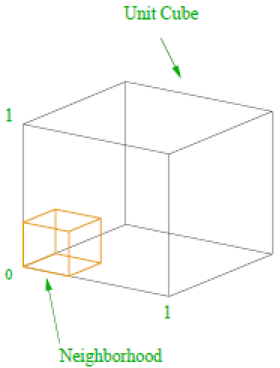
Statistical Inference for Geometric Data

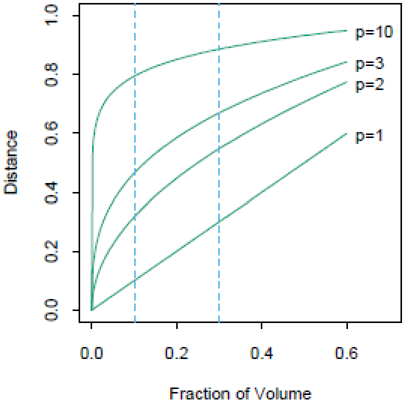
Jisu KIM

Carnege Mellon University

Sep 19, 2018

High dimensional data suffers from the curse of dimensionality.

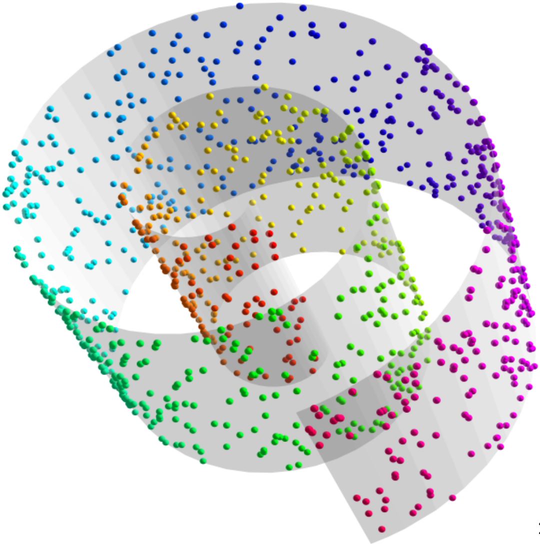




[1]

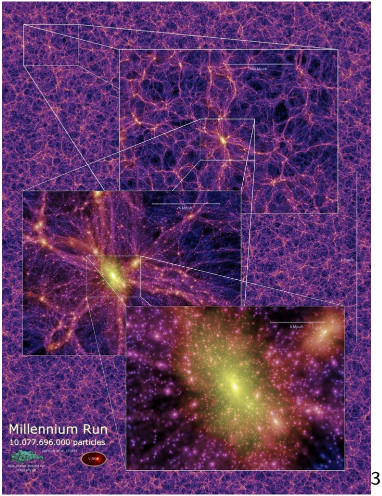
[1] [Hastie et al., 2009, Ch2, Figure 2.6]

The curse of dimensionality is mitigated when there is a low dimensional geometric structure.


[2]

2 http://www.skybluetrades.net/blog/posts/2011/10/30/machine-learning/

Geometric structures in the data provide information.



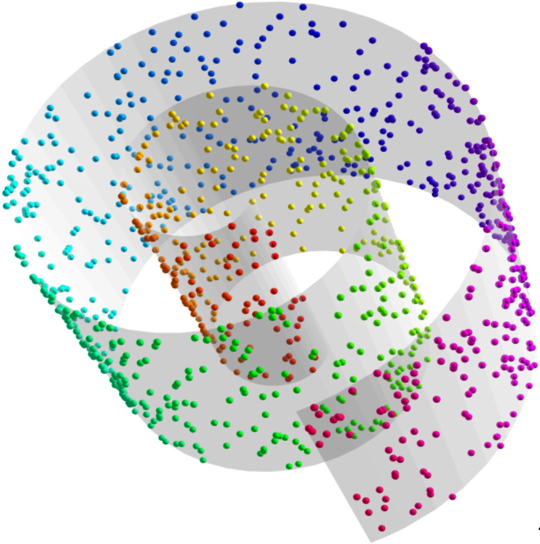
[3]

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

Statistic Inference for Geometric Data is explored.

- Minimax Rates for Geometric Parameters of a Manifold
  - Minimax Rates for Estimating the Dimension of a Manifold (Kim, Rinaldo, Wasserman, 2016)
  - The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory (Aamari, Kim, Chazal, Michel, Rinaldo, Wasserman, 2017)
- Statistical Inference For Homological Features
  - Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)
- Statistical Inference and Computation for Persistent Homology
  - Statistical inference on persistent homology of KDE filtration on rips complex (Shin, Kim, Rinaldo, Wasserman, 2018?)
  - R Package TDA: Statistical Tools for Topological Data Analysis (Fasy, Kim, Lecci, Maria, Milman, Rouvreau, 2014a)

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



[4]

[4] 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 $\hat{\theta}_n$.

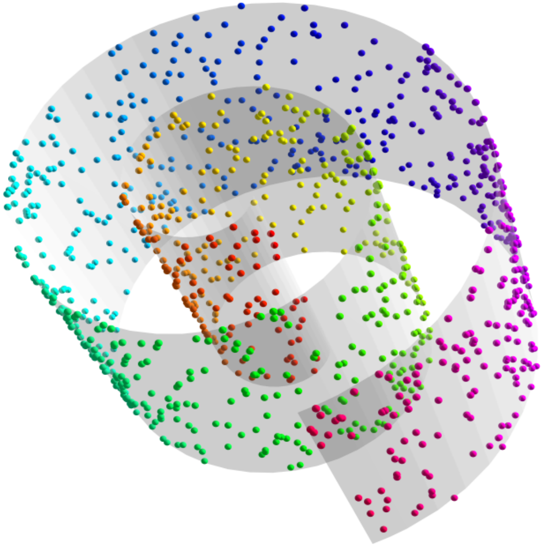
The minimax rate describes the statistical difficulty of estimating a parameter.

- 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, 2016)
- The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory (Aamari, Kim, Chazal, Michel, Rinaldo, Wasserman, 2017)

Manifold learning finds an underlying manifold to reduce dimension.



The intrinsic dimension of a manifold needs to be estimated.

- 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

Theorem

(Proposition 28 and 29)

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

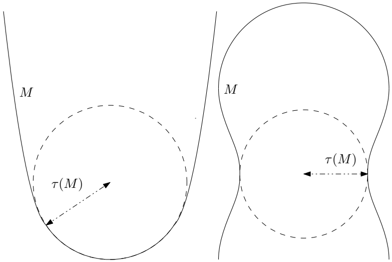
The reach is the maximum radius of a ball that can roll over the manifold.

Definition

The reach of M, denoted by $\tau(M)$, can be defined as

$$\tau(M) = \inf_{a \neq b \in M} \frac{\|b - a\|_2^2}{2d(b - a,\ T_a M)},$$

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 $x, y \in \mathbb{R}$, $\ell(x, y) = \left| \frac{1}{x} - \frac{1}{y} \right|^q$.

An estimator of the reach

The reach of M is

$$\tau(M) = \inf_{a \neq b \in M} \frac{\|b - a\|_2^2}{2d(b - a,\ T_a M)}.$$

Definition

Given observation $X = (X_1, \ldots, X_n)$, we estimate the reach as

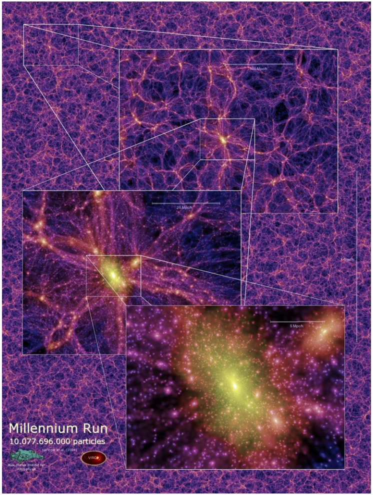
$$\hat{\tau}_n(X) = \inf_{1 \leq i \neq j \leq n} \frac{\|X_j - X_i\|_2^2}{2d(X_j - X_i,\ T_{X_i} M)}.$$

Minimax rate for estimating the reach

Theorem
(Theorem 45 and Proposition 50)

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

Geometric holes in the data provide information.



The number of holes is used to summarize geometrical features.

- Geometrical objects :
  - A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z,
  - , , , , .[5]
- 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 $dim \geq 3$ 

[5] Professor Valerie Ventura

Example : Objects are classified by homologies.

1. β_0 =# of connected components 
2. β_1 =# of loops 

$\beta_0 \setminus \beta_1$	0	1	2
1	C, G, I, J, L, M, N, S, U, V, W, Z,  E, F, T, Y, H, K, X	A, R, D, O, , P, Q	B
2			
3		, 	

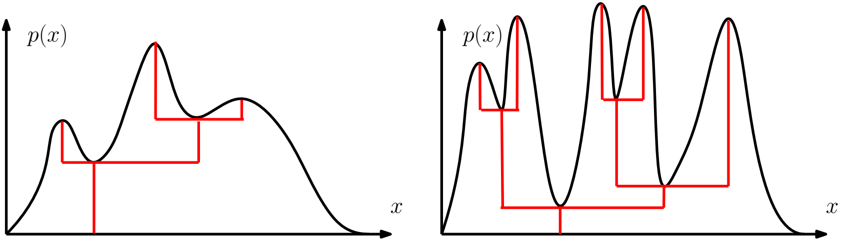
Statistical inference for homological features.

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

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

Definition

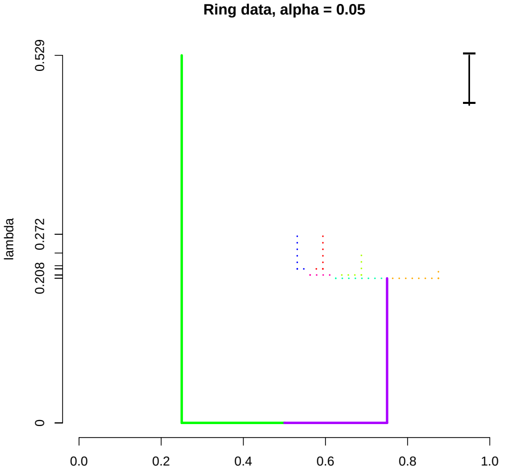
For a density function p, its cluster tree T_p is a function where $T_p(\lambda)$ is the set of connected components of the upper level set $\{x : p(x) \geq \lambda\}$.



A confidence set helps denoising the empirical tree.

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

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



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

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

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

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

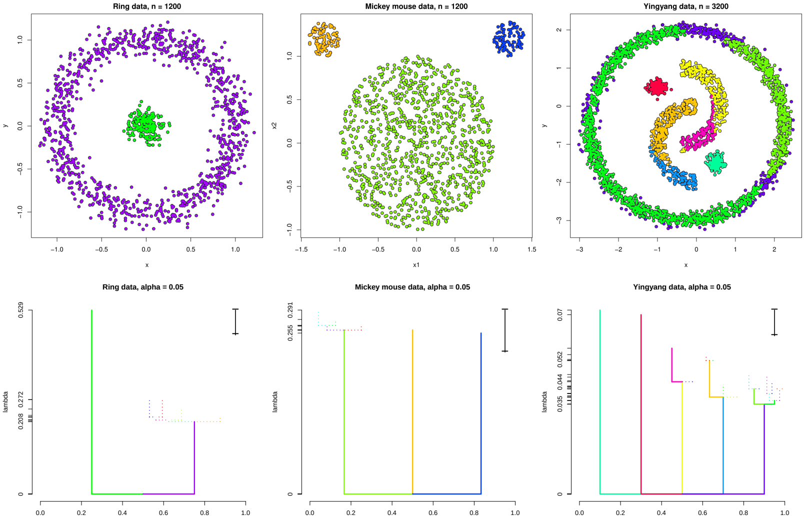
$$\hat{C}_\alpha = \{T : d_\infty(T, T_{\hat{p}_h}) \leq t_\alpha\}.$$

Theorem

(Theorem 57) Above confidence set $\hat{C}_\alpha$ satisfies

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

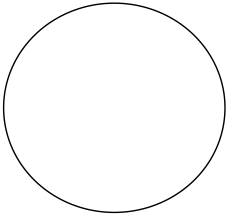
The pruned trees according to the confidence set recover the actual cluster trees.



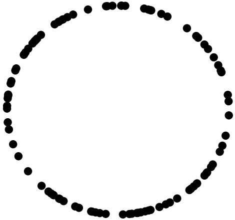
Homology of finite sample is different from homology of underlying manifold, hence it cannot be directly used for the inference.

- 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:

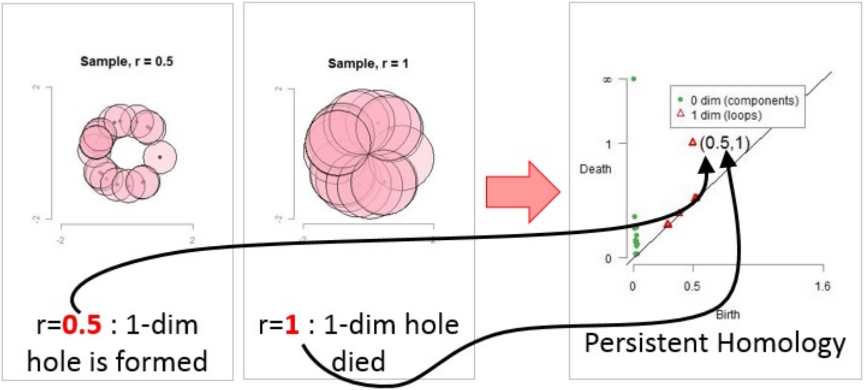
Underlying circle: $\beta_0 = 1$, $\beta_1 = 1$



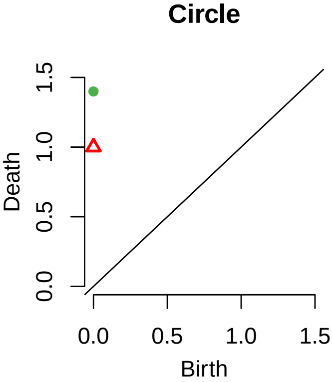
100 samples: $\beta_0 = 100$, $\beta_1 = 0$

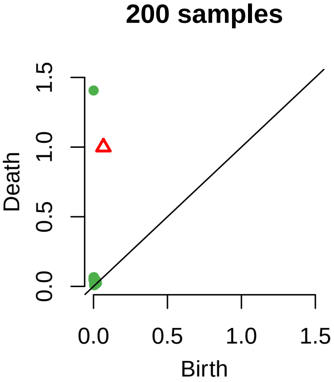


Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.

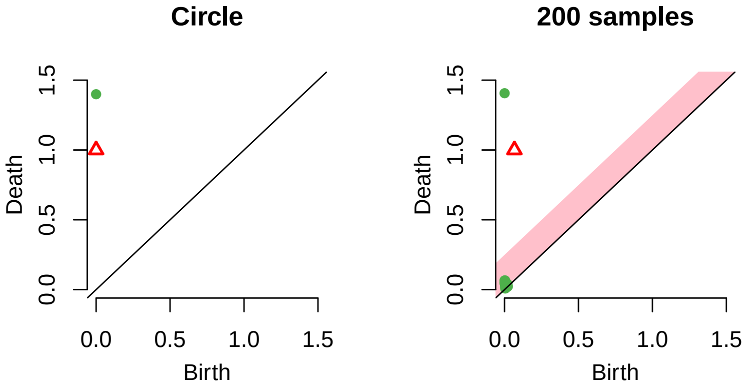


Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.





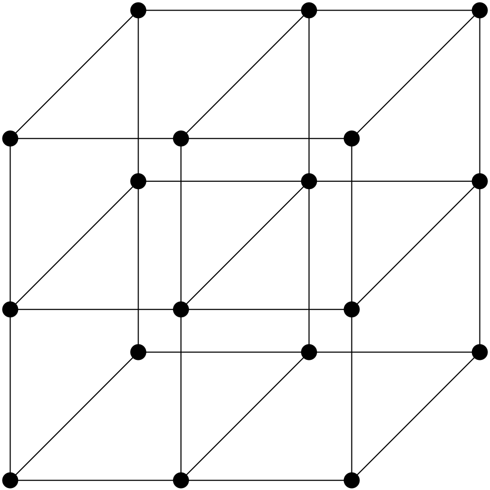
Confidence band for persistent homology separates homological signal from homological noise.



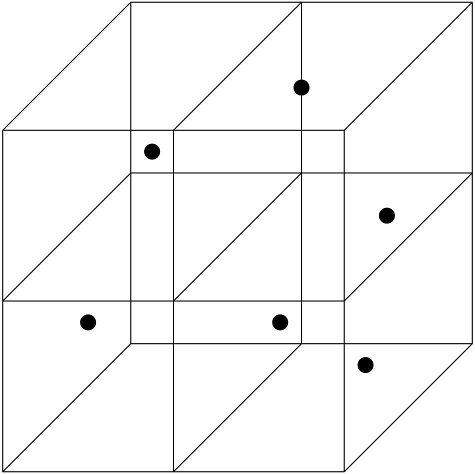
Statistical inference for persistent homology.

- Persistent homology of KDE filtration on rips complex (Shin, Kim, Rinaldo, Wasserman, 2018?)
- R Package TDA: Statistical Tools for Topological Data Analysis (Fasy, Kim, Lecci, Maria, Milman, Rouvreau, 2014a)

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

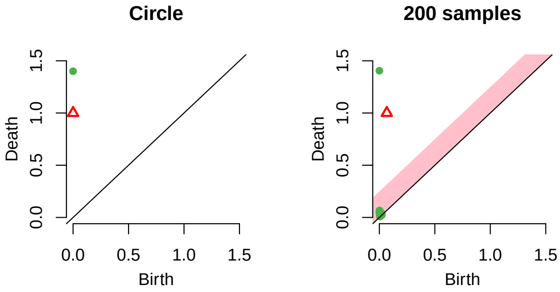


Computing the persistent homology of density function on data points reduces computational complexity.



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

- (Shin, Kim, Rinaldo, Wasserman, 2018?) : extending work from Fasy et al. [2014b], Bobrowski et al. [2014], Chazal et al. [2011].



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

- The kernel density estimator is

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

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

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

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

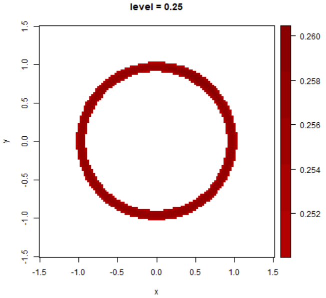
- We are considering the upper level sets of the average kernel density estimator

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

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

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

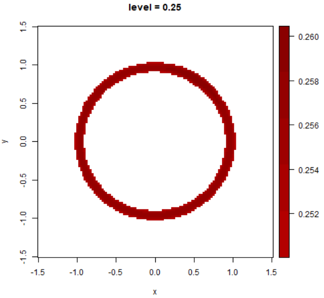


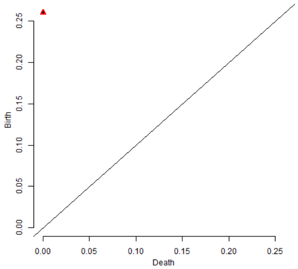
We are targeting the persistent homology of the upper level set of the average kernel density estimator on the support.

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

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

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

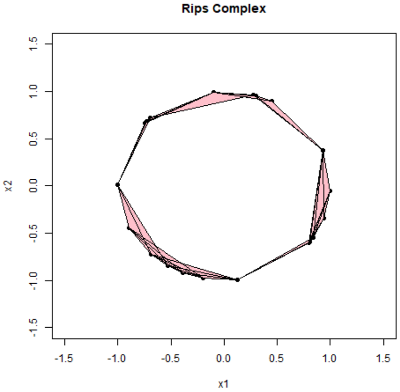




We use the Rips complex to estimate the target persistent homology.

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

$$R(\mathcal{X}, r) = \left\{ [X_{i_1}, \ldots, X_{i_k}] : \ d(X_{i_j}, X_{i_l}) < 2r, \ 1 \leq \forall j \neq l \leq k, \ k = 1, \ldots, n \right\}.$$



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

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

$$R(\mathcal{X}, r) = \left\{[X_{i_1}, \ldots, X_{i_k}] : \, d(X_{i_j}, X_{i_l}) < 2r, \, 1 \leq \forall j \neq l \leq k, \, k = 1, \ldots, n\right\}.$$

- The KDE (kernel density estimator) is

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

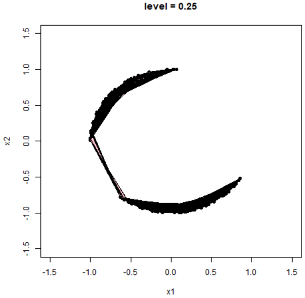
- For $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, we consider the Rips complex generated from the level set of the KDE

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

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

- For $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, we estimate the target level set by the level sets of the KDE on Rips complexes,

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

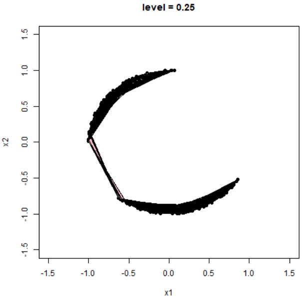


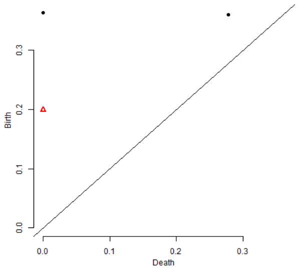
We estimate the target persistent homology by the persistent homology of the KDE filtration on Rips complexes.

- We estimate the target persistent homology by the persistent homology of the level sets of the KDE on Rips complexes,

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

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





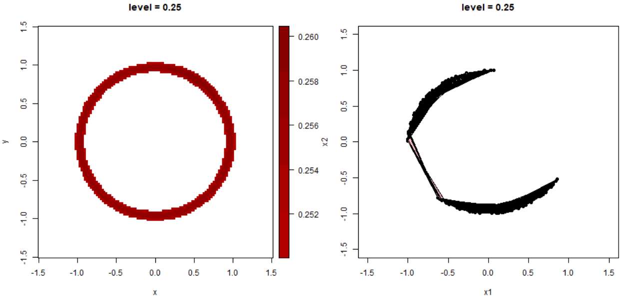
We estimate the target level set by Rips complexes from the KDE level sets.

- We approximate the target level set

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

by the level sets of the KDE on Rips complexes,

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



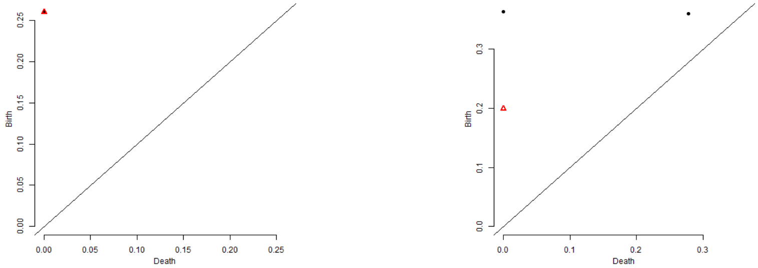
We estimate the target persistent homology by the persistent homology of the KDE filtration on Rips complexes.

- We estimate the target persistent homology

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

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

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



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

Theorem
(Theorem 74)

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

Confidence set

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

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

Confidence set for the persistent homology of the KDE filtration.

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

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

This is a valid confidence set by the following theorem.

Theorem

(Theorem 78)

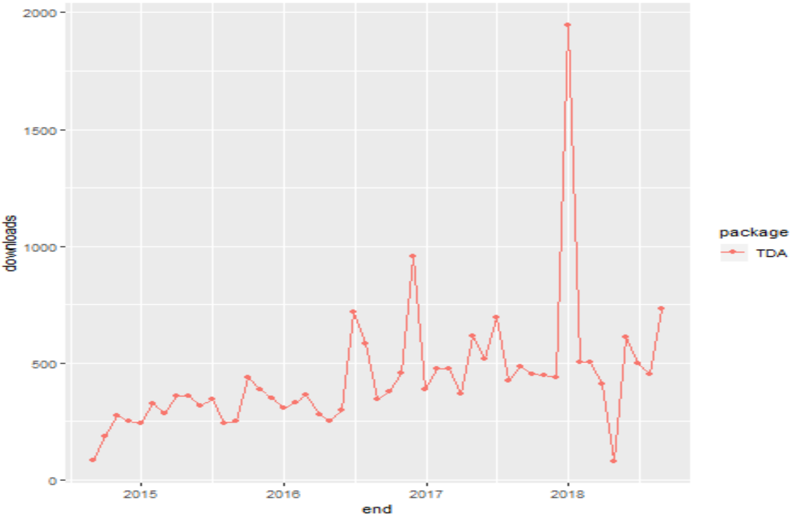
$$\mathbb{P} \left(PH_*^{supp(P)}(p_{h_n}) \in \hat{C}_\alpha \right) \geq 1 - \alpha + o(1).$$

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.

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

- # of downloads (2014-08-18 - 2018-09-19) : 21820



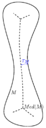
Presented papers

- Jaehyeok Shin, Jisu Kim, Alessandro Rinaldo, Larry Wasserman, Persistent homology of KDE filtration on rips complex
- Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, Larry Wasserman, Estimating the Reach of a Manifold [2017] [arXiv | HAL | BibTeX]
- Jisu Kim, Yen-Chi Chen, Sivaraman Balakrishnan, Alessandro Rinaldo, Larry Wasserman, Statistical Inference for Cluster Trees [2017] [arXiv | NIPS | BibTeX]
- Jisu Kim, Alessandro Rinaldo, Larry Wasserman, Minimax Rates for Estimating the Dimension of a Manifold [2016] [arXiv | BibTeX]
- Brittany T. Fasy, Jisu Kim, Fabrizio Lecci, Clément Maria, David L. Millman, Vincent Rouvreau, Introduction to the R package TDA. [2014] [arXiv | BibTeX]

Other papers

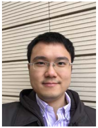
- Kwangho Kim, Jisu Kim, Barnabás Póczos, Distribution Regression in Semi-supervised Learning
- Kwangho Kim, Jisu Kim, Edward H. Kennedy, Causal effects based on distributional distances [2018] [arXiv | BibTeX]
- Kijung Shin, Bryan Hooi, Jisu Kim, Christos Faloutsos, Think before You Discard: Accurate Triangle Counting in Graph Streams with Deletions [2018] [paper | appendix | BibTeX]
- Kijung Shin, Bryan Hooi, Jisu Kim, Christos Faloutsos, DenseAlert: Incremental Dense-Subtensor Detection in Tensor Streams [2017] [paper | appendix | BibTeX]
- Kijung Shin, Bryan Hooi, Jisu Kim, Christos Faloutsos, D-Cube: Dense-Block Detection in Terabyte-Scale Tensors [2017] [paper | appendix | BibTeX]

Thank you!


Eddie Aamari


Sivaraman Balakrishnan


Frédéric Chazal


Yen-Chi Chen

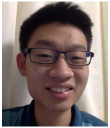

Jessi Cisewski


Christos Faloutsos


Brittany T. Fasy


Stephan Günnemann


Giseon Heo


Bryan Hooi


Edward Kennedy


Kwangho Kim


Jay-Yoon Lee


Clément Maria

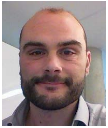

Bertrand Michel

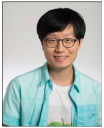

David L. Millman


Jaeseong Oh


Barnabás Póczos

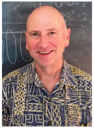

Alessandro Rinaldo


Vincent Rouvreau


Jaehyeok Shin


Kijung Shin


Daren Wang


Larry Wasserman

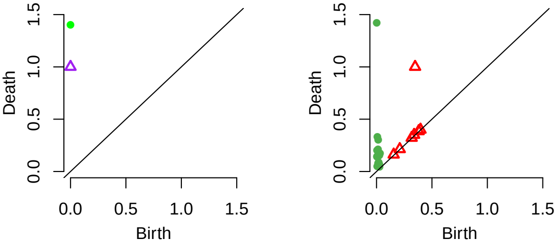
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

where γ ranges over all bijections from D_1 to D_2.



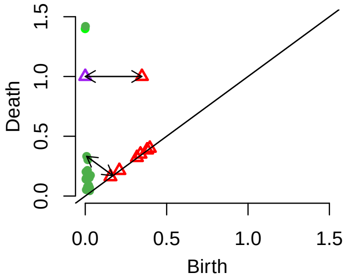
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

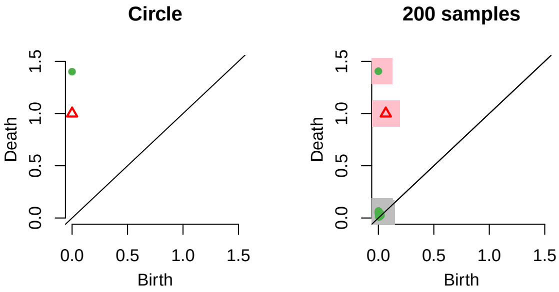
where γ ranges over all bijections from D_1 to D_2.



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

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

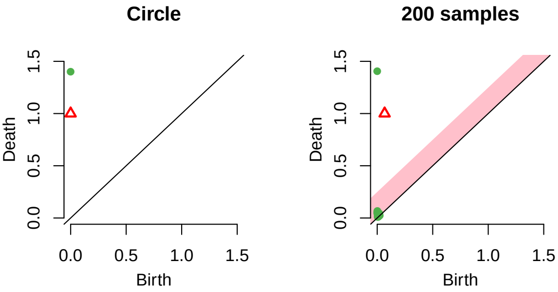
$$\mathbb{P}\left(d_B(Dgm(f_M),\, Dgm(f_X)) \leq c_n\right) \geq 1 - \alpha.$$



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

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

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



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

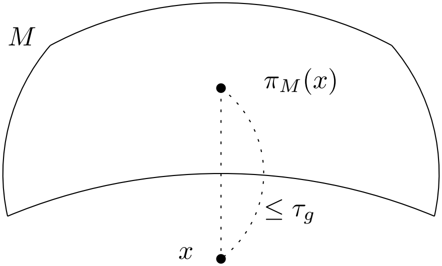
1. Given a sample $X = \{x_1, \ldots, x_n\}$, compute the kernel density estimator $\hat{p}_h$.
2. Draw $X^* = \{x_1^*, \ldots, x_n^*\}$ from $X = \{x_1, \ldots, x_n\}$ (with replacement), and compute $\theta^* = \sqrt{n}||\hat{p}_h^*(x) - \hat{p}_h(x)||_\infty$, where $\hat{p}_h^*$ is the density estimator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

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 $\geq \tau_\ell$.

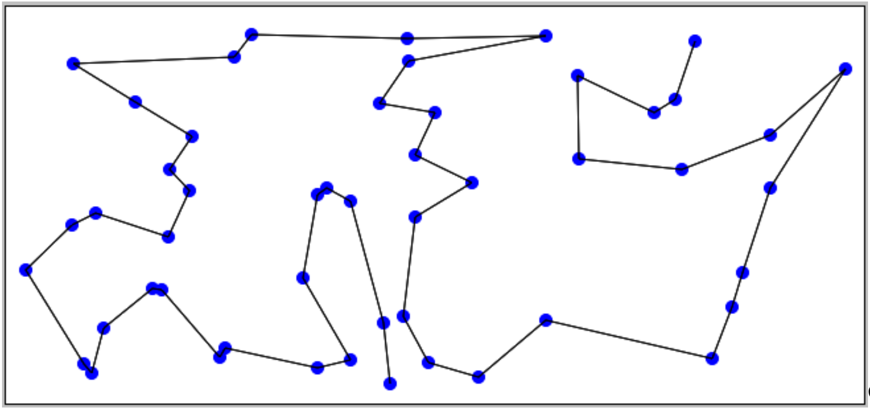
Density is bounded away from ∞ with respect to the uniform measure.

- 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.

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

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

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


[6]

6 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.
-

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

where C is some constant that depends only on K_I, d_1, and m.

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 18) Let $X_1, \cdots, X_n \sim P \in \mathcal{P}^{d_2}_{\kappa_l,\kappa_g,K_p}$, 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

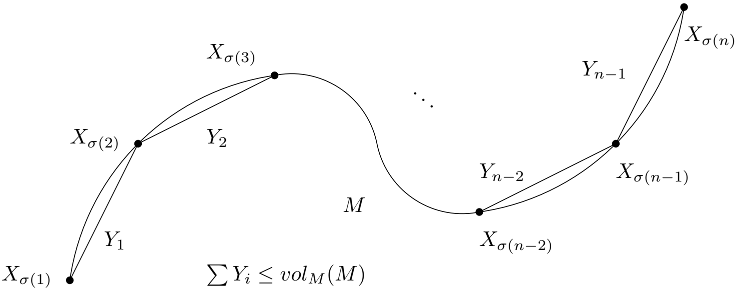
(Lemma 19) Let M be a d_1-dimensional manifold with global reach $\geq \tau_g$ and local reach $\geq \tau_\ell$, and $X_1, \cdots, X_n \in M$. Then there exists C which depends only on m, d_1 and K_l, and there exists $\sigma \in S_n$ such that

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

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

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

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

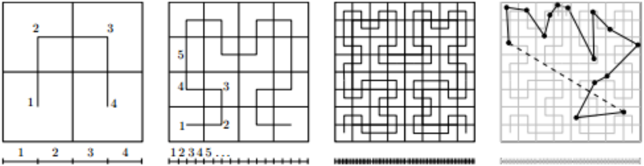


- Global reach$\geq \tau_g$ implies $vol_M(M)$ is bounded.

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.



(a) (b) (c) (d)

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 85, 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.

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

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

Proposition

(Proposition 21) Let $1 \leq d_1 < d_2 \leq m$. *Then*

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

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, $\mathcal{P}_1^{d_1}$ 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.

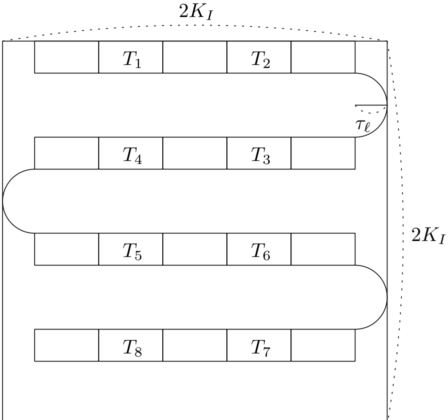
Lemma

(Lemma 22, 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

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \geq \frac{d(\theta_1, \theta_2)}{4} \sup_{Q_i \in co(\mathcal{P}^{d_i})} \int [q_1(x) \wedge q_2(x)] d\nu(x).$$

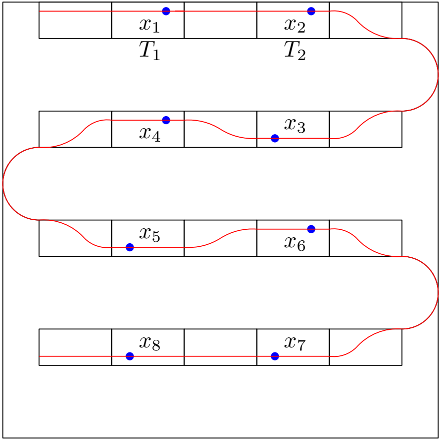
T is constructed so that for any $x = (x_1, \cdots, x_n) \in T$, there exists a d_1-dimensional manifold that satisfies regularity conditions and passes through $x_1, \cdots, x_n$.

- 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 $r < r_x$,

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

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

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

Proposition

(Proposition 26)

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

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

- $$R_n = \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[\ell \left(\hat{\dim}_n(X),\ \dim(P) \right) \right]$$
- 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 $\Omega\left(n^{-2n}\right)$.

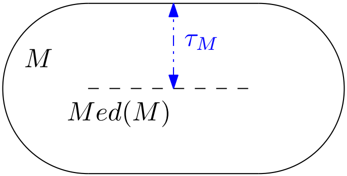
Proposition

(Proposition 28 and 29)

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

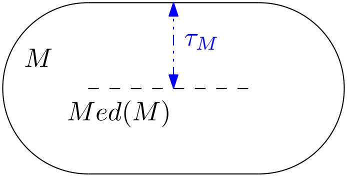
The medial axis of a set M is the set of points that have at least two nearest neighbors on the set M.

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



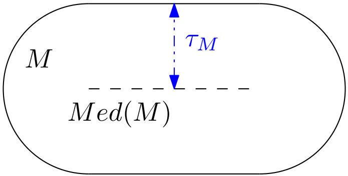
The reach of M, denoted by τ_M, is the minimum distance from $Med(M)$ to M.

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



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

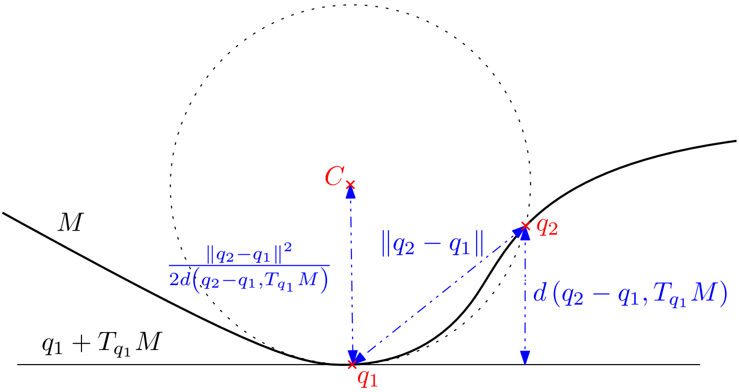
- $$\tau_M = \inf_{x \in Med(M), y \in M} \|x - y\| .$$



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,

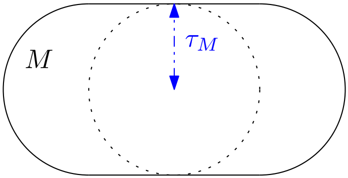
$$\tau_M = \inf_{q_2 \neq q_1 \in M} \frac{\|q_2 - q_1\|^2}{2d(q_2 - q_1, T_{q_1} M)}.$$



The 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,

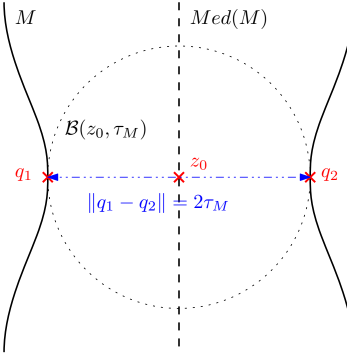
$$\tau_M = \inf_{q_2 \neq q_1 \in M} \frac{\|q_2 - q_1\|^2}{2d(q_2 - q_1, T_{q_1} M)}.$$



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

Definition

(Definition 34) 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 $\|q_1 - q_2\| = 2\tau_M$.

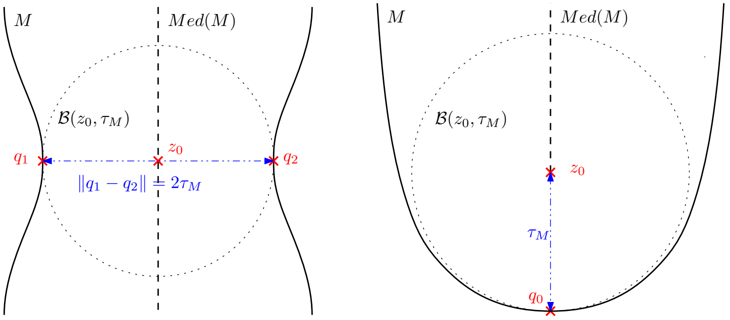


The reach is attained either from the bottleneck (global case) or the area of high curvature (local case).

Theorem

(Theorem 37) At least one of the following two assertions holds:

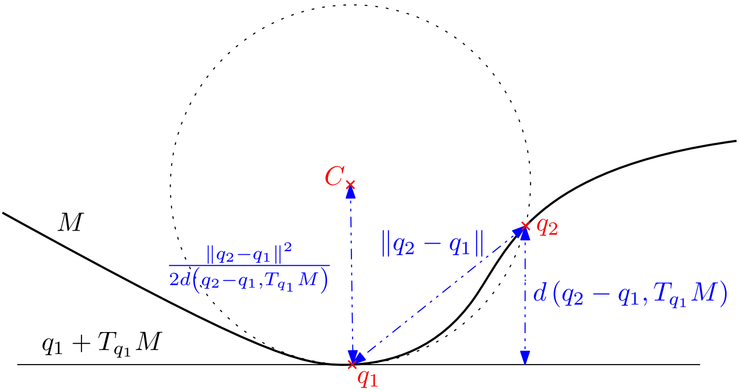
- *(Global Case) M has a bottleneck $(q_1, q_2) \in M^2$.*
- *(Local case) There exists $q_0 \in M$ and an arc-length parametrized γ_0 such that $\gamma_0(0) = q_0$ and $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$.*



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,

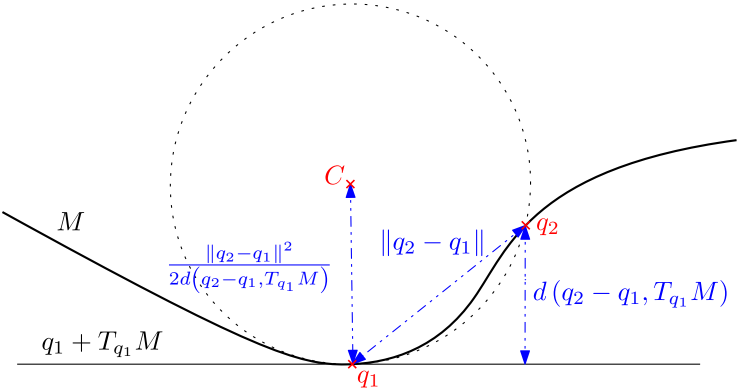
$$\tau_M = \inf_{q \neq p \in M} \frac{\|q - p\|^2}{2d(q - p, T_p M)}.$$



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

- Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a finite point cloud, then the reach estimator $\hat{\tau}$ is a plugin estimator as

$$\hat{\tau}(\mathcal{X}) = \inf_{x_i \neq x_j \in \mathbb{X}} \frac{\|x_j - x_i\|^2}{2d(x_j - x_i, T_{x_i} M)}.$$



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.

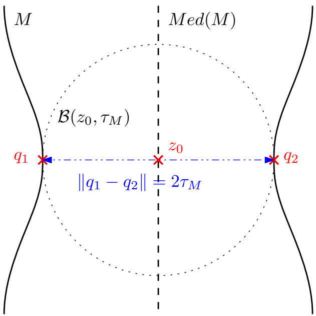
- The risk of the estimator $\hat{\tau}$ is the expected loss the estimator

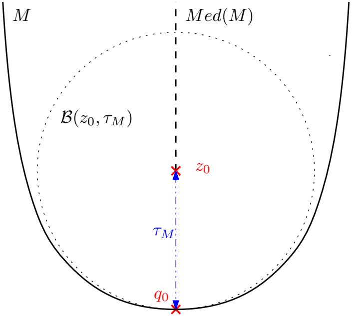
$$\mathbb{E}_{P^{(n)}}\left[\left|\frac{1}{\tau_M} - \frac{1}{\hat{\tau}(\mathcal{X})}\right|^p\right].$$

  - $\mathcal{X} = \{X_1, \ldots, X_n\}$ 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 reach estimator has the risk of $O\left(n^{-\frac{2p}{3d-1}}\right)$.

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



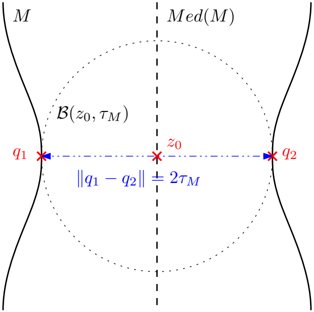


The reach estimator has the maximum risk of $O\left(n^{-\frac{p}{d}}\right)$ for the global case.

Proposition

(Proposition 40) Assume that the support M has a bottleneck. Then,

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

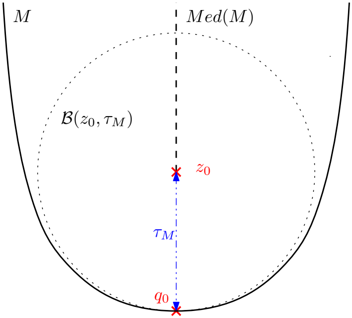


The reach estimator has the maximum risk of $O\left(n^{-\frac{2p}{3d-1}}\right)$ for the local case.

Proposition

(Proposition 44) 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,

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



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.
-
$$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\ell \left(\hat{\tau}_n(\mathcal{X}),\ \tau_M \right) \right].$$
  - $\mathcal{X} = \{X_1, \ldots, X_n\}$ 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 $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^p$, $p \geq 1$.

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.
- $$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n(\mathcal{X})} \right|^p \right].$$
  - $\mathcal{X} = \{X_1, \ldots, X_n\}$ 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 $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^p$, $p \geq 1$.

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

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

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

Theorem
(Theorem 45)

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

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

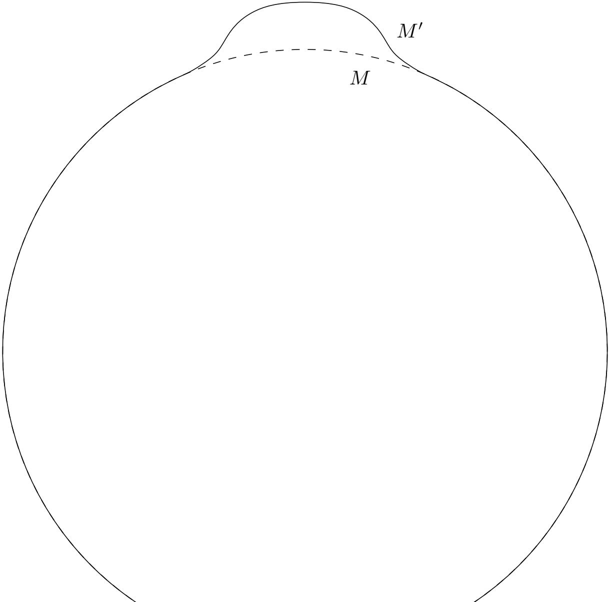
(Lemma 46) Let $P, P' \in \mathcal{P}$ with respective supports M and M'. Then

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

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

- $$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n} \right|^p \right] \gtrsim \left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^p (1 - TV(P, P'))^{2n}.$$
- 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 $\left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^p$ 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 50)

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

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

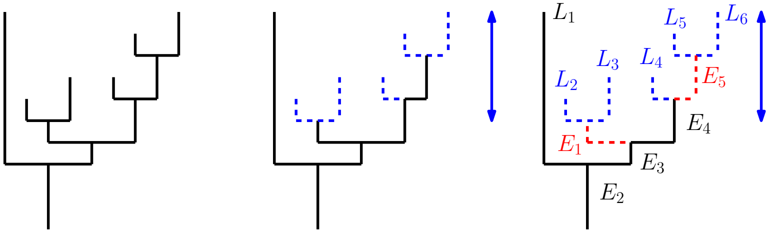
Definition

The l_∞ metric between trees are defined as

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

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

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



Reference

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the Reach of a Manifold. *ArXiv e-prints*, May 2017.

O. Bobrowski, S. Mukherjee, and J. E. Taylor. Topological consistency via kernel estimation. *ArXiv e-prints*, July 2014.

Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Scalar field analysis over point cloud data. *Discrete & Computational Geometry*, 46(4):743–775, 2011.

Brittany T. Fasy, Jisu Kim, Fabrizio Lecci, Clément Maria, David L. Millman, and Vincent Rouvreau. Introduction to the R package TDA. *CoRR*, abs/1411.1830, 2014a. URL `http://arxiv.org/abs/1411.1830`.

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

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009. ISBN 978-0-3878-4857-0. URL `http://statweb.stanford.edu/~tibs/ElemStatLearn/`