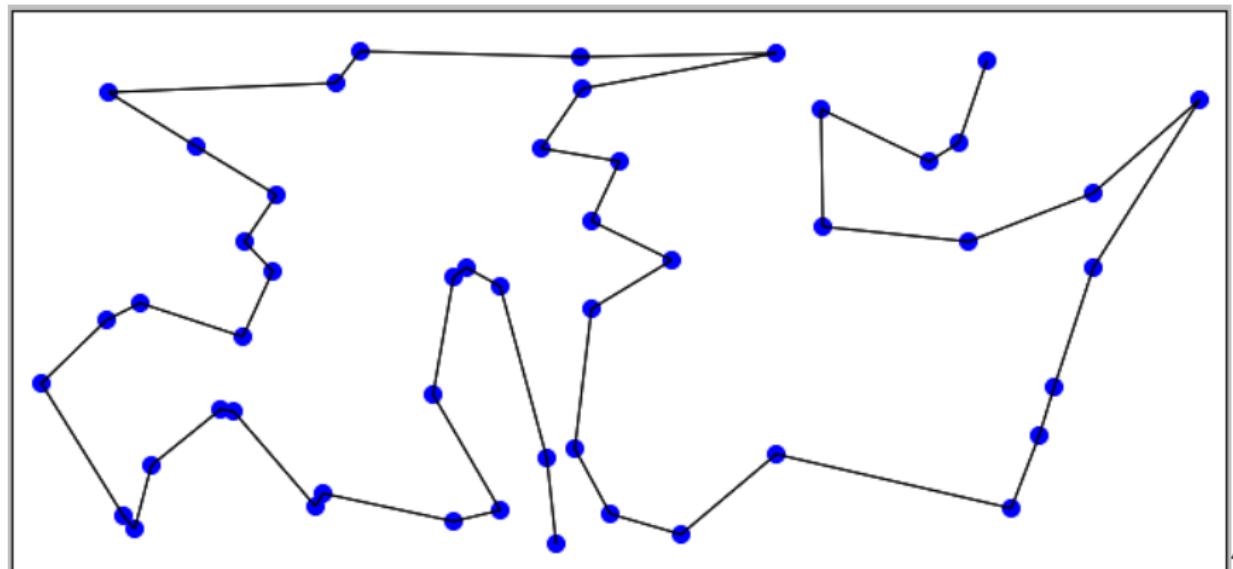
Minimax rate for estimating the dimension: we first consider dimension d_1 vs d_2 .



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

- ▶ $X = (X_1, \dots, X_n)$ 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 $x, y \in \mathbb{R}$,
 $\ell(x, y) = 1(x \neq y)$.

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



4

⁴<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.

Minimax rate for estimating the dimension

Theorem

(*Proposition 16 and 17*)

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

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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

Circular Coordinates and its Applications

Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

References

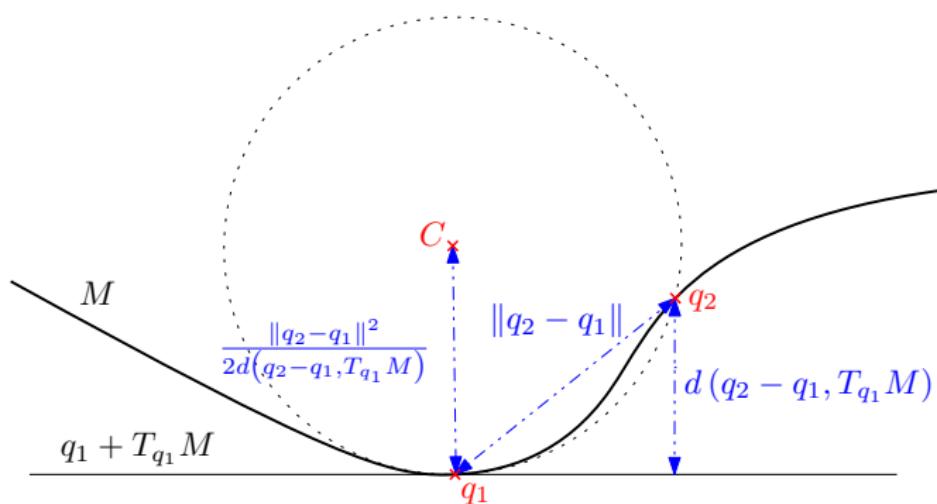
The reach is the maximum radius of a ball that can roll over the manifold.

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



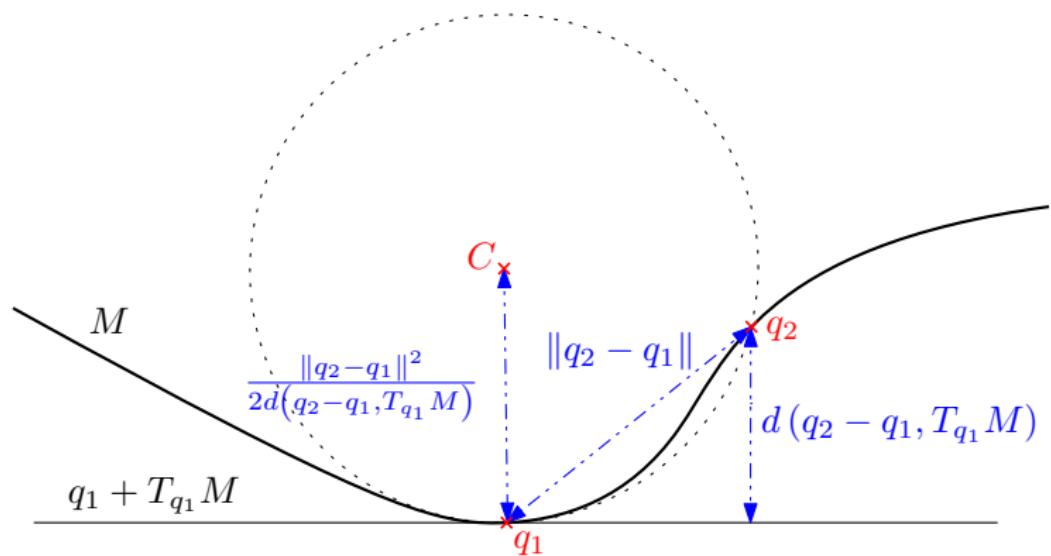
$$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[\left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}_n(X)} \right|^q \right]$$

- ▶ $X = (X_1, \dots, X_n)$ 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 I_q loss function is considered, so for all $x, y \in \mathbb{R}$,
$$\ell(x, y) = \left| \frac{1}{x} - \frac{1}{y} \right|^q.$$

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 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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

Circular Coordinates and its Applications

Computation for Topological Data Analysis

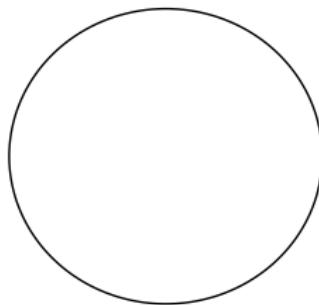
R Package TDA: Statistical Tools for Topological Data Analysis

References

We are interested in estimating the topology of the target space $\mathbb{X} \subset \mathbb{R}^d$ based on samples \mathcal{X} that lies in it or in its proximity.

- ▶ Estimating the topology from samples occurs in: cosmology, time series data, machine learning, etc.

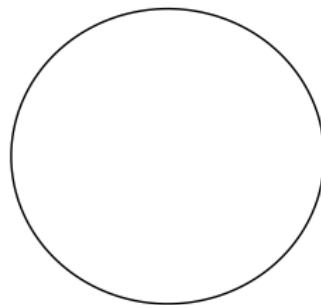
Underlying circle



We are interested in estimating the topology of the target space $\mathbb{X} \subset \mathbb{R}^d$ based on samples \mathcal{X} that lies in it or in its proximity.

- ▶ Estimating the topology from samples occurs in: cosmology, time series data, machine learning, etc.

Underlying circle



20 samples



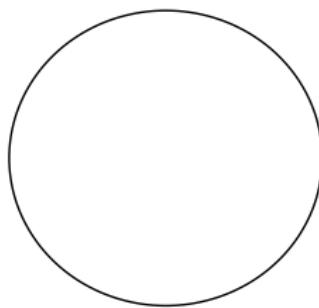
Two topological spaces are homotopy equivalent if they can be continuously deformed into one another.

- ▶ We estimate topology up to homotopy equivalence.
- ▶ A space X is contractible if X is homotopy equivalent to a point.



We estimate the homotopy of the target space $\mathbb{X} \subset \mathbb{R}^d$ based on samples \mathcal{X} that lies in it or in its proximity.

Underlying circle



20 samples

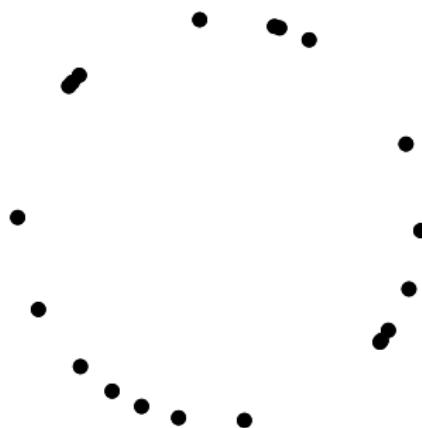


We use the Vietoris-Rips complex to estimate the target space.

- ▶ 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 space.

- ▶ 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

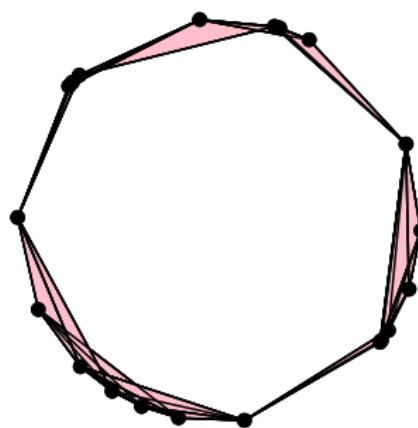


We use the Vietoris-Rips complex to estimate the target space.

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

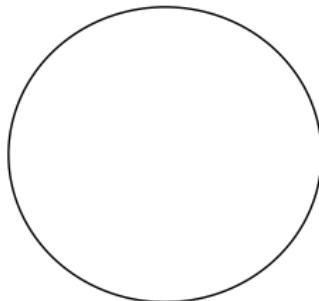
$$\text{Rips}(\mathcal{X}, r) = \{\{x_1, \dots, x_k\} \subset \mathcal{X} : d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k\}.$$

Vietoris-Rips complex



We estimate the homotopy of the target space $\mathbb{X} \subset \mathbb{R}^d$ using the Vietoris-Rips complex on samples \mathcal{X} that lies in it or in its proximity.

Underlying circle

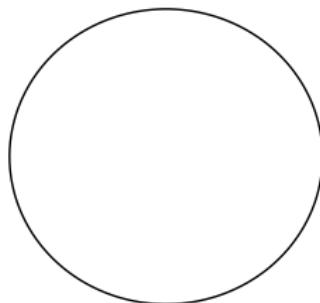


20 samples

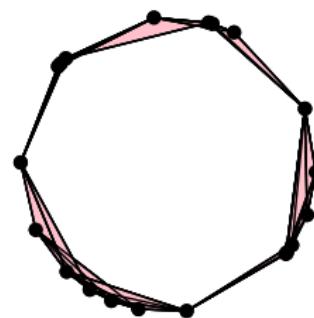


We estimate the homotopy of the target space $\mathbb{X} \subset \mathbb{R}^d$ using the Vietoris-Rips complex on samples \mathcal{X} that lies in it or in its proximity.

Underlying circle



Vietoris–Rips complex



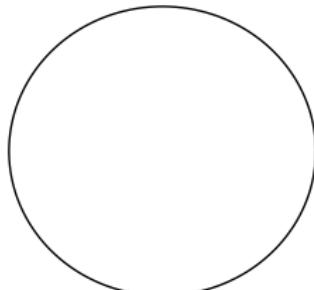
We derive conditions under which the homotopy of the target space $\mathbb{X} \subset \mathbb{R}^d$ is correctly reconstructed using the Vietoris-Rips complex on samples \mathcal{X} .

- ▶ Kim et al. [2020a]: extending Niyogi et al. [2008], Attali et al. [2013]

Theorem

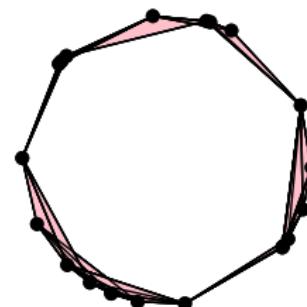
Let $\mathbb{X} \subset \mathbb{R}^d$ be a subset with its reach $\tau > 0$, and let $\mathcal{X} \subset \mathbb{R}^d$ be a closed discrete set of points. If $\frac{d_H(\mathbb{X}, \mathcal{X})}{\tau} \leq C$ where $d_H(\mathbb{X}, \mathcal{X})$ is the Hausdorff distance, then with appropriate choice of r , the Vietoris-Rips complex $\text{Rips}(\mathcal{X}, r)$ is homotopy equivalent to \mathbb{X} , with $C \approx 0.07856 \dots$.

Underlying circle



\approx

Vietoris–Rips complex



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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

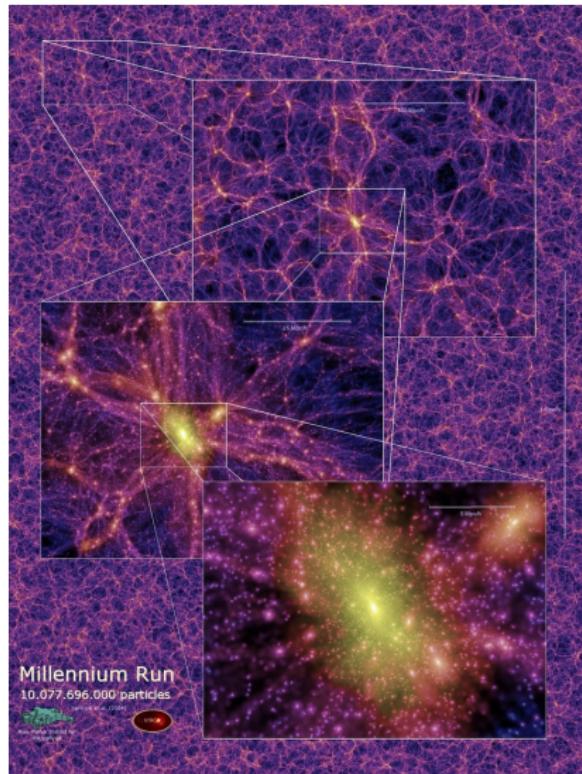
Circular Coordinates and its Applications

Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

References

Topological holes in the data provide information.



The number of holes is used to summarize geometrical features.

- ▶ Geometrical objects :
 - ▶ ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ
 - ▶ A, 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	ㄱ, ㄴ, ㄷ, ㄹ, ㅅ, ㅈ, ㅌ, ㅊ	ㅁ, ㅇ, ㅂ, ㅍ, ㅏ	ㅓ
2	え, 字		
3		𐍈	

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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

Circular Coordinates and its Applications

Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

References

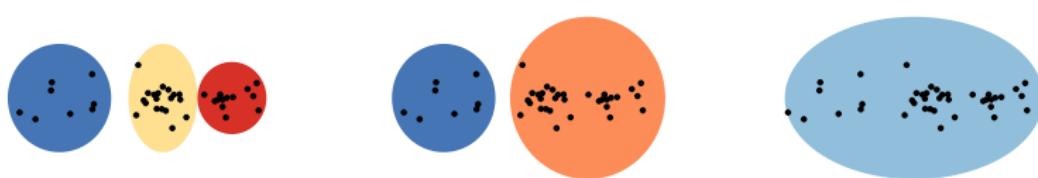
We want to cluster data.

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



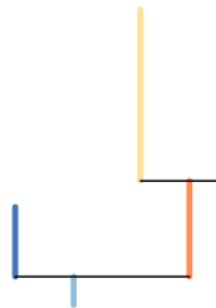
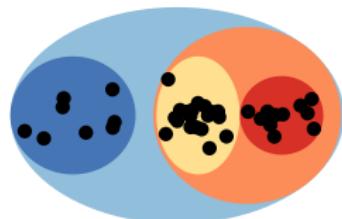
Different clusters can be formed by the desired level of resolution.

- ▶ Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)
- ▶ If you want clusters to describe local and detailed information (high resolution), there will be more clusters with each of smaller sizes.
- ▶ If you want clusters to describe global and rough information (low resolution), there will be less clusters with each of larger sizes.



The network of clusters forms a tree: cluster tree

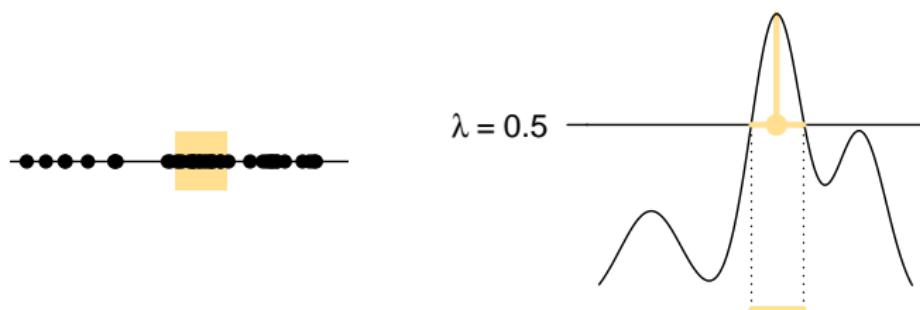
- ▶ Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)
- ▶ Clusters from different levels of resolution have a natural network by inclusion relation.
- ▶ Inclusion network of clusters can be represented as a tree: cluster tree.



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

Definition

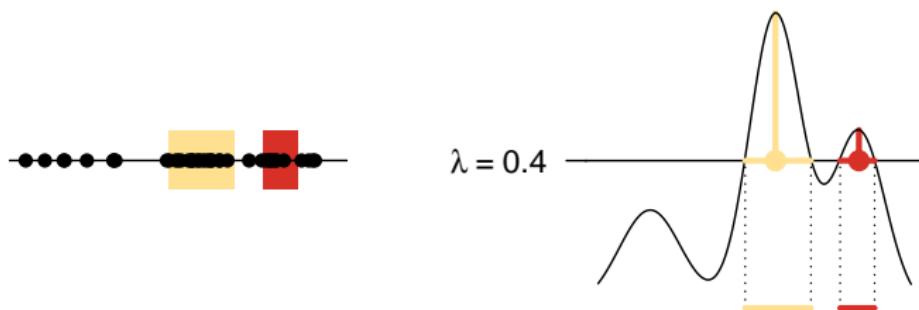
For a density function p , its cluster tree $T_p : \mathbb{R} \rightarrow \mathcal{P}(\mathcal{X})$ is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x \in \mathcal{X} : p(x) \geq \lambda\}$.



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

Definition

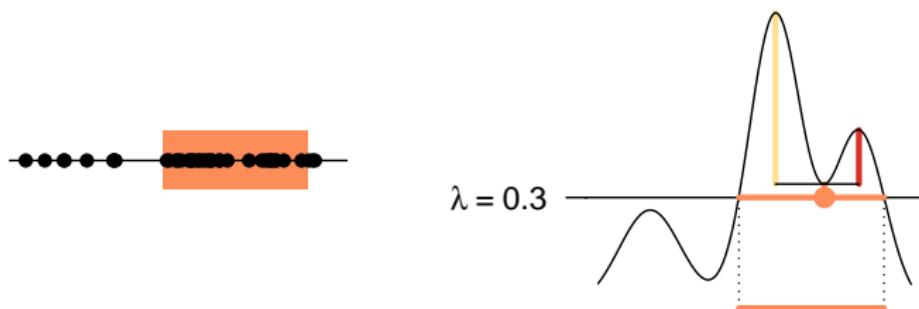
For a density function p , its cluster tree $T_p : \mathbb{R} \rightarrow \mathcal{P}(\mathcal{X})$ is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x \in \mathcal{X} : p(x) \geq \lambda\}$.



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

Definition

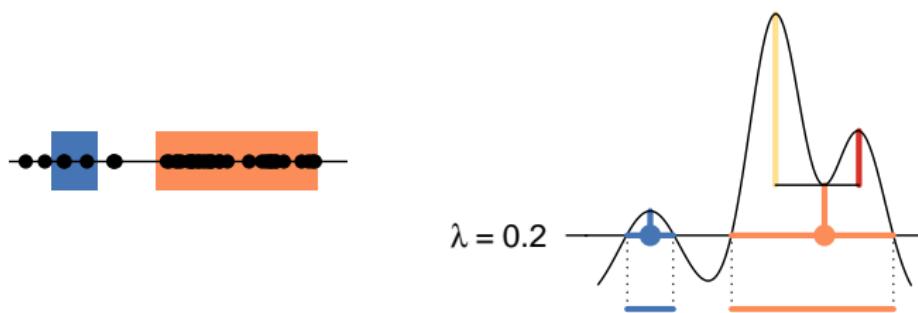
For a density function p , its cluster tree $T_p : \mathbb{R} \rightarrow \mathcal{P}(\mathcal{X})$ is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x \in \mathcal{X} : p(x) \geq \lambda\}$.



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

Definition

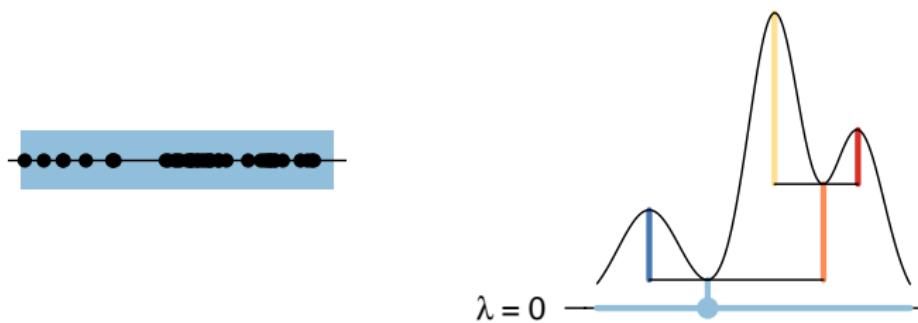
For a density function p , its cluster tree $T_p : \mathbb{R} \rightarrow \mathcal{P}(\mathcal{X})$ is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x \in \mathcal{X} : p(x) \geq \lambda\}$.



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

Definition

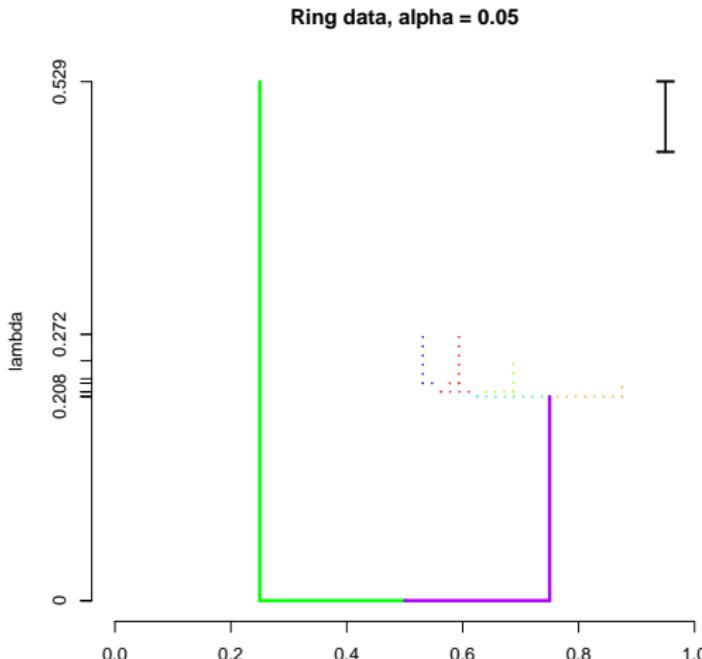
For a density function p , its cluster tree $T_p : \mathbb{R} \rightarrow \mathcal{P}(\mathcal{X})$ is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x \in \mathcal{X} : p(x) \geq \lambda\}$.



A confidence set helps denoising the empirical tree.

- ▶ An asymptotic $1 - \alpha$ confidence set \hat{C}_α 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}_α .

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

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

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

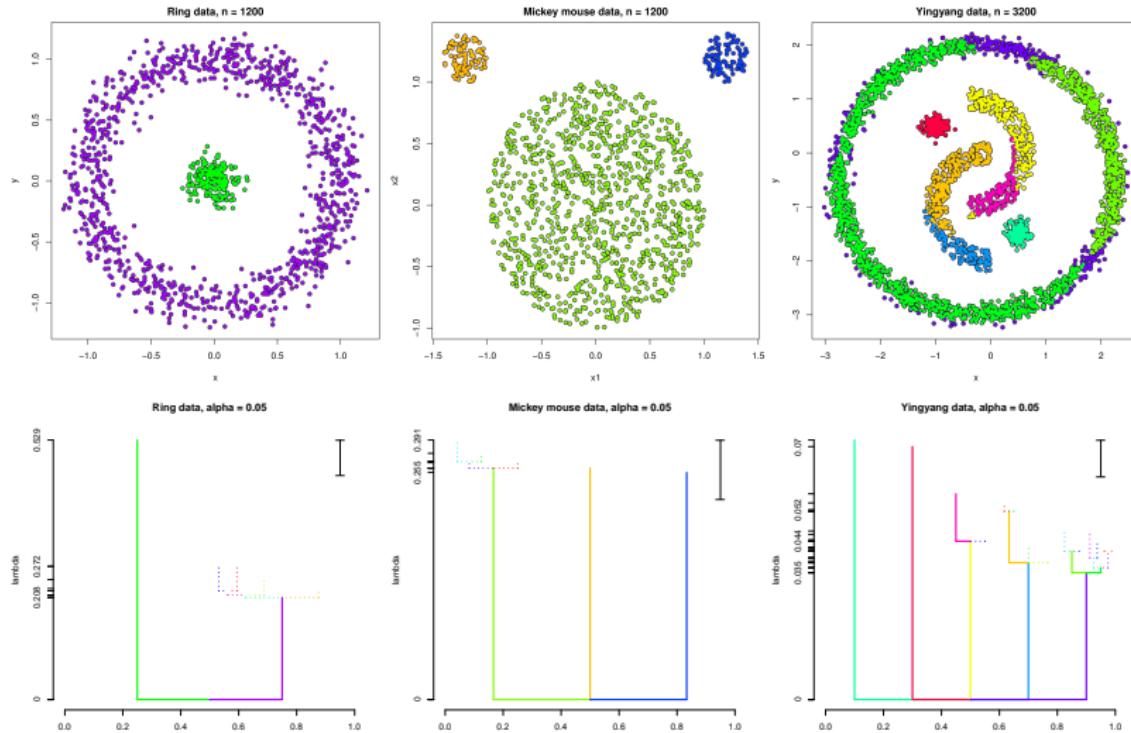
$$\hat{C}_\alpha = \{T : d_\infty(T, T_{\hat{p}_h}) \leq t_\alpha\}.$$

Theorem

(Theorem 3) Above confidence set \hat{C}_α 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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

Circular Coordinates and its Applications

Computation for Topological Data Analysis

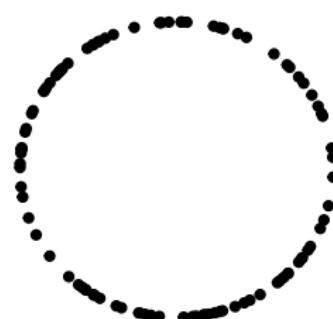
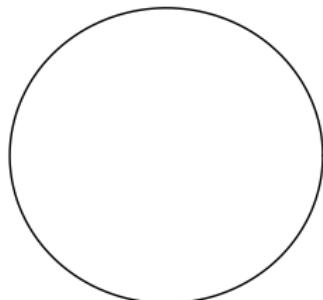
R Package TDA: Statistical Tools for Topological Data Analysis

References

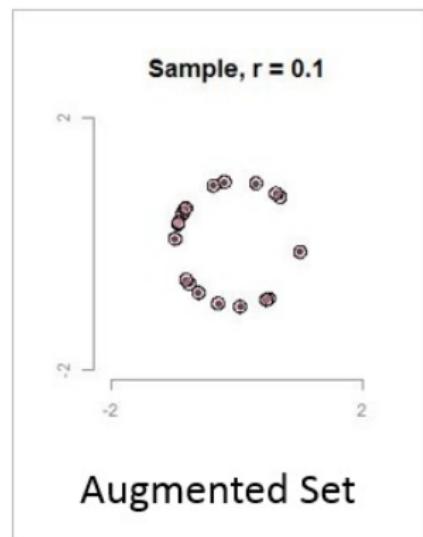
Homology of finite sample is different from homology of underlying manifold, hence it cannot be directly used for the inference.

- ▶ 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:

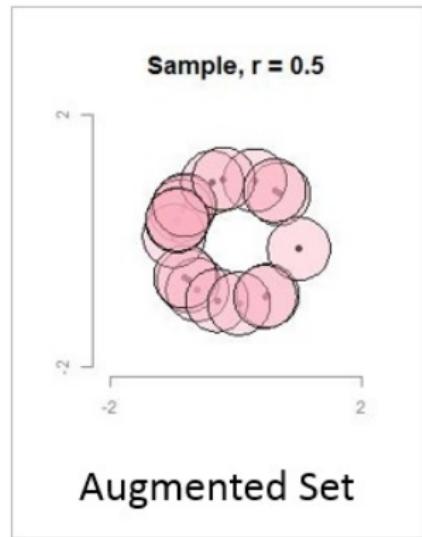
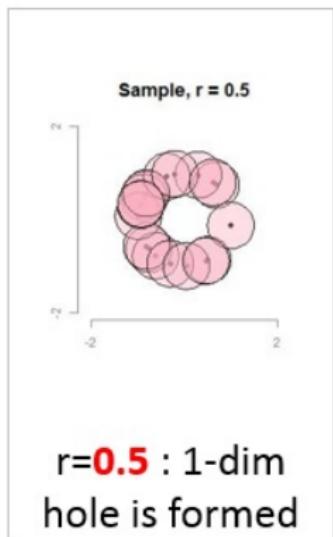
Underlying circle: $\beta_0 = 1, \beta_1 = 1$ 100 samples: $\beta_0 = 100, \beta_1 = 0$



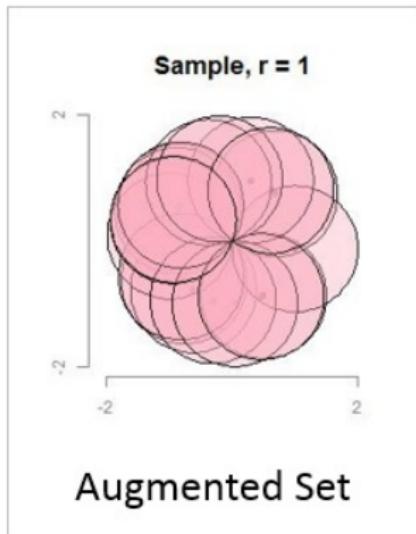
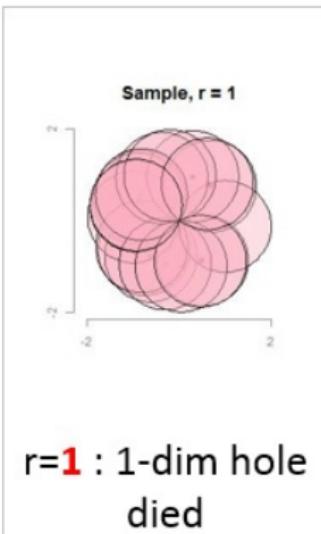
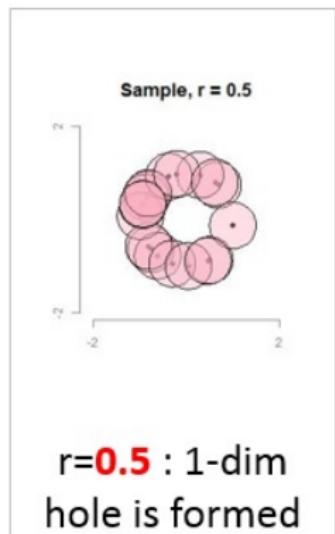
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



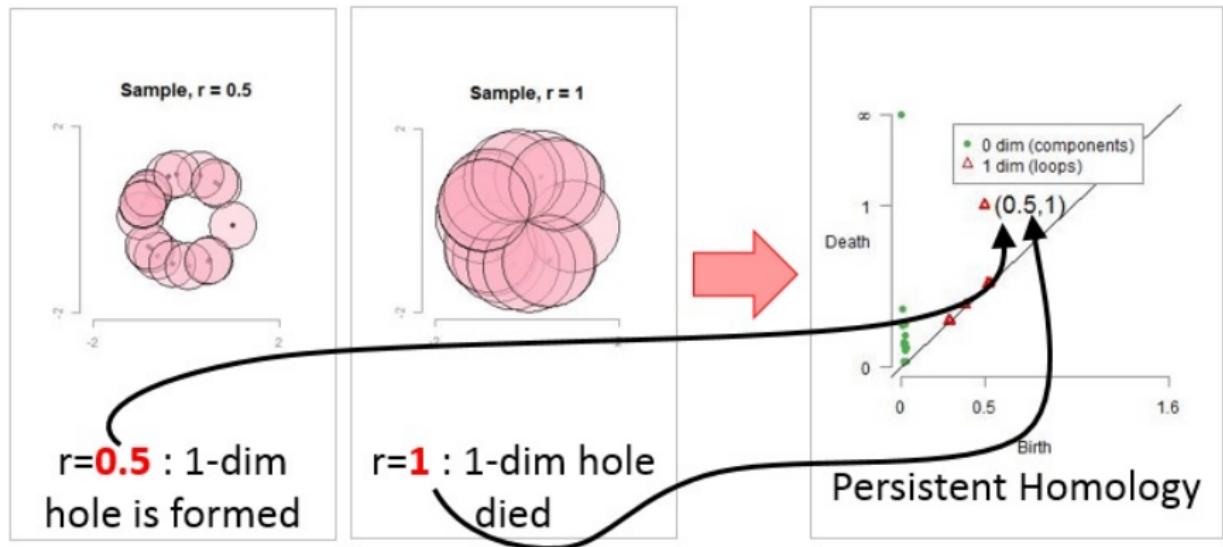
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



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.

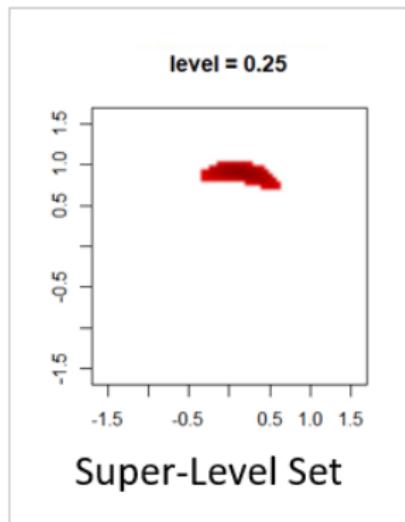


We rely on the kernel density estimator to extract topological information of the underlying distribution.

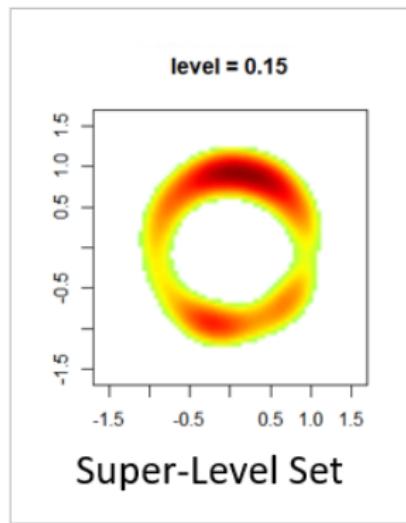
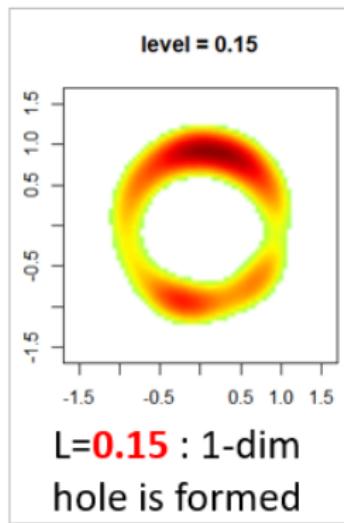
- ▶ The kernel density estimator is

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

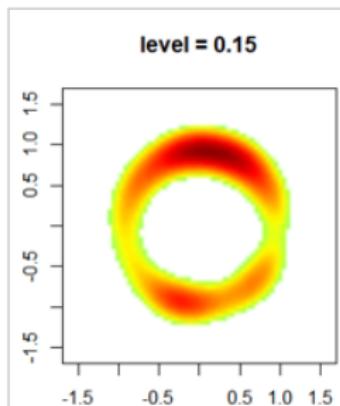
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



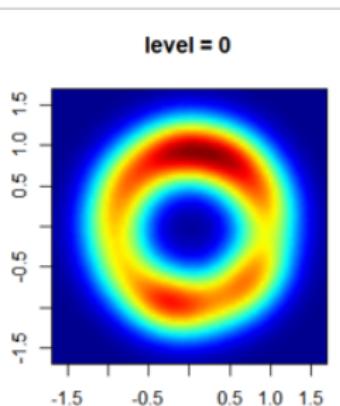
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



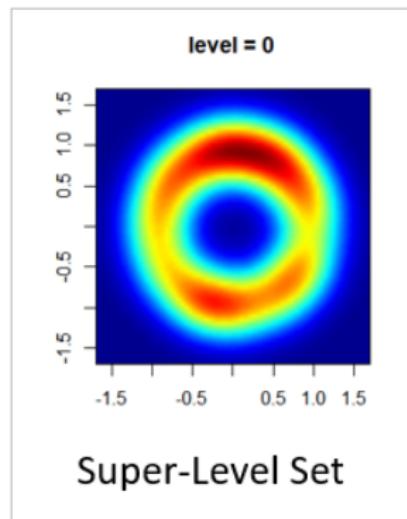
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



L=0.15 : 1-dim
hole is formed

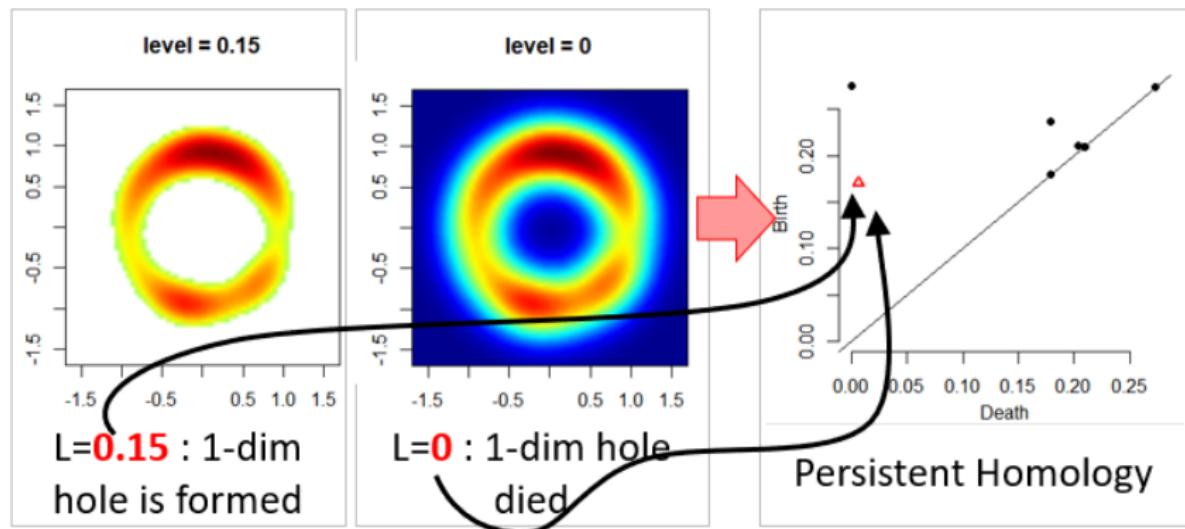


L=0 : 1-dim hole
died



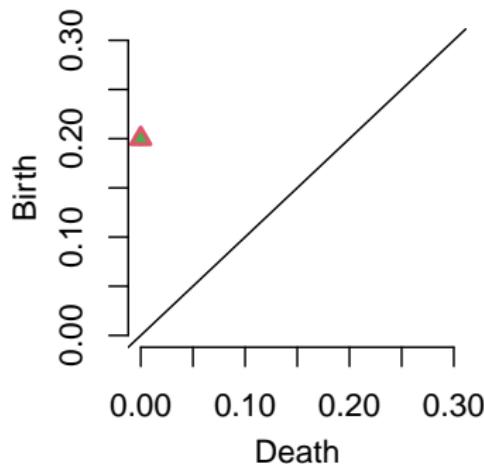
Super-Level Set

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

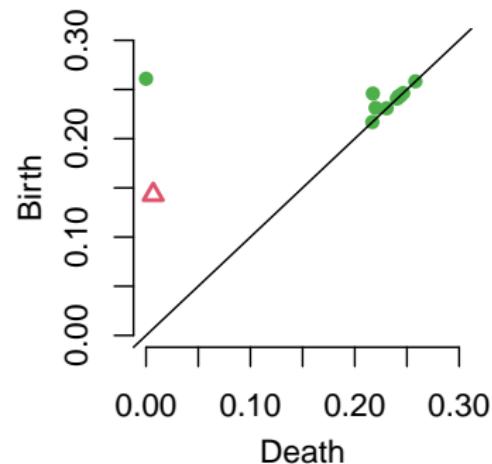


Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.

Circle



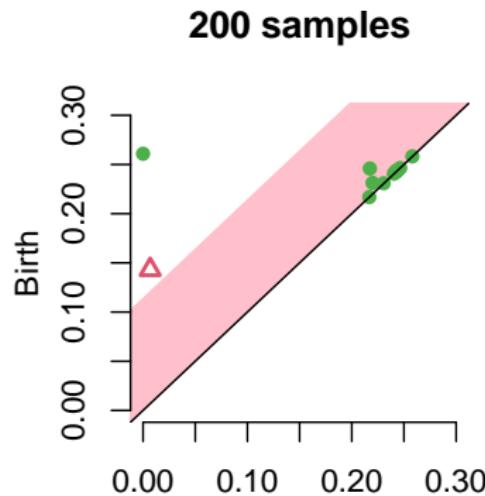
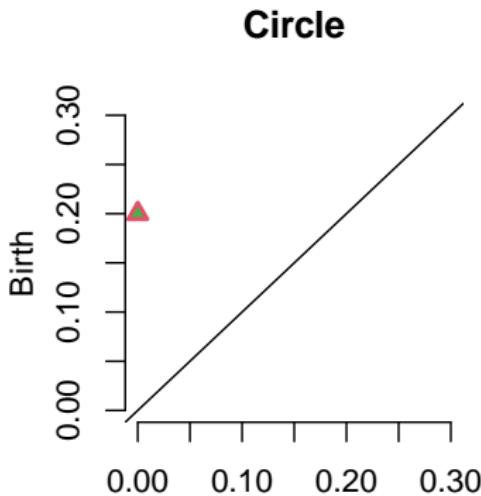
200 samples



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

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

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



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 $\left[\hat{p}_h - \frac{\hat{z}_\alpha}{\sqrt{nh^m}}, \hat{p}_h + \frac{\hat{z}_\alpha}{\sqrt{nh^m}} \right]$.

Statistical inference for persistent homology.

- ▶ Persistent homology of KDE filtration on Vietoris-Rips complex
(Shin, Kim, Rinaldo, Wasserman, 2023?)

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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

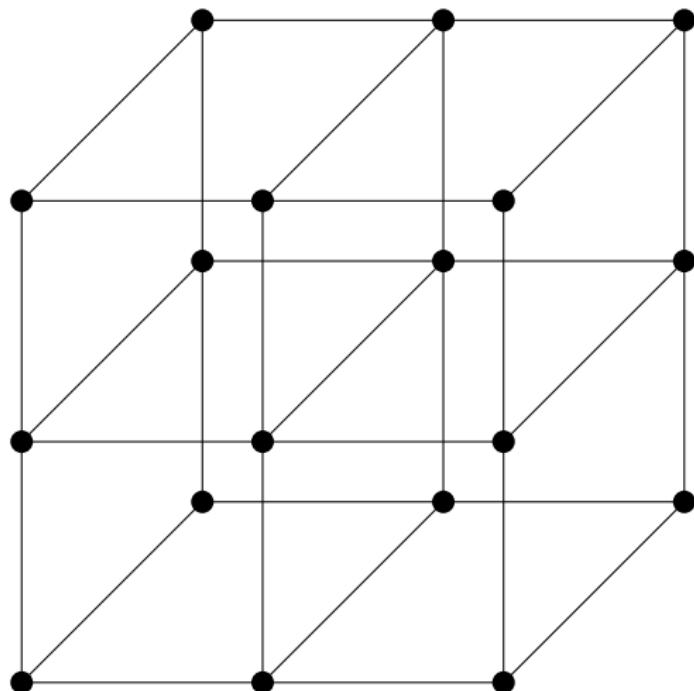
Circular Coordinates and its Applications

Computation for Topological Data Analysis

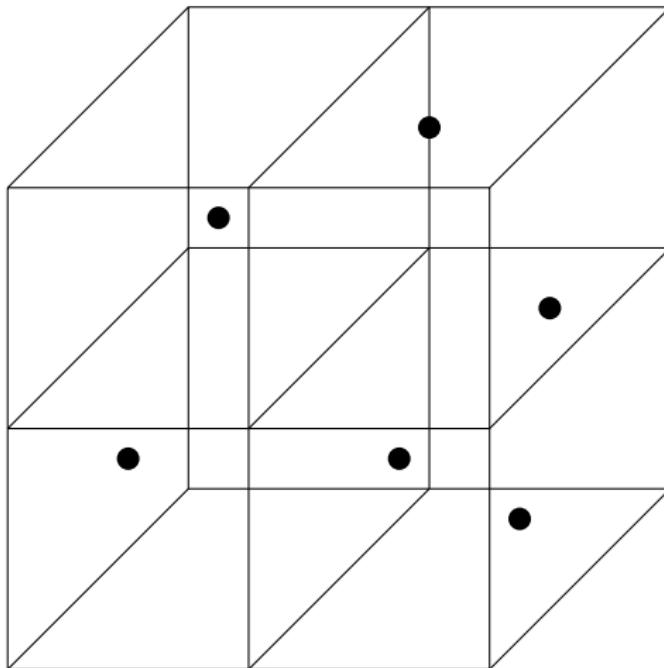
R Package TDA: Statistical Tools for Topological Data Analysis

References

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

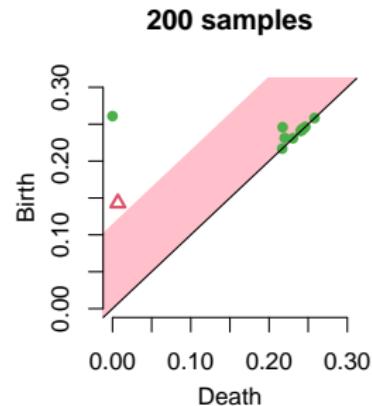
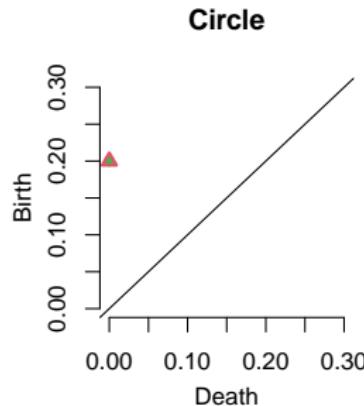


Computing the persistent homology of density function on data points reduces computational complexity.



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].

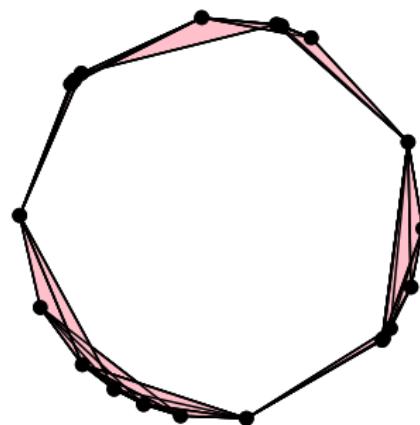


We use the Vietoris-Rips complex to estimate the target persistent homology.

- ▶ 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 estimate the target persistent homology by using the KDE and Vietoris-Rips complexes.

- ▶ 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

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

- ▶ Our persistent homology estimator $PH_*^R(\hat{p}_h, r)$ is built by using the KDE and Vietoris-Rips complexes.

Our persistent homology estimator 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

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

Confidence set

- ▶ An asymptotic $1 - \alpha$ confidence set \hat{C}_α 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 our persistent homology estimator.

- We let the confidence set as the ball centered at $PH_*^R(\hat{p}_{h_n}, r_n)$ and radius \hat{b}_α , i.e.

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

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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

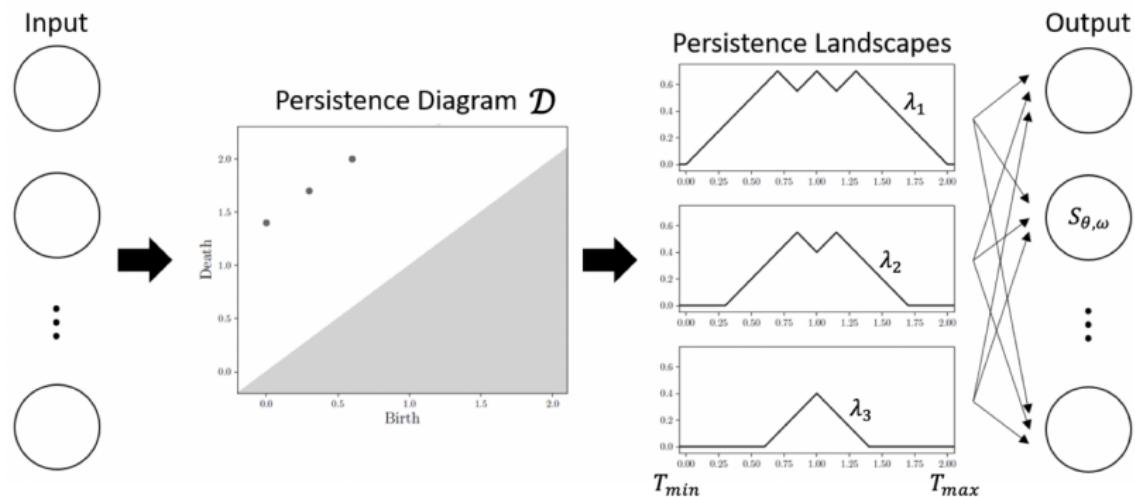
Circular Coordinates and its Applications

Computation for Topological Data Analysis

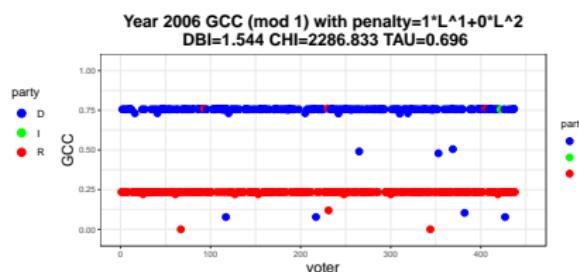
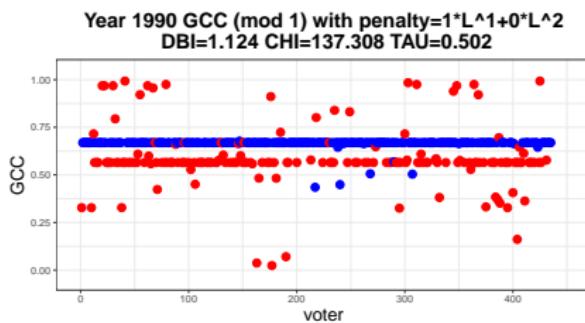
R Package TDA: Statistical Tools for Topological Data Analysis

References

Topological structure is featurized as persistence landscape and circular coordinates to be further applied in statistical or machine learning framework.



Topological structure is featurized as persistence landscape and circular coordinates to be further applied in statistical or machine learning framework.



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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

Circular Coordinates and its Applications

Computation for Topological Data Analysis

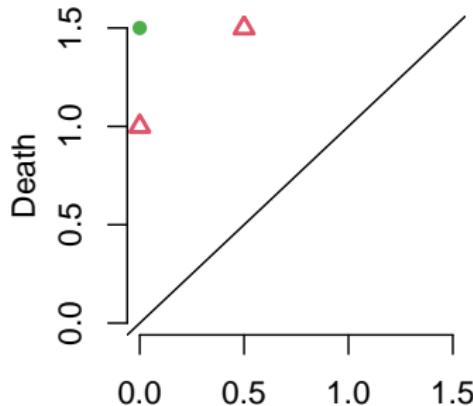
R Package TDA: Statistical Tools for Topological Data Analysis

References

Persistent homology is further summarized and embedded into a Euclidean space or a functional space.

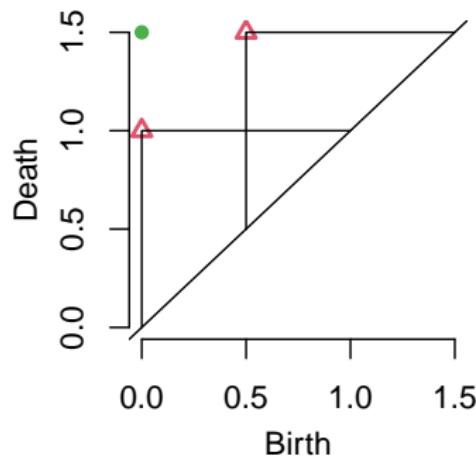
- ▶ The space of the persistent homology is complex, so directly applying in machine learning is difficult.
- ▶ If the persistent homology is further summarized and embedded into a Euclidean space or a functional space, then applying in machine learning becomes much more convenient.
 - ▶ e.g., Persistence Landscape, Persistence Silhouette, Persistence Image

Persistent Homology

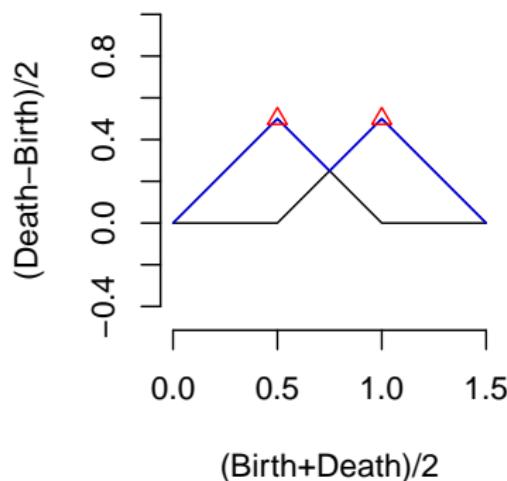


Persistence Landscape is a functional summary of the persistent homology.

Persistent Homology



Landscape

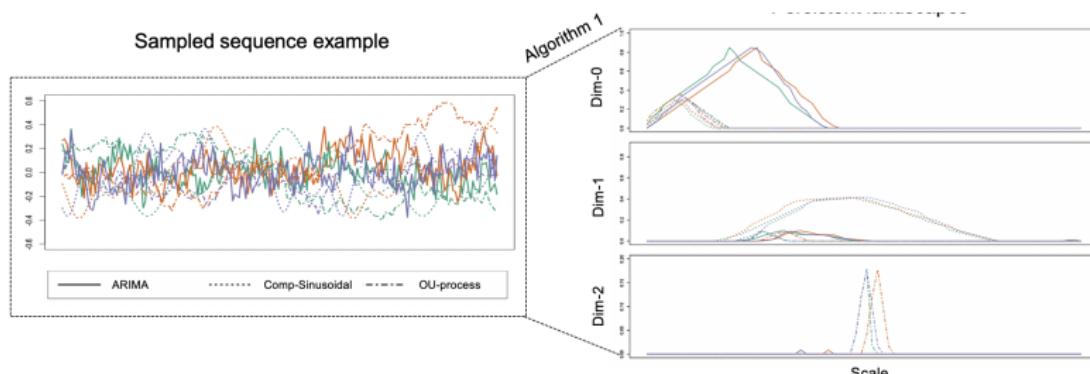


Featurizing using Persistence Landscape

- ▶ Featurization using time-delayed embedding and Persistence Landscape
 - ▶ Time Series Featurization via Topological Data Analysis (Kim, Kim, Rinaldo, Chazal, 2020b)
- ▶ Build topological layer using Persistence Landscape
 - ▶ PLLay: Efficient Topological Layer based on Persistence Landscapes (Kim, Kim, Zaheer, Kim, Chazal, Wasserman, 2020)

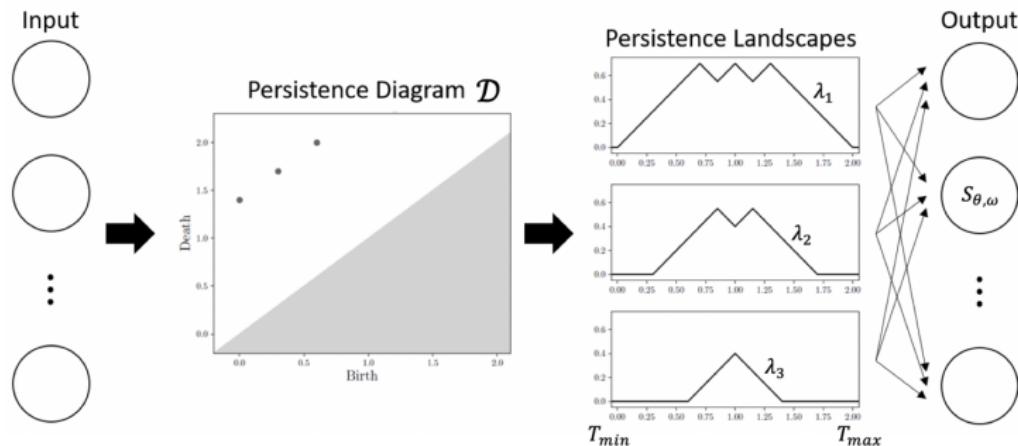
Featurization using time-delayed embedding and Persistence Landscape

1. From time series data $x = \{x_0, \dots, x_N\} \subset \mathbb{R}$, construct the point cloud $X \subset \mathbb{R}^m$ using the time-delayed embedding.
2. Perform PCA(Principal Component Analysis) on X and obtain $X' \subset \mathbb{R}^l$.
3. Construct the Vietoris-Rips filtration $R_{X'}$ and compute the persistence diagram $Dgm(X')$.
4. From $Dgm(X')$, compute the persistence landscape $\lambda : \mathbb{N} \times \mathbb{R} \rightarrow \mathbb{R}$, and vectorize to get $\lambda^K \in \mathbb{R}^K$.
5. Perform PCA on λ^K and get $\lambda^k \in \mathbb{R}^k$.

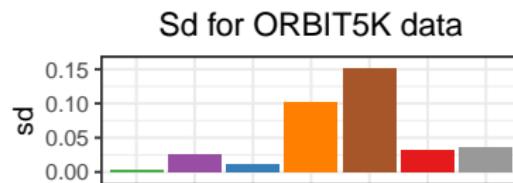
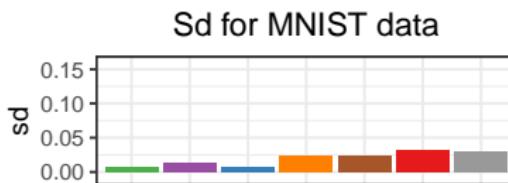
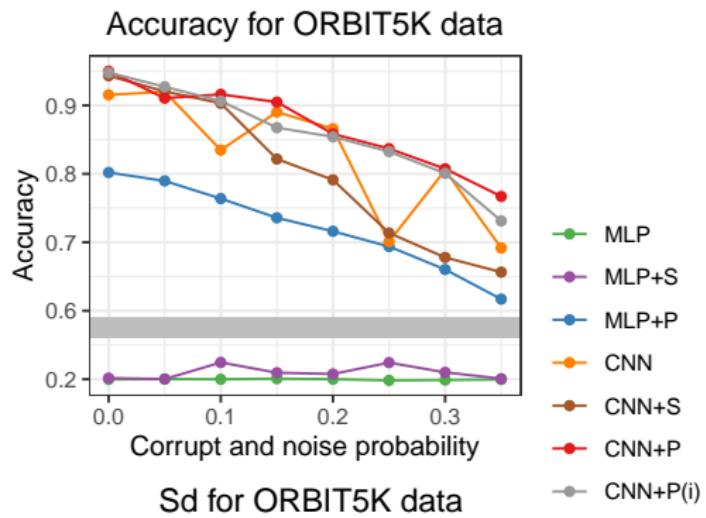
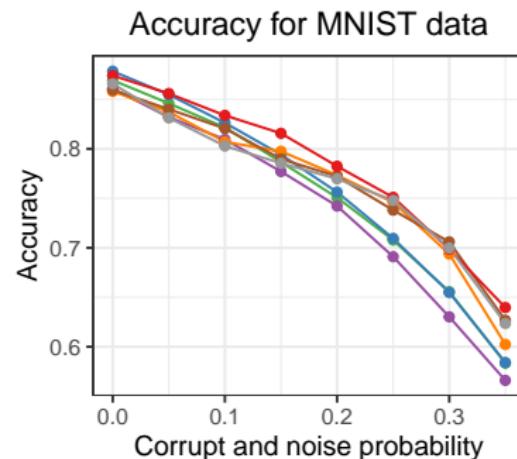


Build topological layer using Persistence Landscape

1. From data X , choose an appropriate simplicial complex K and a function f to compute the Persistence diagram \mathcal{D} .
2. From the persistence diagram \mathcal{D} , compute the persistence landscape $\lambda : \mathbb{N} \times \mathbb{R} \rightarrow \mathbb{R}$.
3. Compute the weighted average function $\bar{\lambda}_\omega(t) := \sum_{k=1}^{K_{\max}} \omega_k \lambda_k(t)$, and vectorize to get $\bar{\Lambda}_\omega \in \mathbb{R}^m$.
4. For a parametrized differentiable map $g_\theta : \mathbb{R}^m \rightarrow \mathbb{R}$, compute $S_{\theta, \omega}(\mathcal{D}) := g_\theta(\bar{\Lambda}_\omega)$.



Build topological layer using Persistence Landscape



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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

Circular Coordinates and its Applications

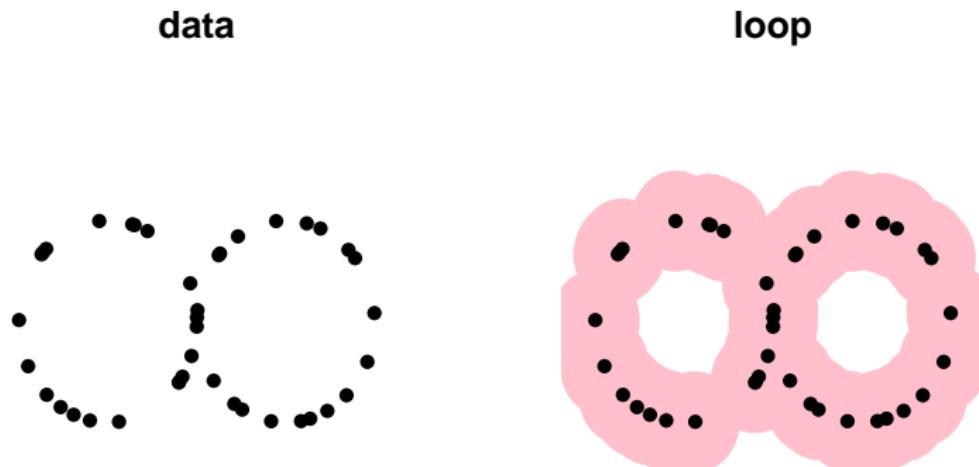
Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

References

Circular coordinates provide topological representations of reduced dimension.

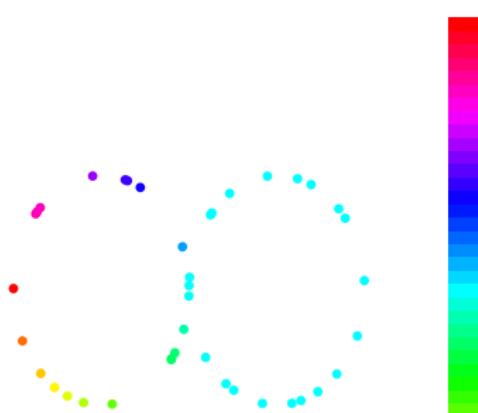
- ▶ Persistent cohomology and circular coordinates (de Silva, Morozov, Vejdemo-Johansson, 2011)
- ▶ Topological Learning for Motion Data via Mixed Coordinates (Vejdemo-Johansson, Pokorny, Skraba, Kragic, 2015)



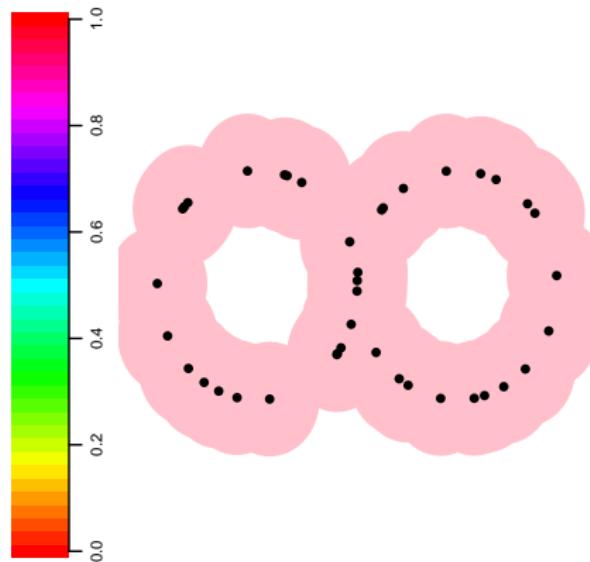
Circular coordinates provide topological representations of reduced dimension.

- circular coordinate is a function that maps from data points X to circle S^1 .

circular coordinates

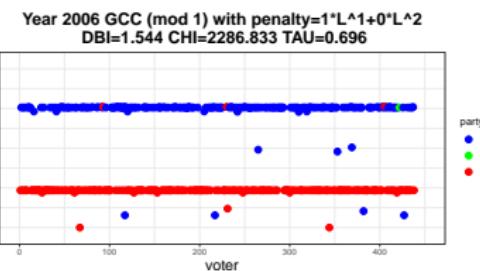
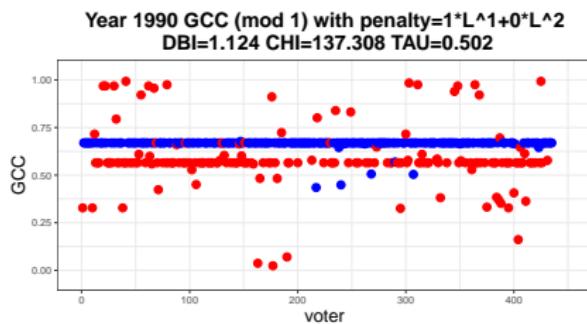


loop



Circular coordinates with generalized penalty better visualizes topological information from data.

- Generalized penalty for circular coordinate representation (Luo, Patania, Kim, Vejdemo-Johansson, 2021)



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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

Circular Coordinates and its Applications

Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

References

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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

Circular Coordinates and its Applications

Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

References

There are many programs for Topological Data Analysis.

- ▶ There are many programs for Topological Data Analysis: e.g., Dionysus, DIPHA, GUDHI, javaPlex, Perseus, PHAT, Ripser, TDA, TDAsstats

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.

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

Inference For Homological Features

Statistical Inference For Homological Features

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization of Topological Data Analysis

Persistence Landscape and its Applications

Circular Coordinates and its Applications

Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

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.
- Dominique Attali, André Lieutier, and David Salinas. Vietoris-Rips complexes also provide topologically correct reconstructions of sampled shapes. *Comput. Geom.*, 46(4):448–465, 2013. ISSN 0925-7721. doi: 10.1016/j.comgeo.2012.02.009. URL <https://doi.org/10.1016/j.comgeo.2012.02.009>.
- O. Bobrowski, S. Mukherjee, and J. E. Taylor. Topological consistency via kernel estimation. *ArXiv e-prints*, July 2014.
- Peter Bubenik. Statistical topological data analysis using persistence landscapes. *arXiv preprint arXiv:1207.6437*, 2012.
- Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primož Skraba. Scalar field analysis over point cloud data. *Discrete & Computational Geometry*, 46(4):743–775, 2011.

References ||

- Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.
- Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Annual Symposium on Computational Geometry*, pages 474–483. ACM, 2014.
- Vin de Silva, Dmitriy Morozov, and Mikael Vejdemo-Johansson. Persistent cohomology and circular coordinates. *Discrete & Computational Geometry*, 45(4):737–759, 2011.
- H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. Applied mathematics. American Mathematical Society, 2010. ISBN 9780821849255. URL
<http://books.google.com/books?id=MDXa6gFRZuIC>.
- 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>.

References III

- 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 1839–1847. Curran Associates, Inc., 2016. URL
<http://papers.nips.cc/paper/6508-statistical-inference-for-cluster-trees.pdf>.
- Jisu Kim, Alessandro Rinaldo, and Larry Wasserman. Minimax Rates for Estimating the Dimension of a Manifold. *ArXiv e-prints*, May 2019.

References IV

- Jisu Kim, Jaehyeok Shin, Frédéric Chazal, Alessandro Rinaldo, and Larry Wasserman. Homotopy Reconstruction via the Cech Complex and the Vietoris-Rips Complex. In Sergio Cabello and Danny Z. Chen, editors, *36th International Symposium on Computational Geometry (SoCG 2020)*, volume 164 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 54:1–54:19, Dagstuhl, Germany, 2020a. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. ISBN 978-3-95977-143-6. doi: 10.4230/LIPIcs.SoCG.2020.54. URL <https://drops.dagstuhl.de/opus/volltexte/2020/12212>.
- Kwangho Kim, Jisu Kim, Alessandro Rinaldo, and Frédéric Chazal. Time series featurization via topological data analysis: an application to cryptocurrency trend forecasting. *CoRR*, abs/1812.02987, 2020b. URL <http://arxiv.org/abs/1812.02987>.
- Kwangho Kim, Jisu Kim, Manzil Zaheer, Joon Sik Kim, Frédéric Chazal, and Larry Wasserman. PLLay: Efficient Topological Layer based on Persistent Landscapes. *arXiv e-prints*, art. arXiv:2002.02778, February 2020.

References V

- Hengrui Luo, Alice Patania, Jisu Kim, and Mikael Vejdemo-Johansson.
Generalized penalty for circular coordinate representation. *Foundations of Data Science*, 3(4):729–767, 2021.
- Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008. ISSN 0179-5376. doi: 10.1007/s00454-008-9053-2. URL <http://dx.doi.org/10.1007/s00454-008-9053-2>.
- Mikael Vejdemo-Johansson, Florian T Pokorny, Primoz Skraba, and Danica Kragic. Cohomological learning of periodic motion. *Applicable algebra in engineering, communication and computing*, 26(1):5–26, 2015.

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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips
Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence
Landscape

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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

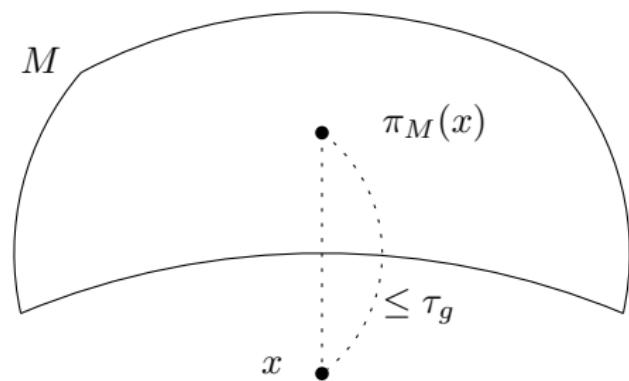
Statistical Inference on Persistence Homology and Persistence
Landscape

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 ∞ with respect to the uniform measure.

- ▶ Distribution P is absolutely continuous to induced Lebesgue measure vol_M , and $\frac{dP}{dvol_M} \leq K_p$ 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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips
Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence
Landscape

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

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

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

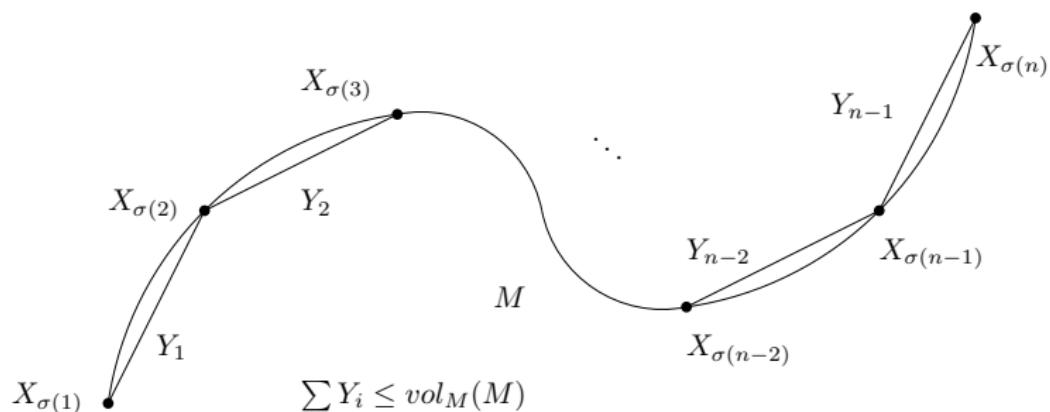
(Lemma 7) Let M be a d_1 -dimensional manifold with global reach $\geq \tau_g$ and local reach $\geq \tau_\ell$, and $X_1, \dots, X_n \in M$. Then there exists C which depends only on m , d_1 and K_I , and there exists $\sigma \in S_n$ such that

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

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 $\text{vol}_M(M)$.

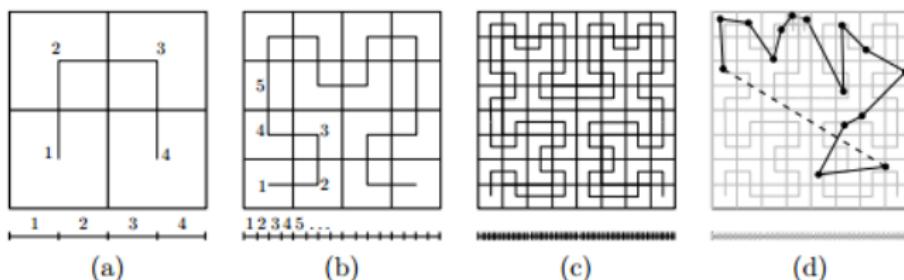


- Global reach $\geq \tau_g$ implies $\text{vol}_M(M)$ is bounded.

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.

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

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

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

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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips
Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence
Landscape

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 $Q_i \in co(\mathcal{P}_i)$, let q_i be density of Q_i with respect to measure ν . Then

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

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

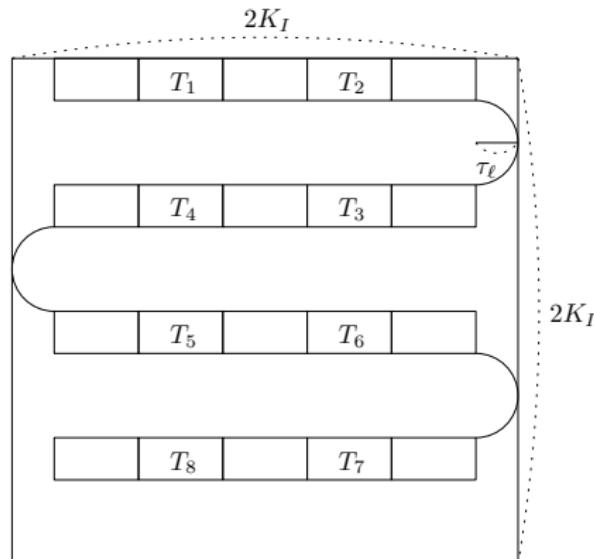
(Lemma 10, Le Cam's Lemma) Let \mathcal{P} be a set of probability measures, and $\mathcal{P}^{d_1}, \mathcal{P}^{d_2} \subset \mathcal{P}$ be such that for all $P \in \mathcal{P}^{d_i}$, $\theta(P) = \theta_i$ for $i = 1, 2$. For any $Q_i \in co(\mathcal{P}_i)$, let q_i be density of Q_i with respect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

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

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 .

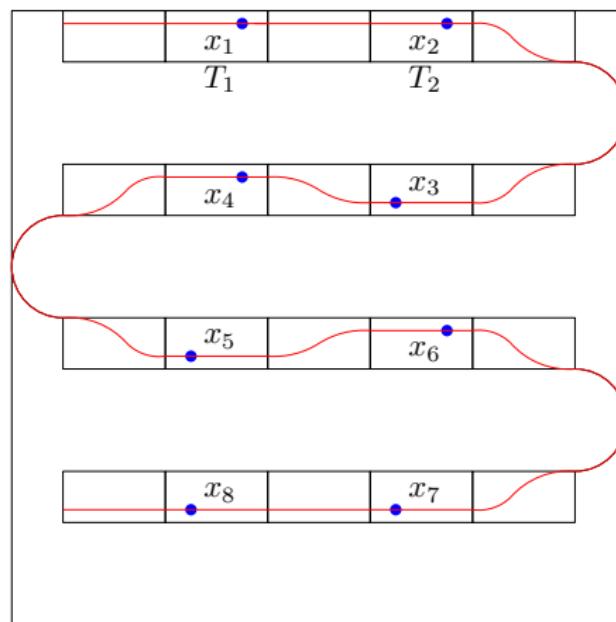
- ▶ T_i 's are cylinder sets in $[-K_I, K_I]^{d_2}$, and then T is constructed as

$T = S_n \prod_{i=1}^n T_i$, 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 $[-K_I, K_I]^{d_2}$.

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 $Q_1 \in co(\mathcal{P}_1^{d_1})$ and $Q_2 \in co(\mathcal{P}_2^{d_2})$ 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 κ_I , K_I , and $r_x > 0$ such that for all $r < r_x$,

$$Q_1(B(x_i, r)) \geq C Q_2(B(x_i, r)).$$

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

Proposition

(*Proposition 14*)

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

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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips
Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence
Landscape

Multinomial Classification and 0 – 1 Loss are Considered.



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

- ▶ 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 $x, y \in \mathbb{R}$, $\ell(x, y) = I(x = y)$.

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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

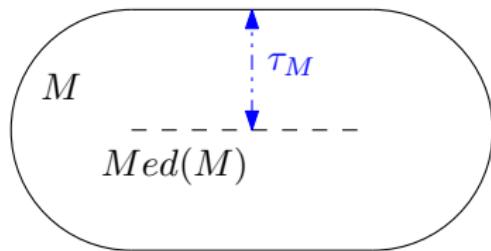
Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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



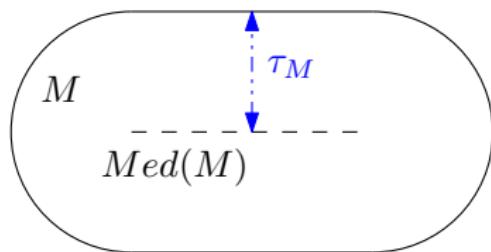
$$Med(M) = \{z \in \mathbb{R}^m : \text{there exists } p \neq q \in M \text{ with } \|p - z\| = \|q - z\| = d(z, M)\}.$$



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



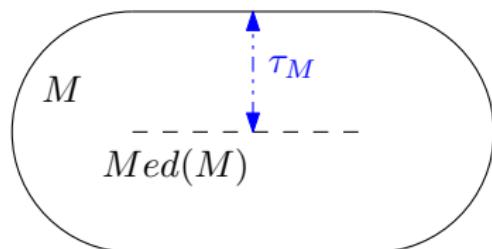
$$\tau_M = \inf_{x \in Med(M), y \in M} \|x - y\|.$$



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



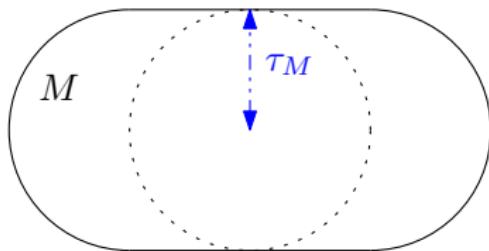
$$\tau_M = \inf_{x \in Med(M), y \in M} \|x - y\|.$$



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

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

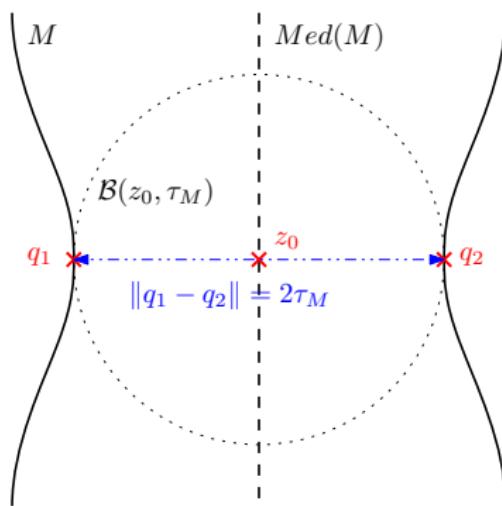
$$\tau_M = \inf_{q_2 \neq q_1 \in M} \frac{\|q_2 - q_1\|^2}{2d(q_2 - q_1, T_{q_1} M)}.$$



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 $z_0 \in Med(M)$ such that $q_1, q_2 \in \mathcal{B}(z_0, \tau_M)$ and $\|q_1 - q_2\| = 2\tau_M$.

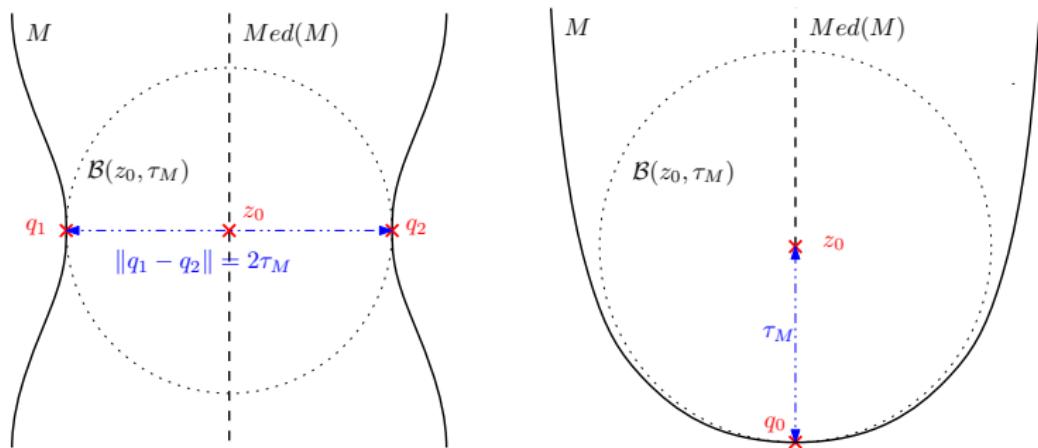


The reach is attained either from the bottleneck (global case) or the area of high curvature (local case).

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 γ_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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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.

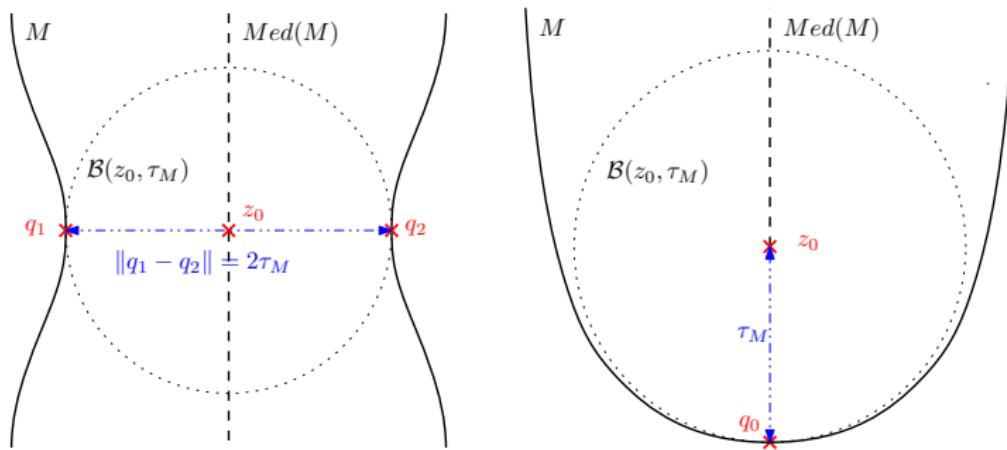
- ▶ The risk of the estimator $\hat{\tau}$ is the expected loss the estimator

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

- ▶ $\mathcal{X} = \{X_1, \dots, X_n\}$ 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.

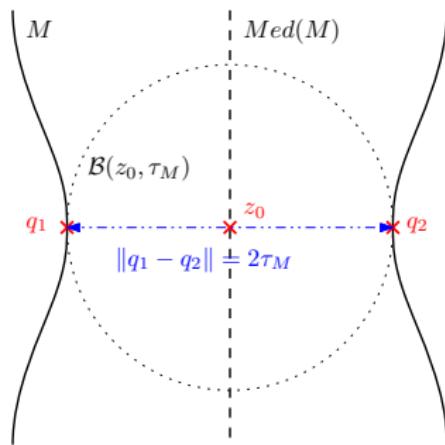


The reach estimator has the maximum risk of $O(n^{-\frac{q}{d}})$ for the global case.

Proposition

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

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

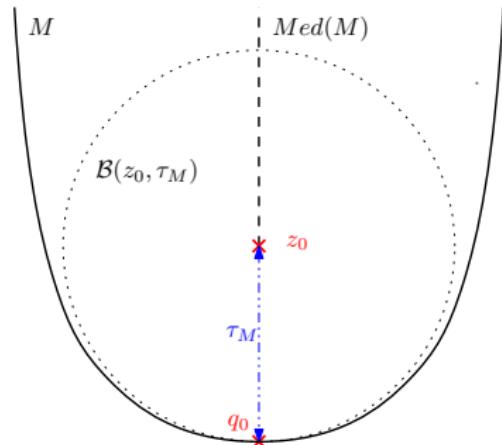


The reach estimator has the maximum risk of $O\left(n^{-\frac{2q}{3d-1}}\right)$ for the local case.

Proposition

(Proposition 4.7) Suppose there exists $q_0 \in M$ and a geodesic γ_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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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 $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^q$, $q \geq 1$.

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

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

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

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

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

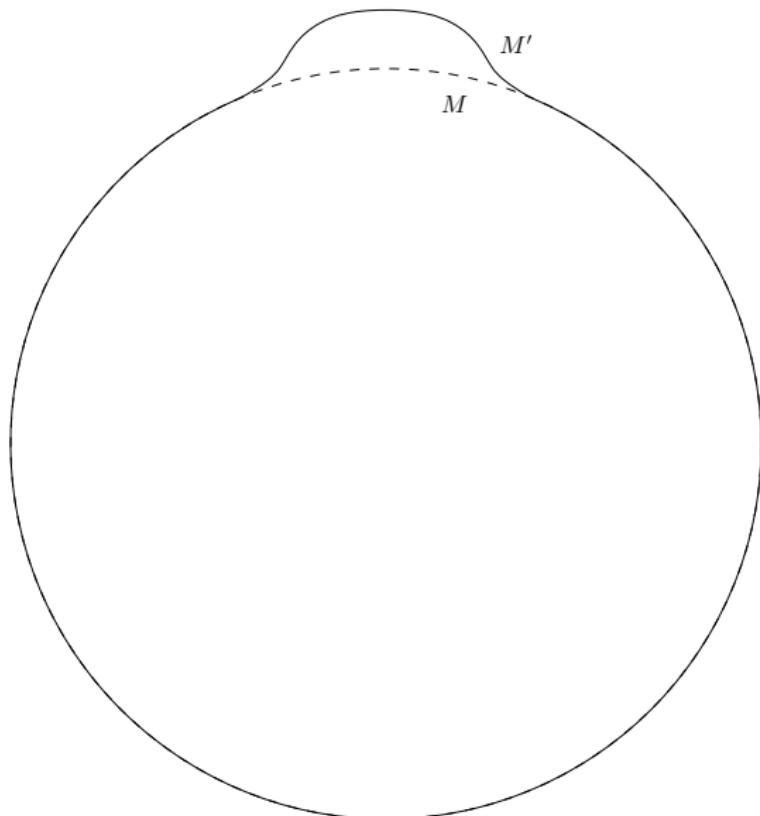
Two distributions P, P' are found so that their reaches differ but they are statistically difficult to distinguish.



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

- ▶ 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.



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 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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

$\check{\text{C}}\text{ech}$ complex is constructed by finding nonempty intersections of balls.

- For $\mathbb{X}, \mathcal{X} \subset \mathbb{R}^d$ and $r > 0$, the restricted $\check{\text{C}}\text{ech}$ complex $\check{\text{C}}\text{ech}_{\mathbb{X}}(\mathcal{X}, r)$ is defined as

$$\check{\text{C}}\text{ech}_{\mathbb{X}}(\mathcal{X}, r) = \left\{ \{x_1, \dots, x_k\} \subset \mathcal{X} : \bigcap_{j=1}^k \mathbb{B}_{\mathbb{X}}(x_j, r) \neq \emptyset \right\},$$

where $\mathbb{B}_{\mathbb{X}}(x, r) = \{y \in \mathbb{X} : \|y - x\| < r\}$ is the restricted ball of radius r centered at x .

```
## Error in plotUnionBalls(X = X[unionBallsSample, ], r = r, xlim =
xlim, : could not find function "plotUnionBalls"
## Error in plot.xy(xy.coords(x, y), type = type, ...): plot.new has not
been called yet
## Error in eval(expr, envir, enclos): object 'unionBallsSample' not
found
```

$\check{\text{C}}\text{ech}$ complex is constructed by finding nonempty intersections of balls.

- For $\mathbb{X}, \mathcal{X} \subset \mathbb{R}^d$ and $r > 0$, the restricted $\check{\text{C}}\text{ech}$ complex $\check{\text{C}}\text{ech}_{\mathbb{X}}(\mathcal{X}, r)$ is defined as

$$\check{\text{C}}\text{ech}_{\mathbb{X}}(\mathcal{X}, r) = \left\{ \{x_1, \dots, x_k\} \subset \mathcal{X} : \bigcap_{j=1}^k \mathbb{B}_{\mathbb{X}}(x_j, r) \neq \emptyset \right\},$$

where $\mathbb{B}_{\mathbb{X}}(x, r) = \{y \in \mathbb{X} : \|y - x\| < r\}$ is the restricted ball of radius r centered at x .

```
## Error in plotUnionBalls(X = X[unionBallsSample, ], r = r, xlim =
xlim, : could not find function "plotUnionBalls"
## Error in plot.xy(xy.coords(x, y), type = type, ...): plot.new has not
been called yet
## Error in eval(expr, envir, enclos): object 'unionBallsSample' not
found
```

\check{C} ech complex is constructed by finding nonempty intersections of balls.

- For $\mathbb{X}, \mathcal{X} \subset \mathbb{R}^d$ and $r > 0$, the restricted \check{C} ech complex $\check{C}ech_{\mathbb{X}}(\mathcal{X}, r)$ is defined as

$$\check{C}ech_{\mathbb{X}}(\mathcal{X}, r) = \left\{ \{x_1, \dots, x_k\} \subset \mathcal{X} : \bigcap_{j=1}^k \mathbb{B}_{\mathbb{X}}(x_j, r) \neq \emptyset \right\},$$

where $\mathbb{B}_{\mathbb{X}}(x, r) = \{y \in \mathbb{X} : \|y - x\| < r\}$ is the restricted ball of radius r centered at x .

```
## Error in plotUnionBalls(X = X[unionBallsSample, ], r = r, xlim =
xlim, : could not find function "plotUnionBalls"
## Error in plot.xy(xy.coords(x, y), type = type, ...): plot.new has not
been called yet
## Error in eval(expr, envir, enclos): object 'unionBallsSample' not
found
```

$\check{\text{C}}\text{ech}$ complex is constructed by finding nonempty intersections of balls.

- ▶ For $\mathbb{X}, \mathcal{X} \subset \mathbb{R}^d$ and $r > 0$, the restricted $\check{\text{C}}\text{ech}$ complex $\check{\text{C}}\text{ech}_{\mathbb{X}}(\mathcal{X}, r)$ is defined as

$$\check{\text{C}}\text{ech}_{\mathbb{X}}(\mathcal{X}, r) = \left\{ \{x_1, \dots, x_k\} \subset \mathcal{X} : \bigcap_{j=1}^k \mathbb{B}_{\mathbb{X}}(x_j, r) \neq \emptyset \right\},$$

where $\mathbb{B}_{\mathbb{X}}(x, r) = \{y \in \mathbb{X} : \|y - x\| < r\}$ is the restricted ball of radius r centered at x .

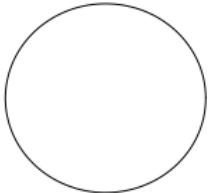
- ▶ The ambient $\check{\text{C}}\text{ech}$ complex $\check{\text{C}}\text{ech}_{\mathbb{R}^d}(\mathcal{X}, r)$ is when $\mathbb{X} = \mathbb{R}^d$.

```
## Error in plotUnionBalls(X = X[unionBallsSample, ], r = r, xlim =
xlim, : could not find function "plotUnionBalls"
## Error in plot.xy(xy.coords(x, y), type = type, ...): plot.new has not
been called yet
## Error in eval(expr, envir, enclos): object 'unionBallsSample' not
found
```

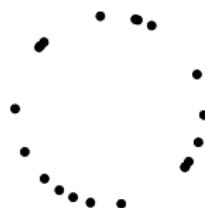
The homotopy of the Vietoris-Rips complex on samples \mathcal{X} can be very different from the homotopy of the target space $\mathbb{X} \subset \mathbb{R}^d$.

- ▶ If r is too small, then $\text{Rips}(\mathcal{X}, r)$ is a set of disconnected points

Underlying circle



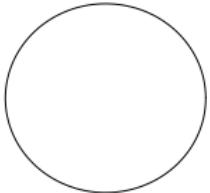
Rips complex, r small



The homotopy of the Vietoris-Rips complex on samples \mathcal{X} can be very different from the homotopy of the target space $\mathbb{X} \subset \mathbb{R}^d$.

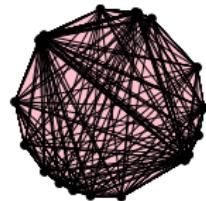
- ▶ If r is too large, then $\text{Rips}(\mathcal{X}, r)$ is homotopy equivalent to a point.

Underlying circle



\neq

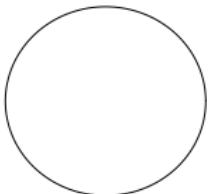
Rips complex, r large



The homotopy of the Vietoris-Rips complex on samples \mathcal{X} can be very different from the homotopy of the target space $\mathbb{X} \subset \mathbb{R}^d$.

- ▶ If r is too small, then $\text{Rips}(\mathcal{X}, r)$ is a set of disconnected points
- ▶ If r is too large, then $\text{Rips}(\mathcal{X}, r)$ is homotopy equivalent to a point.

Underlying circle

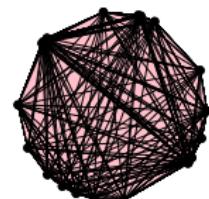


✗

Rips complex, r small



Rips complex, r large



The homotopy of a positive reach set is reconstructed from the ambient Čech complex.

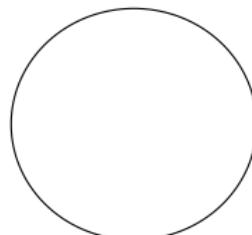
Theorem

Let $\mathbb{X} \subset \mathbb{R}^d$ be a subset with reach $\tau > 0$ and $\mathcal{X} \subset \mathbb{R}^d$ be a closed discrete set of points. Let $r > 0$, $\epsilon := \max_{x \in \mathcal{X}} \{d_{\mathbb{X}}(x)\}$, and $\delta > 0$ be satisfying

$$\mathbb{X} \subset \bigcup_{x \in \mathcal{X}} \mathbb{B}_{\mathbb{R}}(x, \delta) \quad \text{and} \quad r \leq \tau - \epsilon.$$

Then with sufficiently small δ , the ambient Čech complex $\check{\text{Cech}}_{\mathbb{R}^d}(\mathcal{X}, r)$ is homotopy equivalent to \mathbb{X} .

Underlying circle



\approx

```
## Error in eval(expr, envir,
enclos): object 'unionBallsSample'
not found
```

The homotopy of a positive reach set is reconstructed from the Vietoris-Rips complex.

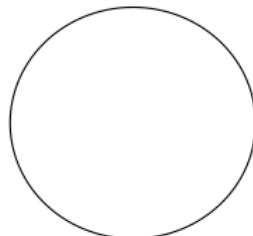
Theorem

Let $\mathbb{X} \subset \mathbb{R}^d$ be a subset with reach $\tau > 0$ and $\mathcal{X} \subset \mathbb{R}^d$ be a closed discrete set of points. Let $r > 0$, $\epsilon := \max_{x \in \mathcal{X}} \{d_{\mathbb{X}}(x)\}$, and $\delta > 0$ be satisfying

$$\mathbb{X} \subset \bigcup_{x \in \mathcal{X}} \mathbb{B}_{\mathbb{R}}(x, \delta) \quad \text{and} \quad r \leq \sqrt{\frac{d+1}{2d}}(\tau - \epsilon).$$

Then with sufficiently small δ , the Vietoris-Rips complex $\text{Rips}(\mathcal{X}, r)$ is homotopy equivalent to \mathbb{X} .

Underlying circle



\approx

Vietoris-Rips complex



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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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

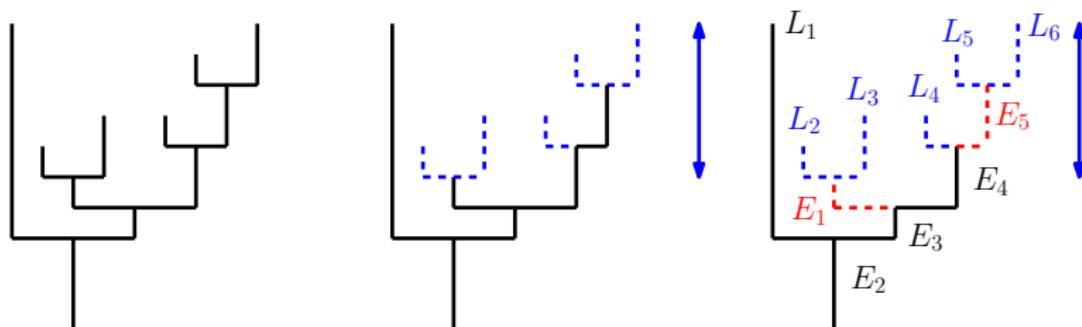
Definition

The ℓ_∞ metric between trees are defined as

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

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

- ▶ We propose two pruning schemes to find trees that are simpler than the empirical tree $T_{\hat{p}_h}$ 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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence
Landscape

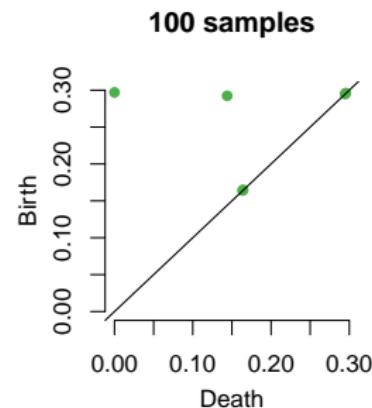
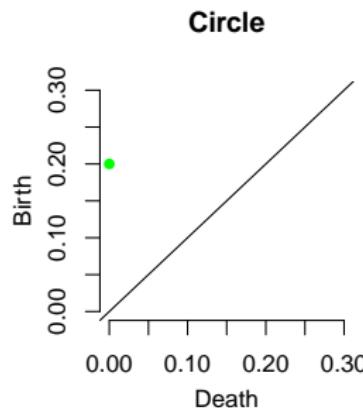
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

where γ ranges over all bijections from D_1 to D_2 .



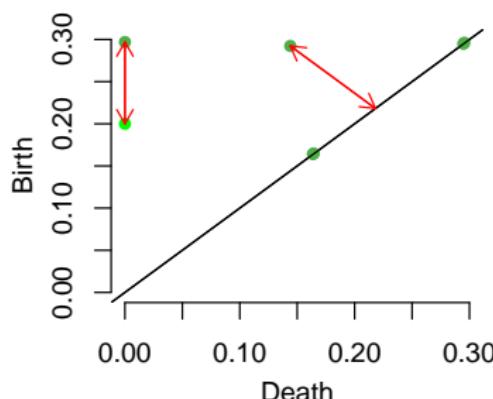
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

where γ ranges over all bijections from D_1 to D_2 .



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

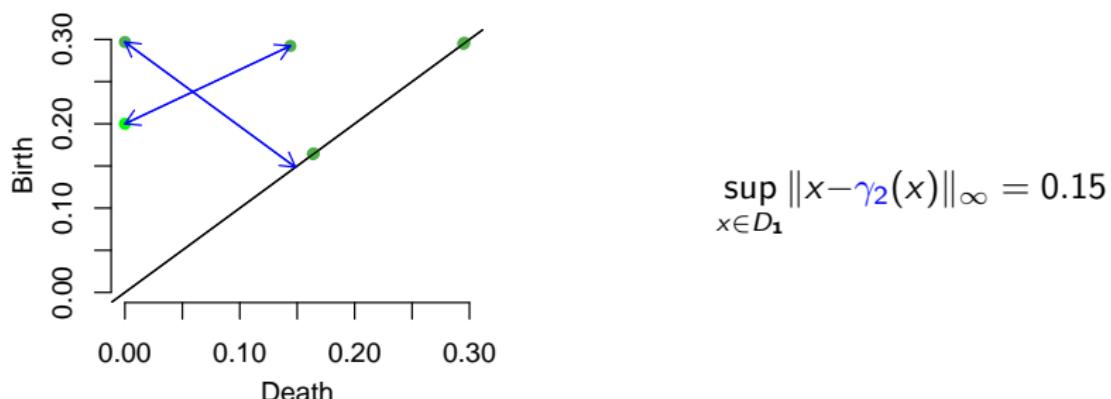
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

where γ ranges over all bijections from D_1 to D_2 .



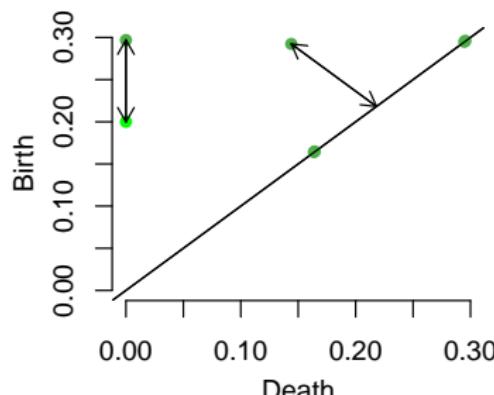
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

where γ ranges over all bijections from D_1 to D_2 .



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

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

Theorem

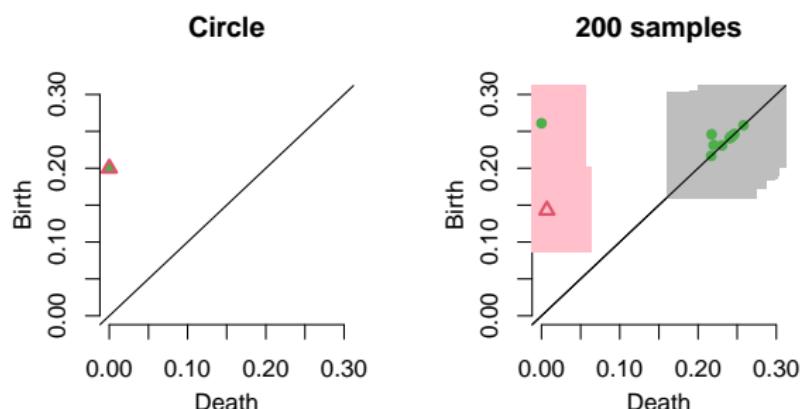
[Edelsbrunner and Harer, 2010][Chazal, de Silva, Glisse, and Oudot, 2012] Let \mathbb{X} be finitely triangulable space and $f, g : \mathbb{X} \rightarrow \mathbb{R}$ be two continuous functions. Let $Dgm(f)$ and $Dgm(g)$ be corresponding persistence diagrams. Then

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

Confidence 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

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



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

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

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

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

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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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

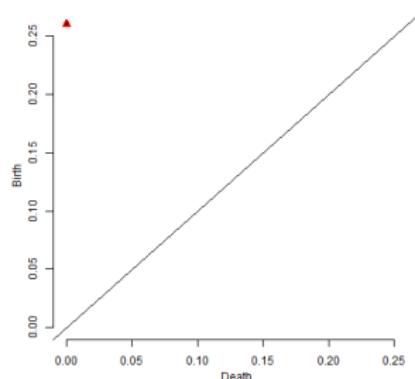
$$\{D_L\}_{L>0}, \text{ where } D_L := \{x \in \text{supp}(P) : p_h(x) \geq L\}.$$

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 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

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

We estimate the target 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 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,

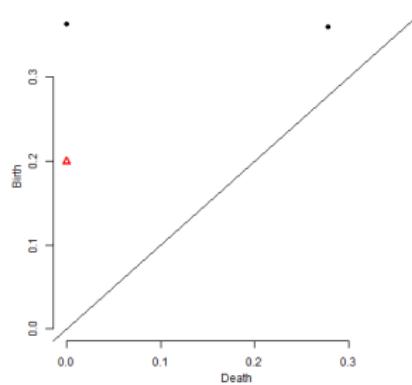
$$\left\{ \text{Rips} \left(\mathcal{X}_{n,L}^{\hat{p}_h}, r \right) \right\}_{L>0}, \text{ where } \mathcal{X}_{n,L}^{\hat{p}_h} = \{X_i \in \mathcal{X}_n : \hat{p}_h(X_i) \geq L\}.$$

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

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

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

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



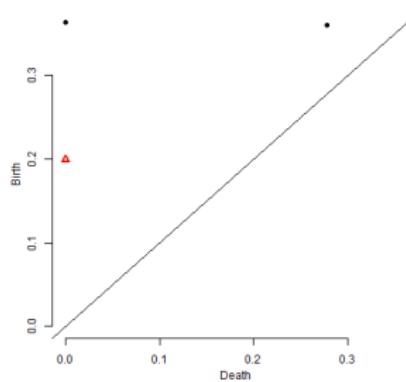
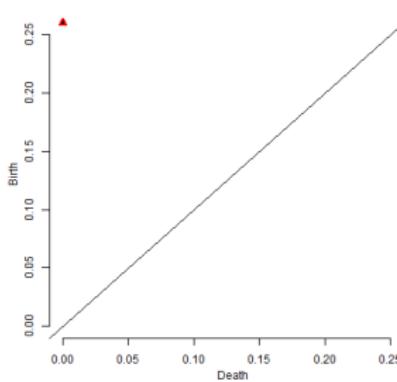
We estimate the target persistent homology by the persistent homology of the KDE filtration on Vietoris-Rips complexes.

- ▶ We estimate the target persistent homology

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

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

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



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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

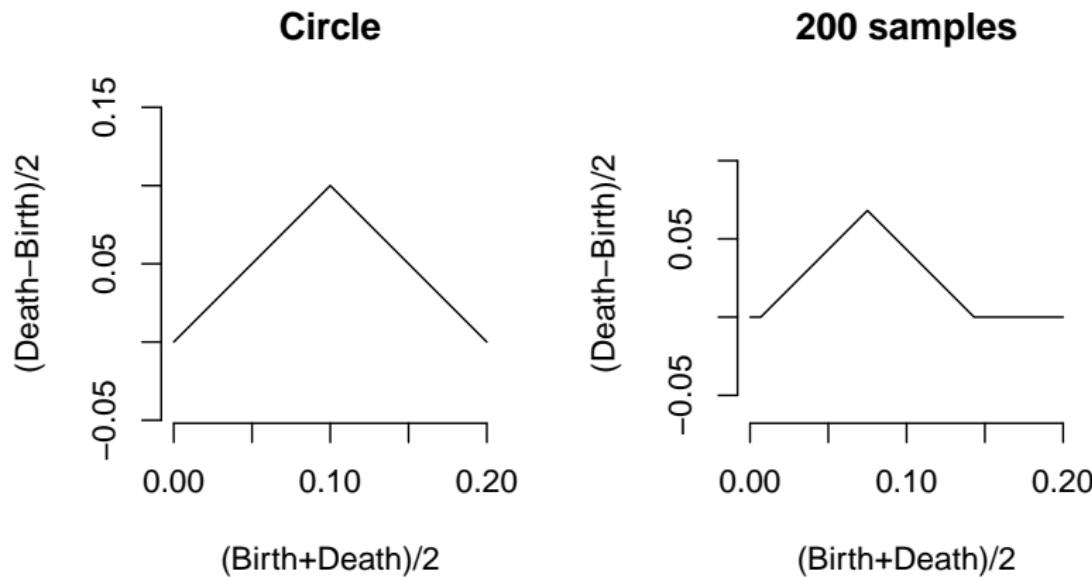
R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

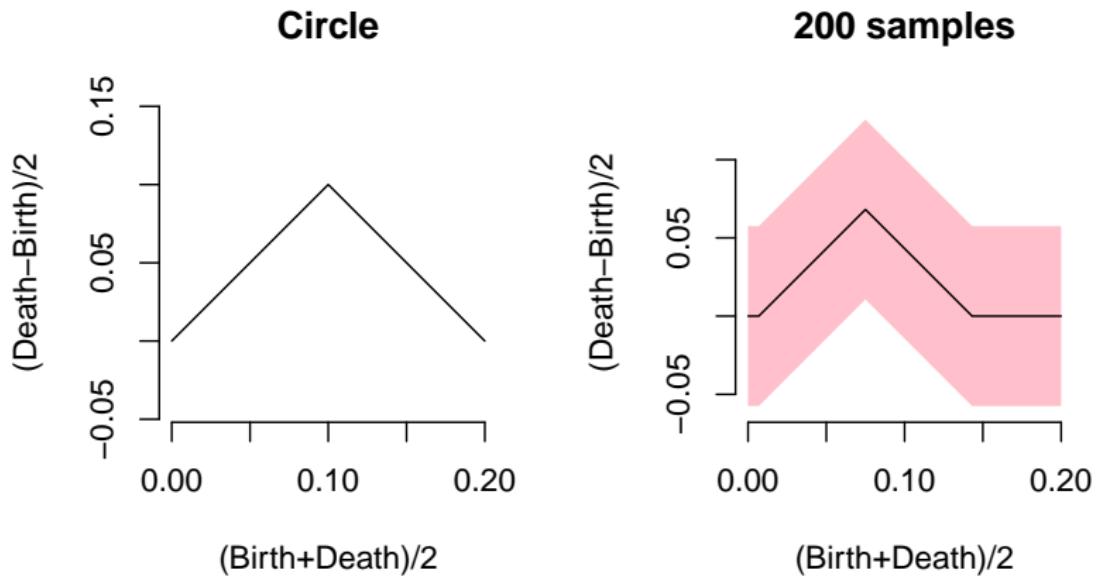
Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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



Confidence band for persistent homology quantifies the randomness of the persistence landscape.

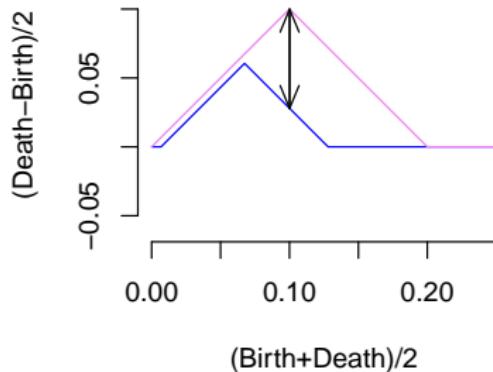


∞ -landscape distance gives a metric on the space of persistence landscapes.

Definition

[Bubenik, 2012] Let D_1, D_2 be multiset of points, and λ_1, λ_2 be corresponding persistence landscapes. ∞ -landscape distance is defined as

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



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

Theorem

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

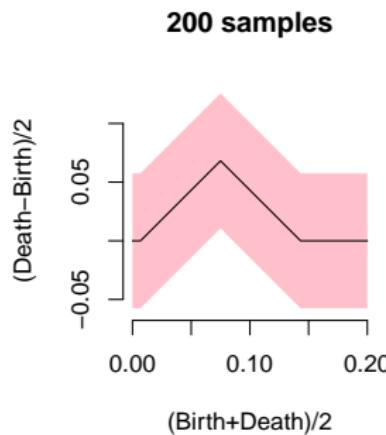
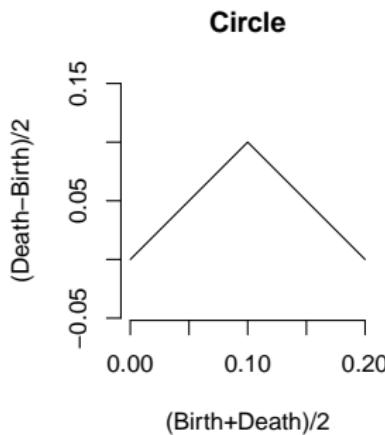
$$\Lambda_\infty(Dgm(f), Dgm(g)) \leq \|f - g\|_\infty.$$

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

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

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

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



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

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

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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence
Landscape

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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

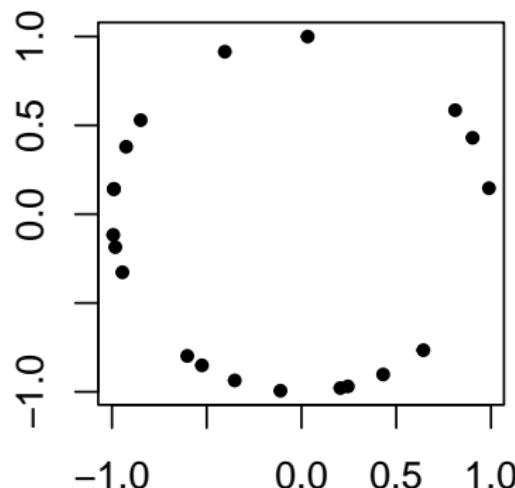
Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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")
```

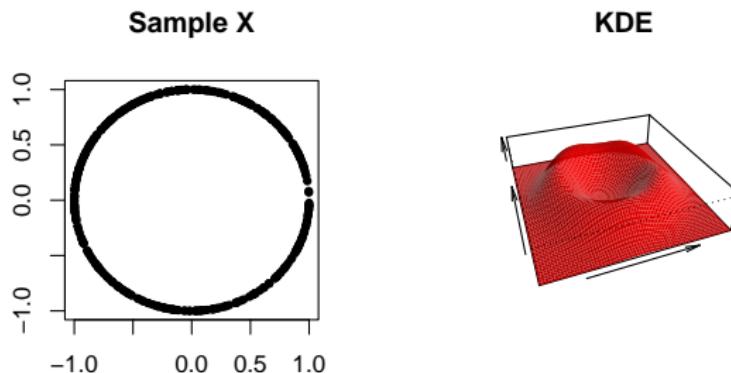
R Package TDA provides KDE function over a grid.

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



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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence
Landscape

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.

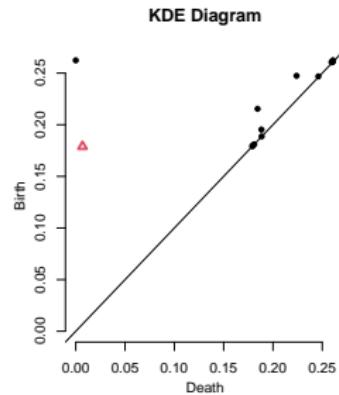
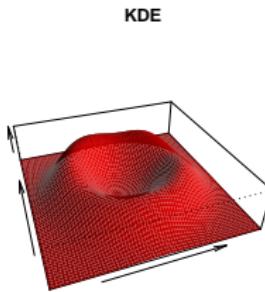
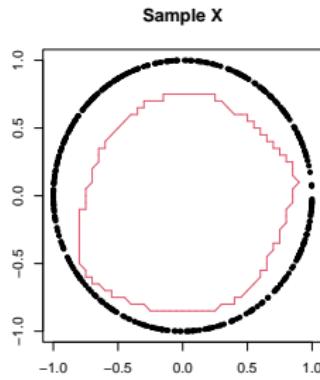
R Package TDA computes Persistent Homology over a grid.

```
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.

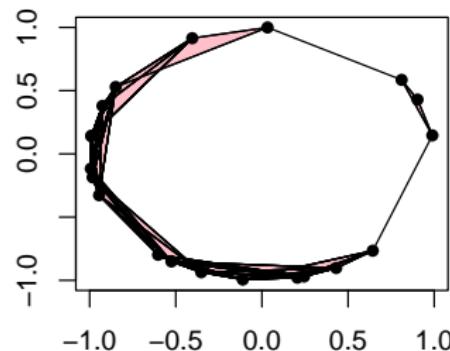
- ▶ The function `gridDiag()` computes the persistent homology of sublevel (and superlevel) sets of the input function.
 - ▶ `gridDiag()` evaluates the real valued input function over a grid.
 - ▶ `gridDiag()` constructs a filtration of simplices using the values of the input function.
 - ▶ `gridDiag()` 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 $2r$ apart, i.e.

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



- ▶ Rips filtration is formed by Rips complexes with gradually increasing r .

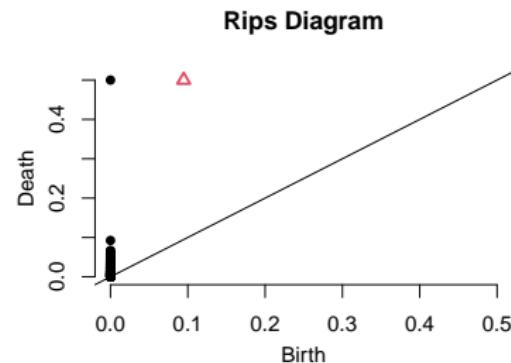
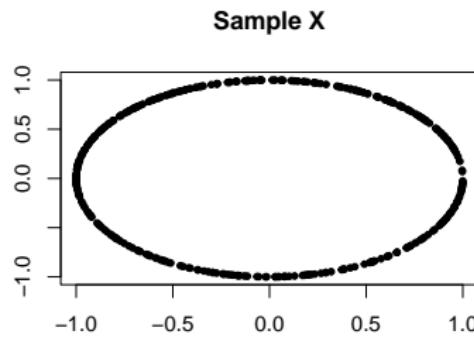
R Package TDA computes Vietoris-Rips Persistent Homology.

- ▶ 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.



R Package TDA computes Persistence Landscape.

- ▶ Let Λ_p be created by tenting each point $p = (x, y) = \left(\frac{b+d}{2}, \frac{d-b}{2}\right)$ representing a birth-death pair (b, d) in the persistence diagram D .
- ▶ The persistence landscape of D is the collection of functions

$$\lambda_k(t) = k \max_p \Lambda_p(t), \quad t \in [0, T], k \in \mathbb{N},$$

where $k \max$ is the k th largest value in the set.

- ▶ The function `landscape()` evaluates the persistence landscape function $\lambda_k(t)$.

```
tseq <- seq(0, 0.2, length = 1000)
Land <- landscape(DiagGrid[["diagram"]], dimension = 1, KK = 1, tseq = tseq)

par(mfrow = c(1,2))
plot(x = DiagGrid[["diagram"]], main = "KDE Diagram")
plot(tseq, Land, type = "l", xlab = "(Birth+Death)/2",
      ylab = "(Death-Birth)/2", asp = 1, axes = FALSE, main = "Landscape")
axis(1); axis(2)
```

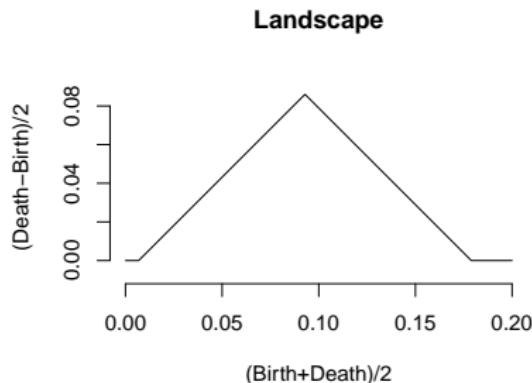
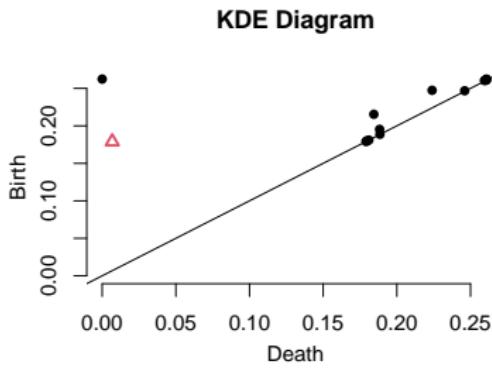
R Package TDA computes Persistence Landscape.

- ▶ Let Λ_p be created by tenting each point $p = (x, y) = (\frac{b+d}{2}, \frac{d-b}{2})$ representing a birth-death pair (b, d) in the persistence diagram D .
- ▶ The persistence landscape of D is the collection of functions

$$\lambda_k(t) = k \max_p \Lambda_p(t), \quad t \in [0, T], k \in \mathbb{N},$$

where $k \max$ is the k th largest value in the set.

- ▶ The function `landscape()` evaluates the persistence landscape function $\lambda_k(t)$.



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

Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex

Statistical Inference for Cluster Trees

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence
Landscape

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

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

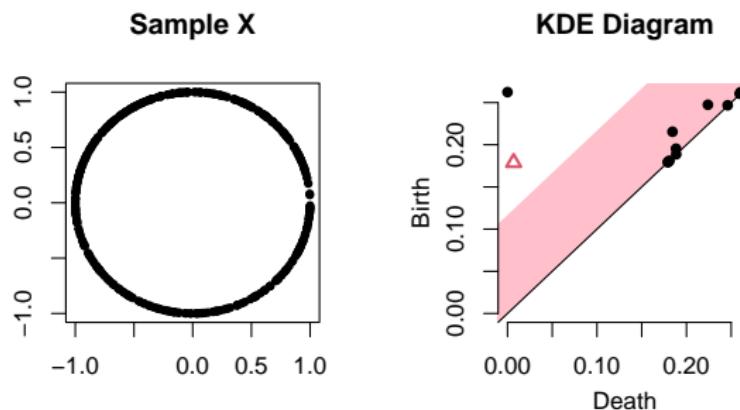
```
bandKDE <- bootstrapBand(X = X, FUN = kde, Grid = Grid, B = 20,
    parallel = FALSE, alpha = 0.1, h = h)
print(bandKDE[["width"]])

##          90%
## 0.05836494
```

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

The $(1 - \alpha)$ bootstrap confidence band for $\mathbb{E}[\hat{p}_h]$ 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")
```



R Package TDA computes the bootstrap confidence band for the persistence landscape.

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

```
par(mfrow = c(1,2))
plot(X, xlab = "", ylab = "", main = "Sample X", pch = 20)
plot(tseq, Land, type = "l", xlab = "(Birth+Death)/2",
      ylab = "(Death-Birth)/2", asp = 1, axes = FALSE, main = "200 samples")
axis(1); axis(2)
polygon(c(tseq, rev(tseq)), c(Land - bandKDE[["width"]],
                                rev(Land + bandKDE[["width"]])), col = "pink", lwd = 1.5,
                                border = NA)
lines(tseq, Land)
```

R Package TDA computes the bootstrap confidence band for the persistence landscape.

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

