

Featurization and Evaluation using Topological Data Analysis

Jisu KIM



The 4th POSTECH MINDS Workshop on
Topological Data Analysis and Machine Learning
2024-02-01

Introduction

Persistent Homology

Featurization using Persistence Landscape

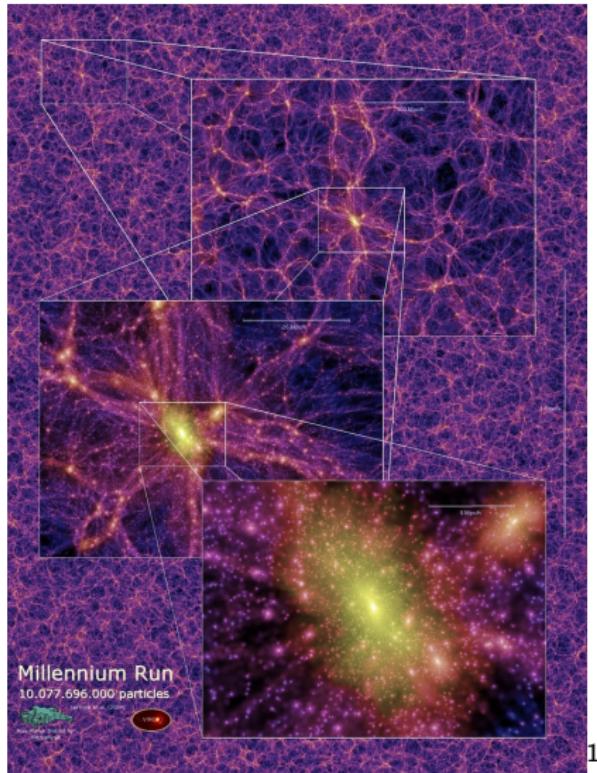
Featurization using Circular Coordinates

Statistical Inference For Homological Features

Evaluation using Confidence of Topological Data Analysis

References

Topological structures in the data provide information.

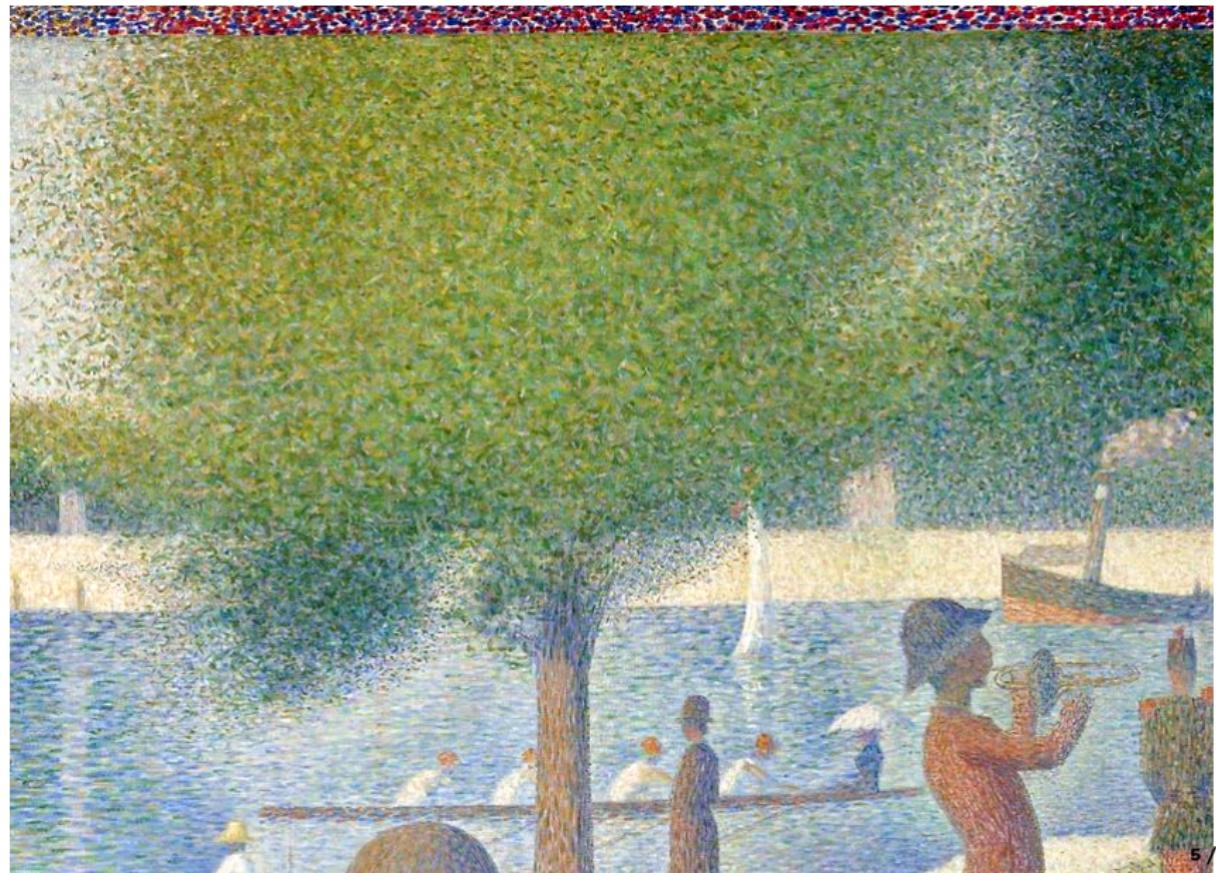


¹http://www.mpa-garching.mpg.de/galform/virgo/millennium/poster_half.jpg

Persistent Homology: observe topological structure with multi resolutions.



Persistent Homology: observe topological structure with multi resolutions.



Persistent Homology: observe topological structure with multi resolutions.



Persistent Homology: observe topological structure with multi resolutions.

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



A (very) rough introduction to Machine Learning

- ▶ For given problem and data, machine learning / deep learning learns a parametrized model.
 - ▶ Given data \mathcal{X} ,
 - ▶ Parametrized model f_θ ,
 - ▶ Loss function \mathcal{L} adapted to a problem,
 - ▶ Machine Learning computes a solution that minimizes the loss function: $\arg \min_\theta \mathcal{L}(f_\theta, \mathcal{X})$.
- ▶ For many cases, computing an explicit formula for the minimizer is impossible or too expensive (e.g. inverting a large matrix). So, we often use gradient descent using $\nabla_\theta \mathcal{L}(f_\theta, \mathcal{X})$:

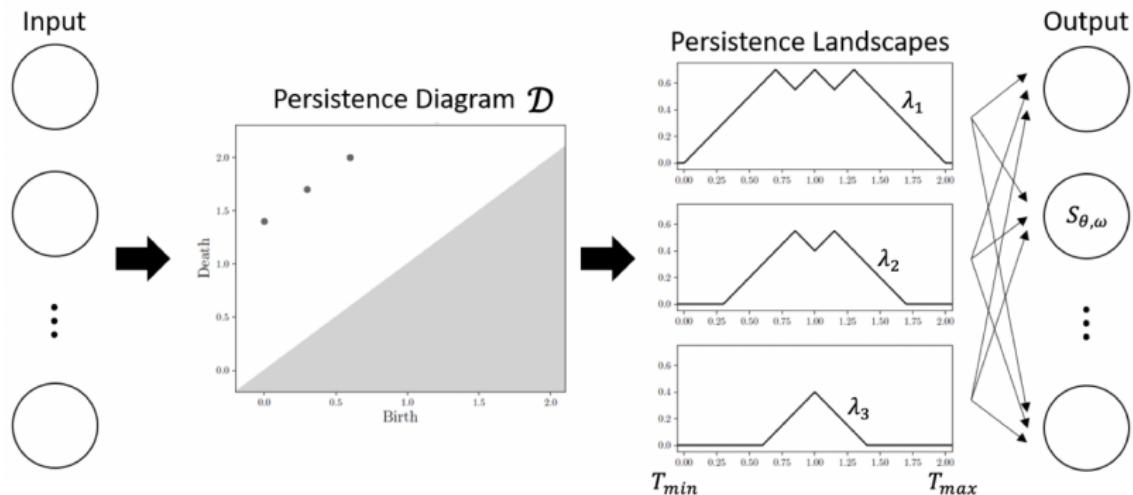
$$\theta_{n+1} = \theta_n - \lambda \nabla_\theta \mathcal{L}(f_\theta, \mathcal{X}).$$

Topological Data Analysis is applied to Machine Learning.

- ▶ A Survey of Topological Machine Learning Methods (Hensel, Moor, Rieck, 2021)
- ▶ Roughly, there are two directions applying Topological Data Analysis (TDA) to Machine Learning:
 - ▶ Make features from TDA to add topological features to data \mathcal{X} : 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 of interest
 - ▶ TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)

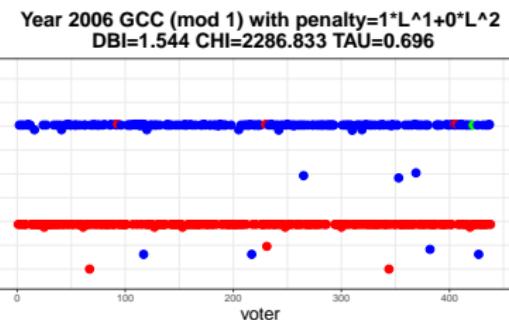
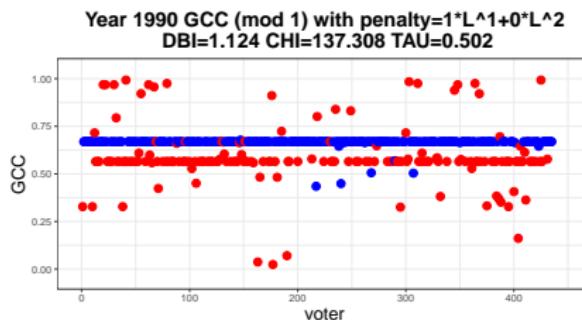
Topological structure is featurized as persistence landscape to be further applied in machine learning framework.

- ▶ Featurization using Persistence Landscape
 - ▶ Efficient Topological Layer based on Persistent Landscapes (Kim, Kim, Zaheer, Kim, Chazal, Wasserman, 2020)



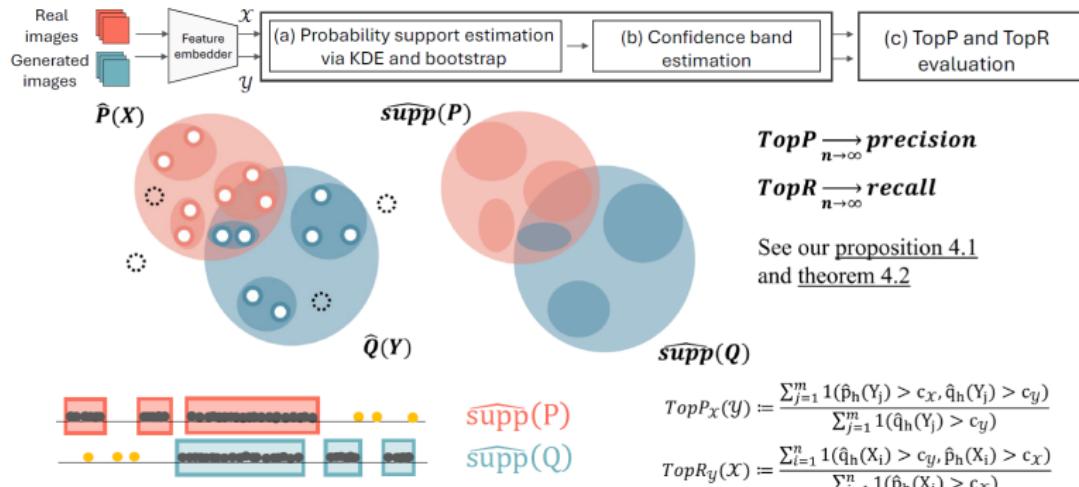
Topological structure is featurized as circular coordinates to be further applied in machine learning framework.

- ▶ Featurization using Circular Coordinates
 - ▶ Generalized penalty for circular coordinate representation (Luo, Patania, Kim, Vejdemo-Johansson, 2021)



Data or Model is evaluated using Topological Data Analysis.

- ▶ Evaluation using Confidence of Topological Data Analysis
 - ▶ TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)



Introduction

Persistent Homology

Featurization using Persistence Landscape

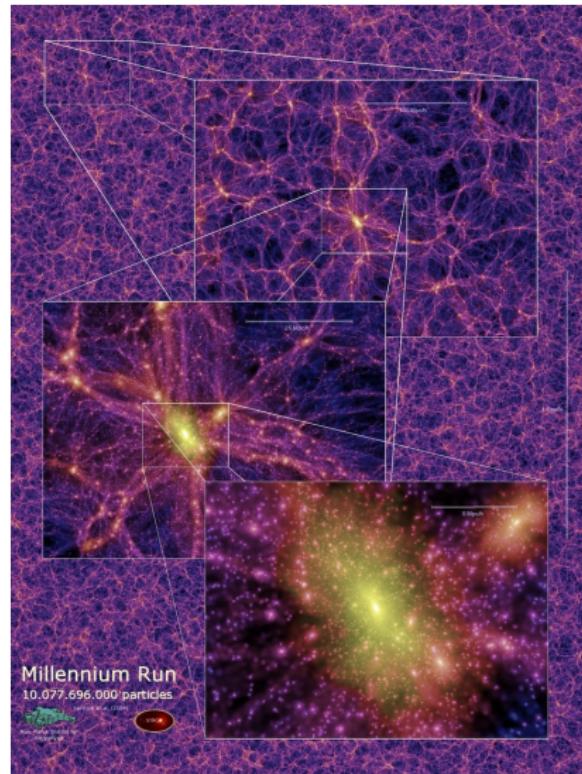
Featurization using Circular Coordinates

Statistical Inference For Homological Features

Evaluation using Confidence of Topological Data Analysis

References

Topological holes in the data provide information.



The number of holes is used to summarize geometrical features.

- ▶ Geometrical objects :
 - ▶ ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ
 - ▶ A, 字, あ
- ▶ The number of holes of different dimensions is considered.
 1. $\beta_0 = \#$ of connected components 
 2. $\beta_1 = \#$ of loops (holes inside 1-dim sphere) 
 3. $\beta_2 = \#$ of voids (holes inside 2-dim sphere) : if $dim \geq 3$ 

Example : Objects are classified by homologies.

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



2. $\beta_1 = \#$ of loops

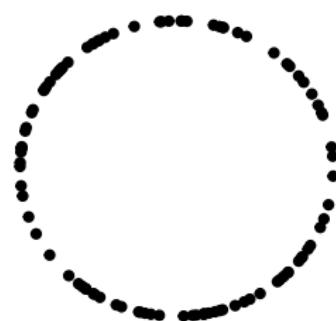
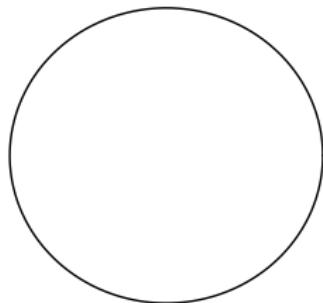


$\beta_0 \setminus \beta_1$	0	1	2
1	ㄱ, ㄴ, ㄷ, ㄹ, ㅅ, ㅈ, ㅌ, ㅊ	ㅁ, ㅇ, ㅂ, ㅍ, ㅏ	ㅓ
2	ㅊ, ㅌ		
3		ㅎ	

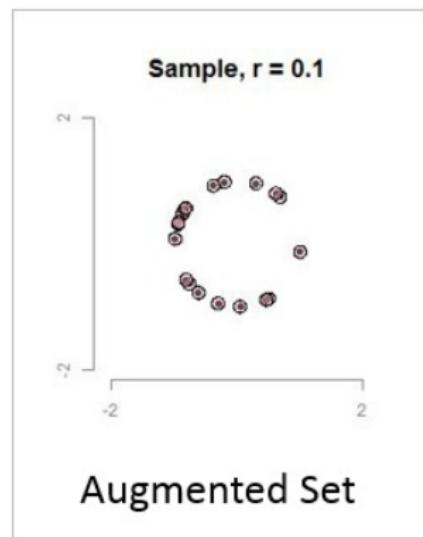
Homology of finite sample is different from homology of underlying manifold, hence it cannot be directly used for the inference.

- ▶ 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:

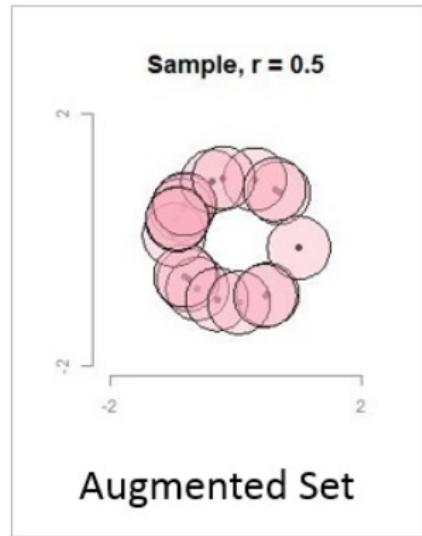
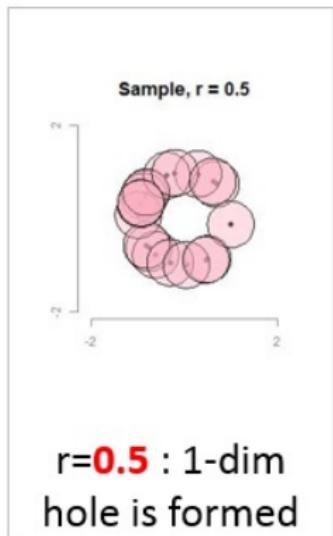
Underlying circle: $\beta_0 = 1, \beta_1 = 1$ 100 samples: $\beta_0 = 100, \beta_1 = 0$



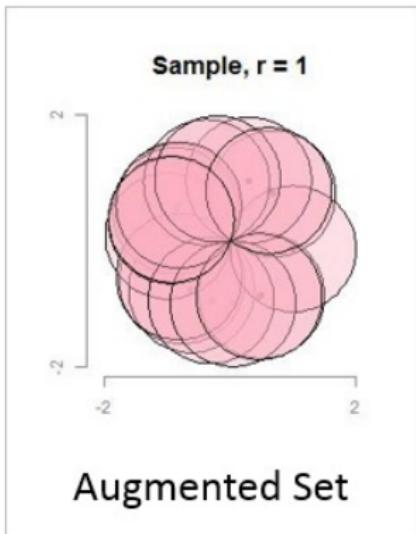
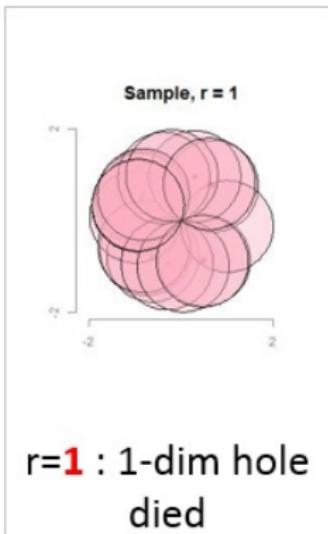
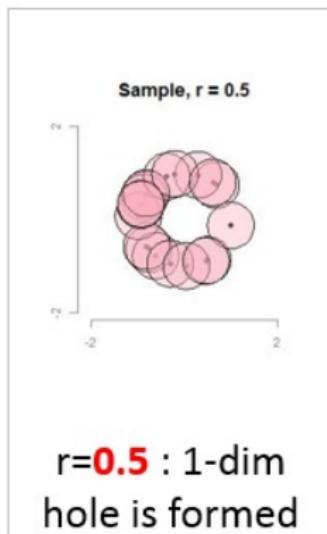
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



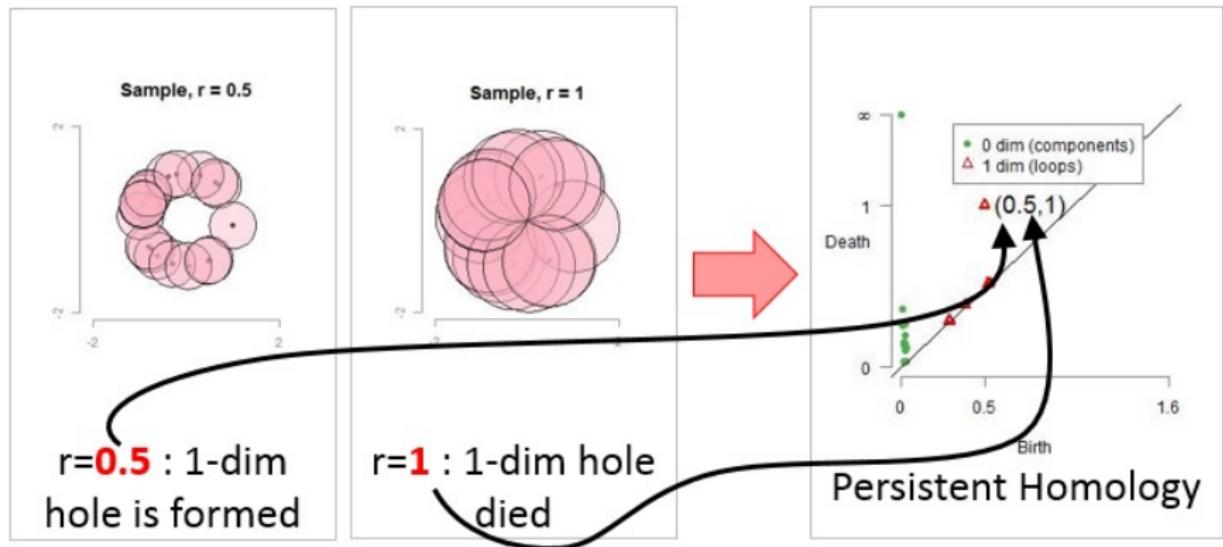
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



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.



Introduction

Persistent Homology

Featurization using Persistence Landscape

Featurization using Circular Coordinates

Statistical Inference For Homological Features

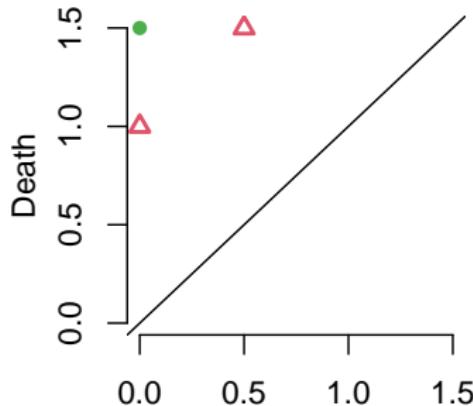
Evaluation using Confidence of Topological Data Analysis

References

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

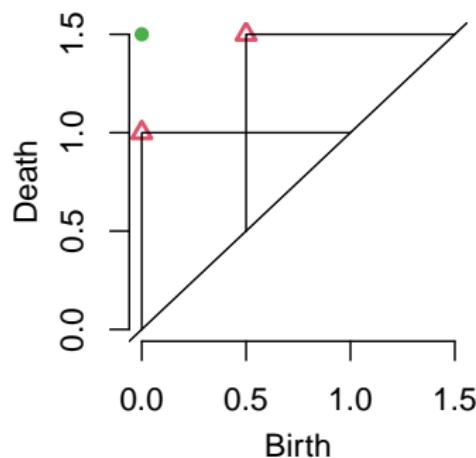
- ▶ The space of the persistent homology is complex, so directly applying in machine learning is difficult.
- ▶ If the persistent homology is further summarized and embedded into a Euclidean space or a functional space, then applying in machine learning becomes much more convenient.
 - ▶ e.g., Persistence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

Persistent Homology

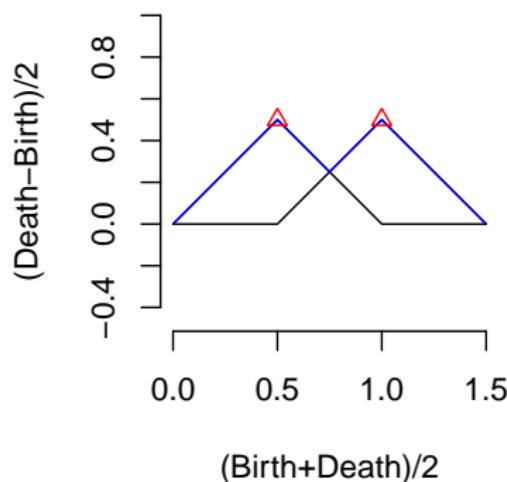


Persistence Landscape is a functional summary of the persistent homology.

Persistent Homology

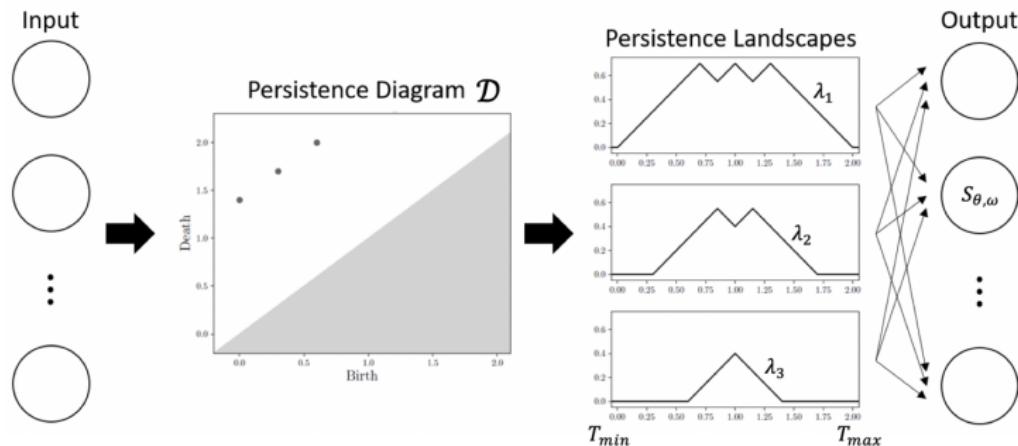


Persistence Landscape



PLay: 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 $S_{\theta, \omega}(\mathcal{D}) := g_\theta(\bar{\Lambda}_\omega)$.



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])

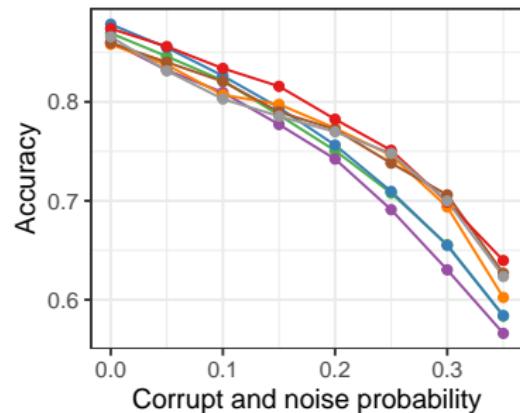
Let $X \sim P$ and P_n be the empirical distribution. Further, let $\mathcal{D}_P, \mathcal{D}_X$ be the persistence diagrams of P, X , respectively. Then

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

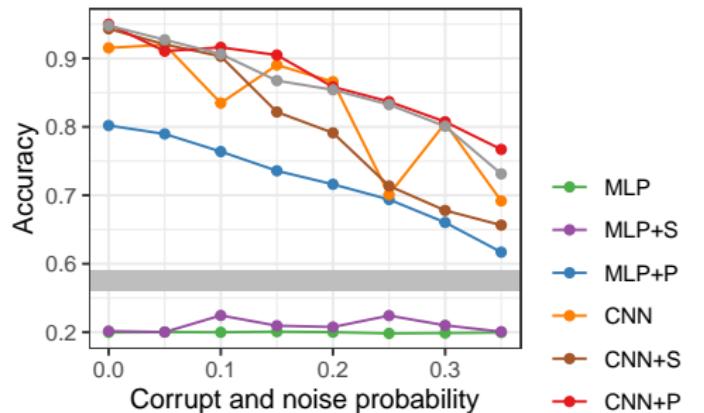
where W_2 is 2-Wasserstein distance.

Experiments

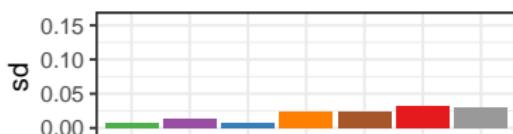
Accuracy for MNIST data



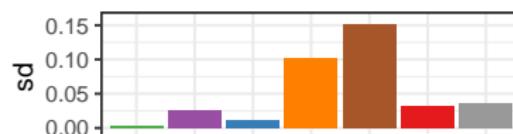
Accuracy for ORBIT5K data



Sd for MNIST data



Sd for ORBIT5K data



Introduction

Persistent Homology

Featurization using Persistence Landscape

Featurization using Circular Coordinates

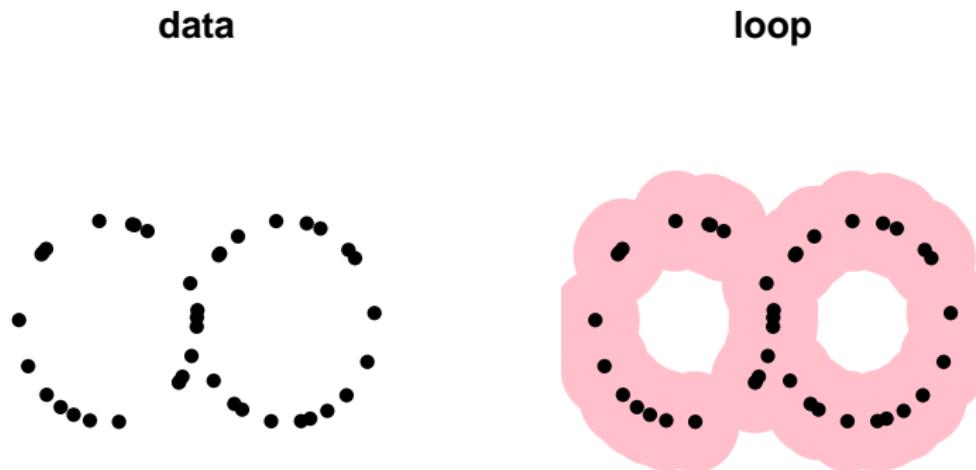
Statistical Inference For Homological Features

Evaluation using Confidence of Topological Data Analysis

References

Circular coordinates provide topological representations of reduced dimension.

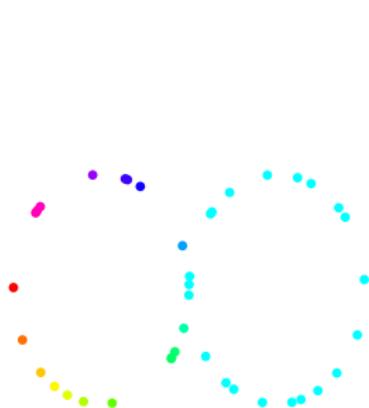
- ▶ Persistent cohomology and circular coordinates (de Silva, Morozov, Vejdemo-Johansson, 2011)
- ▶ Topological Learning for Motion Data via Mixed Coordinates (Vejdemo-Johansson, Pokorny, Skraba, Kragic, 2015)



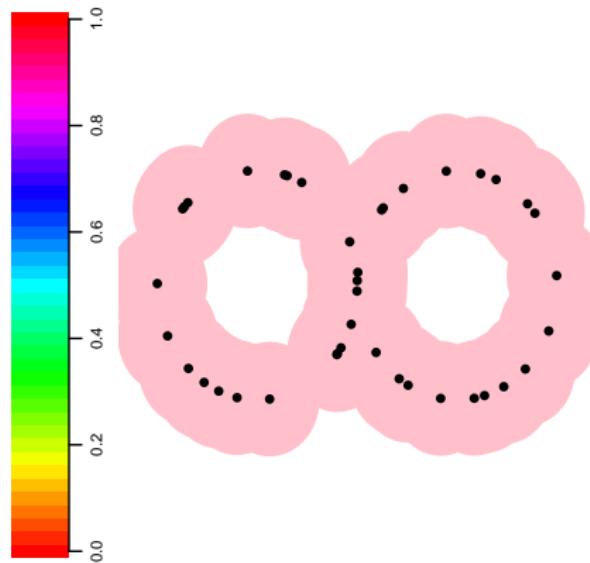
Circular coordinates provide topological representations of reduced dimension.

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

circular coordinates

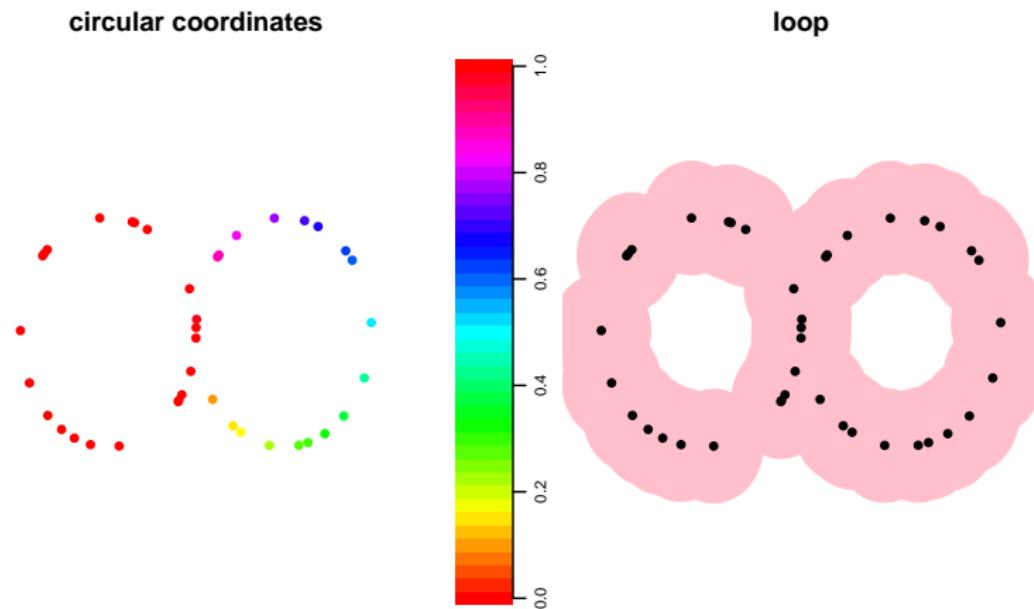


loop



Circular coordinates provide topological representations of reduced dimension.

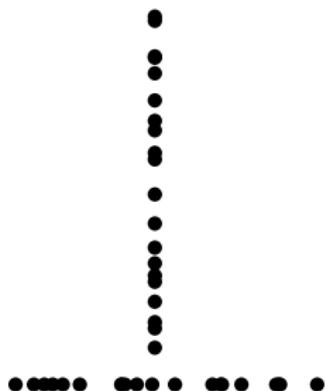
- circular coordinate is a function that maps from data points X to circle S^1 .



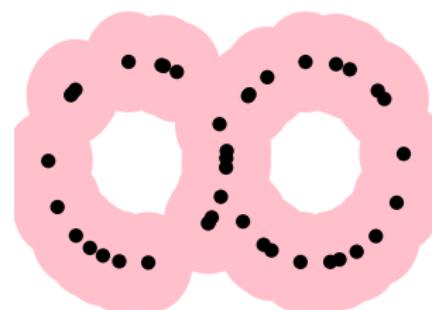
Circular coordinates provide topological representations of reduced dimension.

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

circular coordinates

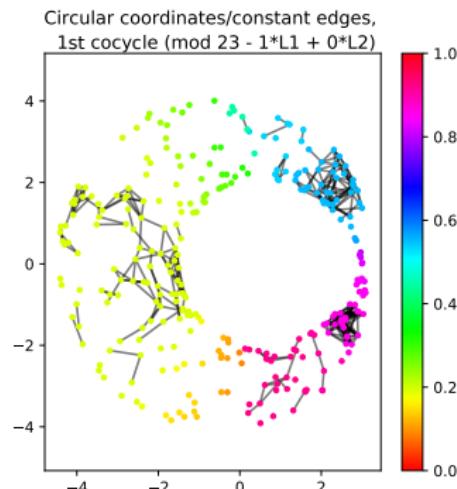
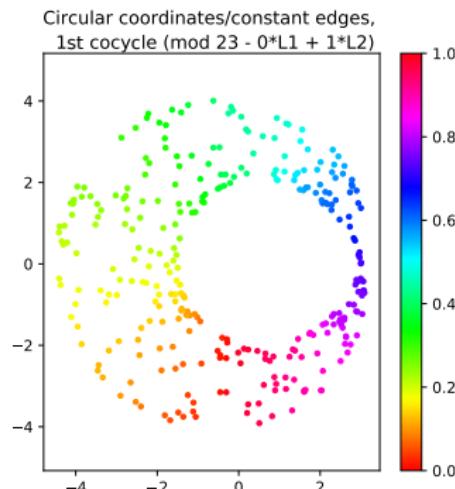


loop



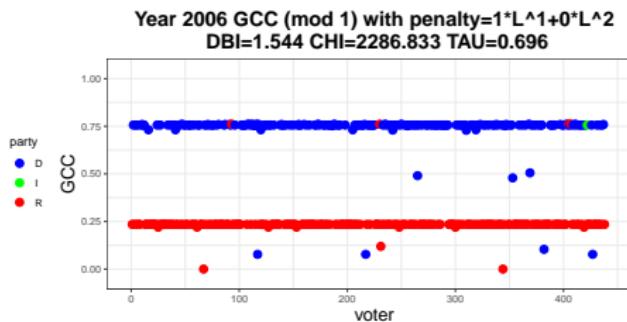
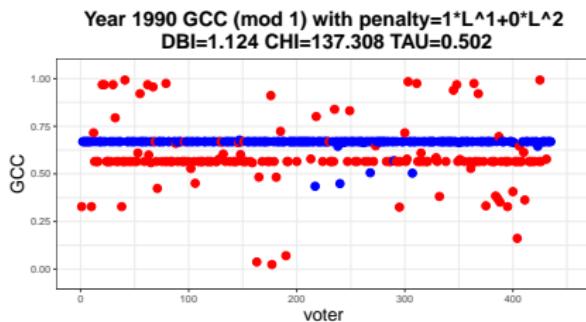
Circular coordinates with generalized penalty better visualizes topological information from data.

- ▶ 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.



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

Persistent Homology

Featurization using Persistence Landscape

Featurization using Circular Coordinates

Statistical Inference For Homological Features

Evaluation using Confidence of Topological Data Analysis

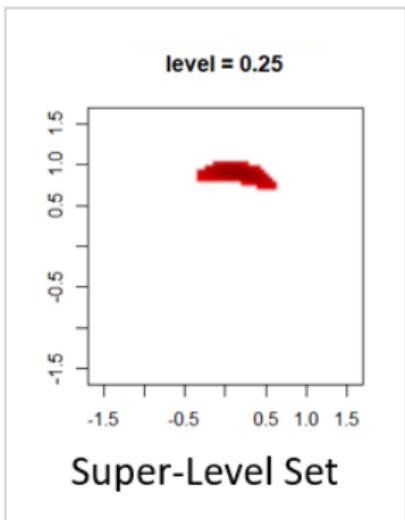
References

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

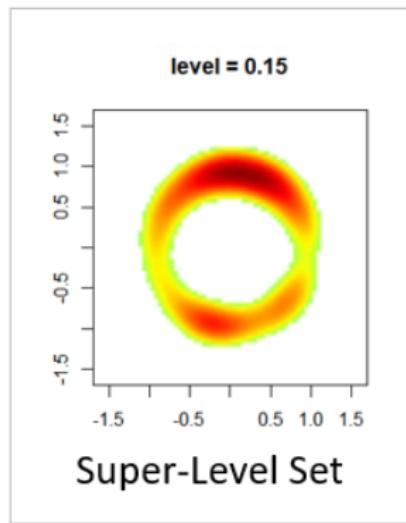
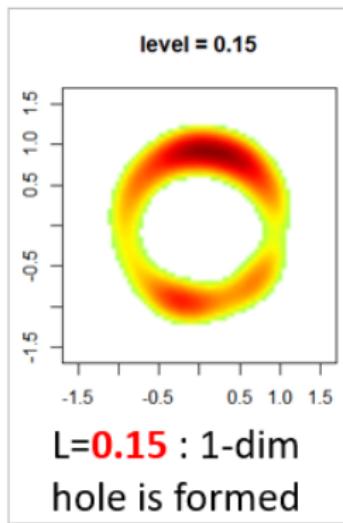
- ▶ The kernel density estimator is

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

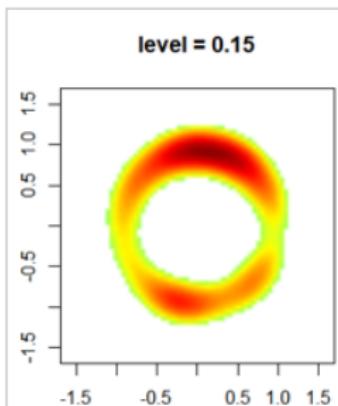
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



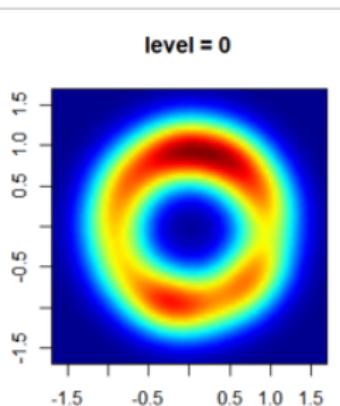
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



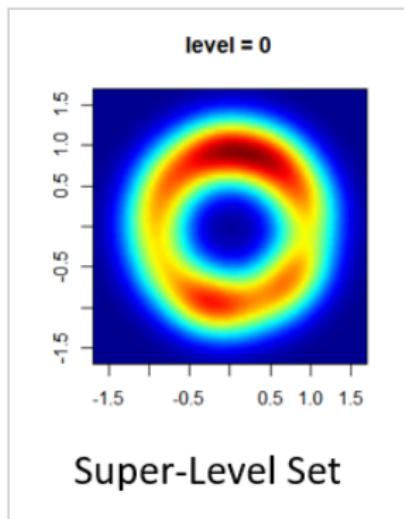
Persistent homology computes homologies on collection of sets, and tracks when topological features are born and when they die.



L=0.15 : 1-dim
hole is formed

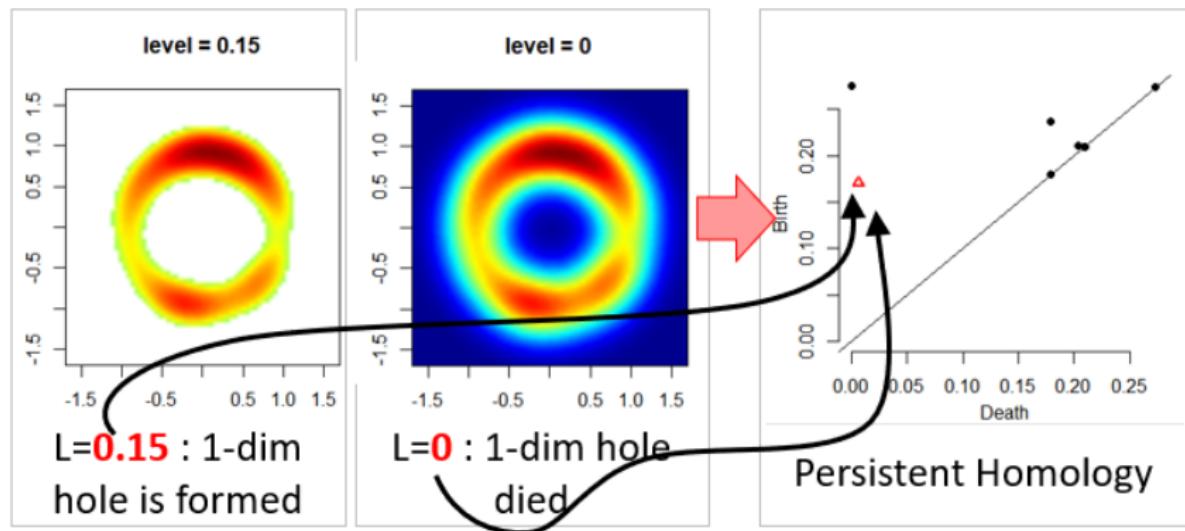


L=0 : 1-dim hole
died



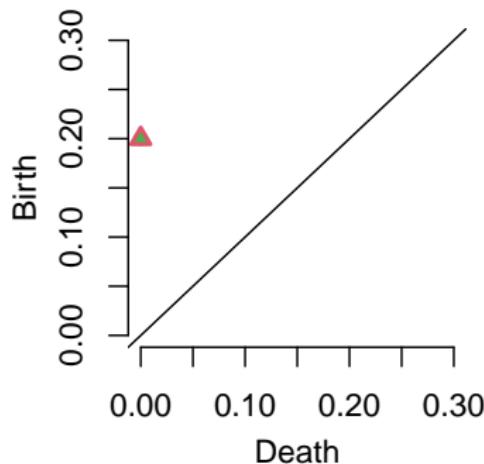
Super-Level Set

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

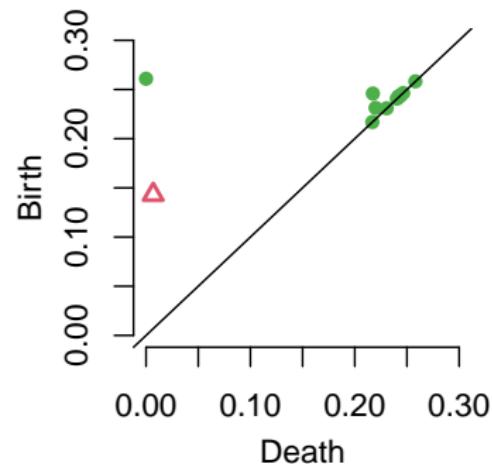


Persistent homology of the underlying manifold can be inferred from persistent homology of finite samples.

Circle



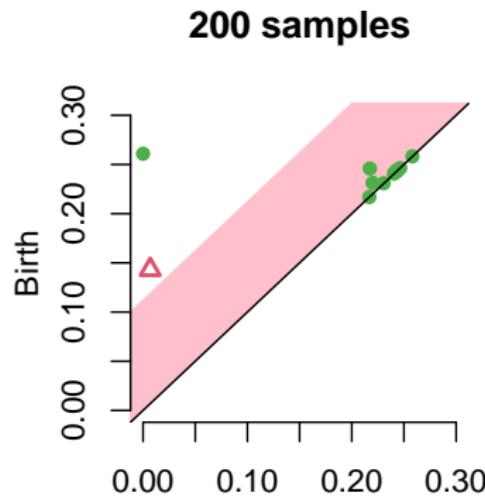
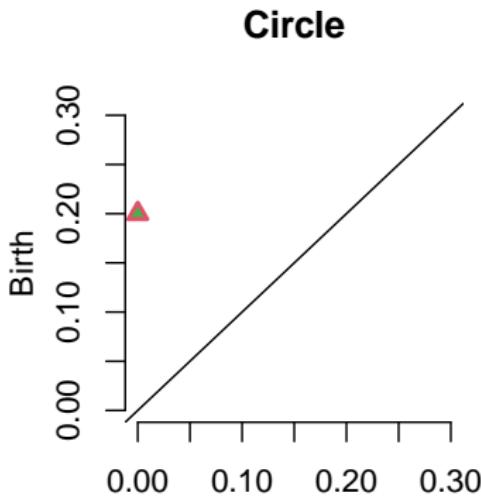
200 samples



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

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

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



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

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

Introduction

Persistent Homology

Featurization using Persistence Landscape

Featurization using Circular Coordinates

Statistical Inference For Homological Features

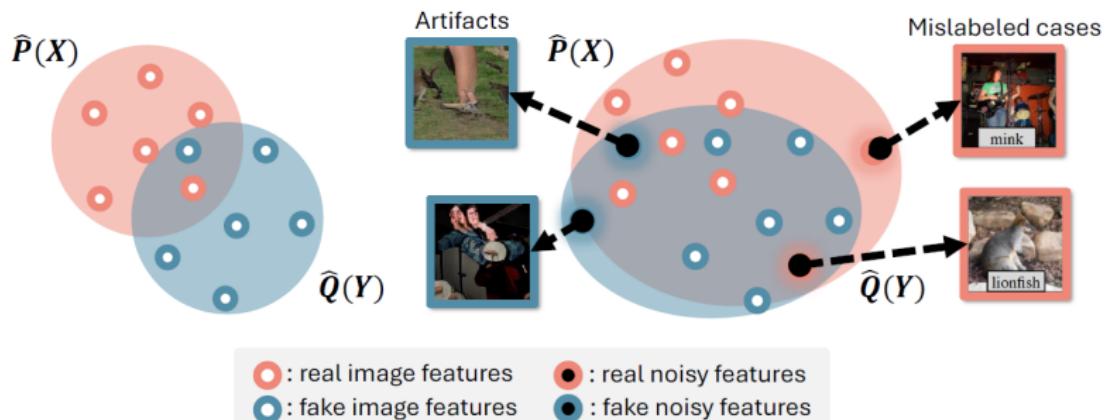
Evaluation using Confidence of Topological Data Analysis

References

Existing evaluation metrics for generative models are vulnerable to noise.

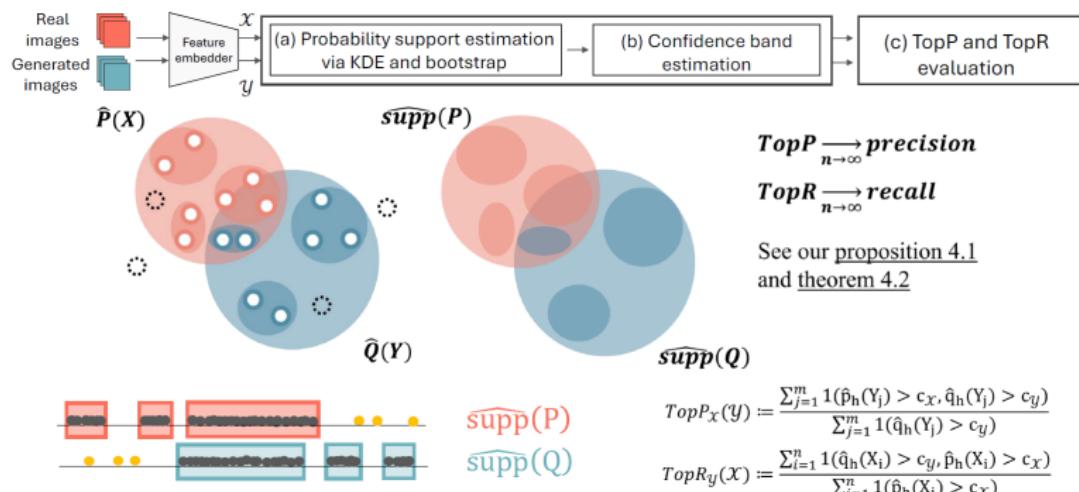
- ▶ TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)
- ▶ To evaluate generative models, metrics compare the support of real image distributions and fake image distributions.
- ▶ Existing evaluation metrics tend to overestimate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

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



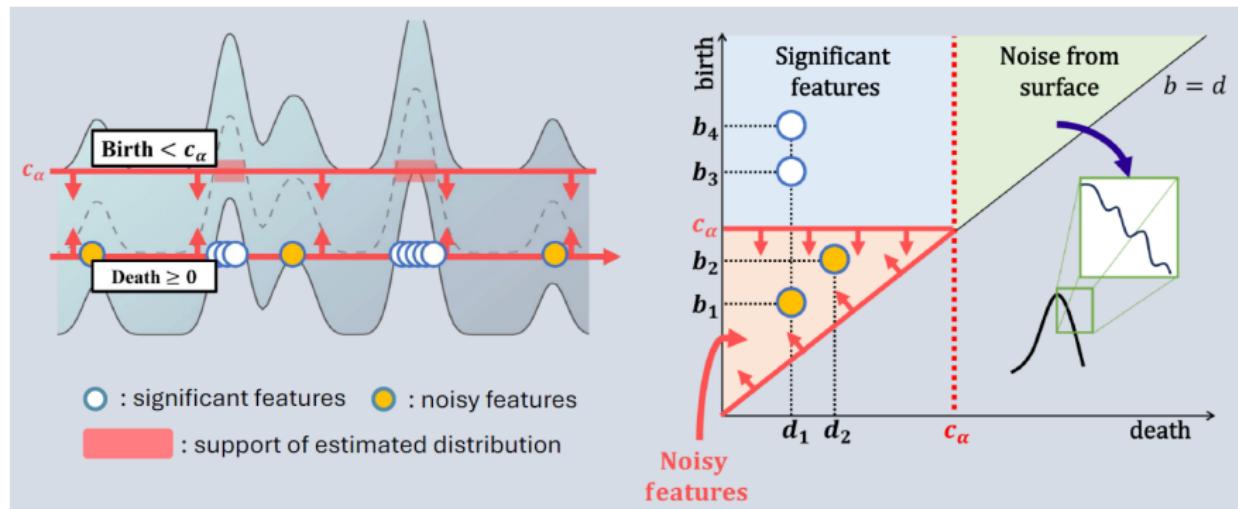
TopP&R robustly evaluates generative models by retaining only topologically and statistically significant features with confidence.

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



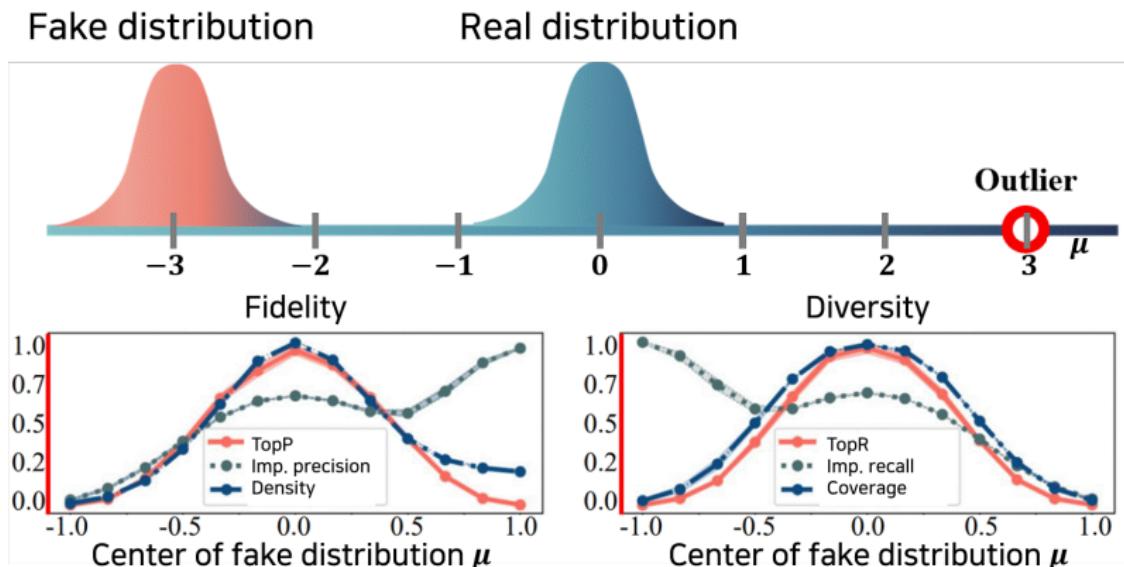
We find threshold c_α that selects statistically and topologically significant features.

- 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

Persistent Homology

Featurization using Persistence Landscape

Featurization using Circular Coordinates

Statistical Inference For Homological Features

Evaluation using Confidence of Topological Data Analysis

References

References |

- Peter Bubenik. Statistical topological data analysis using persistence landscapes. *arXiv preprint arXiv:1207.6437*, 2012.
- Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.
- Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Annual Symposium on Computational Geometry*, pages 474–483. ACM, 2014.
- Vin de Silva, Dmitriy Morozov, and Mikael Vejdemo-Johansson. Persistent cohomology and circular coordinates. *Discrete & Computational Geometry*, 45(4):737–759, 2011.
- H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. Applied mathematics. American Mathematical Society, 2010. ISBN 9780821849255. URL <http://books.google.com/books?id=MDXa6gFRZuIC>.

References ||

- 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, 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.

References III

Mikael Vejdemo-Johansson, Florian T Pokorný, Primoz Skraba, and Danica Kragic. Cohomological learning of periodic motion. *Applicable algebra in engineering, communication and computing*, 26(1):5–26, 2015.

Thank you!

Statistical Inference for Persistent Homology

Featurization using Persistent Homology

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

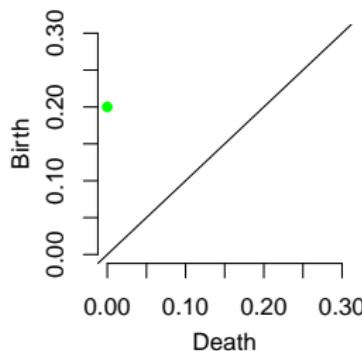
Definition

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

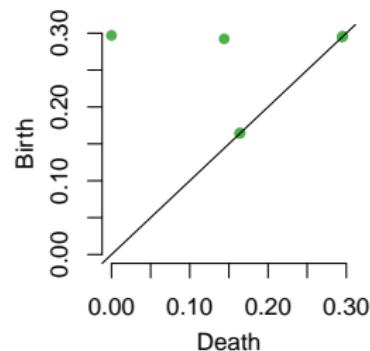
$$d_B(D_1, D_2) = \inf_{\gamma} \sup_{x \in D_1} \|x - \gamma(x)\|_{\infty},$$

where γ ranges over all bijections from D_1 to D_2 .

Circle



100 samples



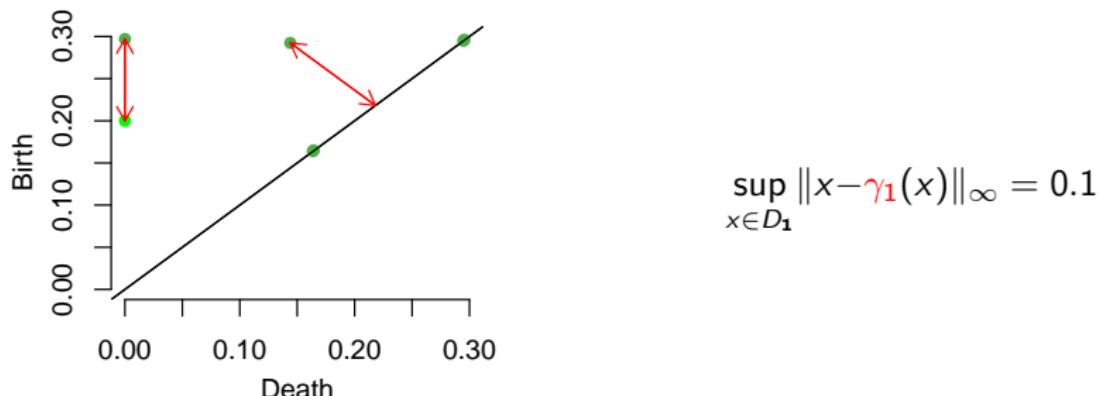
Bottleneck distance gives a metric on the space of persistent homology.

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 .



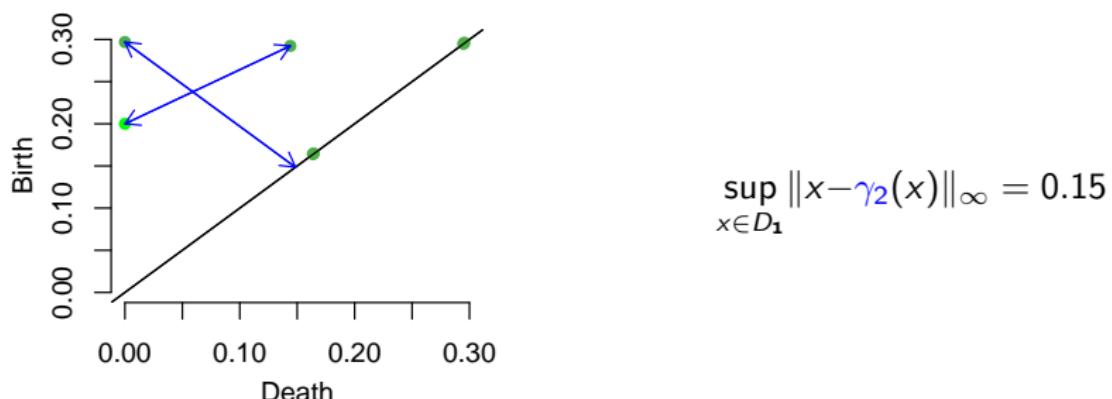
Bottleneck distance gives a metric on the space of persistent homology.

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 .



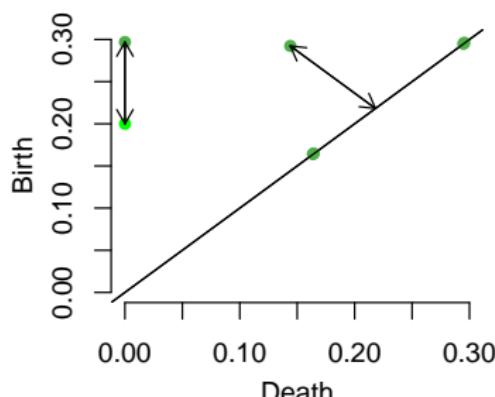
Bottleneck distance gives a metric on the space of persistent homology.

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 .



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

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

Theorem

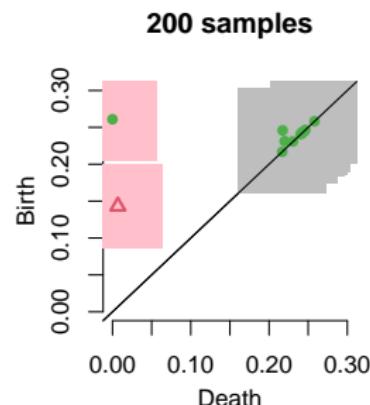
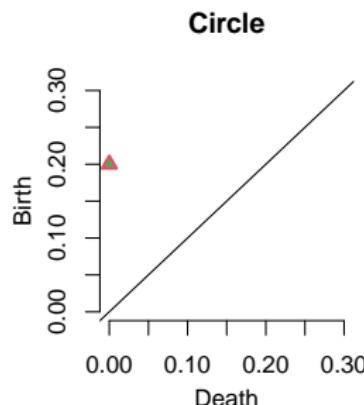
[Edelsbrunner and Harer, 2010][Chazal, de Silva, Glisse, and Oudot, 2012] Let \mathbb{X} be finitely triangulable space and $f, g : \mathbb{X} \rightarrow \mathbb{R}$ be two continuous functions. Let $Dgm(f)$ and $Dgm(g)$ be corresponding persistence diagrams. Then

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

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

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

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



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

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

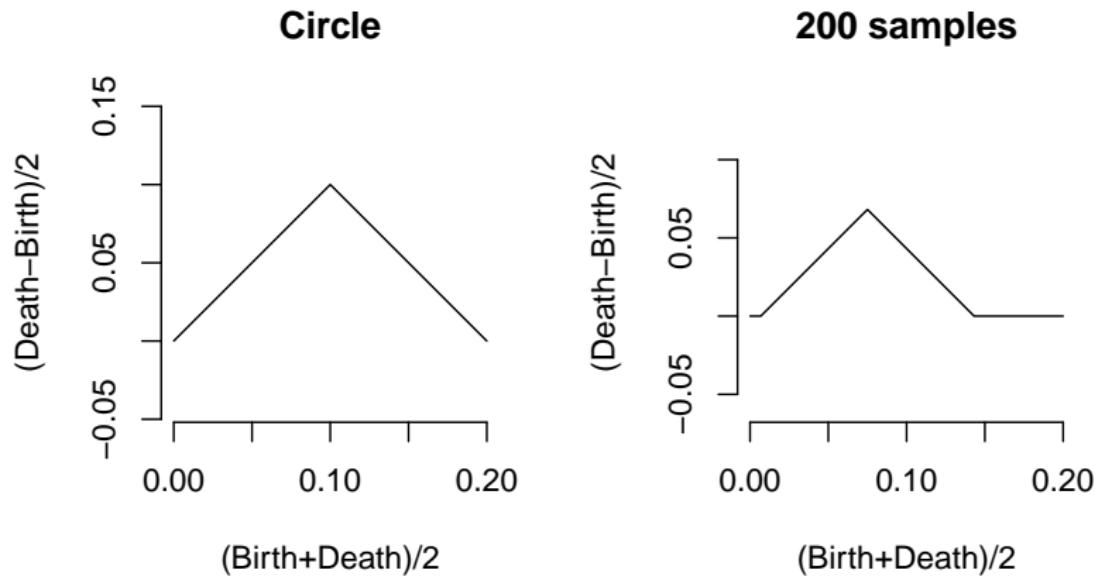
$$\mathbb{P}(d_B(Dgm(f_M), Dgm(f_X)) \leq c_n) \geq \mathbb{P}(\|f_M - f_X\|_\infty \leq c_n) \geq 1 - \alpha,$$

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

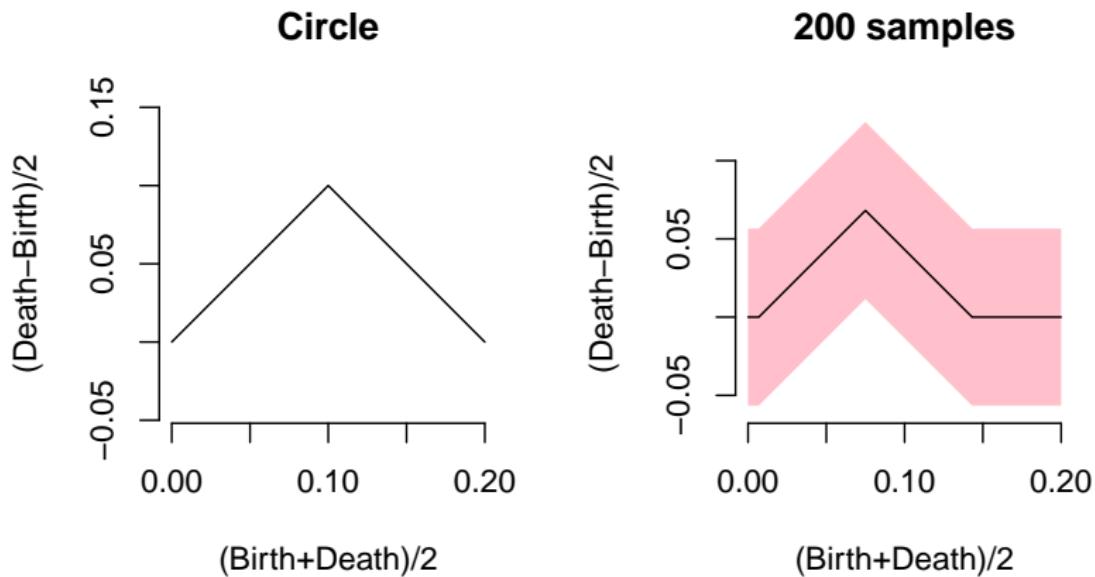
Statistical Inference for Persistent Homology

Featurization using Persistent Homology

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



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

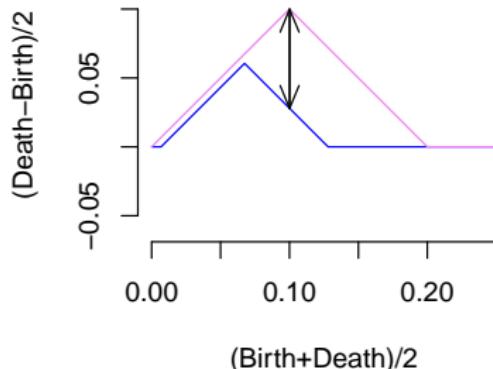


∞ -landscape distance gives a metric on the space of persistence landscapes.

Definition

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

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



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

Theorem

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

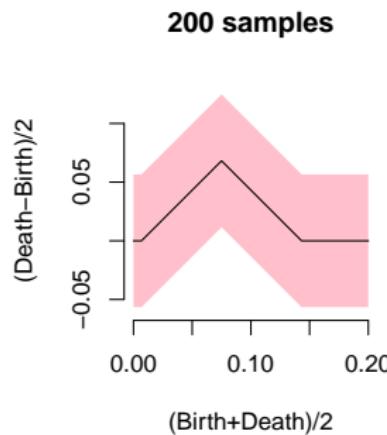
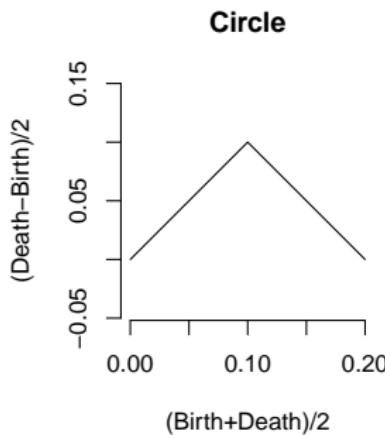
$$\Lambda_\infty(Dgm(f), Dgm(g)) \leq \|f - g\|_\infty.$$

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

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

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

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



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

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