

Introduction to Topological Data Analysis

Jisu KIM



Seoul National University
2024-11-08

Introduction to Topological Data Analysis

Mapper

Homology and Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Data Analysis and Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

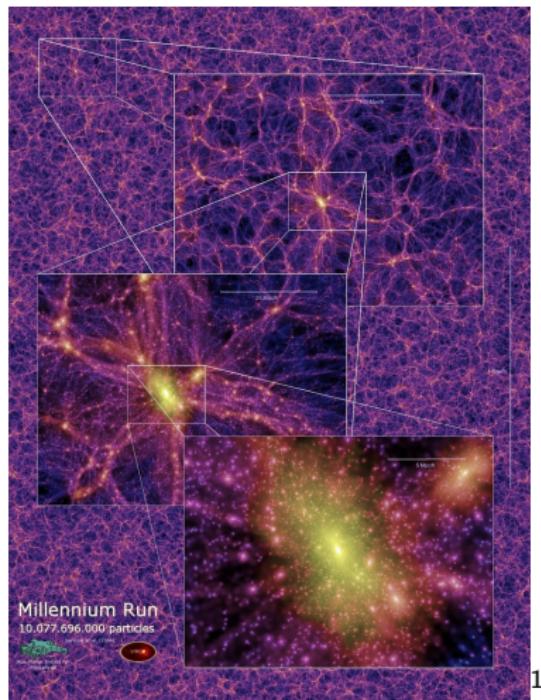
Featurization using Circular Coordinates

Evaluation using Confidence of Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

Reference

Topological structures in the data provide information.



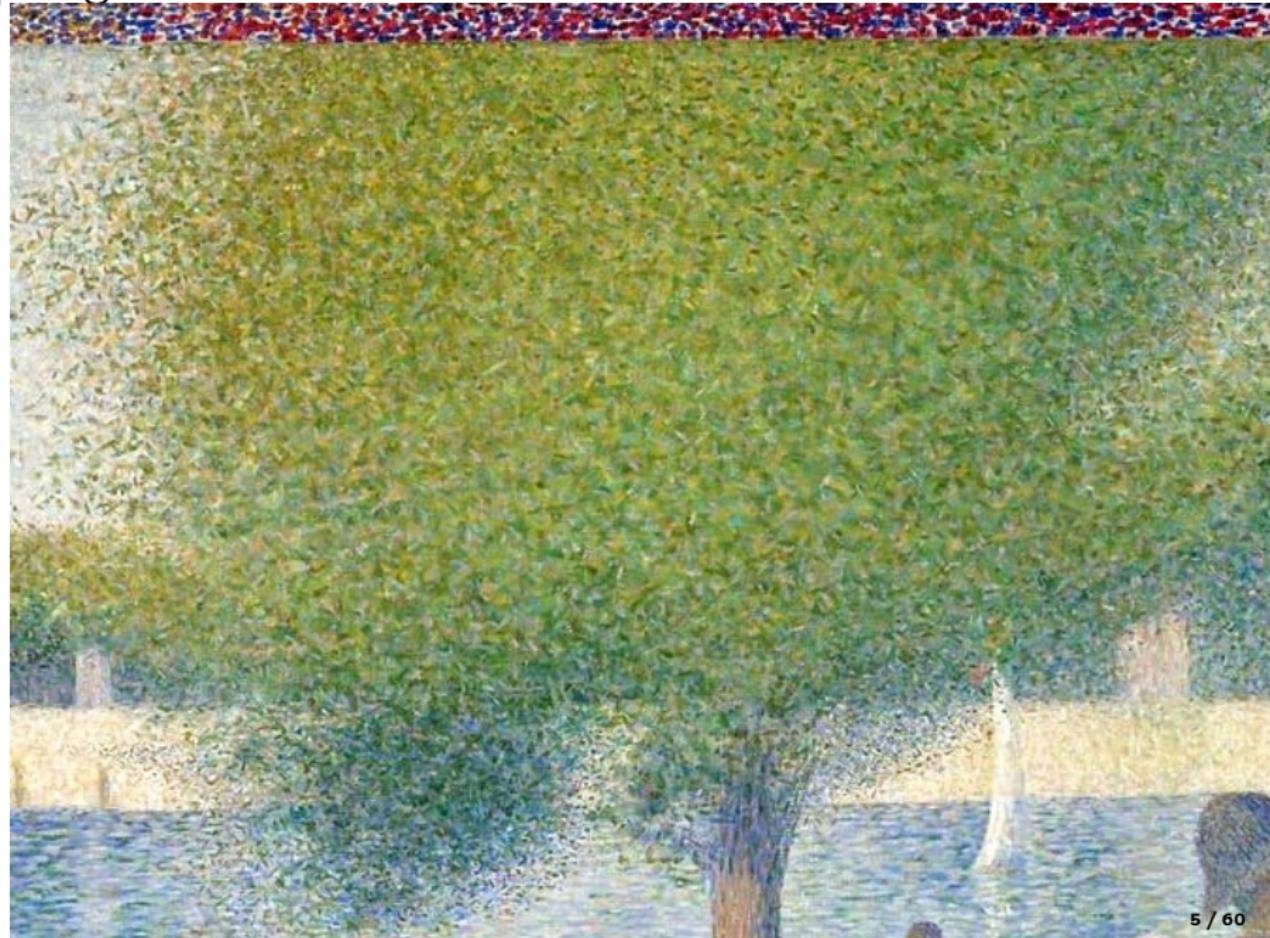
1

¹http://www.mpa-garching.mpg.de/galform/virgo/millennium/poster_half.jpg

Topological structures are observed in different scales.



Topological structures are observed in different scales.



Topological structures are observed in different scales.

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



Statistical Inference for Topological Data Analysis are introduced.

- ▶ 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, 2021)
- ▶ Statistical Inference for Persistent Homology
 - ▶ Confidence sets for persistence diagrams (Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2014b)

Application of Topological Data Analysis to machine learning is introduced.

- ▶ Application of Topological Data Analysis to machine learning
 - ▶ A Survey of Topological Machine Learning Methods (Hensel, Moor, Rieck, 2021)
- ▶ Featurization of Topological Data Analysis
 - ▶ Efficient Topological Layer based on Persistence Landscapes (Kim, Kim, Zaheer, Kim, Chazal, Wasserman, 2020)
 - ▶ Generalized penalty for circular coordinate representation (Luo, Patania, Kim, Vejdemo-Johansson, 2021)
- ▶ Evaluate quality of machine learning model using Topological Data Analysis
 - ▶ TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)
- ▶ R Package TDA: Statistical Tools for Topological Data Analysis
 - ▶ Introduction to the R package TDA (Fasy, Kim, Lecci, Maria, Millman, Rouvreau, 2014a)

Introduction to Topological Data Analysis

Mapper

Homology and Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Data Analysis and Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

Featurization using Circular Coordinates

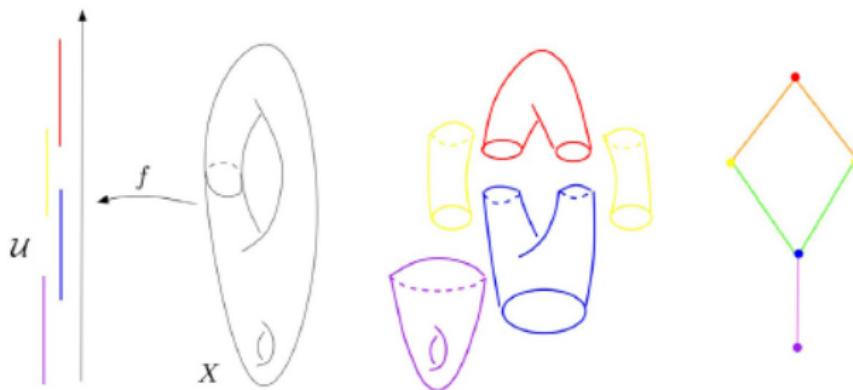
Evaluation using Confidence of Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

Reference

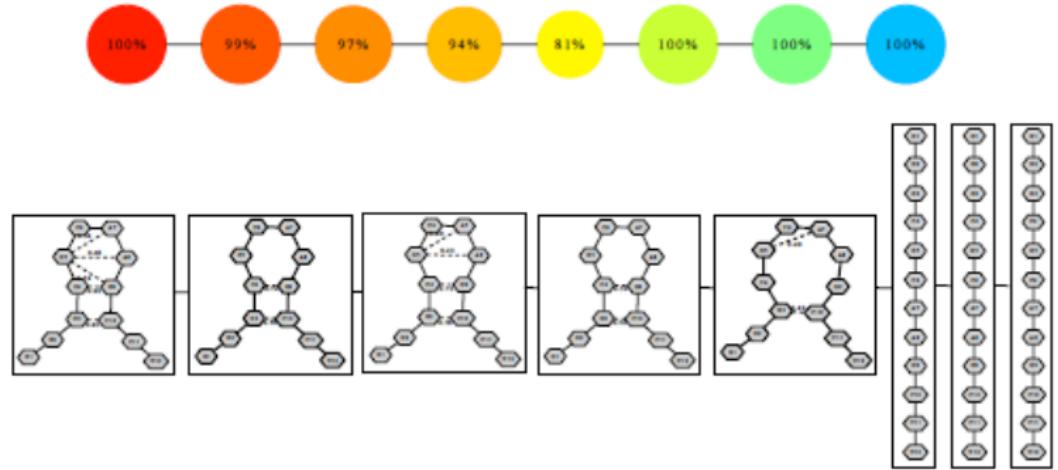
Mapper summarizes and visualizes data as a graph.

- ▶ Given a function $f : X \rightarrow \mathbb{R}^d$ and a cover $\mathcal{U} = \{U_i\}_{i \in I}$ with $\mathbb{R}^d \subset \bigcup_{i \in I} U_i$, we consider a collection of connected components of $f^{-1}(U_i)$.
- ▶ Construct a graph from these connected components \Rightarrow summarizes how data are connected.



Mapper reveals topological structure of data: unfolding pathway

- ▶ Data are conformations of molecules, 760 from unfolding events and 550 from folding events.
- ▶ Mapper of unfolding events has one unfolding pathway.

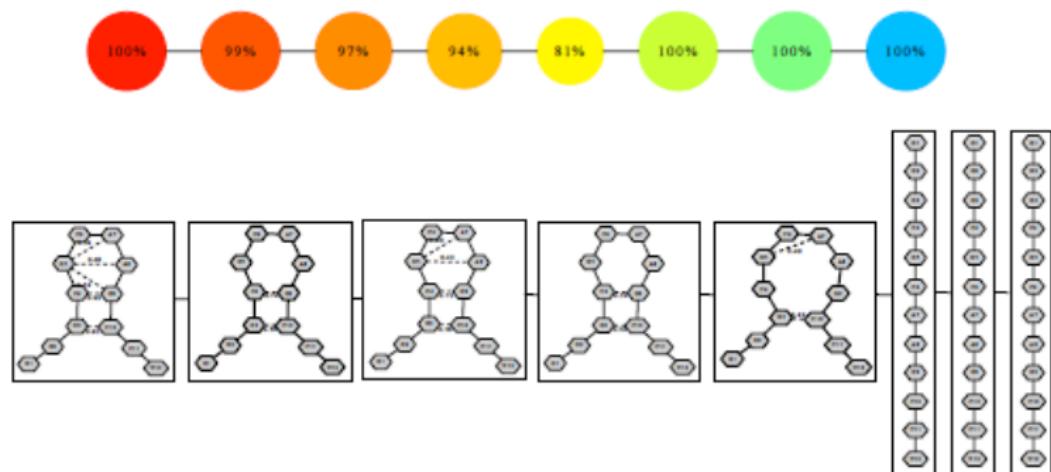


3

³Yao et al. [2009, Figure 3(a)] Unfolding pathway.

Mapper reveals topological structure of data: folding pathway

- ▶ Data are conformations of molecules, 760 from unfolding events and 550 from folding events.
- ▶ Mapper of folding events has two refolding pathways \Rightarrow refolding pathway has two different pathways to follow.



4

⁴Yao et al. [2009, Figure 3(b)] Unfolding pathway.

Introduction to Topological Data Analysis

Mapper

Homology and Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Data Analysis and Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

Featurization using Circular Coordinates

Evaluation using Confidence of Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

Reference

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) 

Example : Objects are classified by homologies.

1. $\beta_0 = \#$ of connected components



2. $\beta_1 = \#$ of loops (holes inside 1-dim sphere)

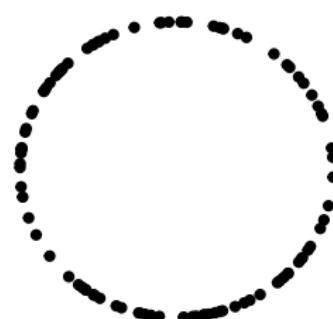
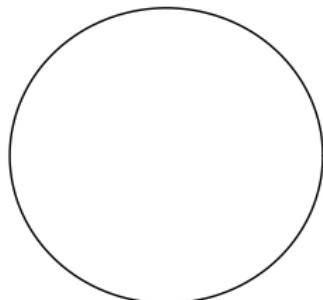


$\beta_0 \setminus \beta_1$	0	1	2
1	ㄱ, ㄴ, ㄷ, ㄹ, ㅅ, ㅈ, ㅌ, ㅊ	ㅁ, ㅇ, ㅂ, ㅍ, ㅏ	ㅓ
2	ㅊ, ㅌ		
3		ㅎ	

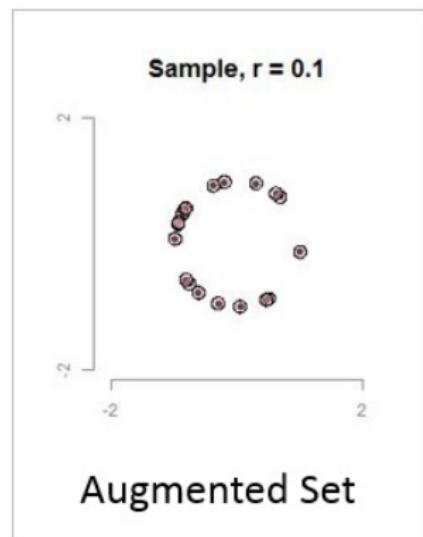
Homology of finite sample is different from homology of underlying manifold, hence it cannot be directly used for the inference.

- ▶ 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:

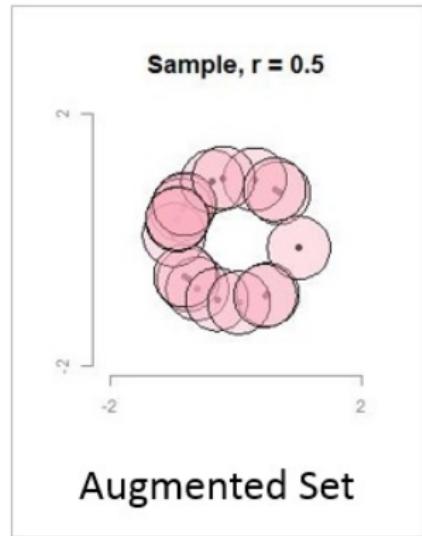
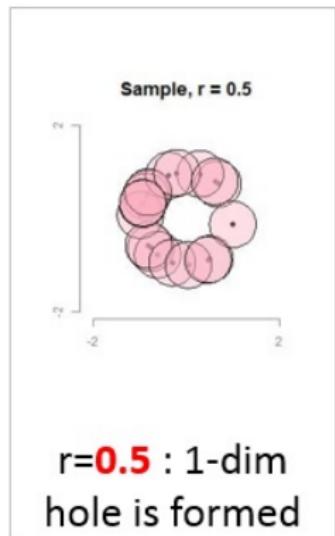
Underlying circle: $\beta_0 = 1$, $\beta_1 = 1$ 100 samples: $\beta_0 = 100$, $\beta_1 = 0$



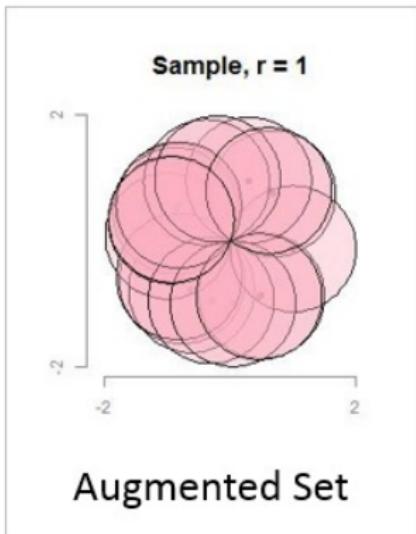
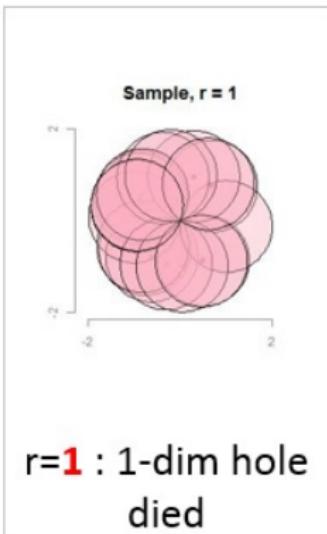
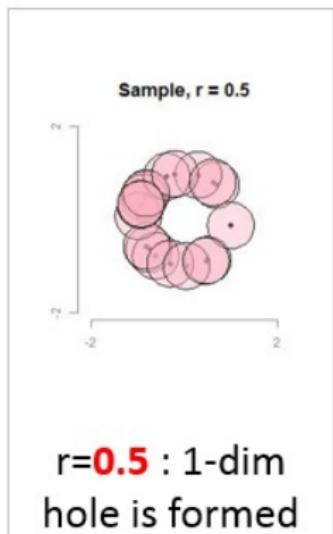
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



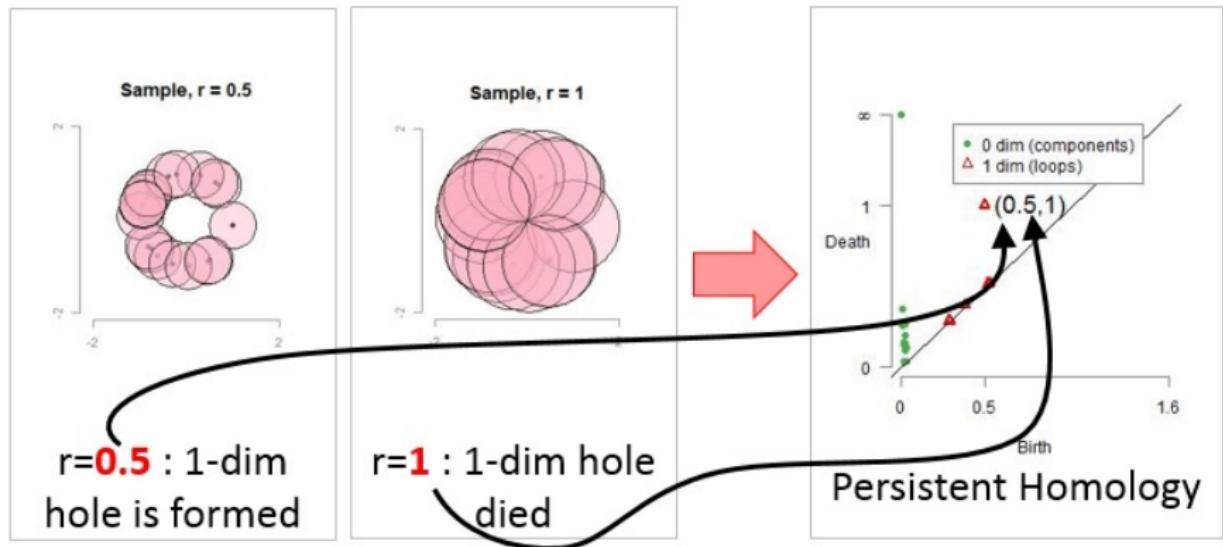
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



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



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



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

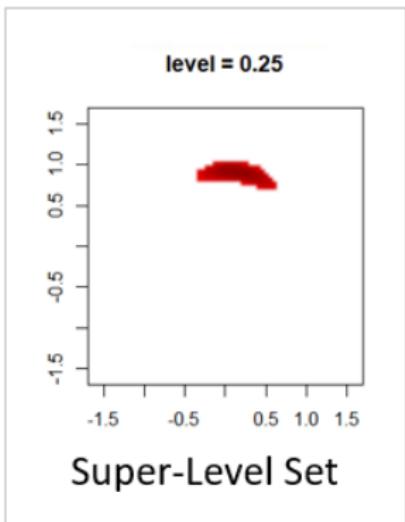
- ▶ The kernel density estimator is

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

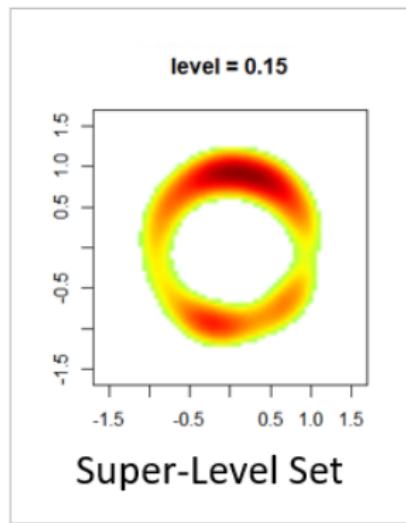
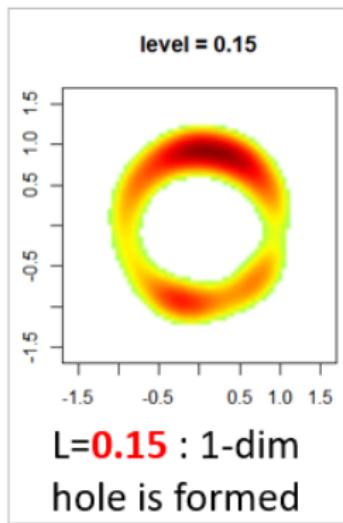
- ▶ We look at superlevel sets of the kernel density estimator as

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

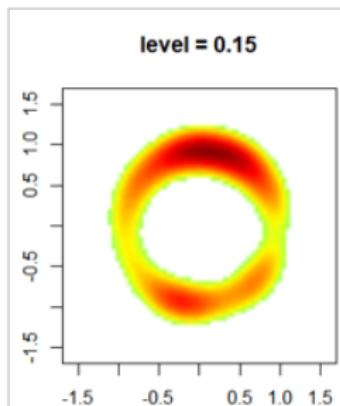
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



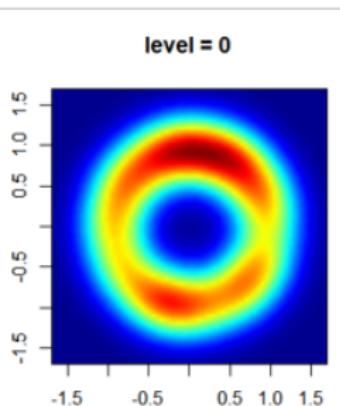
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



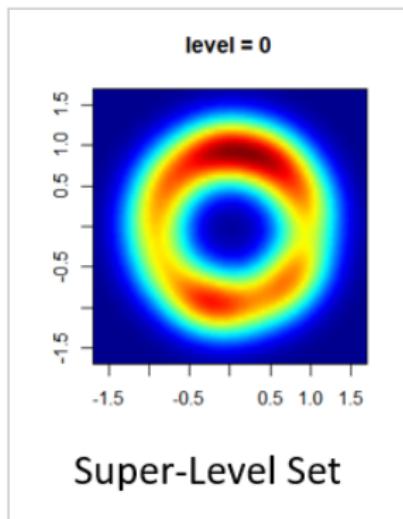
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



L=0.15 : 1-dim
hole is formed

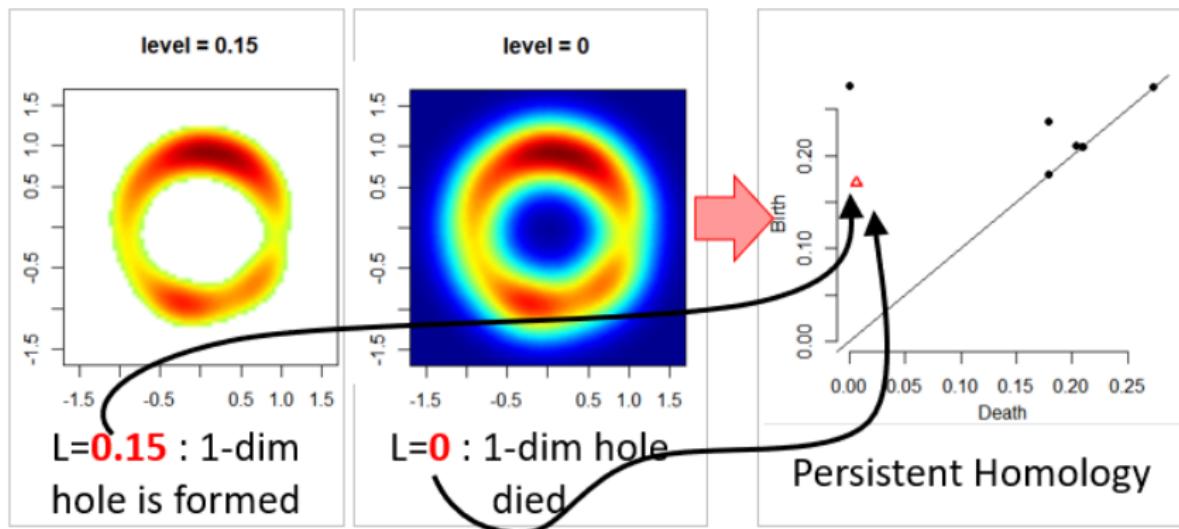


L=0 : 1-dim hole
died



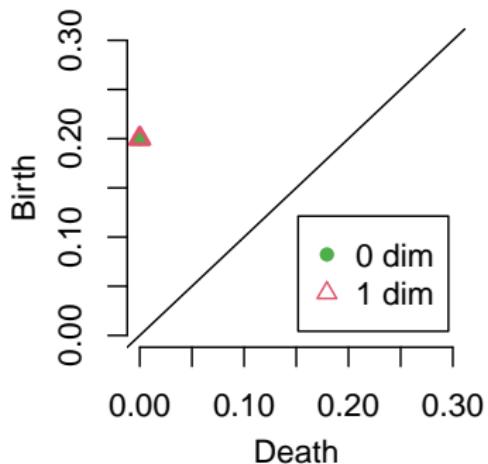
Super-Level Set

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

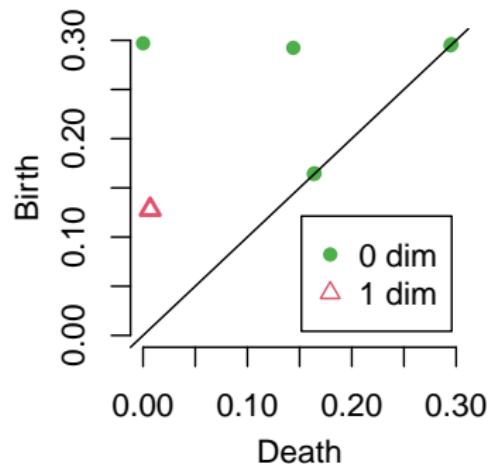


유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.

Circle

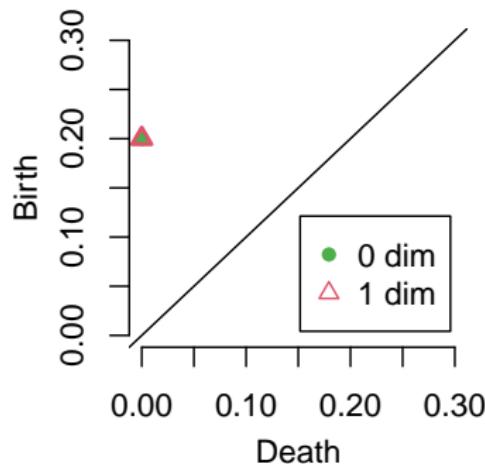


100 samples

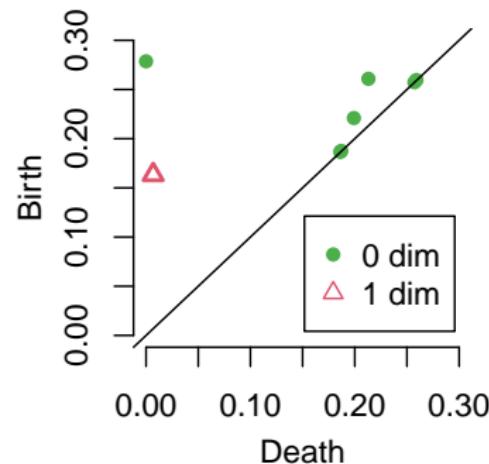


Persistent homology of the underlying structure can be inferred from persistent homology of finite samples.

Circle

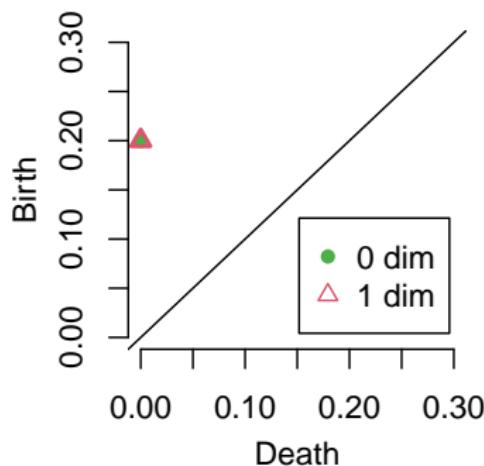


150 samples

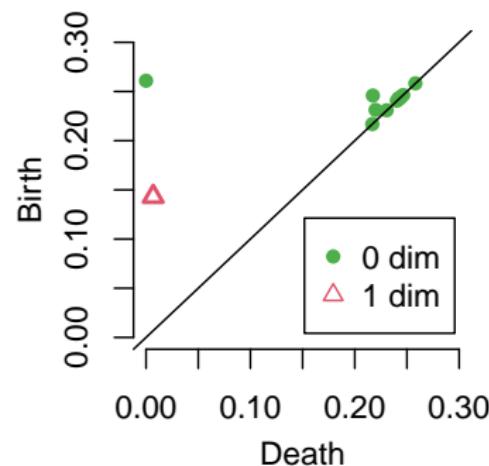


Persistent homology of the underlying structure can be inferred from persistent homology of finite samples.

Circle

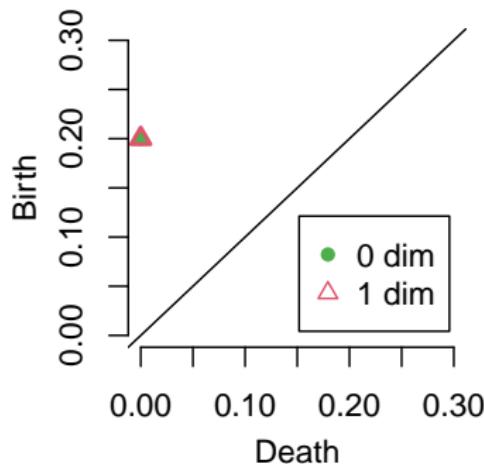


200 samples

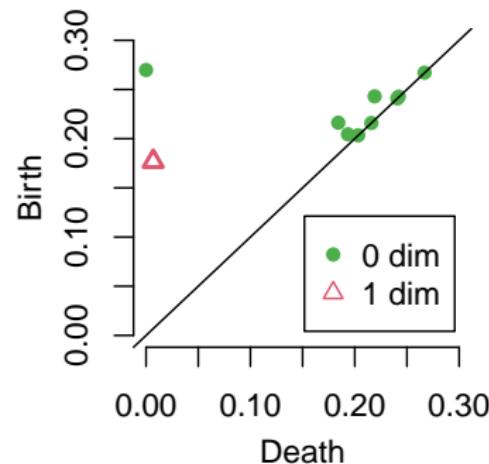


Persistent homology of the underlying structure can be inferred from persistent homology of finite samples.

Circle

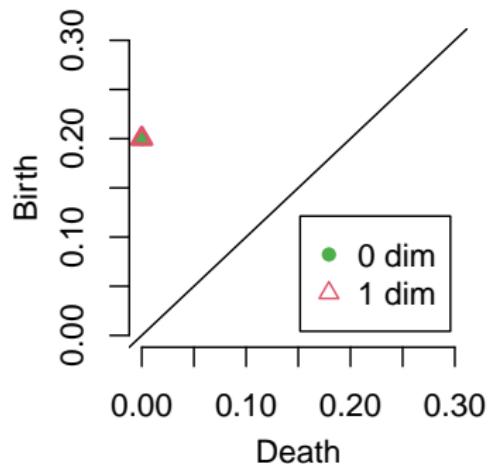


500 samples

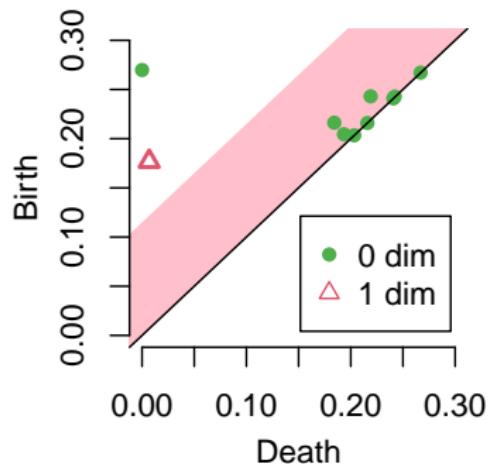


How can we separate statistically significant homology features and noisy homology features?

Circle



500 samples



Introduction to Topological Data Analysis

Mapper

Homology and Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Data Analysis and Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

Featurization using Circular Coordinates

Evaluation using Confidence of Topological Data Analysis

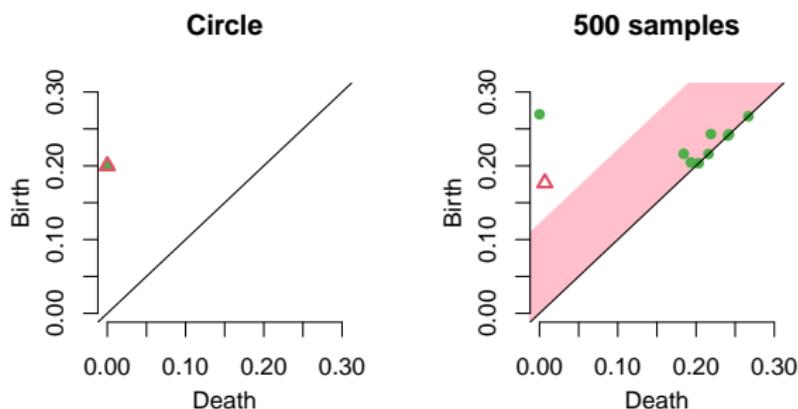
R Package TDA: Statistical Tools for Topological Data Analysis

Reference

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

Let $Dgm(M)$ and $Dgm(X)$ be persistent homologies of the manifold M and the data X , respectively. Given the significance level $\alpha \in (0, 1)$, $(1 - \alpha)$ confidence band $c_n = c_n(X)$ is a random variable satisfying

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



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

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

Introduction to Topological Data Analysis

Mapper

Homology and Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Data Analysis and Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

Featurization using Circular Coordinates

Evaluation using Confidence of Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

Reference

(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

- ▶ A Survey of Topological Machine Learning Methods (Hensel, Moor, Rieck, 2021)
- ▶ Topological Data Analysis (TDA) is applied to Machine Learning in usually two directions:
 - ▶ Make features from TDA, so that the data X is augmented with extra TDA features: more common
 - ▶ PLLay: Efficient Topological Layer based on Persistence Landscapes (Kim, Kim, Zaheer, Kim, Chazal, Wasserman, 2020)
 - ▶ Generalized penalty for circular coordinate representation (Luo, Patania, Kim, Vejdemo-Johansson, 2021)
 - ▶ Evaluate quality of data \mathcal{X} or model f_θ using TDA: recently received attentions
 - ▶ TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)

Introduction to Topological Data Analysis

Mapper

Homology and Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Data Analysis and Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

Featurization using Circular Coordinates

Evaluation using Confidence of Topological Data Analysis

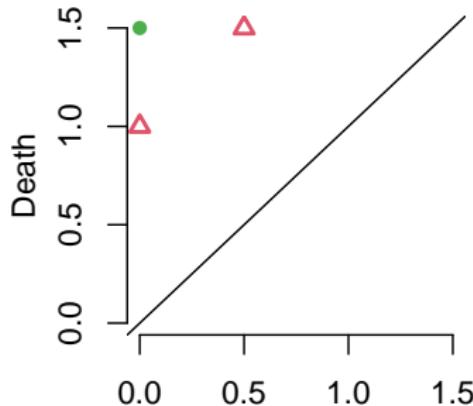
R Package TDA: Statistical Tools for Topological Data Analysis

Reference

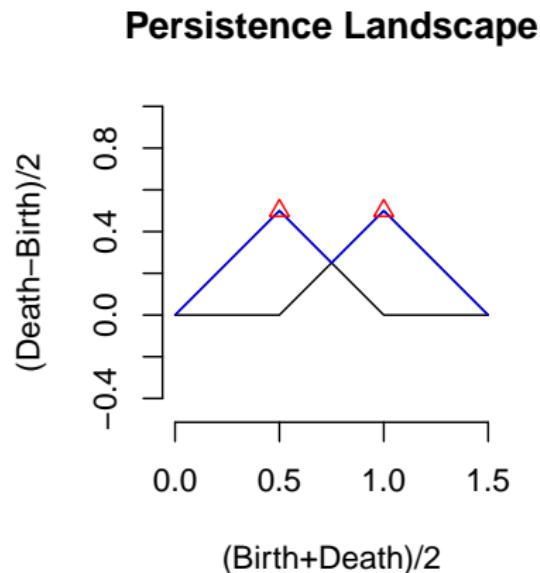
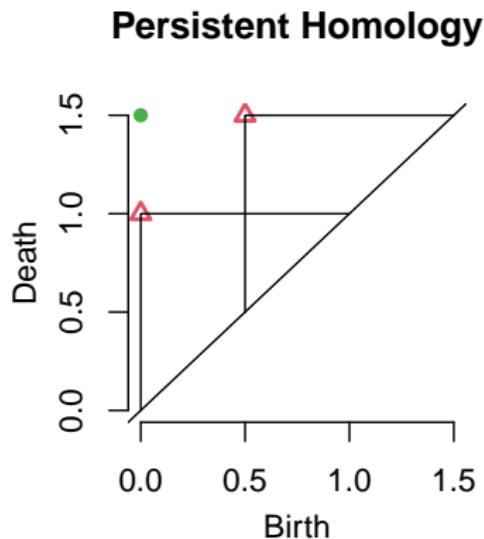
Persistent homology is further summarized and embedded into a Euclidean space or a functional space.

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

Persistent Homology



Persistence Landscape is a functional summary of the persistent homology.

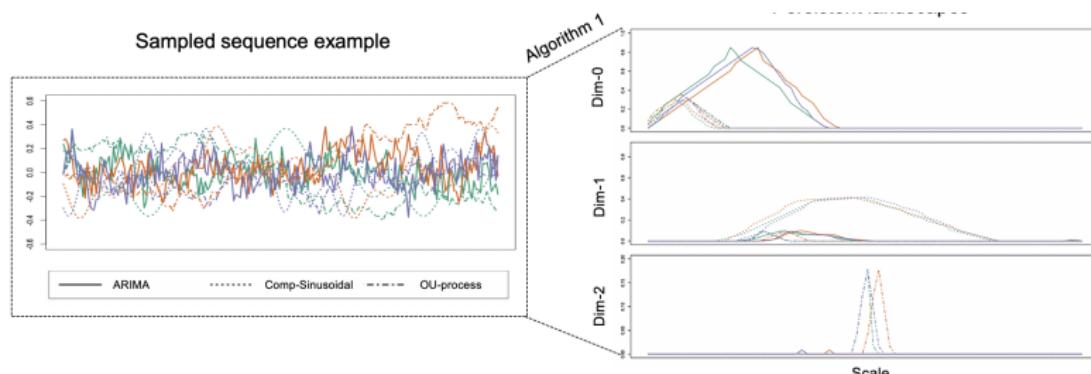


Featurizing using Persistence Landscape

- ▶ 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)

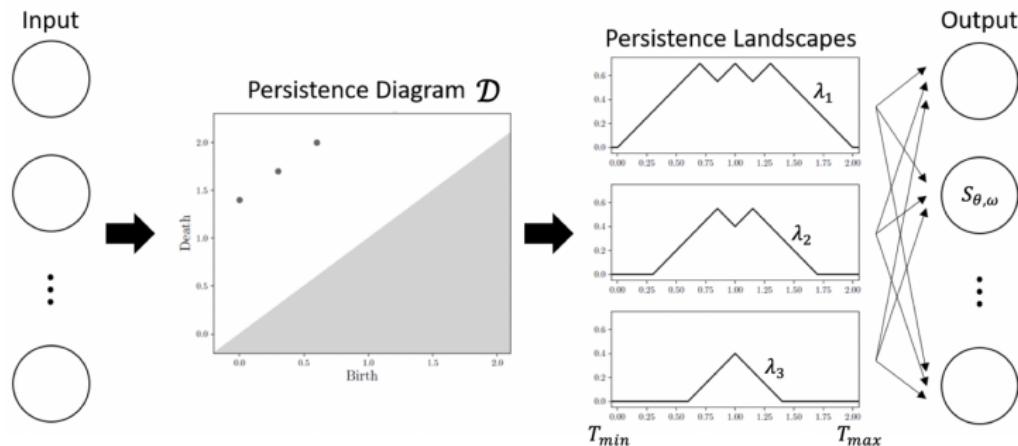
Featurization using time-delayed embedding and Persistence Landscape

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

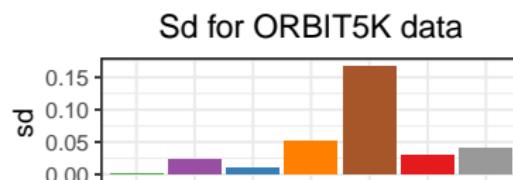
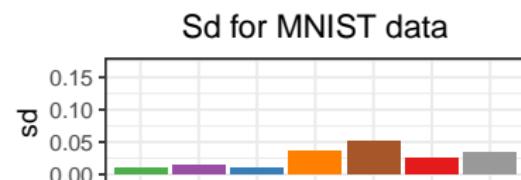
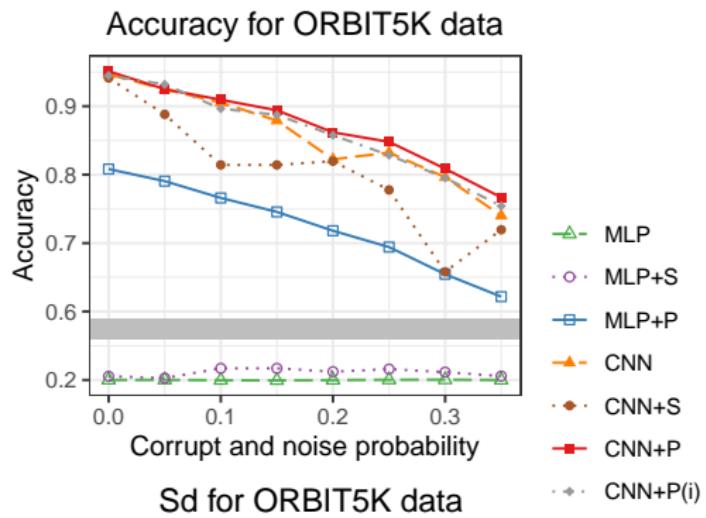
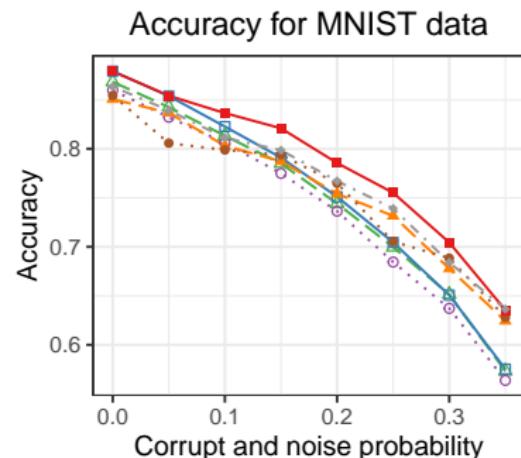


Build topological layer using Persistence Landscape

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



Build topological layer using Persistence Landscape



Introduction to Topological Data Analysis

Mapper

Homology and Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Data Analysis and Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

Featurization using Circular Coordinates

Evaluation using Confidence of Topological Data Analysis

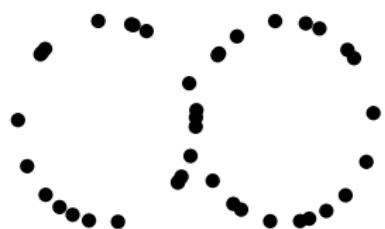
R Package TDA: Statistical Tools for Topological Data Analysis

Reference

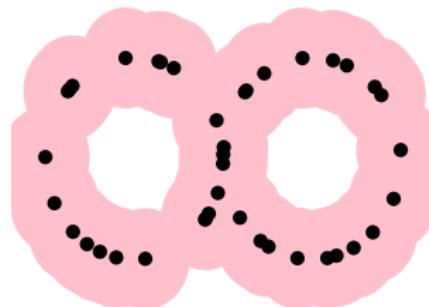
Circular coordinates provide topological representations of reduced dimension.

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

data



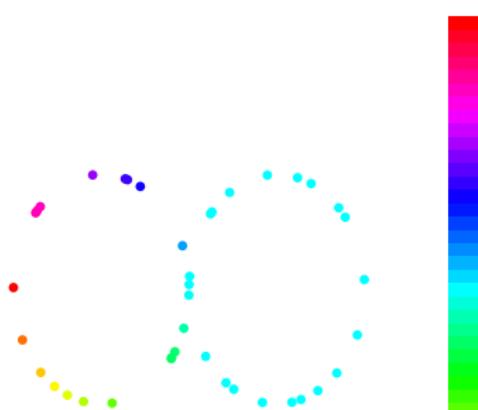
loop



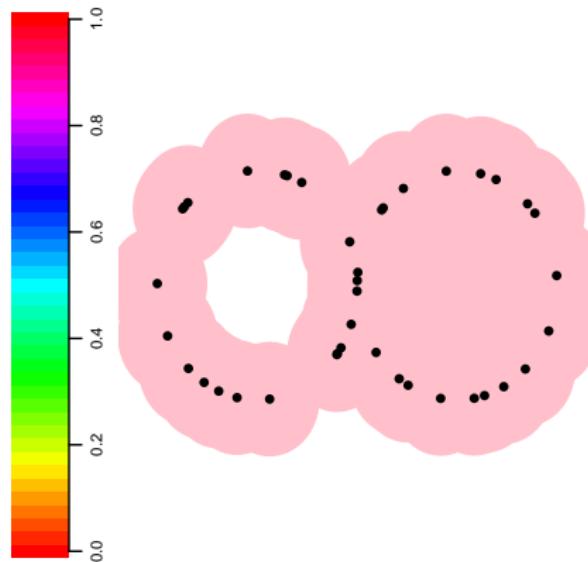
Circular coordinates provide topological representations of reduced dimension.

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

circular coordinates

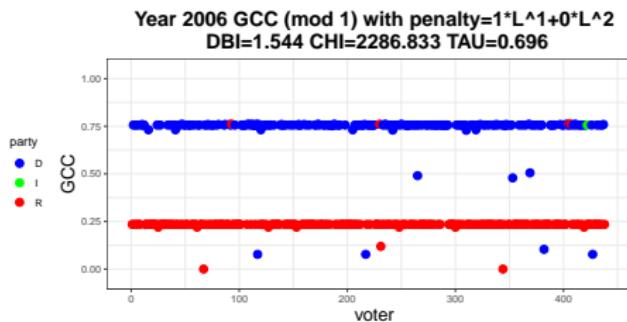
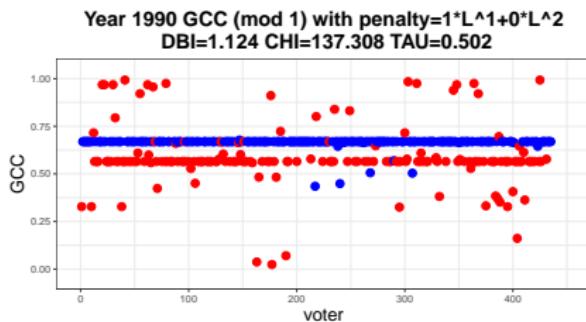


loop



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

- ▶ Generalized penalty for circular coordinate representation (Luo, Patania, Kim, Vejdemo-Johansson, 2021)
- ▶ Voting data in 2006 is more bipolarized than voting data in 1990.



Introduction to Topological Data Analysis

Mapper

Homology and Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Data Analysis and Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

Featurization using Circular Coordinates

Evaluation using Confidence of Topological Data Analysis

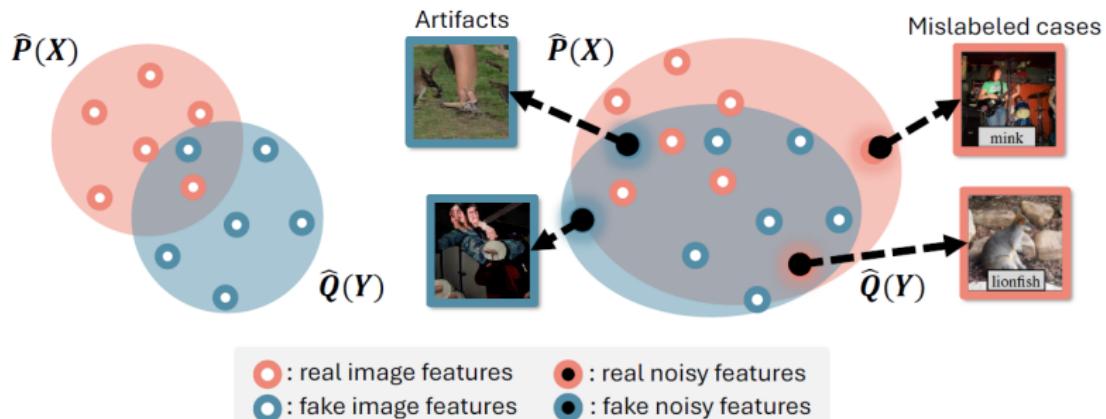
R Package TDA: Statistical Tools for Topological Data Analysis

Reference

Existing evaluation metrics for generative models are vulnerable to noise.

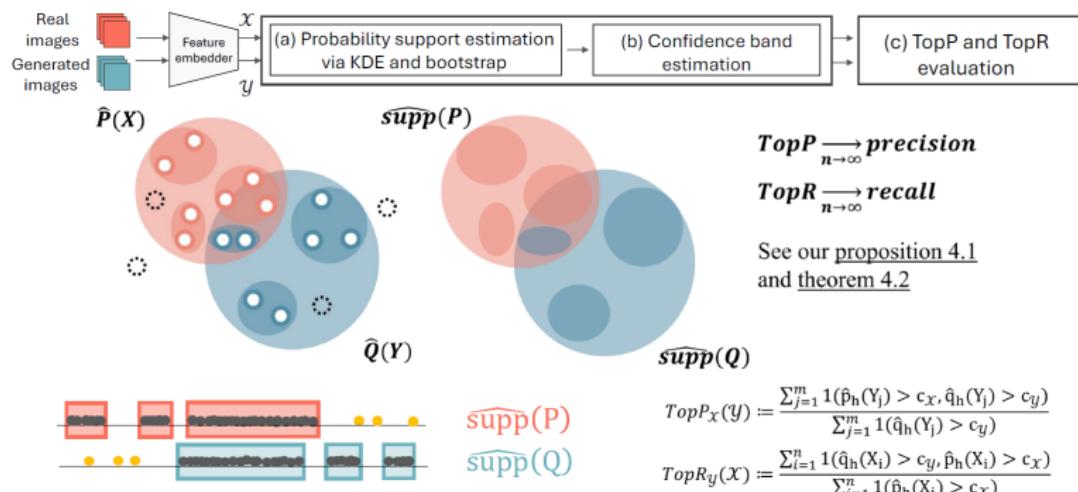
- ▶ TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)
- ▶ To evaluate generative models, metrics compare the support of real image distributions and fake image distributions.
- ▶ Existing evaluation metrics tend to overestimate the support of the data distribution: vulnerable to noise

(1) Ideal estimation of distribution (2) Non-ideal estimation of distribution



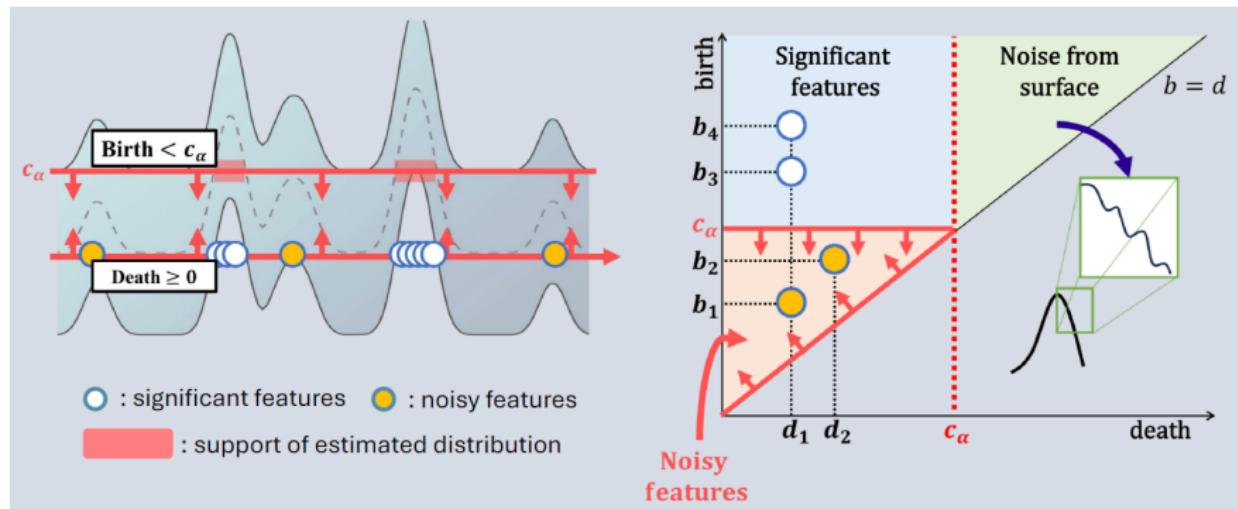
TopP&R robustly evaluates generative models by retaining only topologically and statistically significant features with confidence.

- TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)



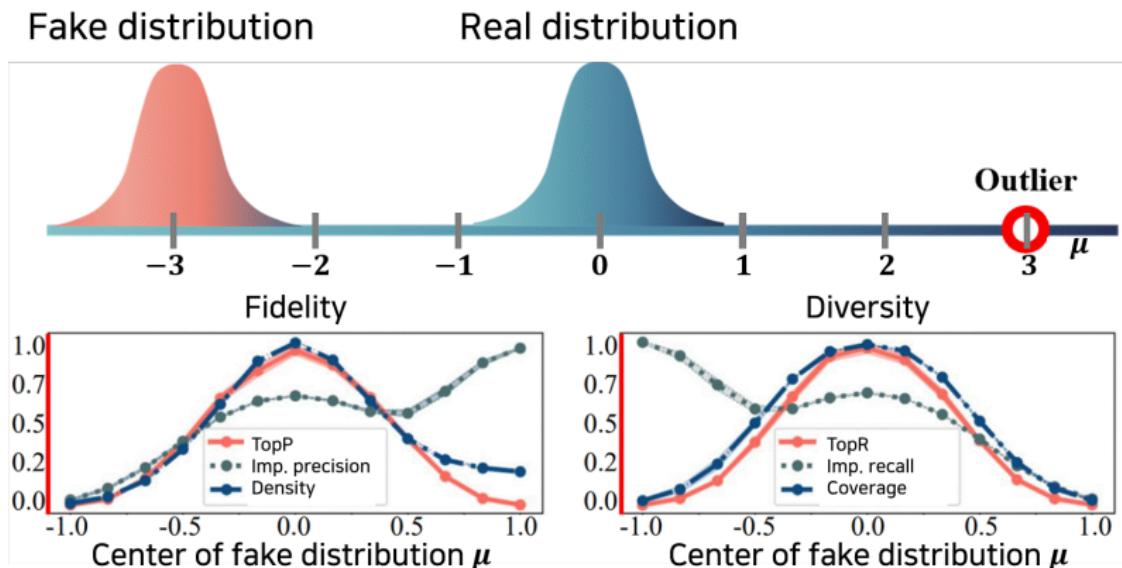
We find threshold c_α that selects statistically and topologically significant features.

- TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)



Experiments

- TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)



Introduction to Topological Data Analysis

Mapper

Homology and Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Data Analysis and Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

Featurization using Circular Coordinates

Evaluation using Confidence of Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

Reference

There are many programs for Topological Data Analysis.

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

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

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

Introduction to Topological Data Analysis

Mapper

Homology and Persistent Homology

Statistical Inference for Persistent Homology

Application of Topological Data Analysis to Data Analysis and Machine Learning

Featurization of Topological Data Analysis using Persistence Landscapes

Featurization using Circular Coordinates

Evaluation using Confidence of Topological Data Analysis

R Package TDA: Statistical Tools for Topological Data Analysis

Reference

Reference |

- Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers Artif. Intell.*, 4:667963, 2021. doi: 10.3389/frai.2021.667963. URL <https://doi.org/10.3389/frai.2021.667963>.
- 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, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance-to-a-measure and kernel distance. *Technical Report*, 2014.
- Herbert Edelsbrunner and John L. Harer. *Computational topology*. American Mathematical Society, Providence, RI, 2010. ISBN 978-0-8218-4925-5. doi: 10.1090/mbk/069. URL <https://doi.org/10.1090/mbk/069>. An introduction.

Reference ||

- 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, 2014b. ISSN 0090-5364. doi: 10.1214/14-AOS1252. URL
<https://doi.org/10.1214/14-AOS1252>.
- Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers Artif. Intell.*, 4:681108, 2021. doi: 10.3389/frai.2021.681108. URL
<https://doi.org/10.3389/frai.2021.681108>.
- Kwangho Kim, Jisu Kim, Alessandro Rinaldo, and Frédéric Chazal. Time series featurization via topological data analysis: an application to cryptocurrency trend forecasting. *CoRR*, abs/1812.02987, 2020. URL
<http://arxiv.org/abs/1812.02987>.

Reference III

- Kwangho Kim, Jisu Kim, Manzil Zaheer, Joon Sik Kim, Frédéric Chazal, and Larry Wasserman. PLLay: Efficient Topological Layer based on Persistent Landscapes. *arXiv e-prints*, art. arXiv:2002.02778, February 2020.
- Pum Jun Kim, Yoojin Jang, Jisu Kim, and Jaejun Yoo. TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models. *arXiv e-prints*, art. arXiv:2306.08013, June 2024. doi: 10.48550/arXiv.2306.08013.
- Hengrui Luo, Alice Patania, Jisu Kim, and Mikael Vejdemo-Johansson. Generalized penalty for circular coordinate representation. *Foundations of Data Science*, 3(4):729–767, 2021.
- Larry Wasserman. Topological data analysis, 2016.
- Yuan Yao, Jian Sun, Xuhui Huang, Gregory R. Bowman, Gurjeet Singh, Michael Lesnick, Leonidas J. Guibas, Vijay S. Pande, and Gunnar Carlsson. Topological methods for exploring low-density states in biomolecular folding pathways. *The Journal of Chemical Physics*, 130(14):144115, 04 2009. ISSN 0021-9606. doi: 10.1063/1.3103496. URL <https://doi.org/10.1063/1.3103496>.

Thank you!

Statistical Inference on Persistent Homology

Featurization of Topological Data Analysis using Persistence Landscapes

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

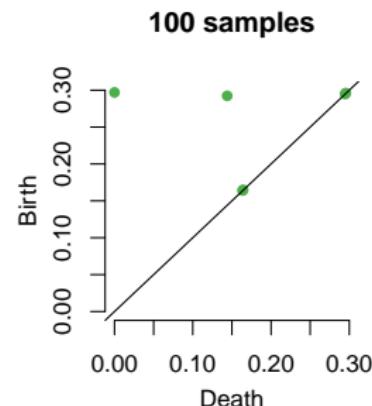
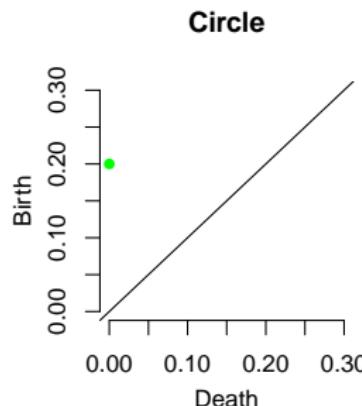
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

where γ ranges over all bijections from D_1 to D_2 .



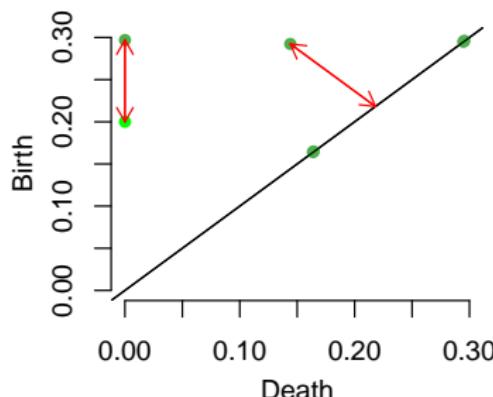
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

where γ ranges over all bijections from D_1 to D_2 .



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

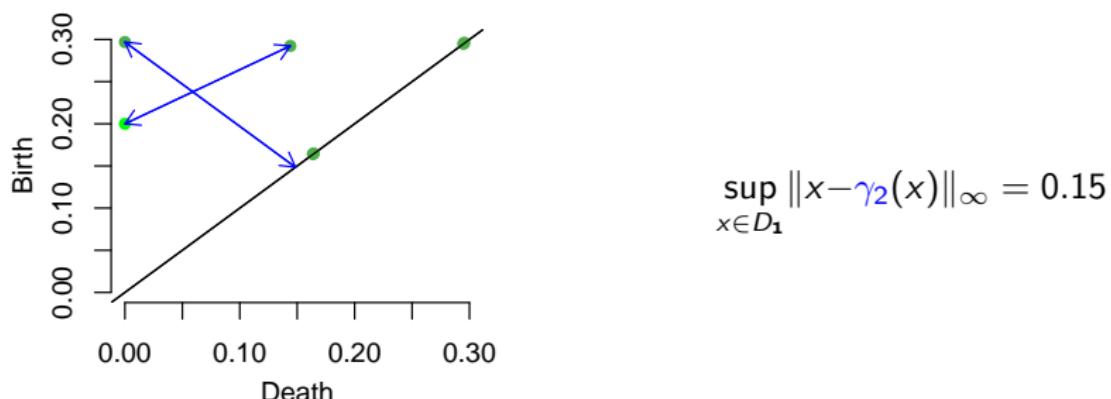
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

where γ ranges over all bijections from D_1 to D_2 .



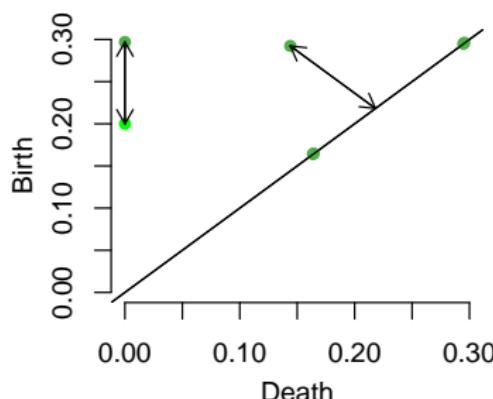
Bottleneck distance gives a metric on the space of persistent homology.

Definition

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

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

where γ ranges over all bijections from D_1 to D_2 .

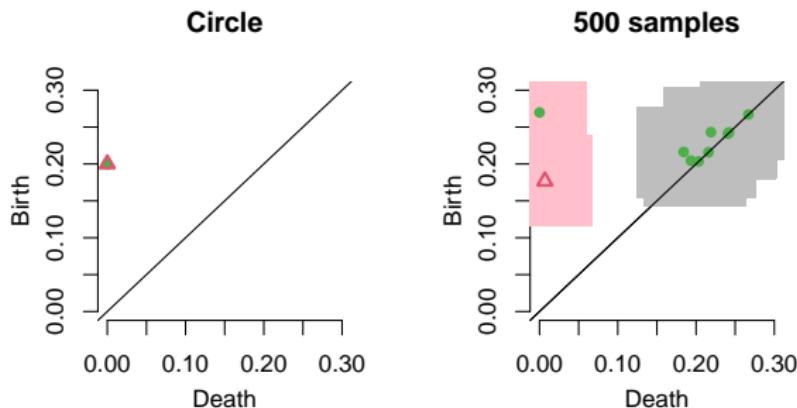


$$\inf_{\gamma} \sup_{x \in D_1} \|x - \gamma(x)\|_\infty = 0.1$$

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

Let $Dgm(M)$ and $Dgm(X)$ be persistent homologies of the manifold M and the data X , respectively. Given the significance level $\alpha \in (0, 1)$, $(1 - \alpha)$ confidence set $\{D \in Dgm : W_\infty(Dgm(X), D) \leq c_n\}$ is a random set satisfying

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



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

Theorem

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

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

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

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

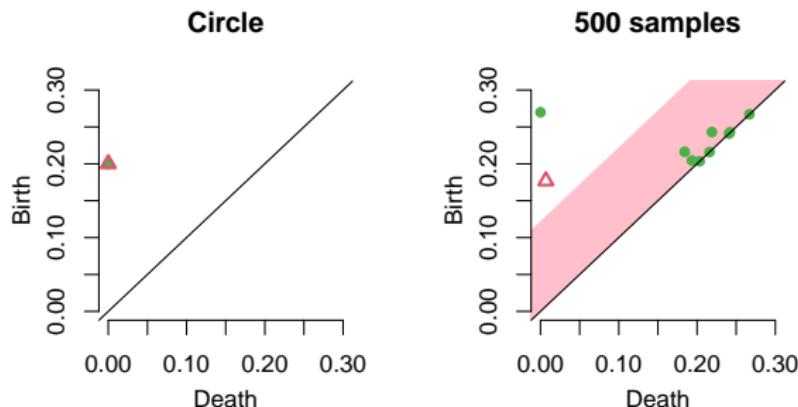
$$\mathbb{P}(W_\infty(Dgm(f_M), Dgm(f_X)) \leq c_n) \geq \mathbb{P}(\|f_M - f_X\|_\infty \leq c_n) \geq 1 - \alpha,$$

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

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

Bootstrap algorithm can be applied to persistent homology.

- ▶ for the case of kernel density estimator in Fasy et al. [2014b],
- ▶ for the case of distance to measure and kernel distance in Chazal et al. [2014].



Statistical Inference on Persistent Homology

Featurization of Topological Data Analysis using Persistence Landscapes

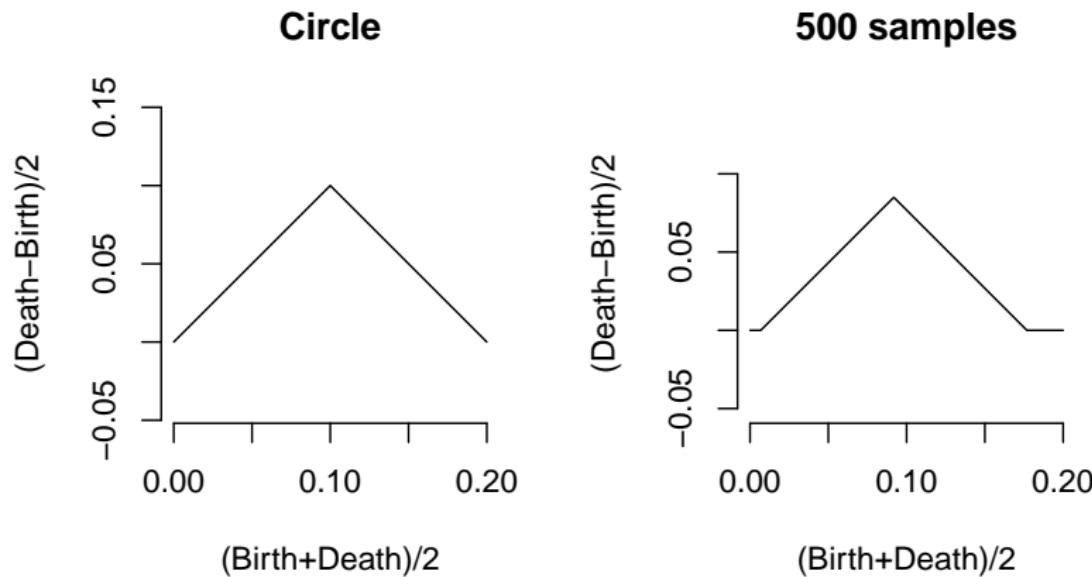
R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

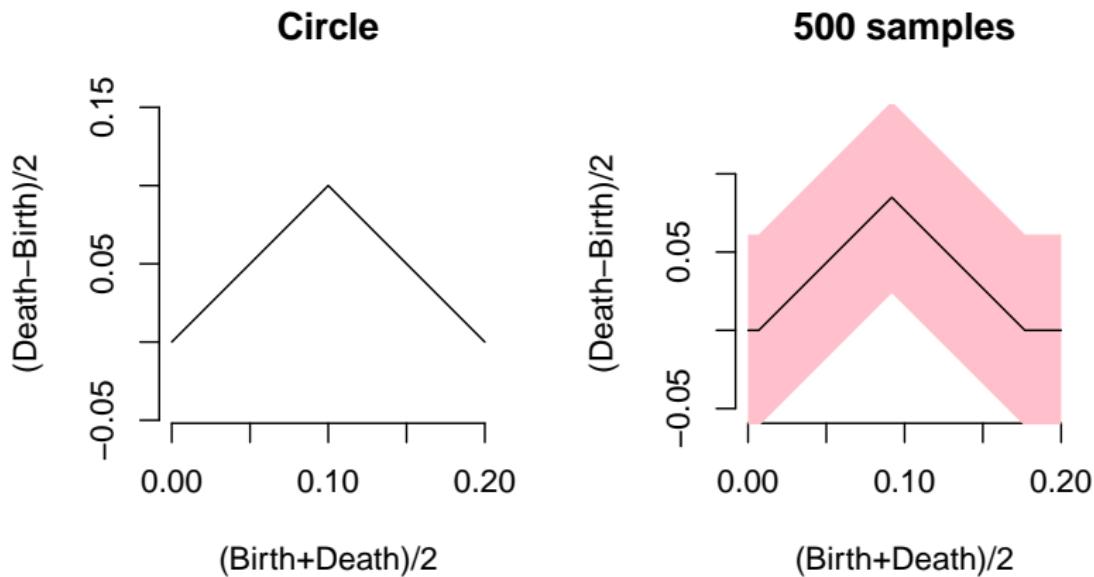
Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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



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

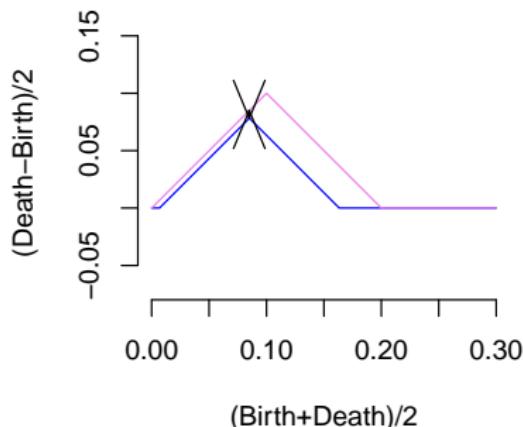


∞ -landscape distance gives a metric on the space of persistence landscapes.

Definition

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

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



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

Theorem

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

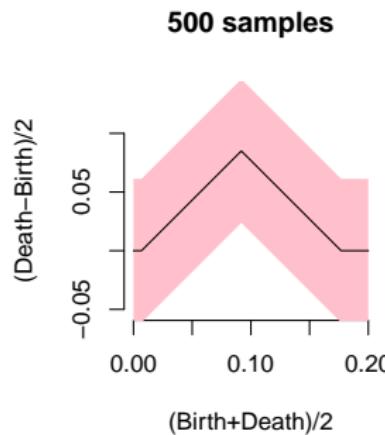
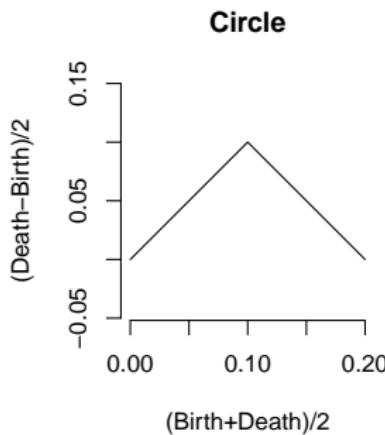
$$\Lambda_\infty(\lambda(f), \lambda(g)) \leq \|f - g\|_\infty.$$

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

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

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

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



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

- ▶ Confidence band for the persistence landscape can be also computed using multiplier bootstrap; see [?].

PLlay is differentiable.

- ▶ A deep learning model learns its parameters by back propagation, which is to apply gradient descent layer-wise.
- ▶ For a deep learning layer to be learnable, it should be differentiable:

Theorem (Theorem 3.1 in Kim et al. [2020])

The PLlay function $S_{\theta, \omega}$ is differentiable with respect to the input data X .

PLlay is stable.

- ▶ PLlay is stable with respect to changes in persistence diagrams:

Theorem (Theorem 4.1 in Kim et al. [2020])

For two persistence diagrams $\mathcal{D}, \mathcal{D}'$,

$$|S_{\theta,\omega}(\mathcal{D}) - S_{\theta,\omega}(\mathcal{D}')| = O(d_B(\mathcal{D}, \mathcal{D}')),$$

where d_B is the bottleneck distance.

PLlay is stable.

- ▶ PLlay is stable with respect to perturbations in input X :

Theorem (Theorem 4.2 in Kim et al. [2020])

Let $X \sim P$ and P_n be the empirical distribution. Further, let $\mathcal{D}_P, \mathcal{D}_X$ be the persistence diagrams of P, X , respectively. Then

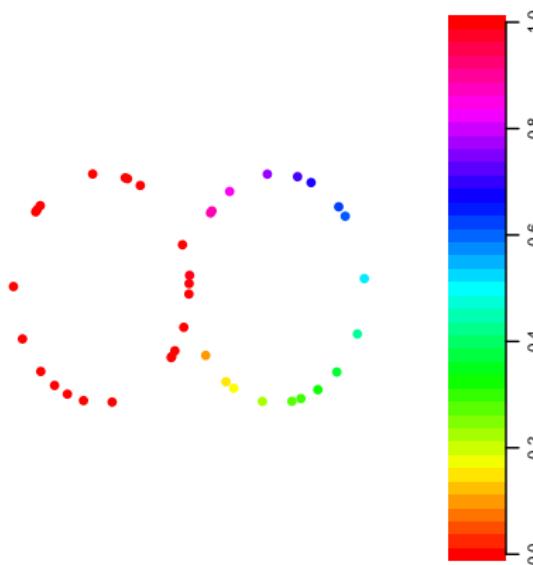
$$|S_{\theta, \omega}(\mathcal{D}_X) - S_{\theta, \omega}(\mathcal{D}_P)| = O(W_2(P_n, P)),$$

where W_2 is 2-Wasserstein distance.

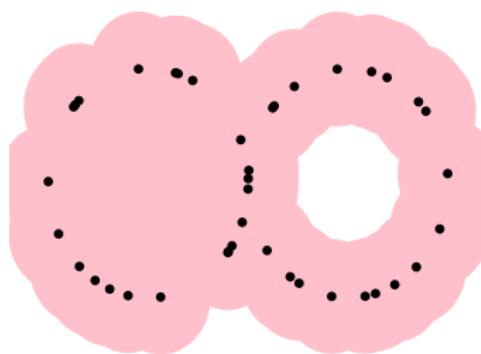
Circular coordinates provide topological representations of reduced dimension.

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

circular coordinates



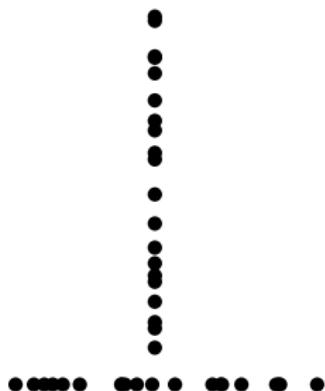
loop



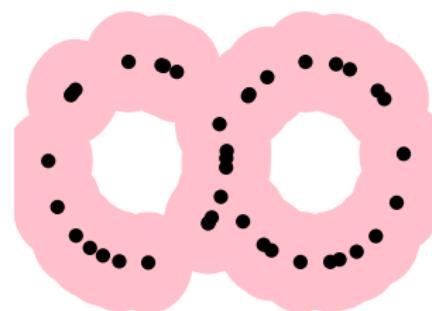
Circular coordinates provide topological representations of reduced dimension.

- ▶ circular coordinate is a function that maps from data points X to torus $\mathbb{T}^k = (S^1)^k$.

circular coordinates

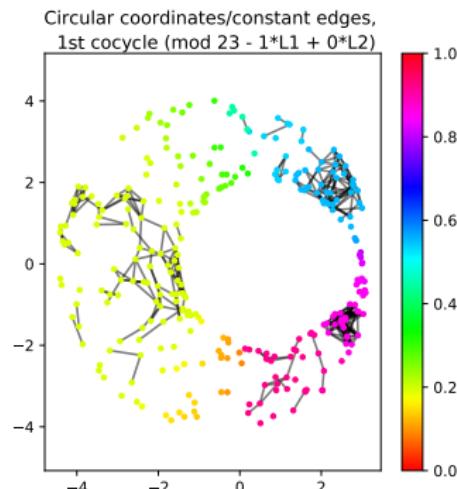
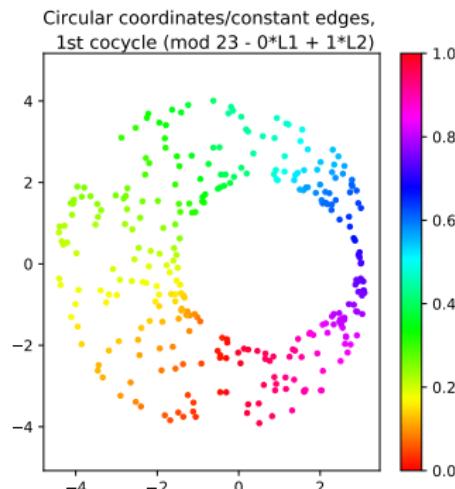


loop



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

- ▶ Generalized penalty for circular coordinate representation (Luo, Patania, Kim, Vejdemo-Johansson, 2021)
- ▶ When computing circular coordinates, we solve an optimization problem.
- ▶ We switch L_2 loss by L_1 loss for circular coordinate values to change more abruptly: better visualizes topological information from data.



Statistical Inference on Persistent Homology

Featurization of Topological Data Analysis using Persistence Landscapes

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

Statistical Inference on Persistent Homology

Featurization of Topological Data Analysis using Persistence Landscapes

R Package TDA: Statistical Tools for Topological Data Analysis
Sample on manifolds, Distance Functions, and Density Estimators

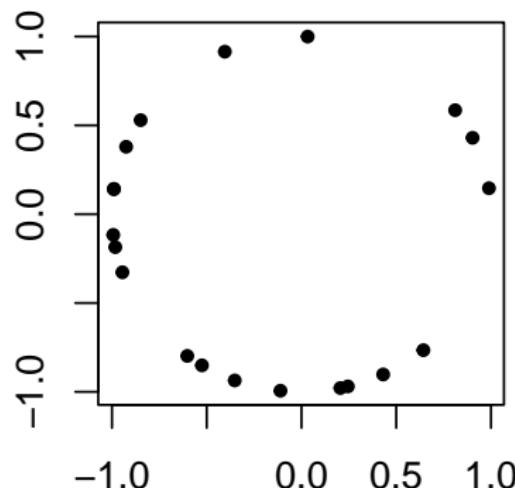
Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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

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

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



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

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

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

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

R Package TDA provides KDE function over a grid.

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

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

where h is a smoothing parameter.

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

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

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

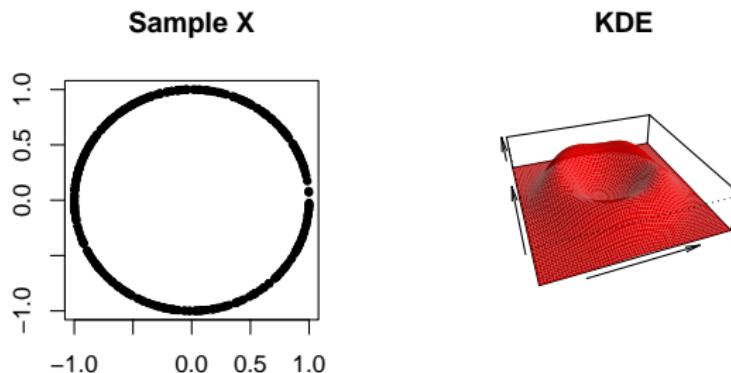
R Package TDA provides KDE function over a grid.

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

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

where h is a smoothing parameter.

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



Statistical Inference on Persistent Homology

Featurization of Topological Data Analysis using Persistence Landscapes

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

R Package TDA computes Persistent Homology over a grid.

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

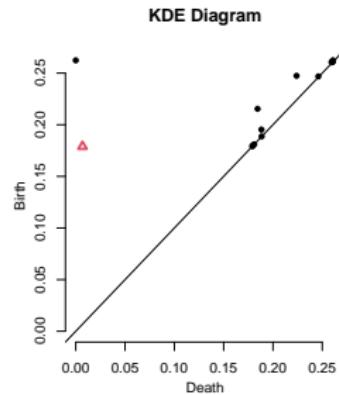
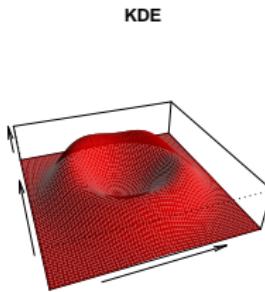
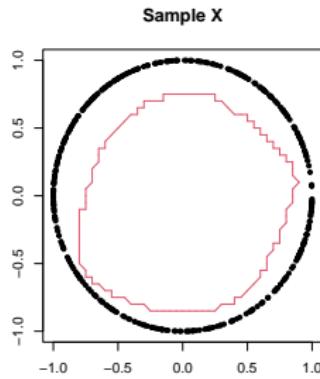
R Package TDA computes Persistent Homology over a grid.

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

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

R Package TDA computes Persistent Homology over a grid.

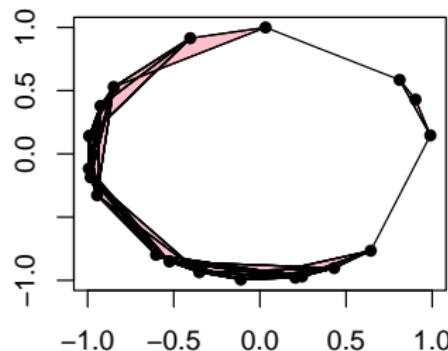
- ▶ The function `gridDiag()` computes the persistent homology of sublevel (and superlevel) sets of the input function.
 - ▶ `gridDiag()` evaluates the real valued input function over a grid.
 - ▶ `gridDiag()` constructs a filtration of simplices using the values of the input function.
 - ▶ `gridDiag()`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



R Package TDA computes Vietoris-Rips Persistent Homology.

- ▶ Vietoris-Rips complex consists of simplices whose pairwise distances of vertices are at most $2r$ apart, i.e.

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



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

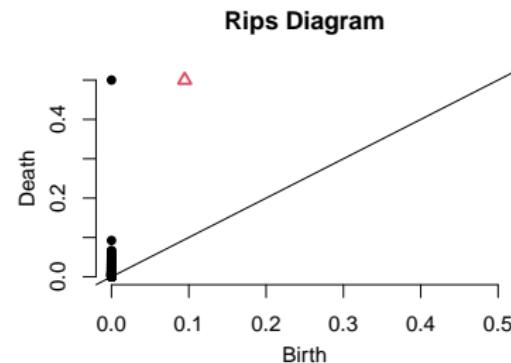
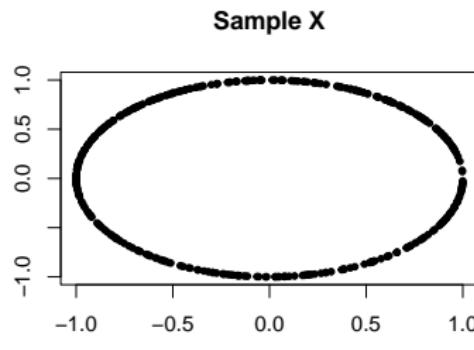
R Package TDA computes Vietoris-Rips Persistent Homology.

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

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

R Package TDA computes Vietoris-Rips Persistent Homology.

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



R Package TDA computes Persistence Landscape.

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

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

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

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

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

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

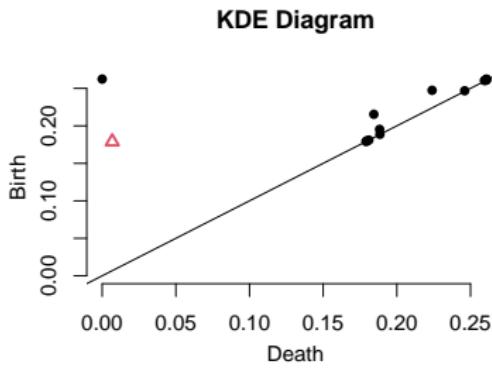
R Package TDA computes Persistence Landscape.

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

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

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

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



Statistical Inference on Persistent Homology

Featurization of Topological Data Analysis using Persistence Landscapes

R Package TDA: Statistical Tools for Topological Data Analysis

Sample on manifolds, Distance Functions, and Density Estimators

Persistent Homology and Persistence Landscape

Statistical Inference on Persistence Homology and Persistence Landscape

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

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

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

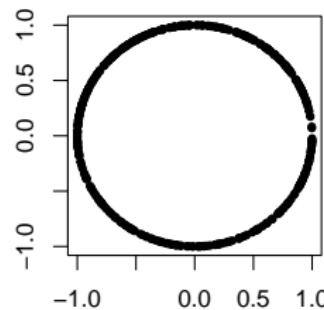
##           90%
## 0.05836494
```

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

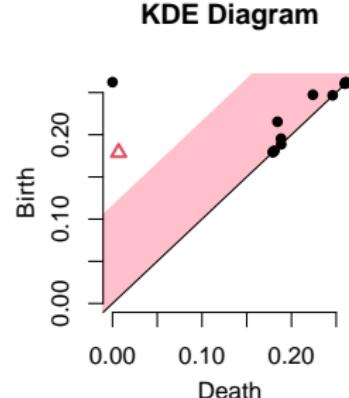
The $(1 - \alpha)$ bootstrap confidence band for $\mathbb{E}[\hat{p}_h]$ is used as the confidence band for the persistent homology.

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

Sample X



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 = "500 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.

