Statistical Inference For Topological Data Analysis and its application to Machine Learning

Jisu KIM

Inria

Kyungpook National University

2021-01-19

Introduction

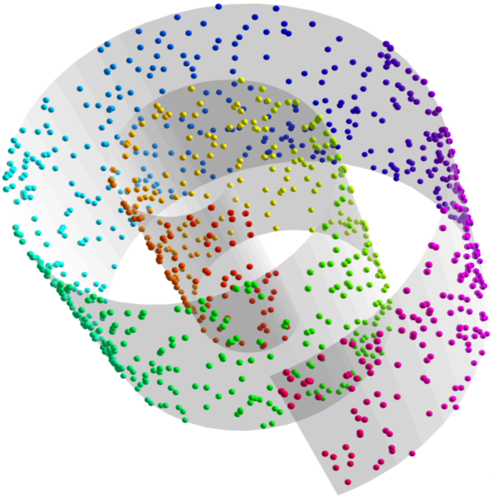
Topological Data Analysis: Persistent Homology

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

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

Computation for Topological Data Analysis
- R Package TDA: Statistical Tools for Topological Data Analysis

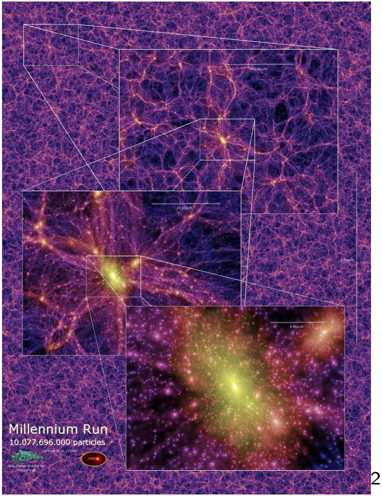
The curse of dimensionality from the high dimensional data is mitigated when there is a low dimensional geometric and topological structure.



[1]

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

Topological structures in the data provide information.



[2]

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

Statistic Inference for Topological Data Analysis and application to Machine Learning are explored.

- 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, 2014b)
  - Statistical inference on persistent homology of KDE filtration on Vietoris-Rips complex (Shin, Kim, Rinaldo, Wasserman, 2021?)
- 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)
- Computation for Topological Data Analysis
  - R Package TDA: Statistical Tools for Topological Data Analysis (Fasy, Kim, Lecci, Maria, Milman, Rouvreau, 2014a)

Introduction

Topological Data Analysis: Persistent Homology

Statistical Inference for Persistent Homology

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

Application of Topological Data Analysis to Machine Learning

- Featurization of Topological Data Analysis using Persistence Landscapes

Computation for Topological Data Analysis

- R Package TDA: Statistical Tools for Topological Data Analysis

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.

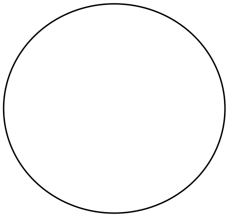
1. $\beta_0 = \#$ of connected components ●
2. $\beta_1 = \#$ of loops ○

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

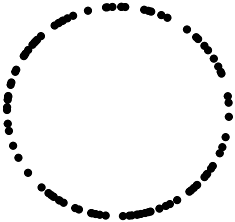
Homology of finite sample is different from homology of underlying manifold, hence it cannot be directly used for the inference.

- 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:

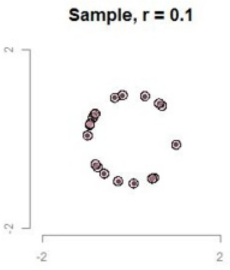
Underlying circle: $\beta_0 = 1$, $\beta_1 = 1$



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

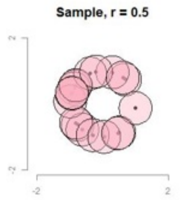


Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.

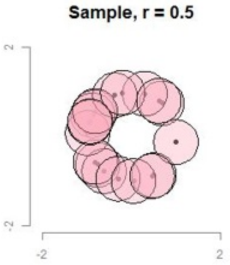


Augmented Set

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

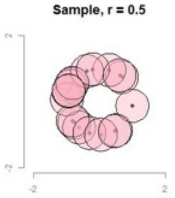


r=**0.5** : 1-dim hole is formed

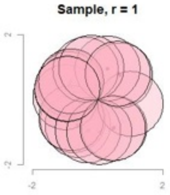


Augmented Set

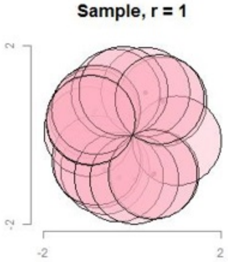
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



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

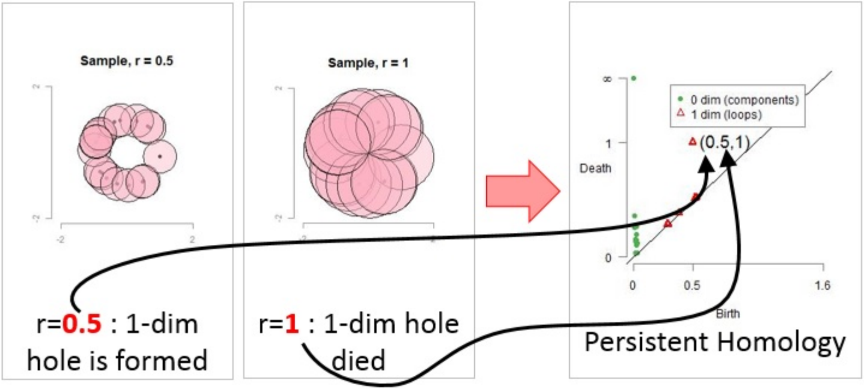


r=**1** : 1-dim hole died



Augmented Set

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



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

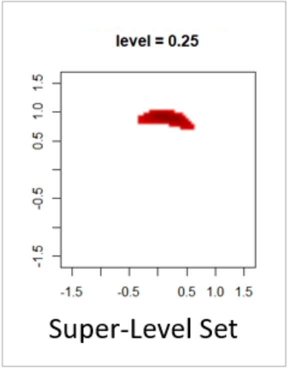
- The kernel density estimator is

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

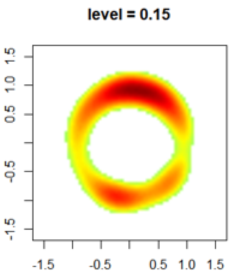
- We look at superlevel sets of the kernel density estimator as

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

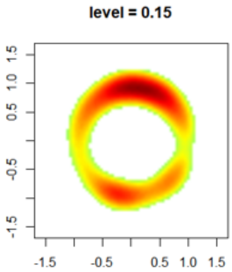
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



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

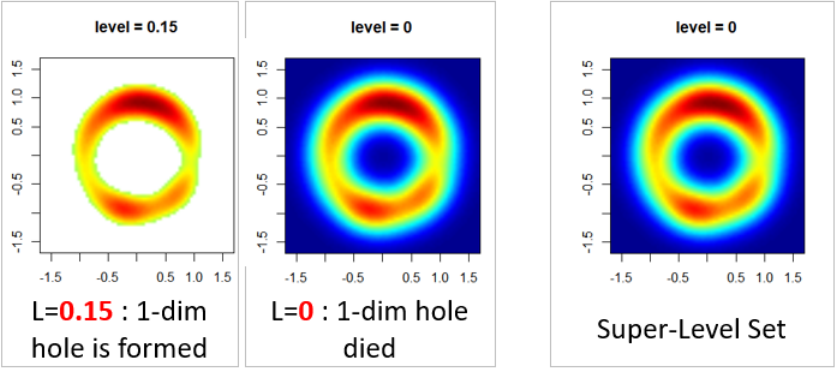


L=0.15 : 1-dim hole is formed

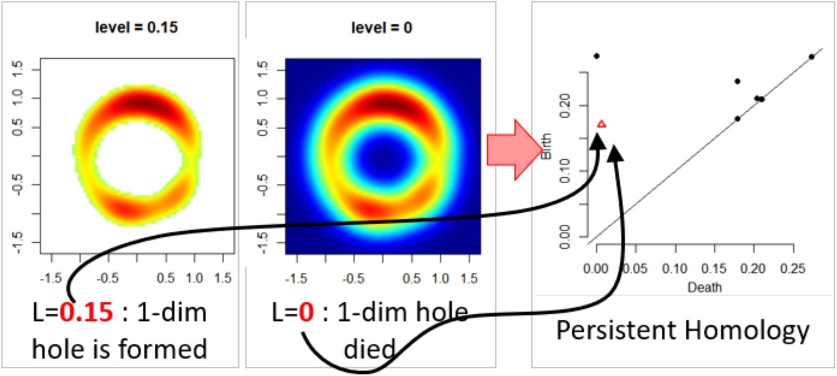


Super-Level Set

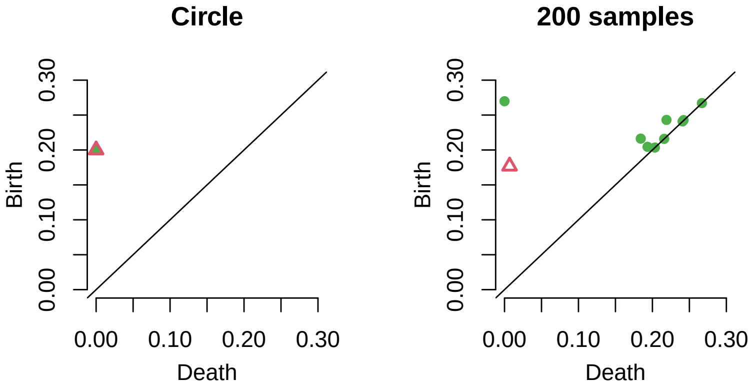
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



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.



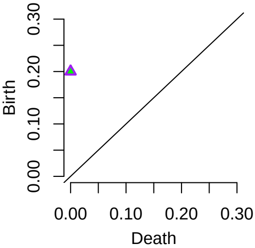
Bottleneck distance gives a metric on the space of persistent homology.

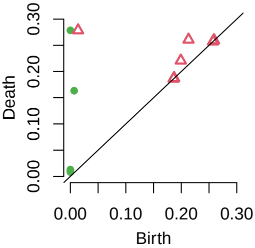
Definition

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.





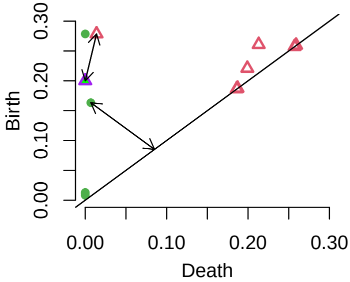
Bottleneck distance gives a metric on the space of persistent homology.

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.



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

Topological Data Analysis: Persistent Homology

Statistical Inference for Persistent Homology

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

Application of Topological Data Analysis to Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

Computation for Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

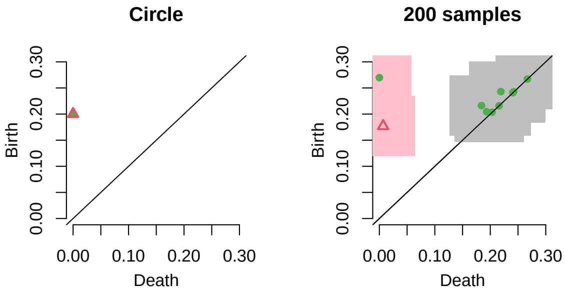
Statistical inference for persistent homology.

- Confidence sets for persistence diagrams (Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2014b)
- Persistent homology of KDE filtration on Vietoris-Rips complex (Shin, Kim, Rinaldo, Wasserman, 2021?)

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

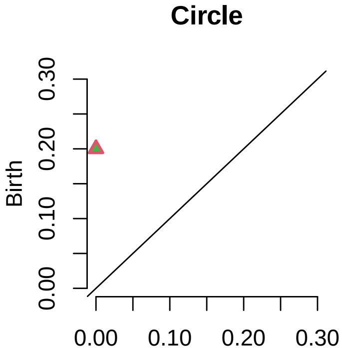
$$\mathbb{P}\left(Dgm(f_M) \in \{\mathcal{D} : d_B(\mathcal{D}, Dgm(f_X)) \leq c_n\}\right) \geq 1 - \alpha.$$

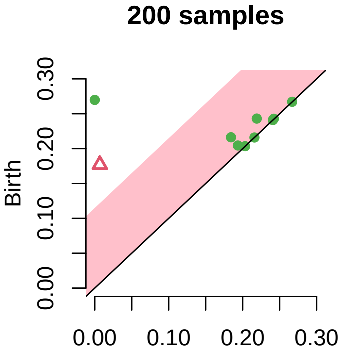


Confidence band for persistent homology separates homological signal from homological noise.

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

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





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

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

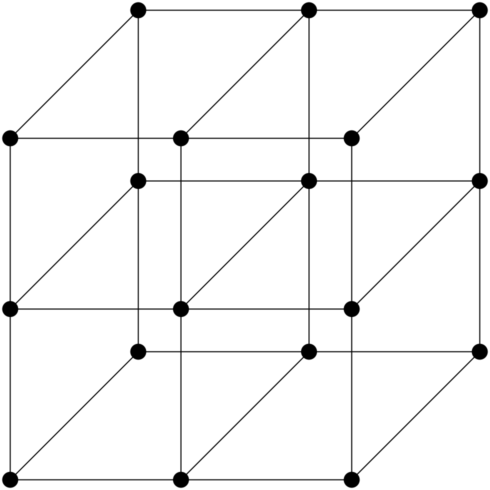
$$\mathbb{P}(d_B(Dgm(f_M), Dgm(f_X)) \leq c_n) \geq \mathbb{P}(||f_M - f_X||_\infty \leq c_n) \geq 1 - \alpha,$$

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

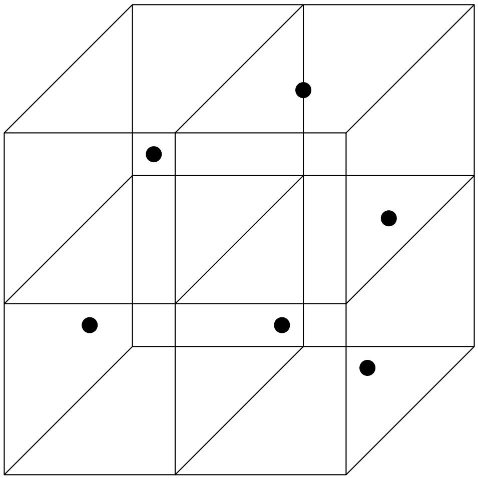
Confidence band for the persistent homology can be computed using the bootstrap algorithm.

1. Given a sample $X = \{x_1, \ldots, x_n\}$, 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 $\theta^* = \sqrt{nh^m}||\hat{p}_h^*(x) - \hat{p}_h(x)||_\infty$, where $\hat{p}_h^*$ is the density estimator computed using X^*.
3. Repeat the previous step B times to obtain $\theta_1^*, \ldots, \theta_B^*$
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 $\left[\hat{p}_h - \frac{\hat{z}_\alpha}{\sqrt{nh^m}}\,,\, \hat{p}_h + \frac{\hat{z}_\alpha}{\sqrt{nh^m}}\right]$.

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

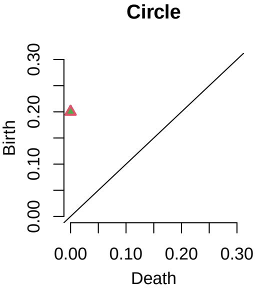


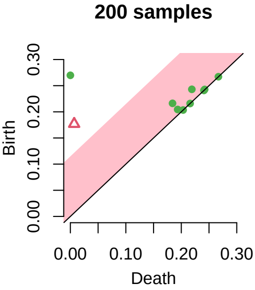
Computing the persistent homology of density function on data points reduces computational complexity.



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

- (Shin, Kim, Rinaldo, Wasserman, 2021?) : extending work from Fasy et al. [2014b], Bobrowski et al. [2014], Chazal et al. [2011].

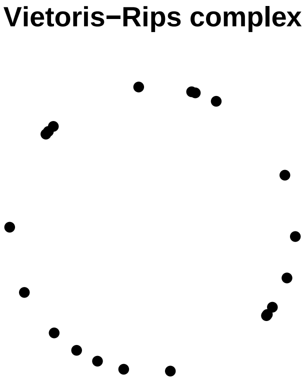




We use the Vietoris-Rips complex to estimate the target persistent homology.

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

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

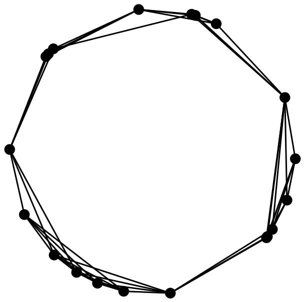


We use the Vietoris-Rips complex to estimate the target persistent homology.

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

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

Vietoris−Rips complex

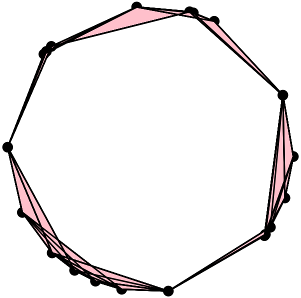


We use the Vietoris-Rips complex to estimate the target persistent homology.

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

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

Vietoris−Rips complex



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

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

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

- The KDE (kernel density estimator) is

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

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

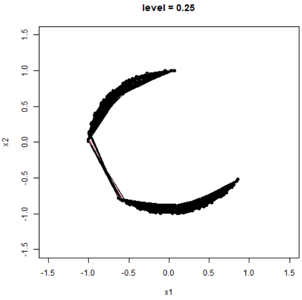
$$\left\{ \mathrm{Rips}\left(\mathcal{X}_{n,L}^{\hat{p}_h}, r \right) \right\}_{L>0}, \text{ where } \mathcal{X}_{n,L}^{\hat{p}_h} = \{ X_i \in \mathcal{X}_n : \, \hat{p}_h(X_i) \geq L \}.$$

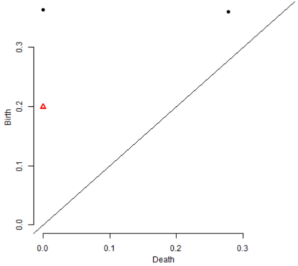
We estimate the target persistent homology by the persistent homology of the KDE filtration on Vietoris-Rips complexes.

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

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

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





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

Theorem

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

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

Confidence set

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

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

Confidence set for the persistent homology of the KDE filtration.

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

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

  This is a valid confidence set by the following theorem.

Theorem
(Theorem 20)

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

Introduction

Topological Data Analysis: Persistent Homology

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

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

Computation for Topological Data Analysis
- R Package TDA: Statistical Tools for Topological Data Analysis

(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 $\nabla_\theta \mathcal{L}(f_\theta, \mathcal{X})$:

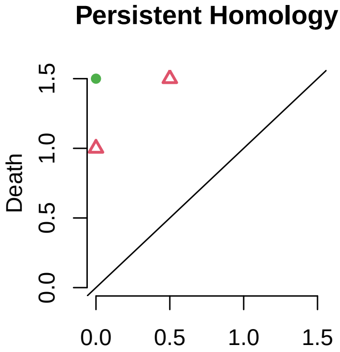
$$\theta_{n+1} = \theta_n - \lambda \nabla_\theta \mathcal{L}(f_\theta, \mathcal{X}).$$

Application of Topological Data Analysis to Machine Learning

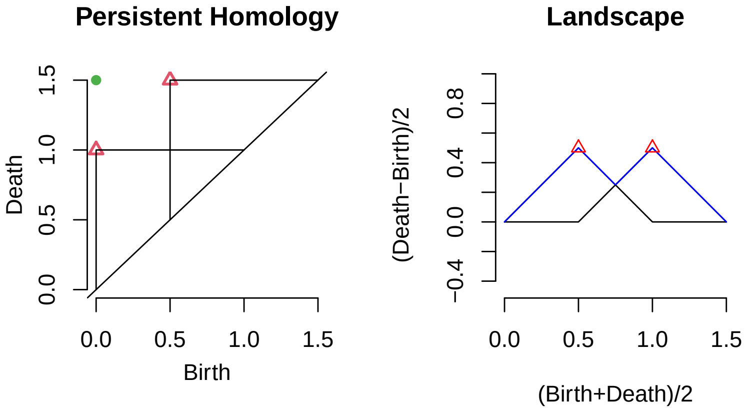
- Application of Topological Data Analysis to Machine Learning is usually in two directions:
  - using TDA as features, so that the data X is augmented with extra TDA features : more common
  - Loss function $\mathcal{L}$ is accompanied with topological loss terms : recently received attentions

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.



Featurizing using Persistence Landscape

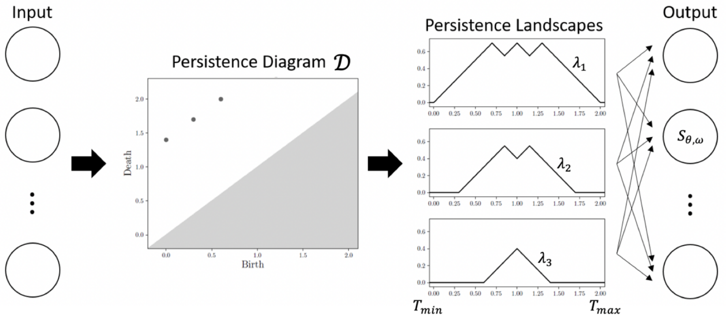
- Featurization using time-delayed embedding and Persistence Landscape
  - Time Series Featurization via Topological Data Analysis (Kim, Kim, Rinaldo, Chazal, 2020)
- Build topological layer using Persistence Landscape
  - PLLay: Efficient Topological Layer based on Persistence Landscapes (Kim, Kim,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

Featurization using time-delayed embedding and Persistence Landscape

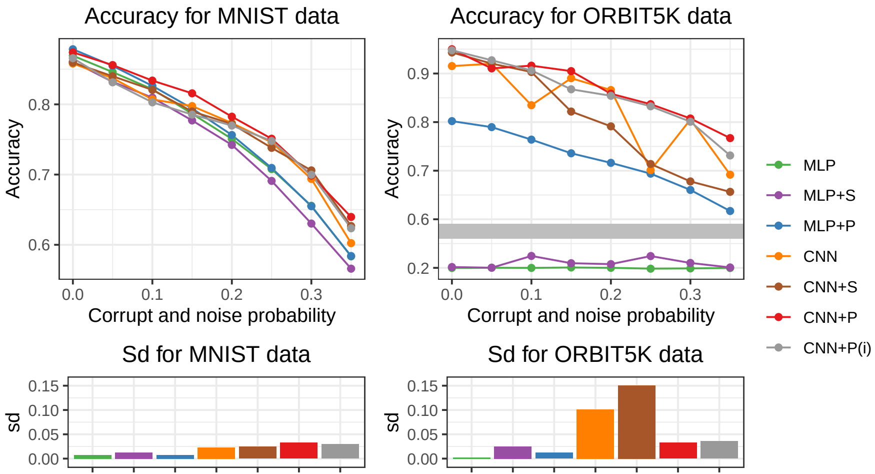
1. From time series data $x = \{x_0, \cdots, x_N\} \subset \mathbb{R}$, construct the point cloud $X \subset \mathbb{R}^m$ using the time-delayed embedding.
2. Perform PCA(Principal Component Analysis) on X and obtain $X^\ell \subset \mathbb{R}^l$.
3. Construct the Vietoris-Rips filtration R_{X^l} and 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



Introduction

Topological Data Analysis: Persistent Homology

Statistical Inference for Persistent Homology

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

Application of Topological Data Analysis to Machine Learning

- Featurization of Topological Data Analysis using Persistence Landscapes

Computation for Topological Data Analysis

- R Package TDA: Statistical Tools for Topological Data Analysis

There are many programs for Topological Data Analysis.

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

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

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

Thank you!

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

References

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

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

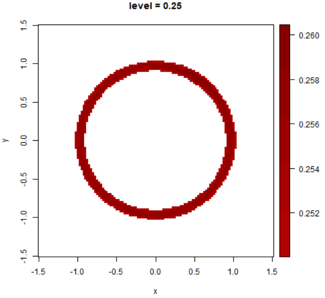
- We are considering the upper level sets of the average kernel density estimator

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

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

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

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

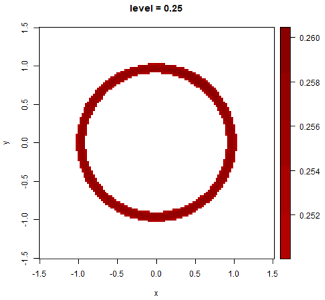


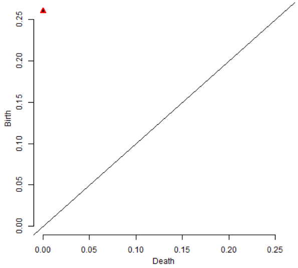
We are targeting the persistent homology of the upper level set of the average kernel density estimator on the support.

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

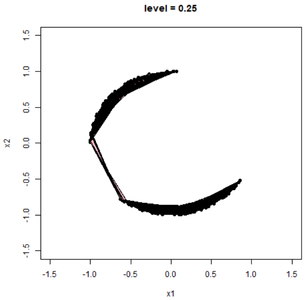




We estimate the target level set by considering the Vietoris-Rips complex generated from the level set of the KDE.

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

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



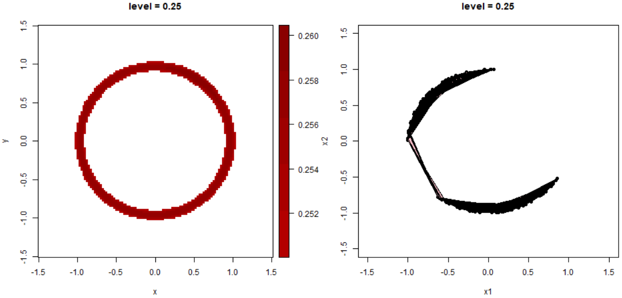
We estimate the target level set by Vietoris-Rips complexes from the KDE level sets.

- We approximate the target level set

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

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

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



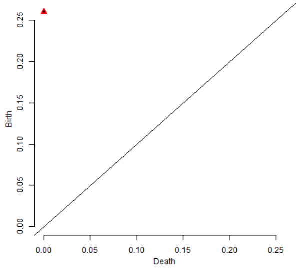
We estimate the target persistent homology by the persistent homology of the KDE filtration on Vietoris-Rips complexes.

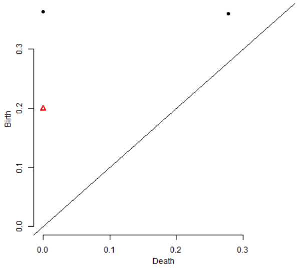
- We estimate the target persistent homology

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

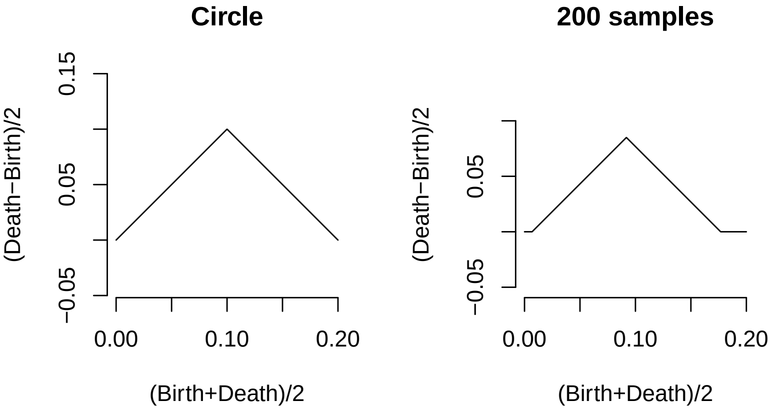
by the persistent homology of the KDE filtration on Vietoris-Rips complexes,

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

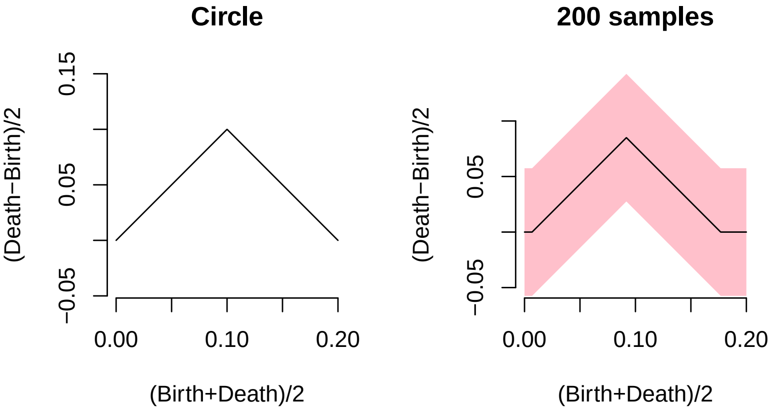




Persistence Landscape of the underlying manifold can be inferred from Persistence Landscape of finite samples.



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

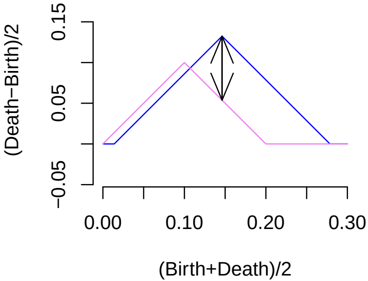


∞-landscape distance gives a metric on the space of persistence landscapes.

Definition

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

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



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

Theorem

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

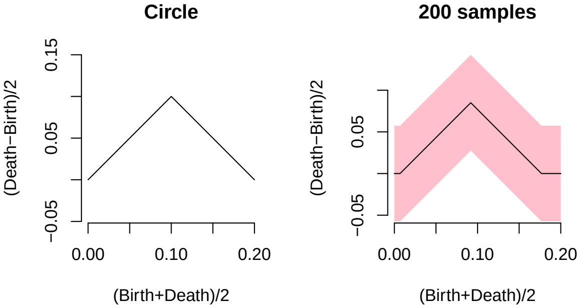
$$\Lambda_\infty(Dgm(f),\, Dgm(g)) \leq \|f - g\|_\infty .$$

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

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

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

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



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

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

Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

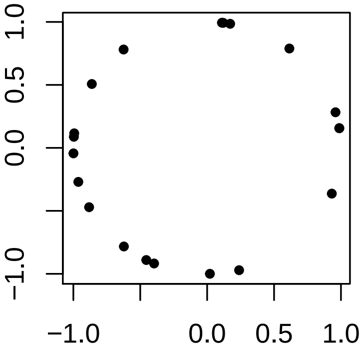
Statistical Inference on Persistence Homology and Persistence Landscape

References

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

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

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



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

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

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

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

R Package TDA provides KDE function over a grid.

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

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

where h is a smoothing parameter.
The function kde() computes the KDE function $\hat{p}_h$ on a grid of points.

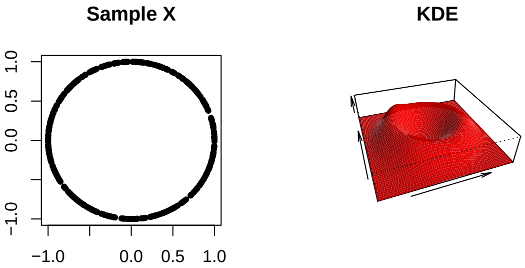
```
h <- 0.3
KDE <- kde(X = X, Grid = Grid, h = h)

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

R Package TDA provides KDE function over a grid.

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

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

where h is a smoothing parameter.
The function kde() computes the KDE function $\hat{p}_h$ on a grid of points.



Statistical Inference for Persistent Homology

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

Featurization using Persistent Homology

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

References

R Package TDA computes Persistent Homology over a grid.

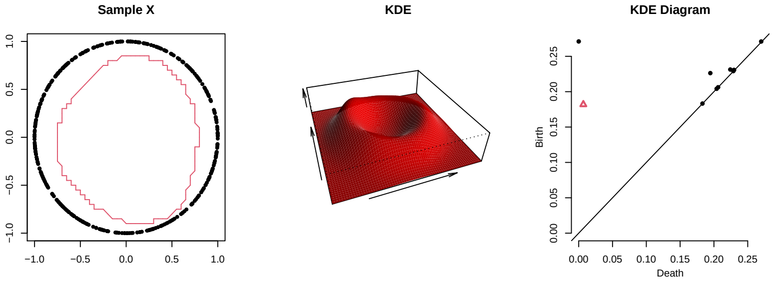
- The function gridDiag() computes the persistence diagram of sublevel (and superlevel) sets of the input function.
  - gridDiag() evaluates the real valued input function over a grid.
  - gridDiag() constructs a filtration of simplices using the values of the input function.
  - gridDiag() computes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

R Package TDA computes Persistent Homology over a grid.

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

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

R Package TDA computes Persistent Homology over a grid.

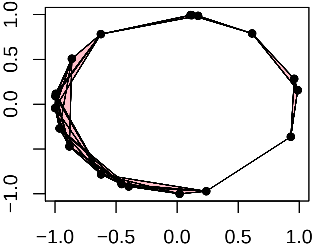
- The function gridDiag() computes the persistent homology of sublevel (and superlevel) sets of the input function.
  - gridDiag() evaluates the real valued input function over a grid.
  - gridDiag() constructs a filtration of simplices using the values of the input function.
  - gridDiag() computes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



R Package TDA computes Vietoris-Rips Persistent Homology.

- Vietoris-Rips complex consists of simplices whose pairwise distances of vertices are at most ϵ apart, i.e.

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



- Rips filtration is formed by Rips complices with gradually increasing ϵ.

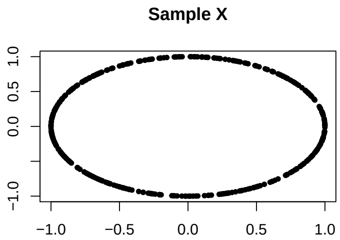
R Package TDA computes Vietoris-Rips Persistent Homology.

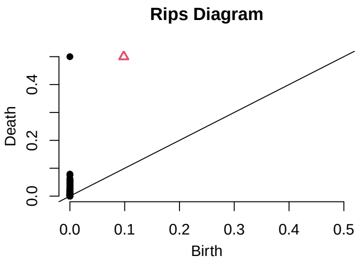
- The function ripsDiag() computes the persistence diagram of the Rips filtration built on top of a point cloud.
  - ripsDiag() constructs the Vietoris-Rips filtration using the data points.
  - ripsDiag() computes the persistent homology of the Vietoris-Rips filtration.
- The user can choose to compute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

```r
DiagRips <- ripsDiag(X = X, maxdimension = 1, maxscale = 0.5,
    library = c("GUDHI", "Dionysus"), location = TRUE)

par(mfrow = c(1,2))
plot(X, xlab = "", ylab = "", main = "Sample X", pch = 20)
plot(x = DiagRips[["diagram"]], main = "Rips Diagram")
```

R Package TDA computes Vietoris-Rips Persistent Homology.

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





R Package TDA computes Persistence Landscape.

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

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

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

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

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

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

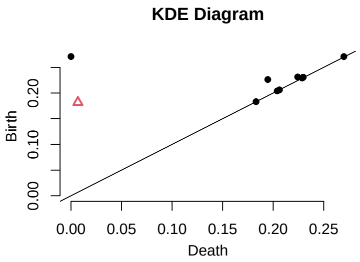
R Package TDA computes Persistence Landscape.

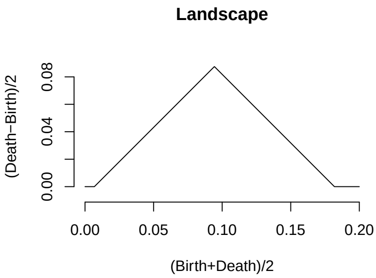
- Let Λ_p be created by tenting each point $p = (x, y) = \left(\frac{b+d}{2}, \frac{d-b}{2}\right)$ representing a birth-death pair (b, d) in the persistence diagram D.
- The persistence landscape of D is the collection of functions

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

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

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





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

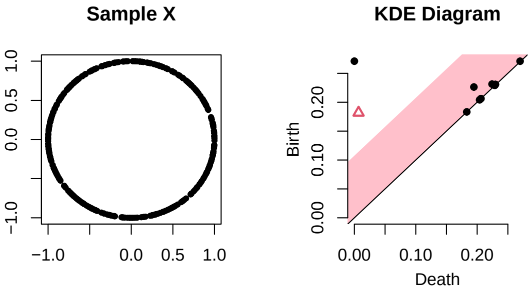
The function bootstrapBand() computes $(1-\alpha)$ bootstrap confidence band for $\mathbb{E}[\hat{p}_h]$.

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

##       90%
## 0.0537502
```

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

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

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



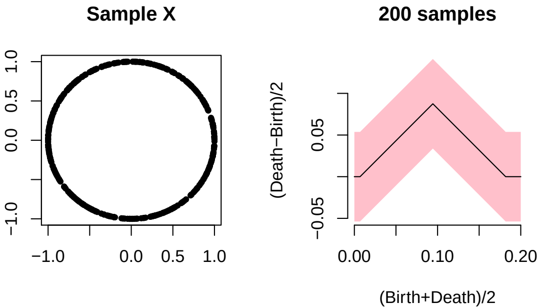
R Package TDA computes the bootstrap confidence band for the persistence landscape.

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

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

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

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



References

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

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *arXiv preprint arXiv:1207.6437*, 2012.

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

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

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Annual Symposium on Computational Geometry*, pages 474–483. ACM, 2014.

H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. Applied mathematics. American Mathematical Society, 2010. ISBN 9780821849255. URL `http://books.google.com/books?id=MDXa6gFRZuIC`.

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