Statistical Inference and application to Machine Learning
For Geometric and Topological Data

Jisu KIM

KAIST

2021-02-05

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

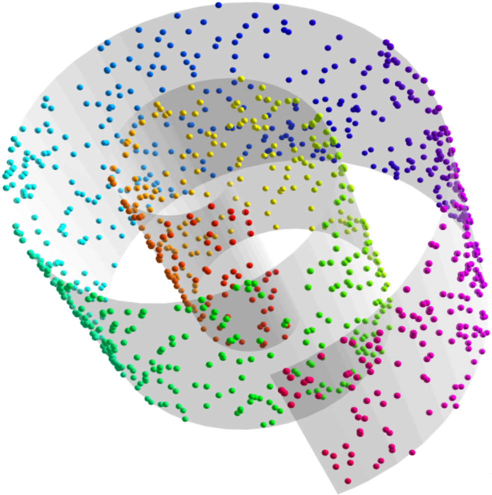
Topological Data Analysis: Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Machine Learning

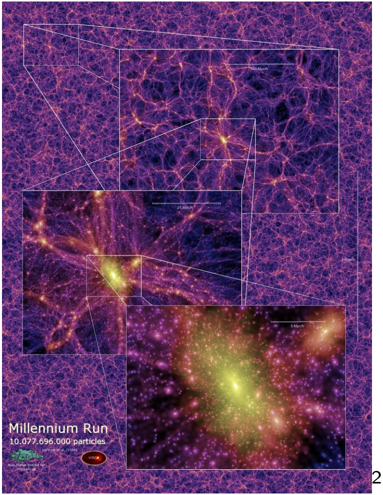
- Featurization of Topological Data Analysis using Persistence Landscapes

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]

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

Statistical Inference and application to Machine Learning For Geometric and Topological Data are 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)
- General Introduction to Topological Data Analysis
  - Computational Topology: An Introduction (Edelsbrunner, Harer, 2010)
  - Topological Data Analysis (Wasserman, 2016)
  - An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists (Chazal, Michel, 2017)
- Statistical Inference for Persistent Homology
  - Confidence sets for persistence diagrams (Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2014)
- Application of Topological Data Analysis to Machine Learning
  - Time Series Featurization via Topological Data Analysis (Kim, Kim, Rinaldo, Chazal, 2020)
  - Efficient Topological Layer based on Persistence Landscapes (Kim, Kim, Zaheer, Kim, Chazal, Wasserman, 2020)

Introduction

Minimax Rates for Geometric Parameters of a Manifold

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

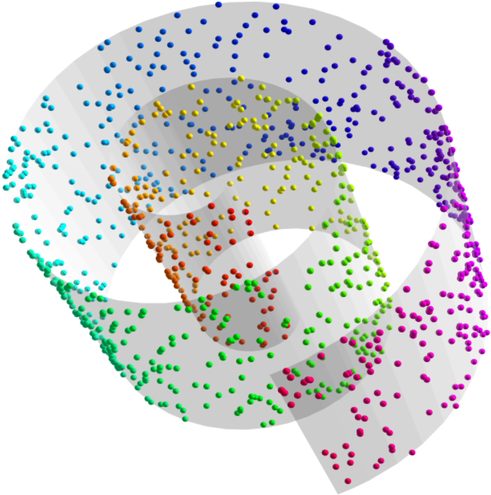
Topological Data Analysis: Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Machine Learning

- Featurization of Topological Data Analysis using Persistence Landscapes

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.
- $$\sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[\ell \left(\hat{\theta}_n(X),\ \theta(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$ 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.
- $$R_n = \inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[\ell \left(\hat{\theta}_n(X),\ \theta(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$ 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

Topological Data Analysis: Persistent Homology

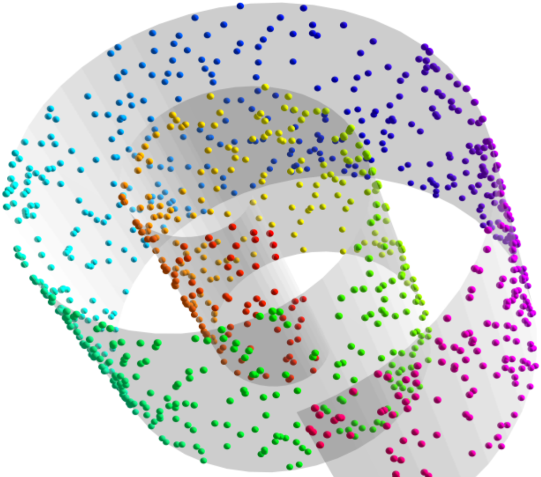
Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Machine Learning

- Featurization of Topological Data Analysis using Persistence Landscapes

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

- $$R_n = \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[1 \left(\hat{\dim}_n(X) \neq \dim(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$ 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 $x, y \in \mathbb{R}$, $\ell(x, y) = 1(x \neq y)$.

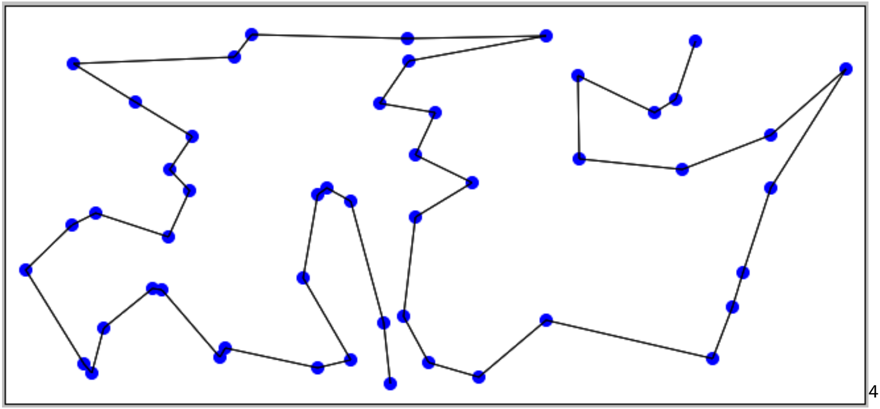
Minimax rate for estimating the dimension: we first consider dimension d_1 vs d_2.

- $$R_n = \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[1 \left(\hat{\dim}_n(X) \neq \dim(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$ is drawn from a fixed distribution P, where P is contained in set of distributions $\mathcal{P} = \mathcal{P}^{d_1} \cup \mathcal{P}^{d_2}$, where $\mathcal{P}^d$ is a set of d-dimensional distributions..
  - estimator $\hat{\dim}_n$ is any function of data X.
  - $0 - 1$ loss function is considered, so for all $x, y \in \mathbb{R}$, $\ell(x, y) = 1(x \neq y)$.

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

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

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



[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 dimesion is higher, length of TSP path is likely to be longer.
- $$\hat{\dim}_n(X) = d_1 \iff$$
$$\min_{\sigma \in S_n} \sum_{i=1}^{n-1} \|X_{\sigma(i+1)} - X_{\sigma(i)}\|_{\mathbb{R}^m}^{\boldsymbol{d_1}} \leq C,$$

  where C is some constant.

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

Proposition

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

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

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

Lemma

(Lemma 10, Le Cam's Lemma) Let $\mathcal{P}$ be a set of probability measures, and $\mathcal{P}^{\mathbf{d_1}}, \mathcal{P}^{\mathbf{d_2}} \subset \mathcal{P}$ be such that for all $P \in \mathcal{P}^{\mathbf{d}_i}$, $\theta(P) = \theta_i$ 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

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

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

Proposition

(Proposition 14)

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

Minimax rate for estimating the dimension

Theorem

(Proposition 16 and 17)

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

- 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

- Topological Data Analysis: Persistent Homology

- Statistical Inference for Persistent Homology

- Application of Topological Data Analysis to Machine Learning
  - Featurization of Topological Data Analysis using Persistence Landscapes

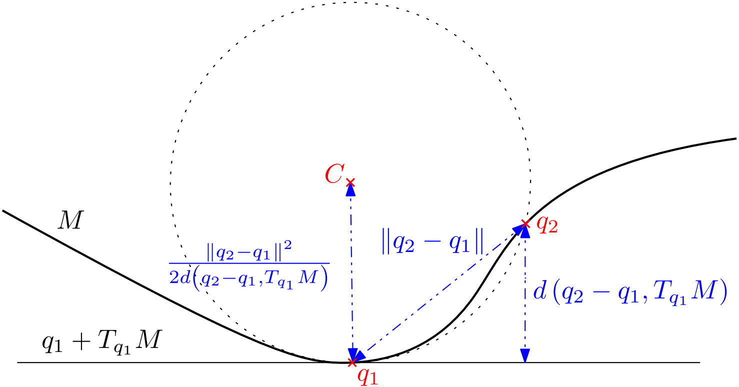
The reach is the maximum radius of a ball that can roll over the manifold.

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 paramter in:
  - Dimension estimation
  - Homology inference
  - Volume estimation
  - Manifold clustering
  - Diffusion maps

Minimax rate for estimating the reach

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

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

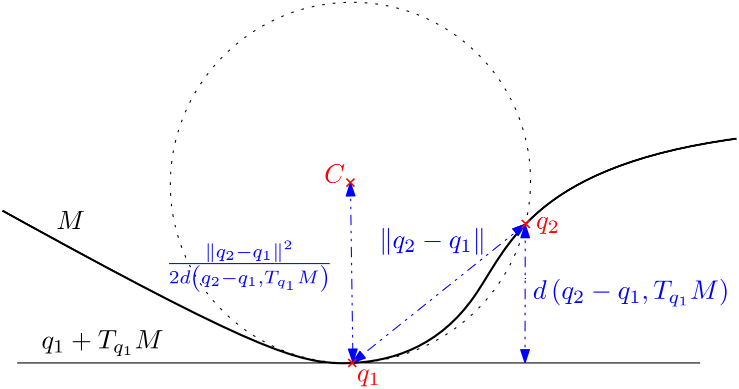
The maximum risk of our estimator provides an upper bound on the minimax rate.

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

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

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

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


C
M
$\frac{\|q_2 - q_1\|^2}{2d\left(q_2 - q_1, T_{q_1}M\right)}$
$\|q_2 - q_1\|$
q_2
$d\left(q_2 - q_1, T_{q_1}M\right)$
$q_1 + T_{q_1}M$
q_1

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

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

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

Proposition

(Proposition 5.6)

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

Minimax rate for estimating the reach

Theorem

(Theorem 5.1 and Proposition 5.6)

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

- 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
- Topological Data Analysis: Persistent Homology
- Statistical Inference for Persistent Homology
- Application of Topological Data Analysis to Machine Learning
    - Featurization of Topological Data Analysis using Persistence Landscapes

The number of holes is used to summarize topological 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 $dim \geq 3$

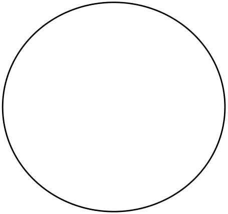
Example : Objects are classified by homologies.

1. $\beta_0 =\#$ of connected components ●
2. $\beta_1 =\#$ of loops ○

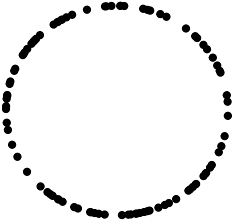
$\beta_0 \setminus \beta_1$	0	1	2
1	ㄱ, ㄴ, ㄷ, ㄹ, ㅅ, ㅈ, ㅋ, ㅌ	ㅁ, ㅇ, ㅂ, ㅍ, A	あ
2	ㅊ, 字		
3		ㅎ	

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:

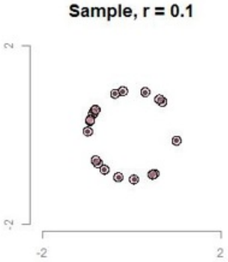


Underlying circle: $\beta_0 = 1$, $\beta_1 = 1$



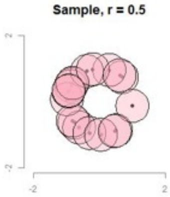
100 samples: $\beta_0 = 100$, $\beta_1 = 0$

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

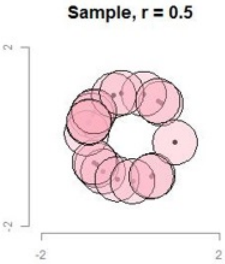

Sample, r = 0.1

Augmented Set

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

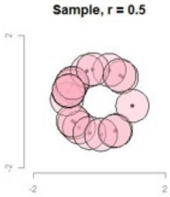


r=**0.5** : 1-dim hole is formed

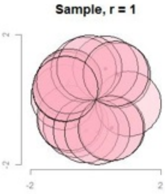


Augmented Set

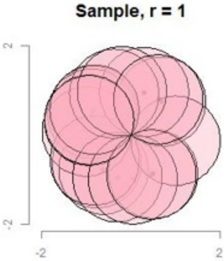
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



r=**0.5** : 1-dim hole is formed

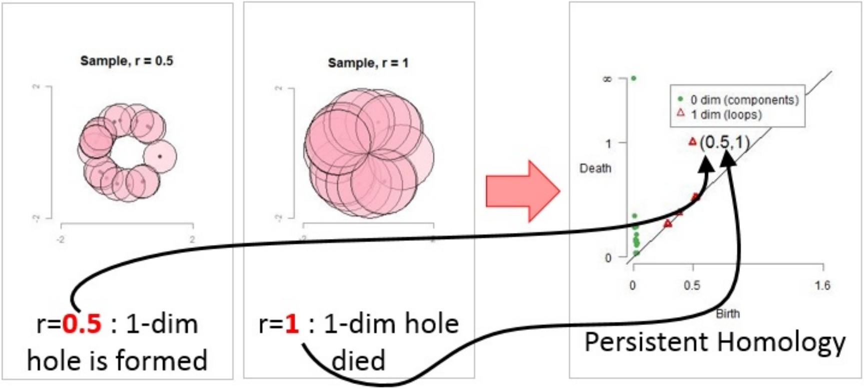


r=**1** : 1-dim hole died



Augmented Set

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



r=**0.5** : 1-dim hole is formed

r=**1** : 1-dim hole died

Persistent Homology

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

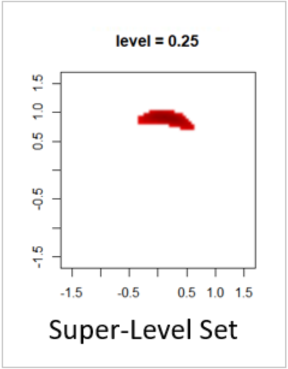
- The kernel density estimator is

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

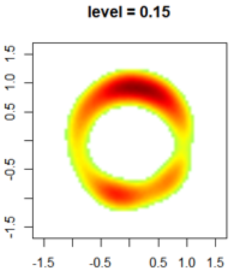
- We look at superlevel sets of the kernel density estimator as

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

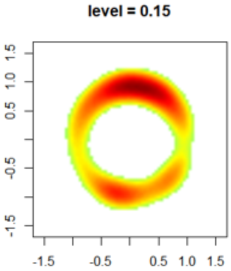
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



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

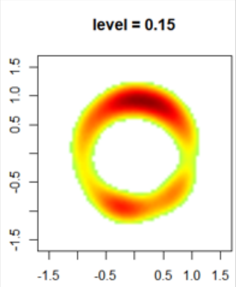


L=0.15 : 1-dim hole is formed

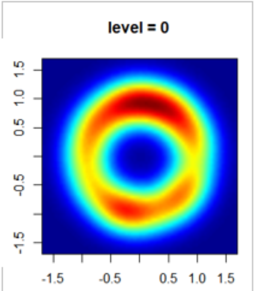


Super-Level Set

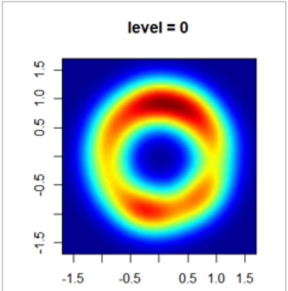
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



L=**0.15** : 1-dim hole is formed

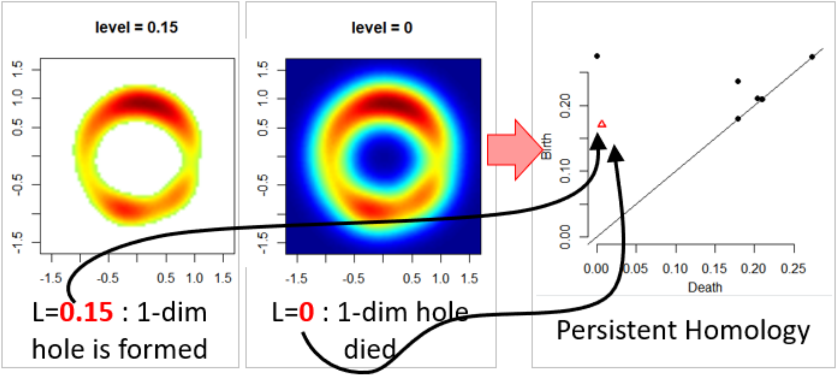


L=**0** : 1-dim hole died



Super-Level Set

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

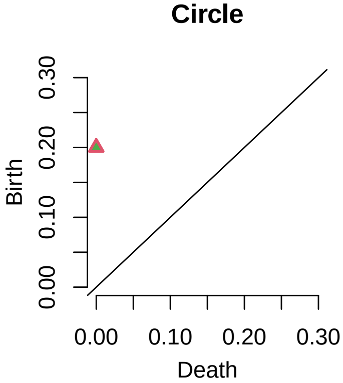


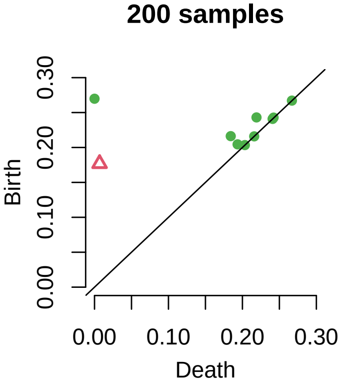
L=0.15 : 1-dim hole is formed

L=0 : 1-dim hole died

Persistent Homology

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





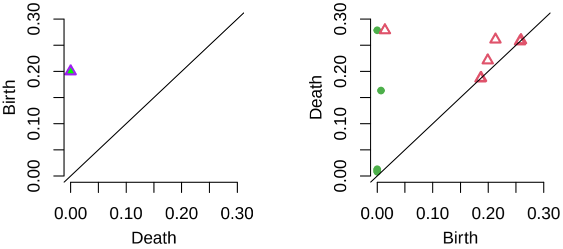
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

$$d_B(D_1,\, D_2) = \inf_{\gamma} \sup_{x \in D_\mathbf{1}} \|x - \gamma(x)\|_\infty,$$

where γ ranges over all bijections from D_1 to D_2.



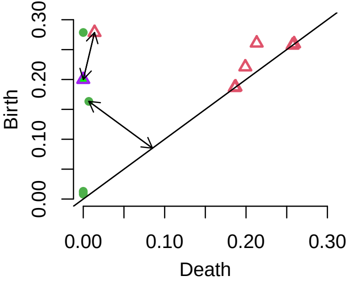
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

$$d_B(D_1,\, D_2) = \inf_{\gamma} \sup_{x \in D_\mathbf{1}} \|x - \gamma(x)\|_\infty,$$

where γ ranges over all bijections from D_1 to D_2.



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

Theorem

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

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

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

Topological Data Analysis: Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Machine Learning

- Featurization of Topological Data Analysis using Persistence Landscapes

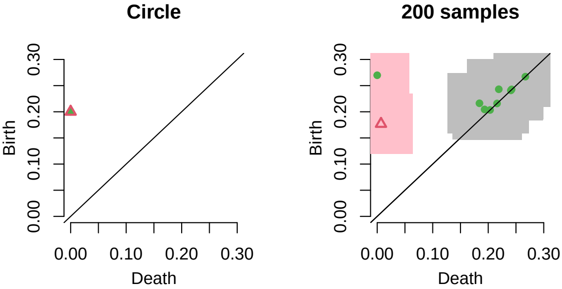
Statistical inference for persistent homology.

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

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

Let M be a compact manifold, and $X = \{X_1, \cdots, X_n\}$ be n samples. Let f_M and f_X be corresponding functions whose persistent homology is of interest. Given the significance level $\alpha \in (0,1)$, $(1-\alpha)$ confidence band $c_n = c_n(X)$ is a random variable satisfying

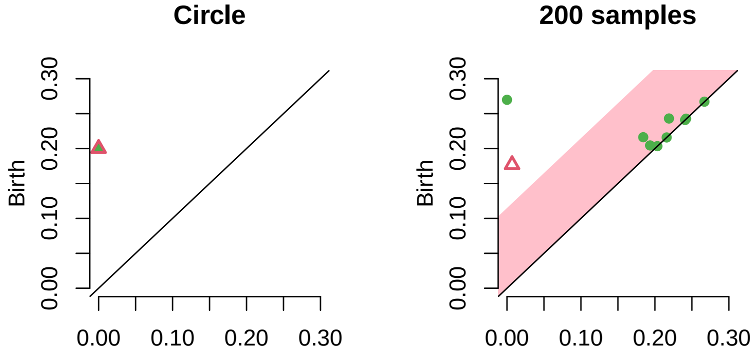
$$\mathbb{P}\left(Dgm(f_M) \in \{\mathcal{D} : d_B(\mathcal{D}, Dgm(f_X)) \leq c_n\}\right) \geq 1 - \alpha.$$



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

Let M be a compact manifold, and $X = \{X_1, \cdots, X_n\}$ be n samples. Let f_M and f_X be corresponding functions whose persistent homology is of interest. Given the significance level $\alpha \in (0,1)$, $(1-\alpha)$ confidence band $c_n = c_n(X)$ is a random variable satisfying

$$\mathbb{P}\left(d_B(Dgm(f_M),\, Dgm(f_X)) \le c_n\right) \ge 1 - \alpha.$$



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

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

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

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

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

1. Given a sample $X = \{x_1, \dots, x_n\}$, compute the kernel density estimator $\hat{p}_h$.
2. Draw $X^* = \{x_1^*, \dots, x_n^*\}$ from $X = \{x_1, \dots, x_n\}$ (with replacement), and compute $\theta^* = \sqrt{nh^m}||\hat{p}_h^*(x) - \hat{p}_h(x)||_\infty$, where $\hat{p}_h^*$ is the density estimator computed using X^*.
3. Repeat the previous step B times to obtain $\theta_1^*, \dots, \theta_B^*$
4. Compute $\hat{z}_\alpha = \inf \left\{ q : \frac{1}{B} \sum_{j=1}^{B} I(\theta_j^* \geq q) \leq \alpha \right\}$
5. The $(1-\alpha)$ confidence band for $\mathbb{E}[p_h]$ is $\left[\hat{p}_h - \frac{\hat{z}_\alpha}{\sqrt{nh^m}} \, , \, \hat{p}_h + \frac{\hat{z}_\alpha}{\sqrt{nh^m}} \right]$.

Introduction

Minimax Rates for Geometric Parameters of a Manifold

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

Topological Data Analysis: Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Machine Learning

- Featurization of Topological Data Analysis using Persistence Landscapes

(Very rough) sketch to Machine Learning

- For a given task and data, Machine Learning / Deep Learning fits a parametrized model.
  - Given data X,
  - Parametrized model f_θ,
  - Loss function $\mathcal{L}$ tailored to the task,
  - Machine Learning minimizes $\arg\min_\theta \mathcal{L}(f_\theta, \mathcal{X})$.
- Many cases, getting explicit formula for $\arg\min_\theta \mathcal{L}(f_\theta, \mathcal{X})$ is impossible or too costly (e.g., inverting a large scale matrix). So, gradient descent is used with the $\nabla_\theta \mathcal{L}(f_\theta, \mathcal{X})$:

$$\theta_{n+1} = \theta_n - \lambda \nabla_\theta \mathcal{L}(f_\theta, \mathcal{X}).$$

Application of Topological Data Analysis to Machine Learning

- Application of Topological Data Analysis to Machine Learning is usually in two directions:
  - using TDA as features, so that the data X is augmented with extra TDA features : more common
  - Loss function $\mathcal{L}$ is accompanied with topological loss terms : recently received attentions

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

Topological Data Analysis: Persistent Homology

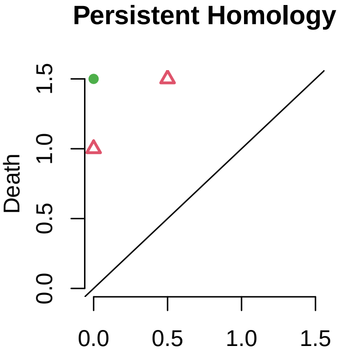
Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Machine Learning

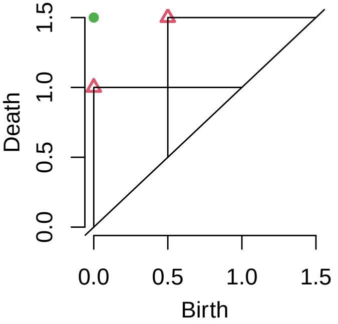
- Featurization of Topological Data Analysis using Persistence Landscapes

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

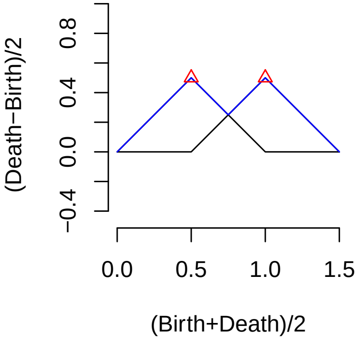
- The space of the persistent homology is complex, so directly applying in machine learning is difficult.
- If the persistent homology is further summarized and embedded into a Euclidean space or a functional space, then applying in machine learning becomes much more convenient.
    - e.g., Persistence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



Persistence Landscape is a functional summary of the persistent homology.



Persistent Homology



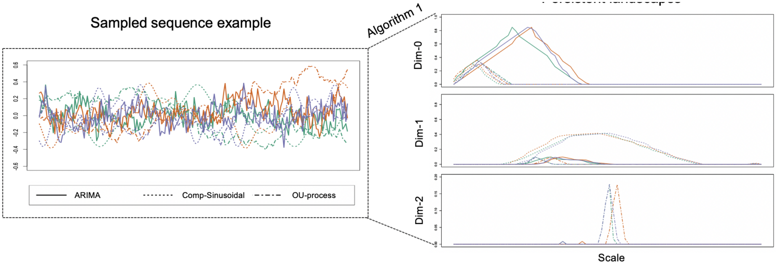
Landscape

Featurizing using Persistence Landscape

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

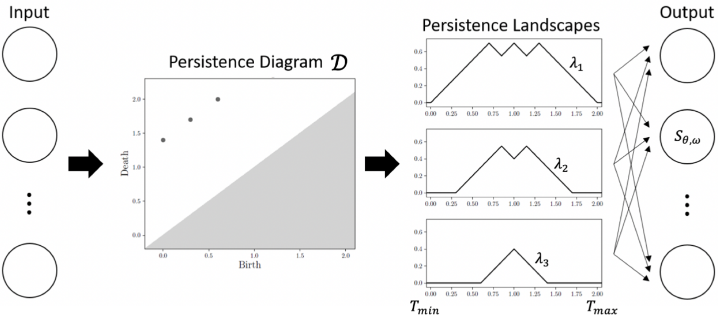
Featurization using time-delayed embedding and Persistence Landscape

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

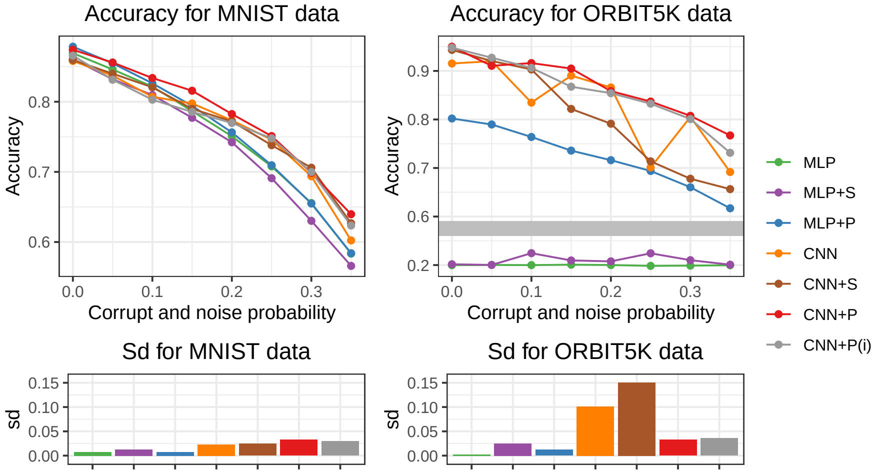


Build topological layer using Persistence Landscape

1. From data X, choose an appropriate simplicial complex K and a function f to compute the Persistece diagram $\mathcal{D}$.
2. From the persistence diagram $\mathcal{D}$, compute the persistence landscape $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$.
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 $g_\theta : \mathbb{R}^m \to \mathbb{R}$, compute $S_{\theta,\omega}(\mathcal{D}) := g_\theta(\bar{\Lambda}_\omega)$.



Build topological layer using Persistence Landscape



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

References

Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions
- Upper Bound
- Lower Bound
- Upper Bound and Lower Bound for General Case

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

- Reach and its Geometry
- Reach estimator and its analysis
- Minimax Estimates

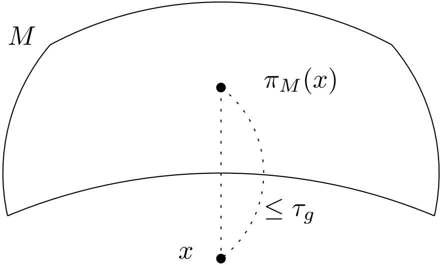
References

The supporting manifold M is assumed to be bounded.

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

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

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



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

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

Density is bounded away from ∞ 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.
- $\mathcal{P}^d_{\kappa_l,\kappa_g,K_p}$ 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

References

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

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

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

- Based on the following lemma:

Lemma

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

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

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

- Based on following lemma:

Lemma

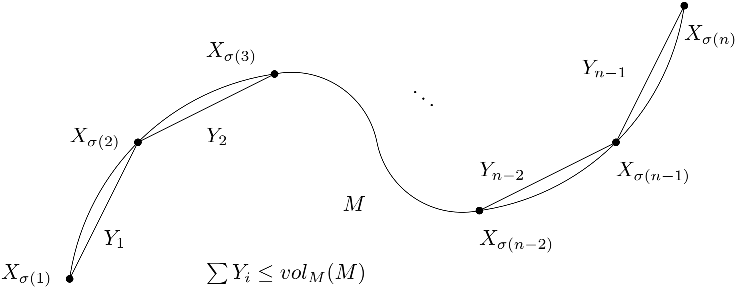
(Lemma 7) Let M be a d_1-dimensional manifold with global reach $\geq \tau_g$ and local reach $\geq \tau_\ell$, and $X_1, \cdots, X_n \in M$. Then there exists C which depends only on m, d_1 and K_l, and there exists $\sigma \in S_n$ such that

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

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

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

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

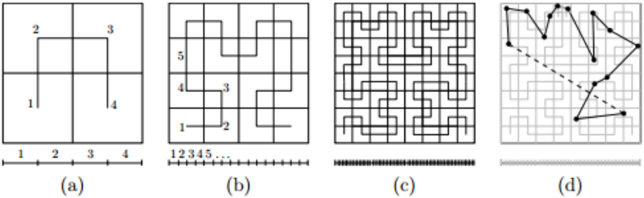


- Global reach$\geq \tau_g$ implies $vol_M(M)$ is bounded.

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

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

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



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

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

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

Lemma

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

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

Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions
- Upper Bound
- Lower Bound
- Upper Bound and Lower Bound for General Case

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

- Reach and its Geometry
- Reach estimator and its analysis
- Minimax Estimates

References

A subset $T \subset [-K_I, K_I]^n$ and set of distributions $\mathcal{P}_1^{d_1}$, $\mathcal{P}_2^{d_2}$ are found so that, whenever $X = (X_1, \cdots, X_n) \in T$, 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, $\mathcal{P}_1^{\boldsymbol{d_1}}$ and $\mathcal{P}_2^{\boldsymbol{d_2}}$ 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.

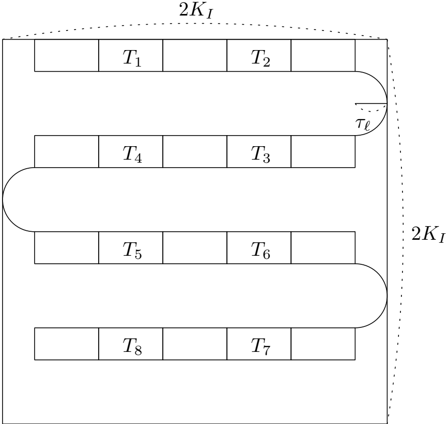
Lemma

(Lemma 10, Le Cam's Lemma) Let $\mathcal{P}$ be a set of probability measures, and $\mathcal{P}^{\mathbf{d_1}}, \mathcal{P}^{\mathbf{d_2}} \subset \mathcal{P}$ 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

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

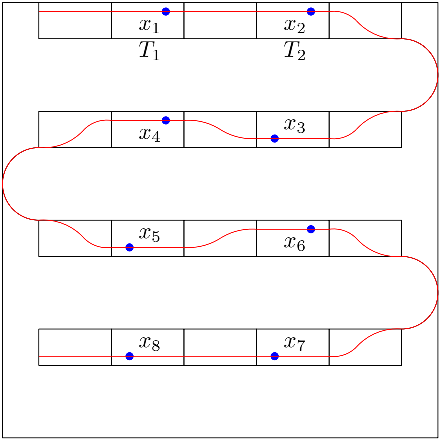
T is constructed so that for any $x = (x_1, \cdots, x_n) \in T$, there exists a d_1-dimensional manifold that satisfies regularity conditions and passes through $x_1, \cdots, x_n$.

- 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 $x = (x_1, \cdots, x_n) \in T$, there exists a d_1-dimensional manifold that satisfies regularity conditions and passes through $x_1, \cdots, x_n$.

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



$\mathcal{P}_1^{d_1}$ is constructed as set of distributions that are supported on manifolds that passes through $x_1, \cdots, x_n$ for $x = (x_1, \cdots, x_n) \in T$, and $\mathcal{P}_2^{d_2}$ is a singleton set consisting of the uniform distirbution 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 $x \in \operatorname{int} T$, there exists $C > 0$ that depends only on κ_I, K_I, and $r_x > 0$ such that for all $r < r_x$,

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

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

References

Multinary Classification and $0 - 1$ Loss are Considered.

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

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

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

References

Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions
- Upper Bound
- Lower Bound
- Upper Bound and Lower Bound for General Case

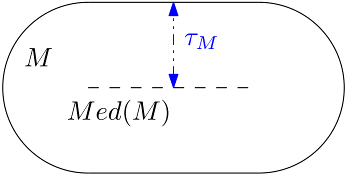
The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory

- Reach and its Geometry
- Reach estimator and its analysis
- Minimax Estimates

References

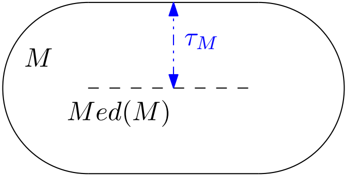
The medial axis of a set M is the set of points that have at least two nearest neighbors on the set M.

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



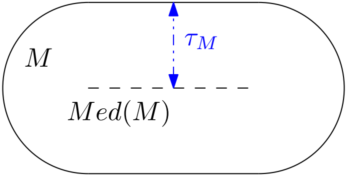
The reach of M, denoted by τ_M, is the minimum distance from $Med(M)$ to M.

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



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

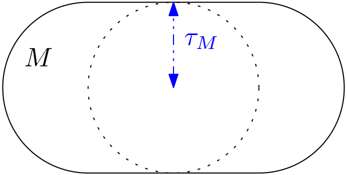
- $$\tau_M = \inf_{x \in Med(M), y \in M} \|x - y\|.$$



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,

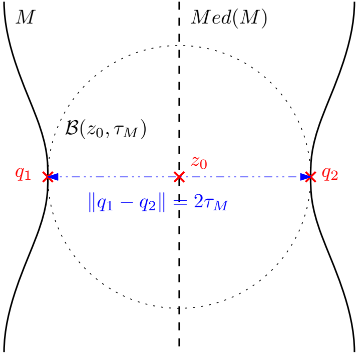
$$\tau_M = \inf_{q_2 \neq q_1 \in M} \frac{\|q_2 - q_1\|^2}{2d(q_2 - q_1, T_{q_1}M)}.$$



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 $\|q_1 - q_2\| = 2\tau_M$.

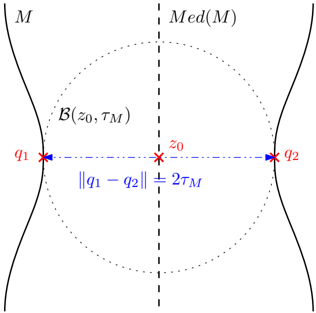


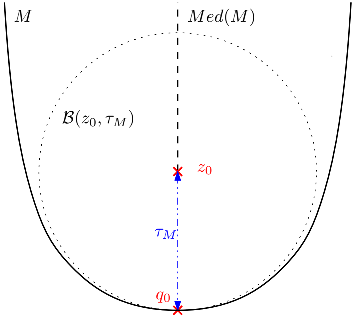
The reach is attained either from the bottleneck (global case) or the area of high curvature (local case).

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

References

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

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

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

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

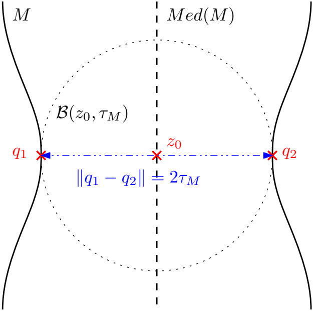
- The risk of the estimator $\hat{\tau}$ is the expected loss the estimator

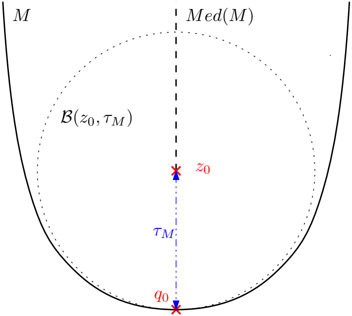
$$\mathbb{E}_{P^{(n)}}\left[\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathcal{X})}\right|^q\right].$$

  - $\mathcal{X} = \{X_1, \ldots, X_n\}$ 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 $O\left(n^{-\frac{\mathbf{2q}}{\mathbf{3d-1}}}\right)$ for the local case.



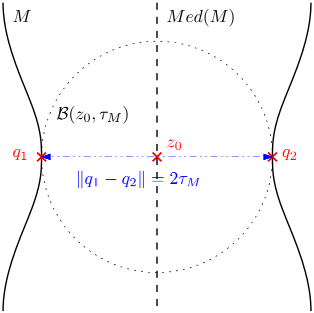


The reach estimator has the maximum risk of $O\left(n^{-\frac{q}{d}}\right)$ for the global case.

Proposition

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

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

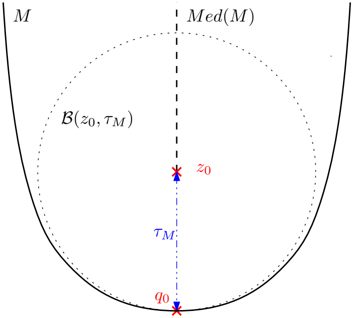


The reach estimator has the maximum risk of $O\left(n^{-\frac{2q}{3d-1}}\right)$ for the local case.

Proposition

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

References

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

- Minimax rate is the risk of an estimator that performs best in the worst case, as a function of sample size.
- $$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\ell \left(\hat{\tau}_n(\mathcal{X}),\ \tau_M \right) \right].$$
  - $\mathcal{X} = \{X_1, \ldots, X_n\}$ 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 $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^q$, $q \geq 1$.

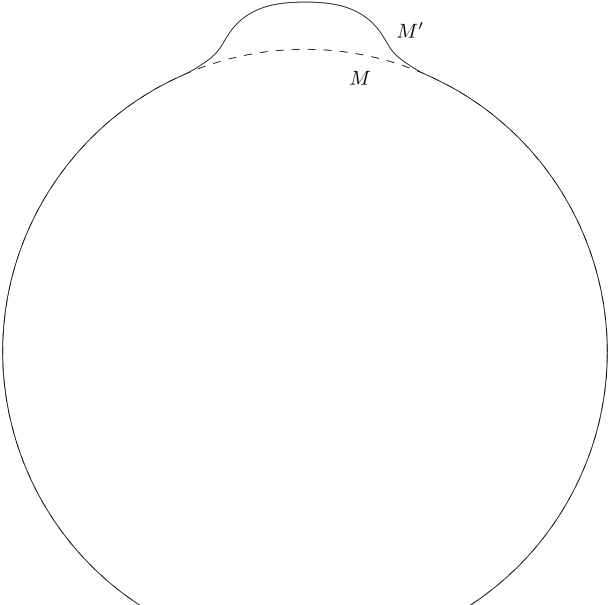
The statistical difficulty of the reach estimation problem is analyzed by the minimax rate.

- Minimax rate is the risk of an estimator that performs best in the worst case, as a function of sample size.
- $$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n(\mathcal{X})} \right|^q \right].$$
  - $\mathcal{X} = \{X_1, \ldots, X_n\}$ is drawn from a fixed distribution P with its support M, where P is contained in set of distributions $\mathcal{P}$.
  - An estimator $\hat{\tau}_n$ is any function of data $\mathcal{X}$.
  - The loss function used is $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^q$, $q \geq 1$.

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

- $$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n} \right|^q \right] \gtrsim \left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^q (1 - TV(P, P'))^{2n} .$$
- The lower bound measures how hard it is to tell whether the data is from distributions with different reaches.
- P and P' are found so that $\left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^q$ is large while $(1 - TV(P, P'))^{2n}$ is small.

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



Minimax Rates for Estimating the Dimension of a Manifold

- Regularity conditions
- Upper Bound
- Lower Bound
- Upper Bound and Lower Bound for General Case

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

- Reach and its Geometry
- Reach estimator and its analysis
- Minimax Estimates

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.

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

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 Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *Ann. Statist.*, 42(6):2301–2339, 12 2014. doi: 10.1214/14-AOS1252. URL `http://dx.doi.org/10.1214/14-AOS1252`.

Jisu Kim, Alessandro Rinaldo, and Larry Wasserman. Minimax Rates for Estimating the Dimension of a Manifold. *ArXiv e-prints*, May 2019.

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, 2020. URL `http://arxiv.org/abs/1812.02987`