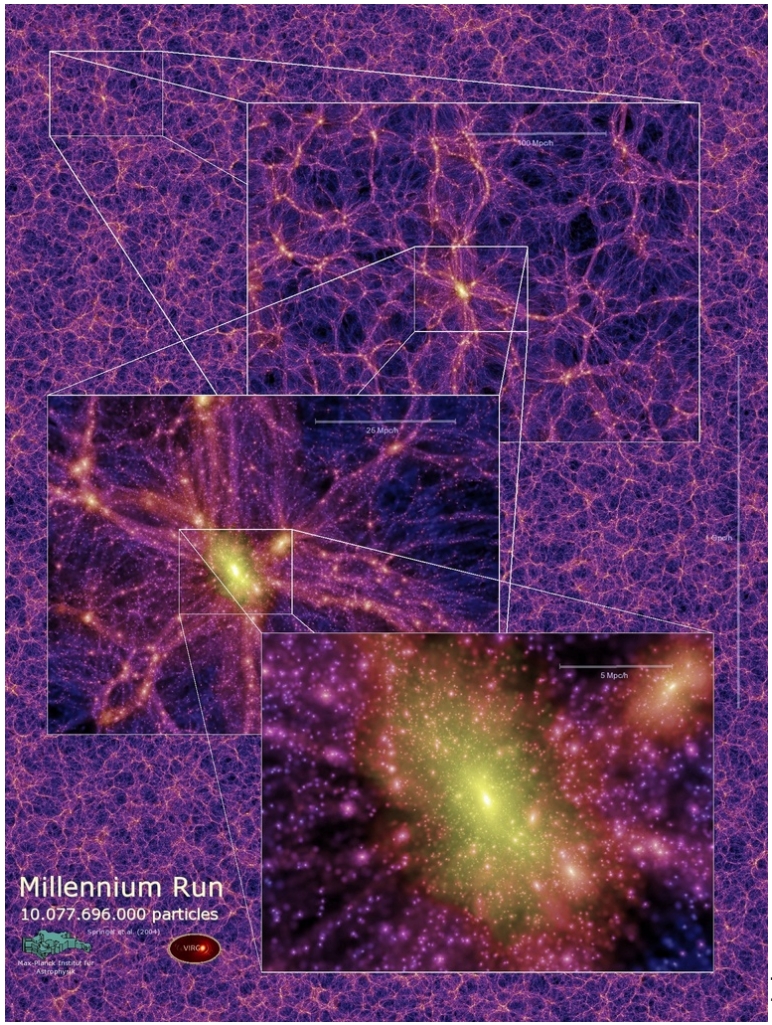
위상학적 자료 분석(Topological Data Analysis)의 통계적 추정 및 기계학습에의 응용

김지수

2023년 한국통계학회 동계학술대회
2023-12-01

위상학적 자료 분석(Topological Data Analysis) 소개

자료의 위상학적 구조로부터 정보를 얻을 수 있습니다.



[1]

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

위상학적 자료 분석(Topological Data Analysis) 및 통계적 추정 및 기계학습에의 응용을 소개합니다.

- 위상학적 자료 분석(Topological Data Analysis) 소개
  - 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)
- 호몰로지(homology)를 통계적으로 추정하기
  - Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)
- Persistent Homology를 통계적으로 추정하기
  - Confidence sets for persistence diagrams (Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2014)
- 위상학적 자료 분석(Topological Data Analysis)을 기계학습에 응용
  - A Survey of Topological Machine Learning Methods (Hensel, Moor, Rieck, 2021)
  - Efficient Topological Layer based on Persistence Landscapes (Kim, Kim, Zaheer, Kim, Chazal, Wasserman, 2020)

구멍의 개수로 기하학적 대상들을 분류할 수 있습니다.

- 기하학적 대상들:
  - ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ
  - A, 字, あ
- 여러 차원에서 구멍들의 개수들을 각각 고려합니다.
  1. β_0 =연결된 성분의 개수 
  2. β_1 =고리(1차원 구의 구멍)의 개수 
  3. β_2 =2차원 구의 구멍의 개수 

예제: 대상들을 호몰로지에 따라 분류합니다.

1. β_0 =연결된 성분의 개수 
2. β_1 =고리의 개수 

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

호몰로지(homology)를 통계적으로 추정합니다.

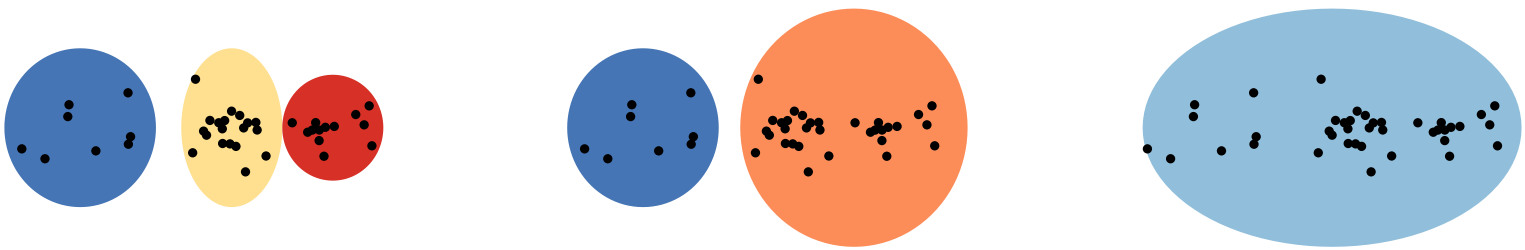
- Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)

자료를 군집으로 묶고자 합니다.

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

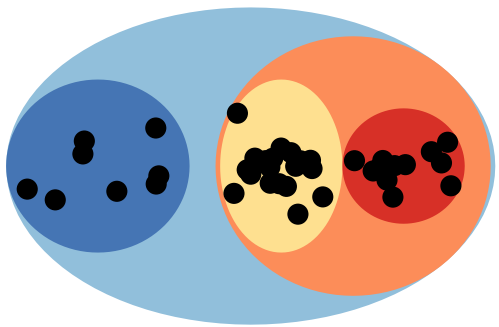
원하는 해상도에 따라 다른 군집이 생길 수 있습니다.

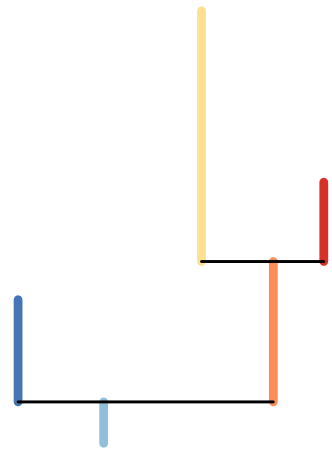
- Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)
- 국소적(local)이고 상세한 정보를 묘사하고 싶으면 (높은 해상도), 작은 규모의 많은 군집이 생깁니다.
- 대역적(global)이고 개략적인 정보를 묘사하고 싶으면 (낮은 해상도), 큰 규모의 적은 군집이 생깁니다.



군집들의 네트워크가 나무를 형성합니다: 군집 나무 (cluster tree)

- Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)
- 다른 수준의 해상도로부터 얻어지는 군집들은 포함 관계에 의해 자연스러운 네트워크가 생깁니다.
- 포함 관계 네트워크는 나무로 표현될 수 있습니다: 군집 나무(cluster tree)

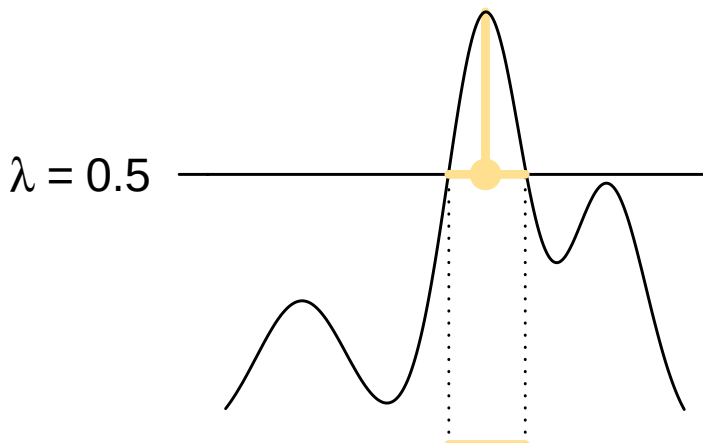




군집 나무(Cluster Tree)는 고밀도 군집들의 계층 구조입니다.

Definition

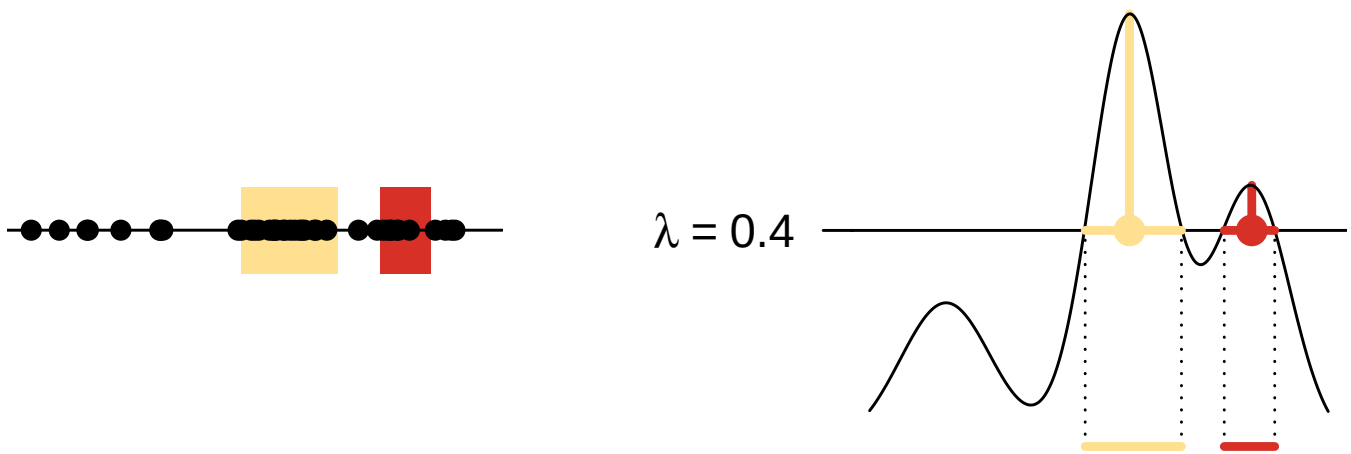
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



군집 나무(Cluster Tree)는 고밀도 군집들의 계층 구조입니다.

Definition

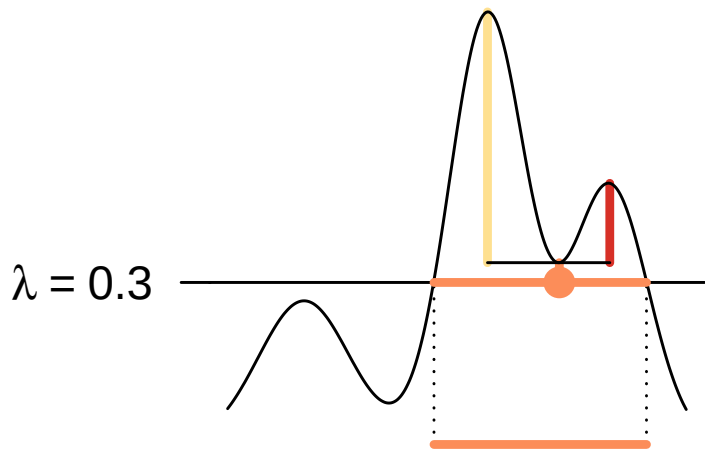
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



군집 나무(Cluster Tree)는 고밀도 군집들의 계층 구조입니다.

Definition

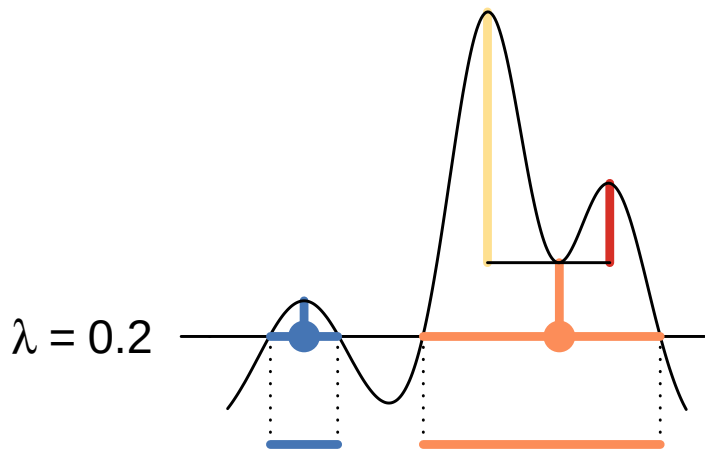
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



군집 나무(Cluster Tree)는 고밀도 군집들의 계층 구조입니다.

Definition

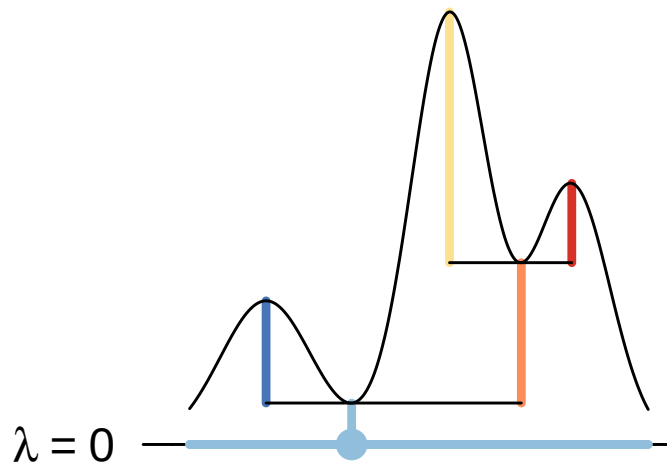
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



군집 나무(Cluster Tree)는 고밀도 군집들의 계층 구조입니다.

Definition

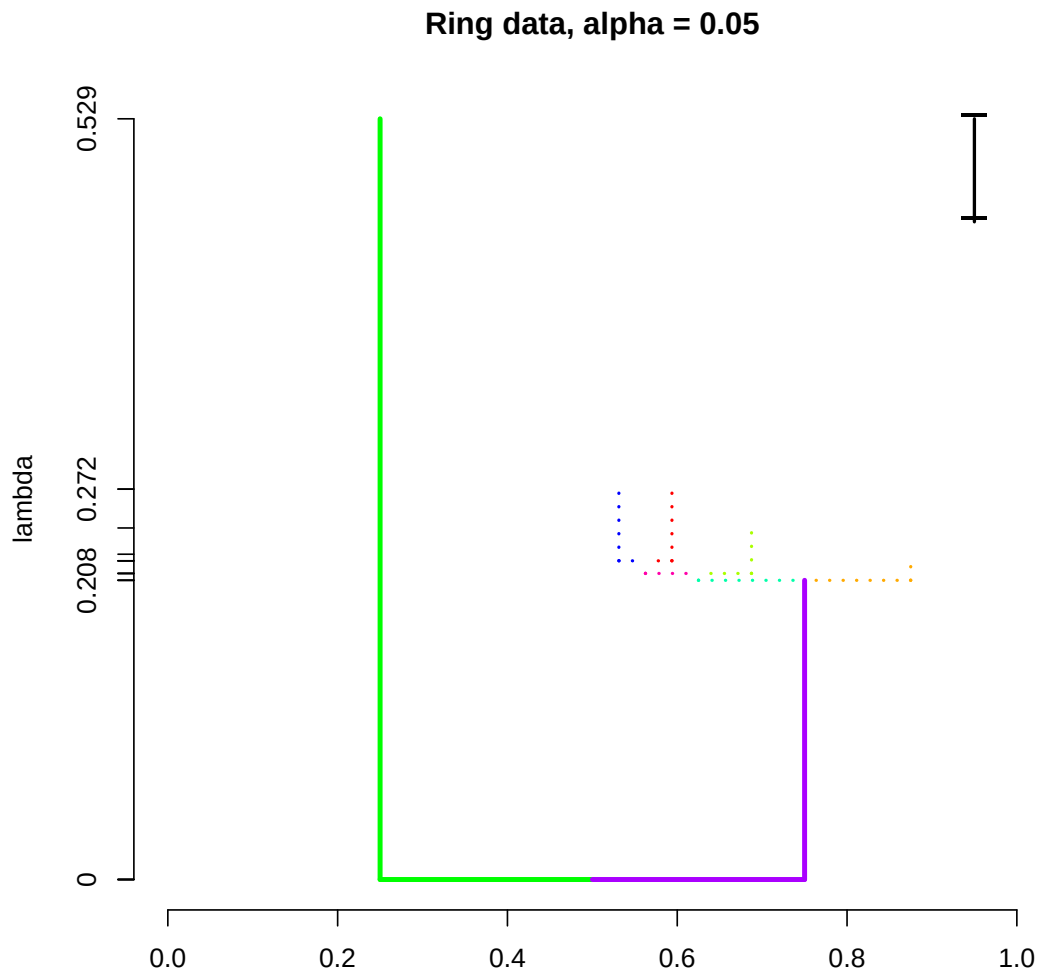
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



신뢰집합은 경험적 군집 나무에서 잡음을 줄이는 데에 도움을 줍니다.

- 점근적 $1-\alpha$ 신뢰집합 C_α는 다음을 만족하는 군집 나무들의 집합입니다:

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



$1-\alpha$ 신뢰집합 C_α는 부트스트랩으로 계산할 수 있습니다.

- $T_{\hat{p}_h}$ 를 핵밀도추정(kernel density estimator) $\hat{p}_h$에서 계산된 군집나무로 놓습니다. 이 때,

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

그리고 신뢰집합을 $T_{\hat{p}_h}$을 중심으로 하고 반지름이 t_α인 공으로 정의합니다, 즉,

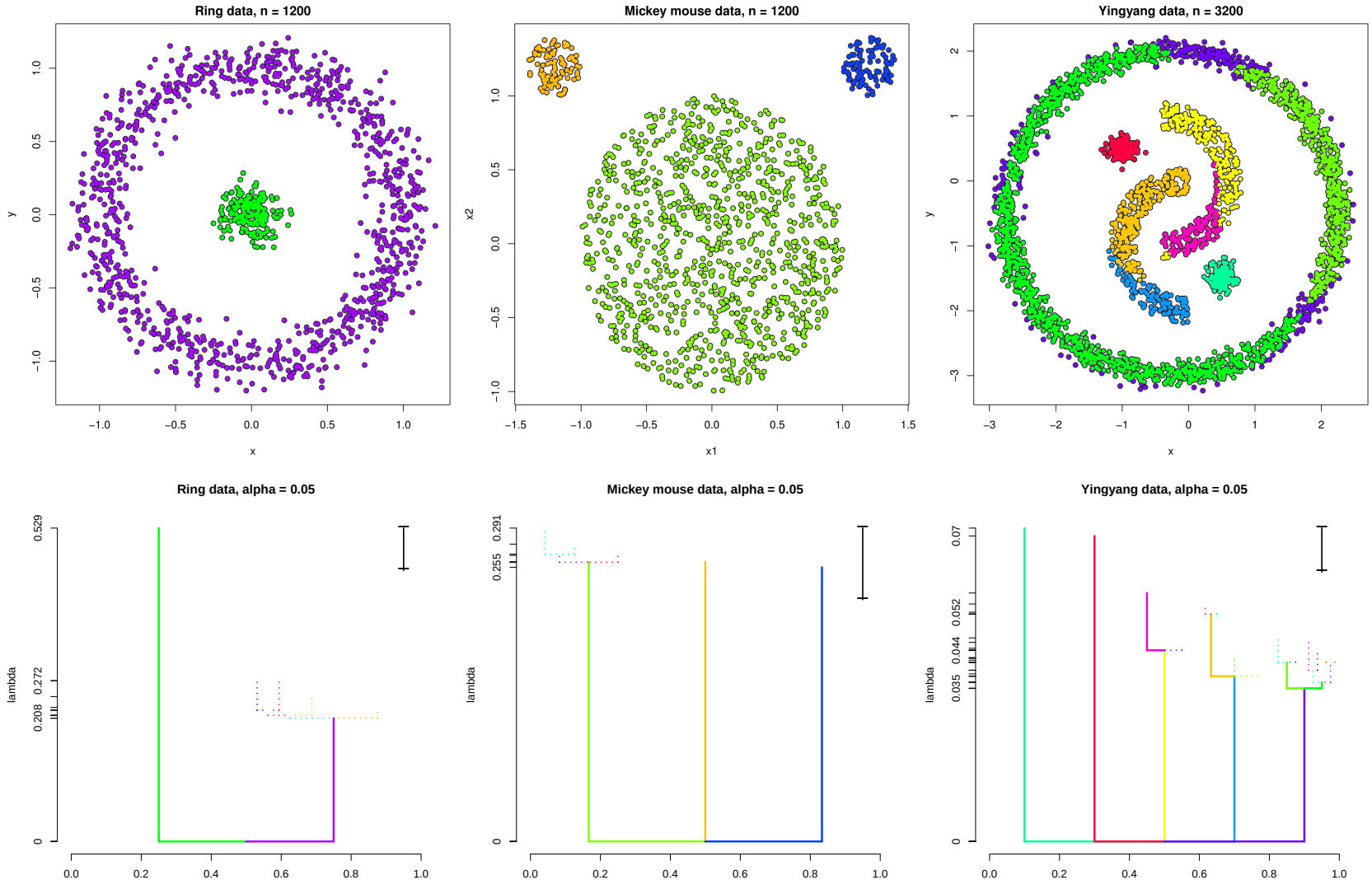
$$\hat{C}_\alpha = \{T : d_\infty(T, T_{\hat{p}_h}) \leq t_\alpha\}.$$

Theorem

(Theorem 3) 위의 신뢰집합 $\hat{C}_\alpha$ 은 다음을 만족합니다:

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

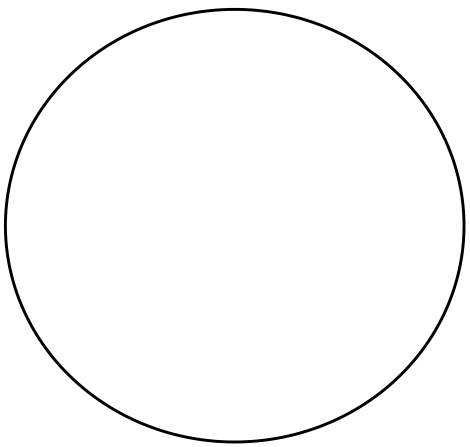
신뢰집합을 이용하여 가지치기한 군집나무로 실제 군집나무를 찾을 수 있습니다.



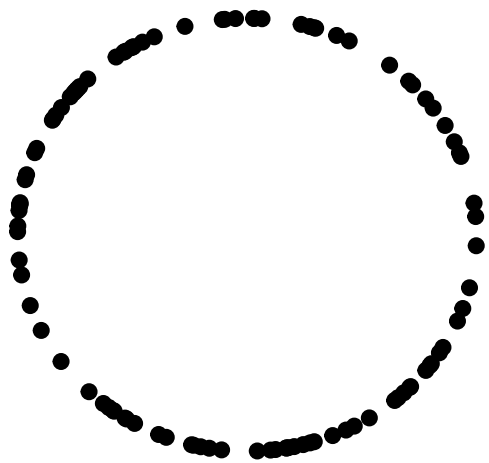
유한한 자료의 호몰로지는 기저 구조의 호몰로지와 다르기 때문에, 유한한 자료로 직접 기저 구조의 호몰로지를 추정할 수는 없습니다.

- 자료를 분석할 때, 기저 구조의 특성을 자료의 특성으로부터 추정할 수 있는 로버스트한 특성을 선호합니다.
- 호몰로지는 로버스트하지 않습니다:

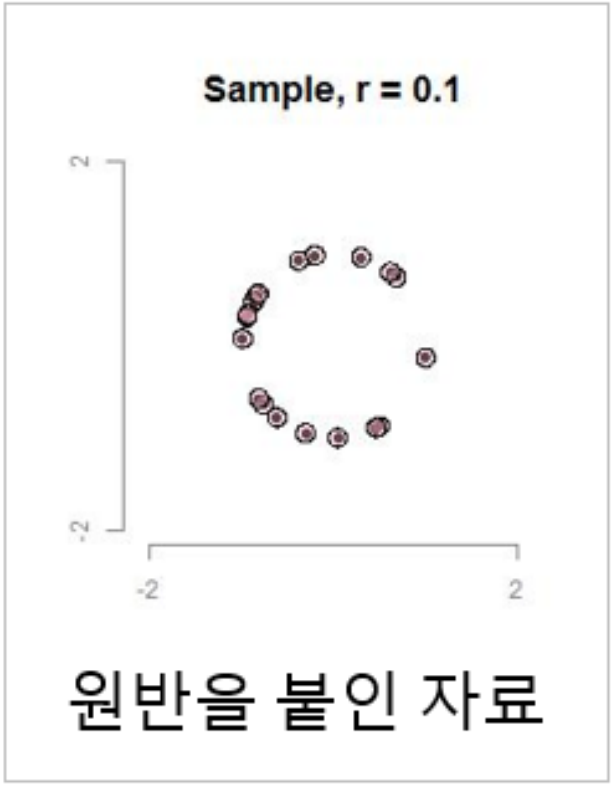
Underlying circle: $\beta_0 = 1, \beta_1 = 1$



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

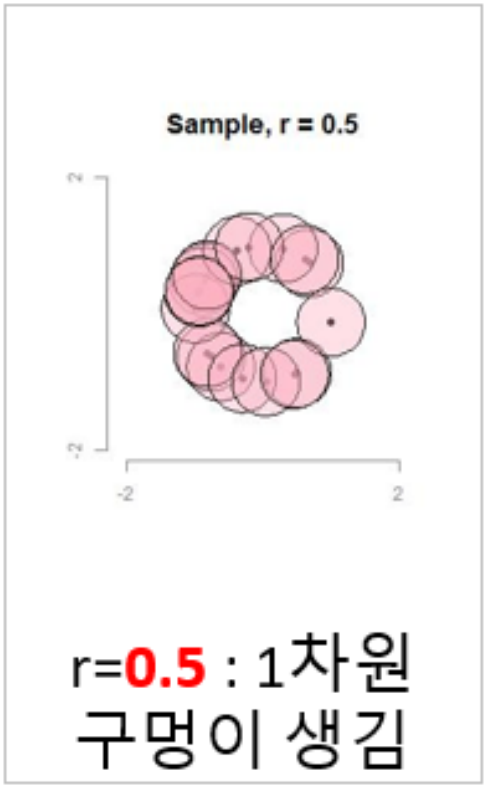


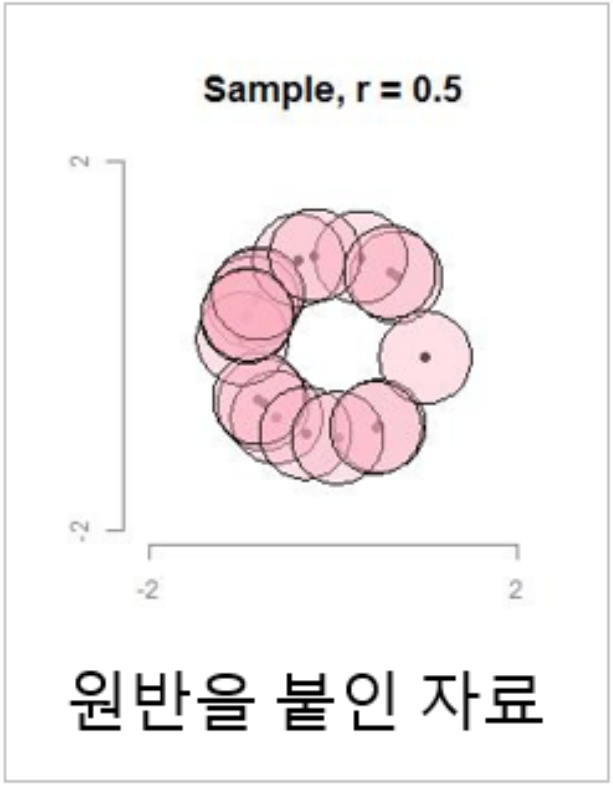
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



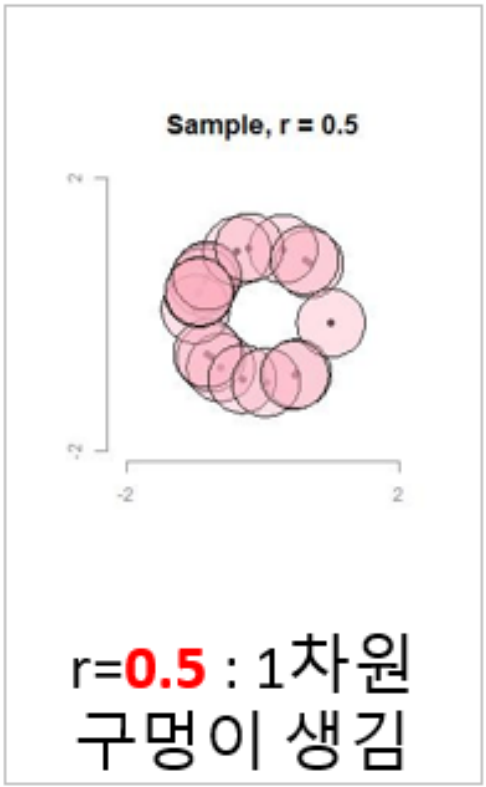
원반을 붙인 자료

Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.

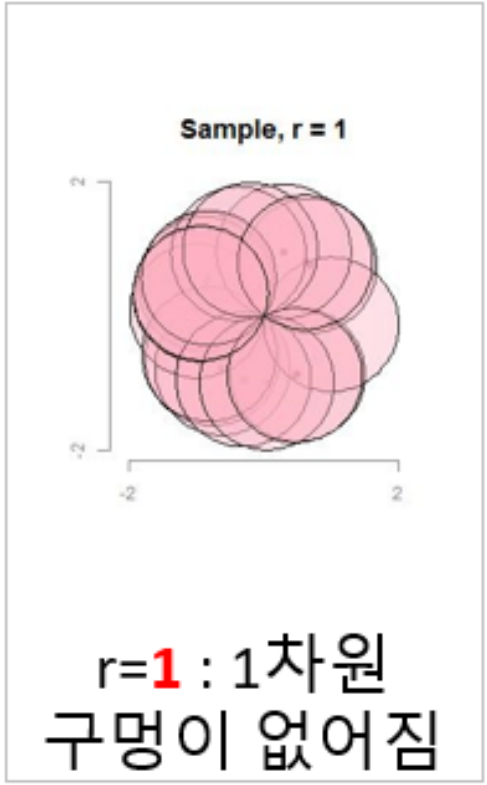




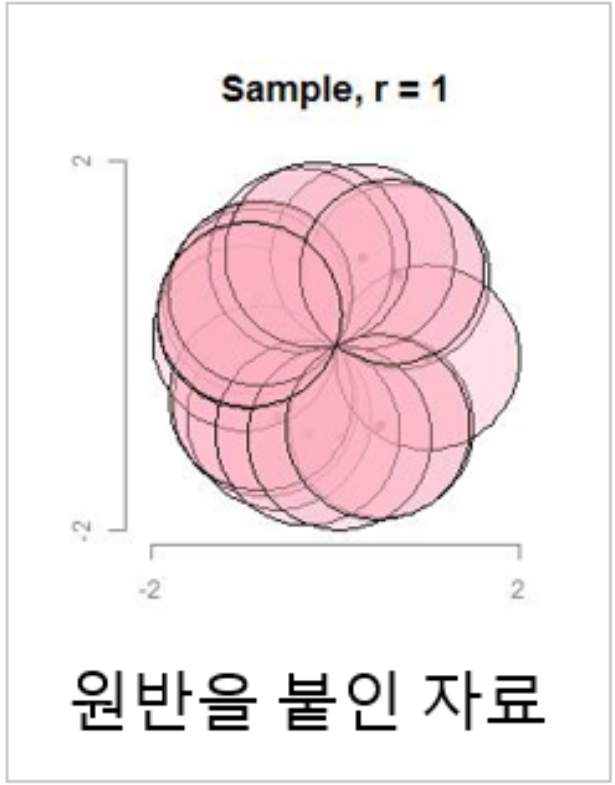
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



r=**0.5** : 1차원 구멍이 생김

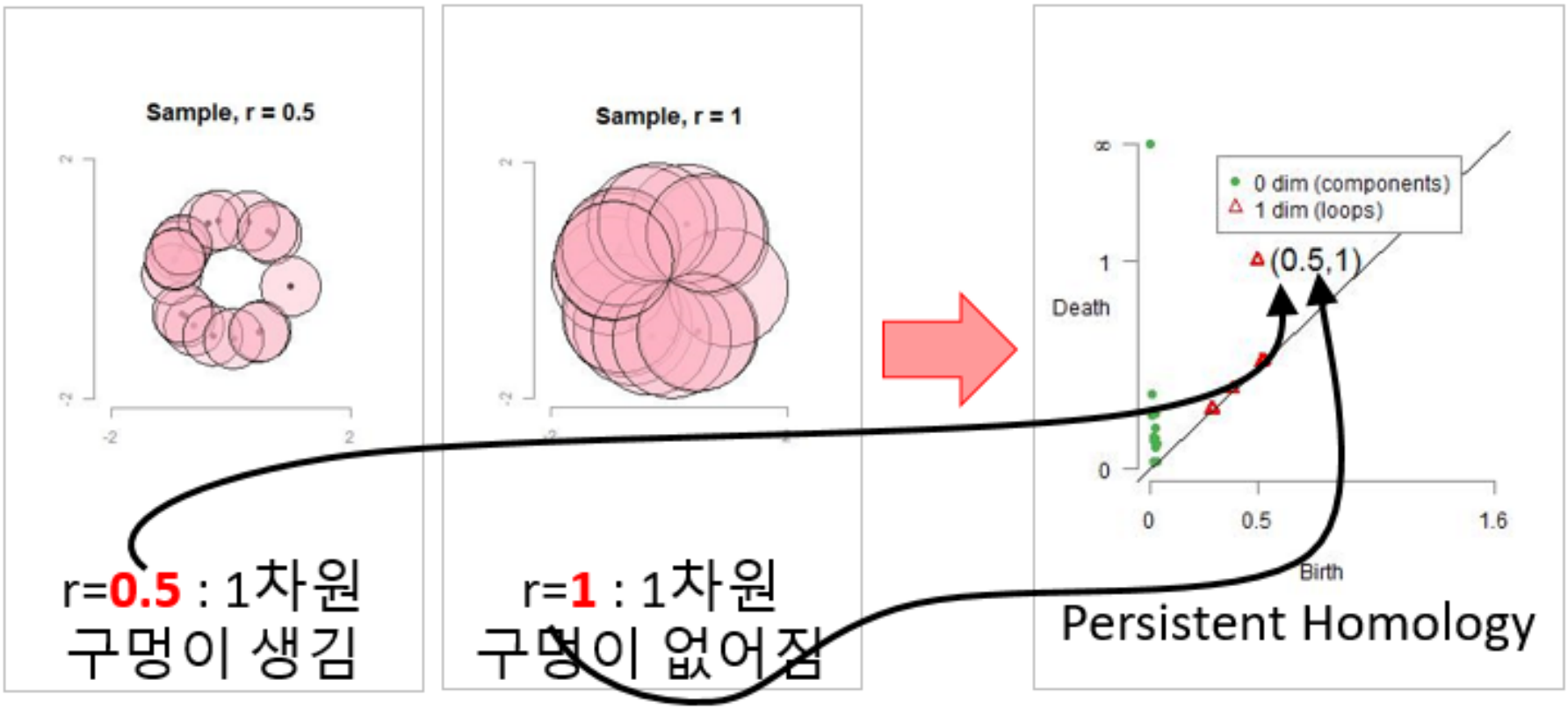


r=**1** : 1차원 구멍이 없어짐



원반을 붙인 자료

Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.

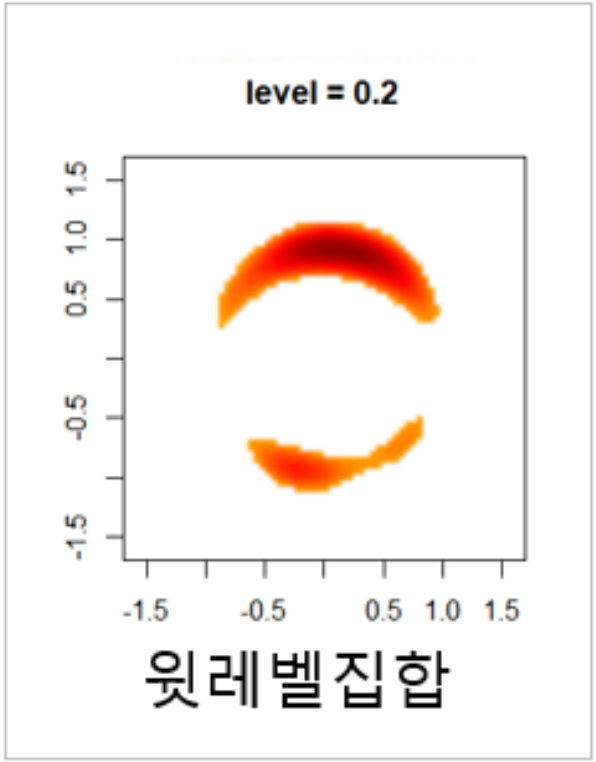


기저 구조의 위상학적 정보를 추출하는 데에 핵밀도추정(kernel density estimator)을 사용합니다.

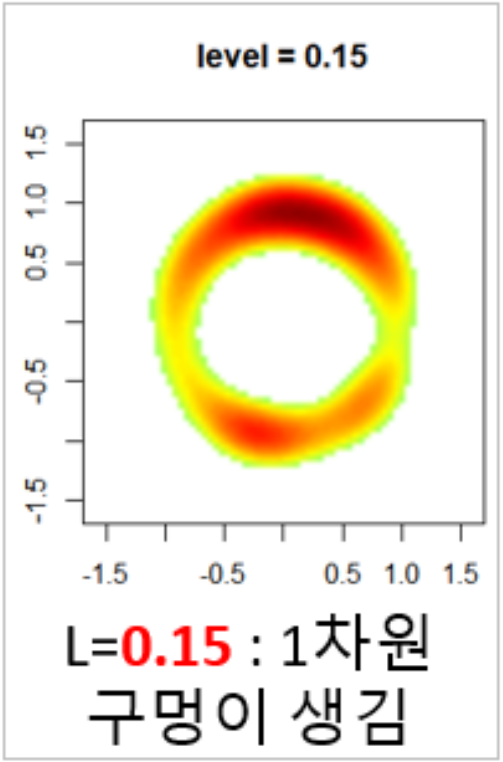
- 핵밀도추정(kernel density estimator)은 다음과 같습니다:

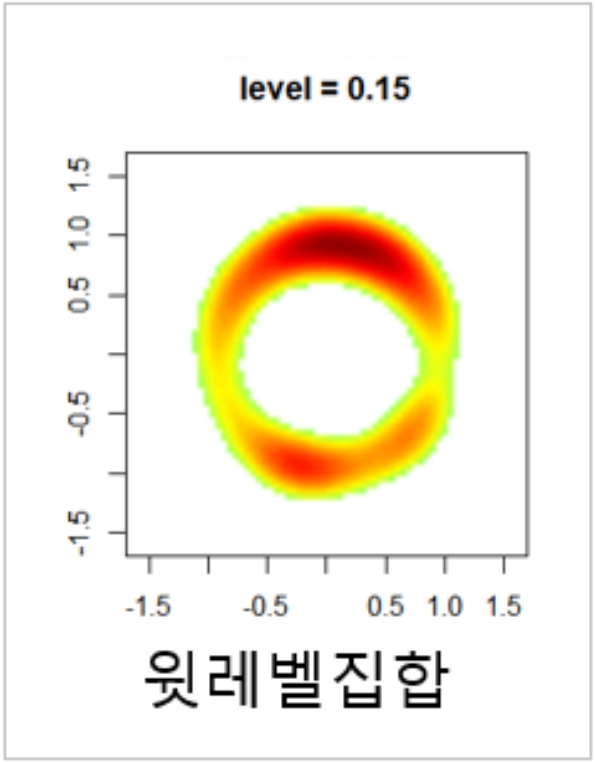
$$\hat{p}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.

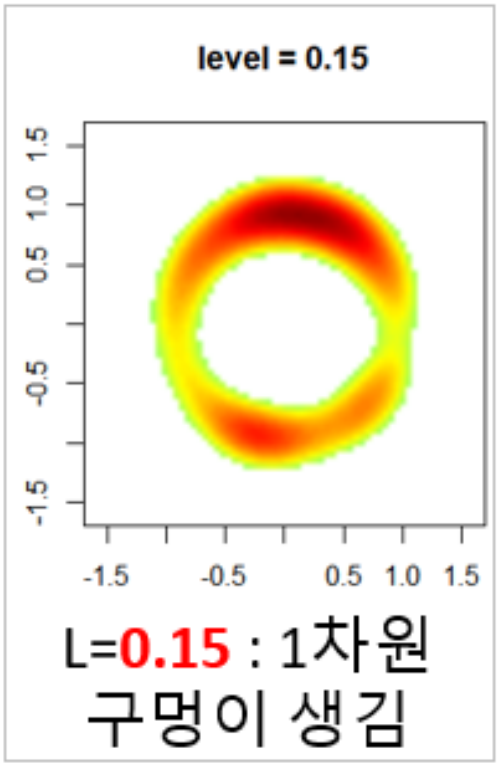


Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.

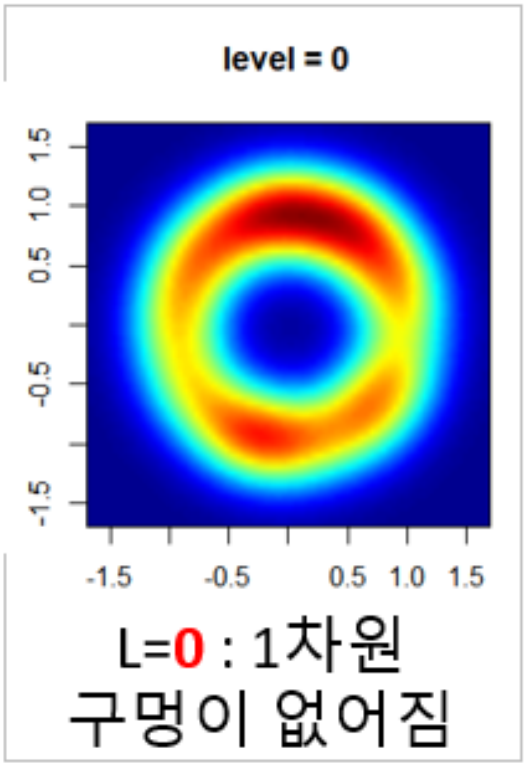




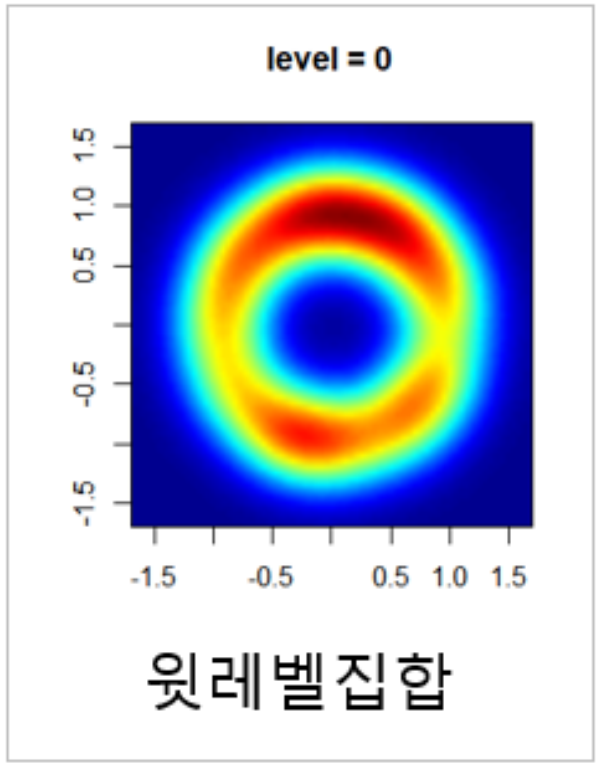
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



L=**0.15** : 1차원 구멍이 생김

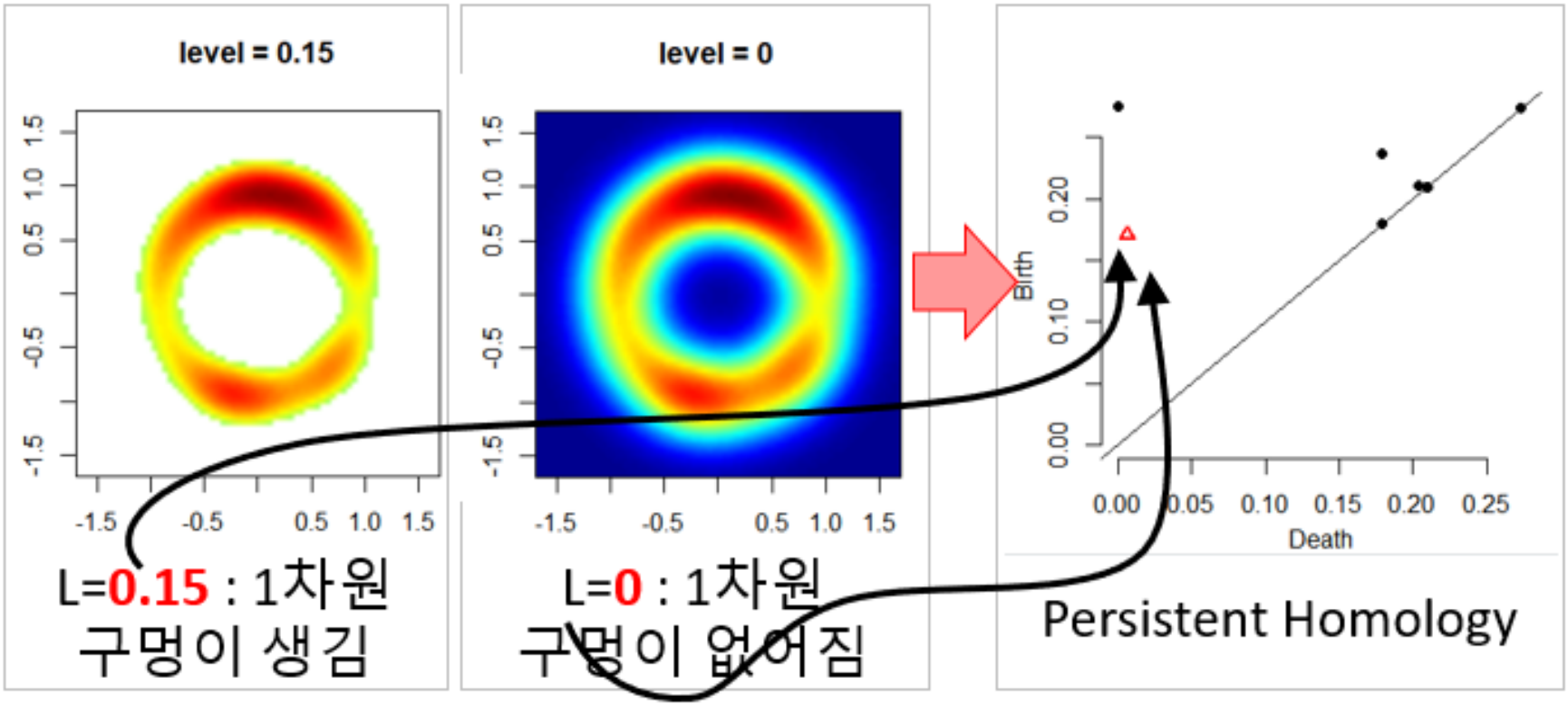


L=**0** : 1차원 구멍이 없어짐

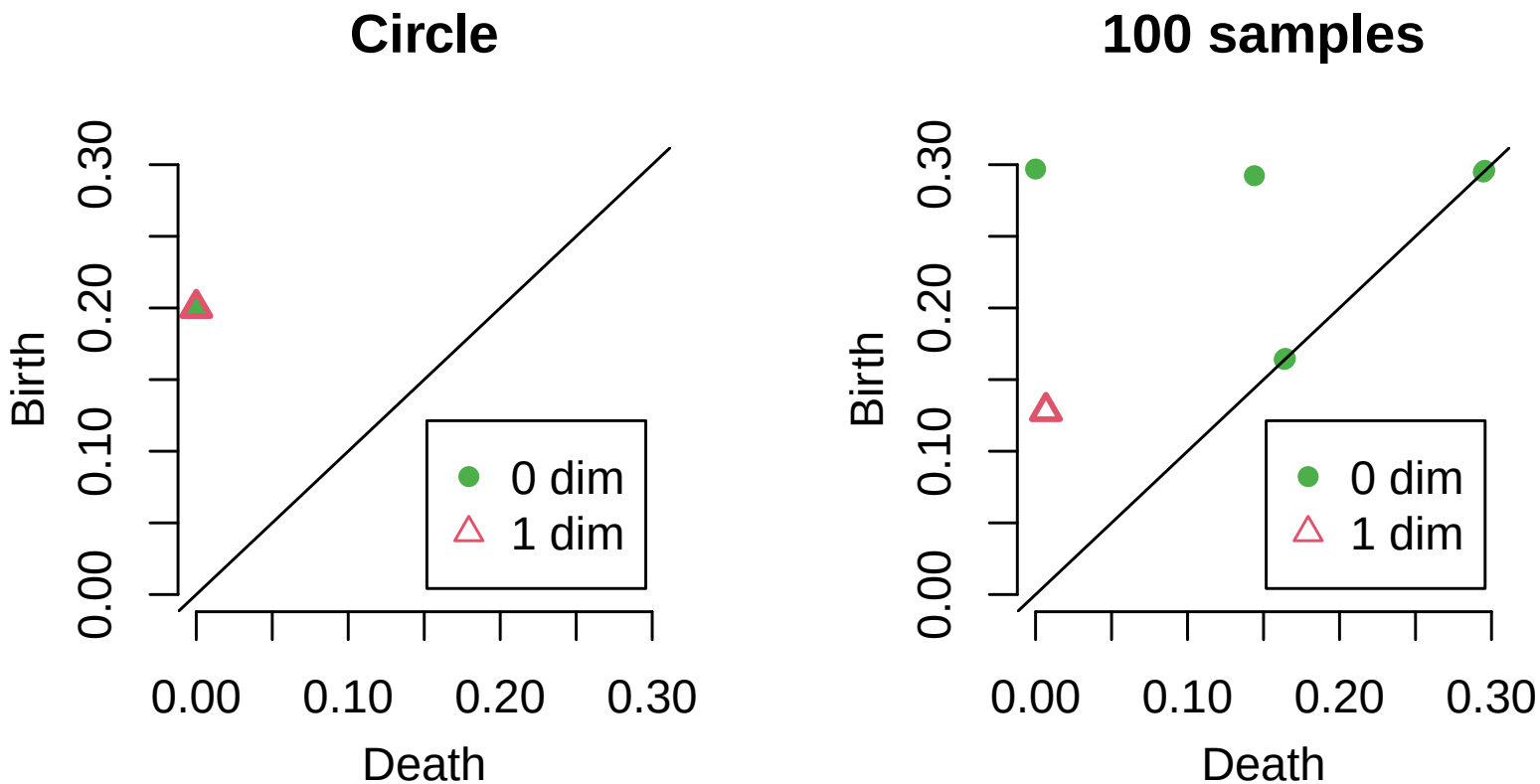


윗레벨집합

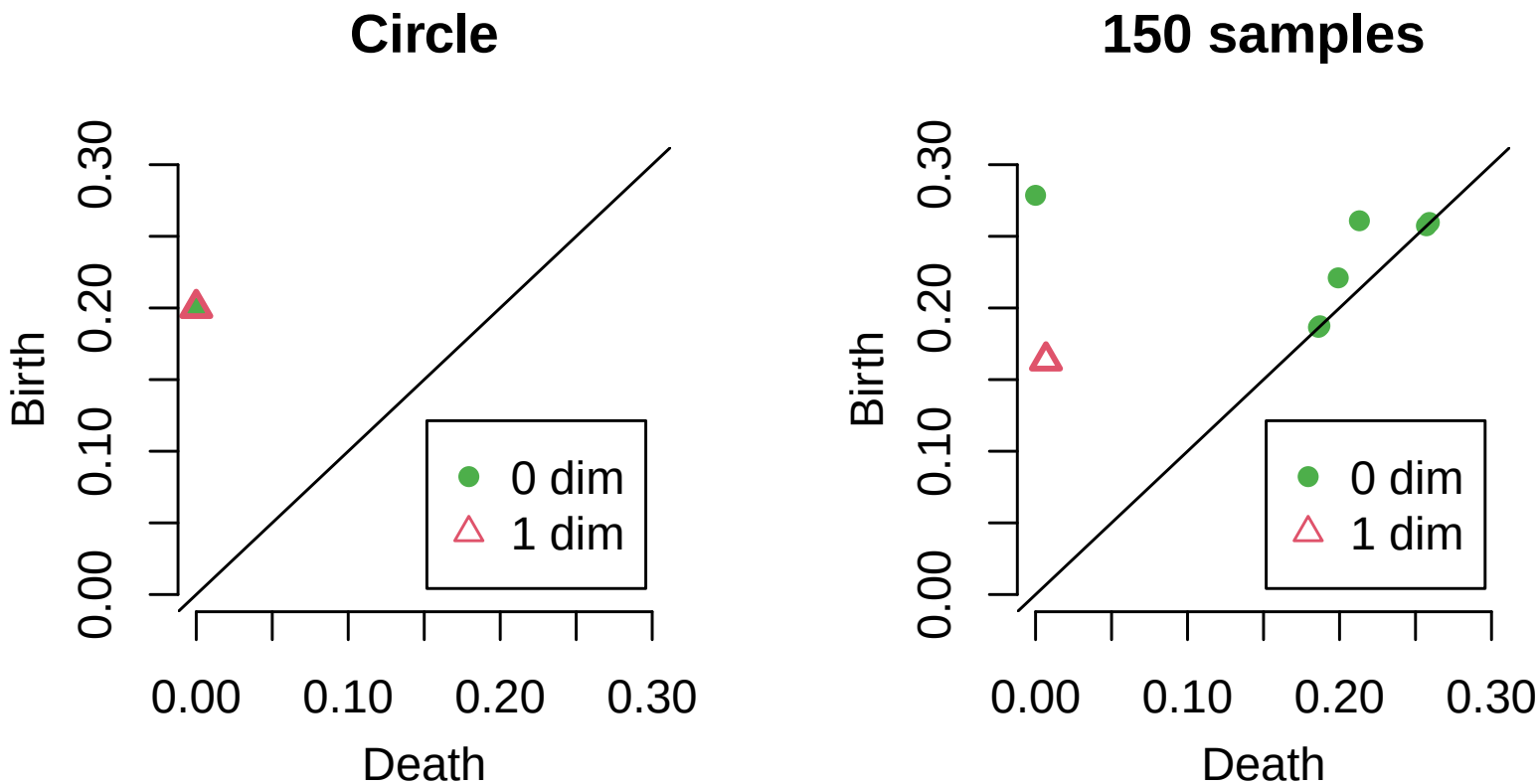
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



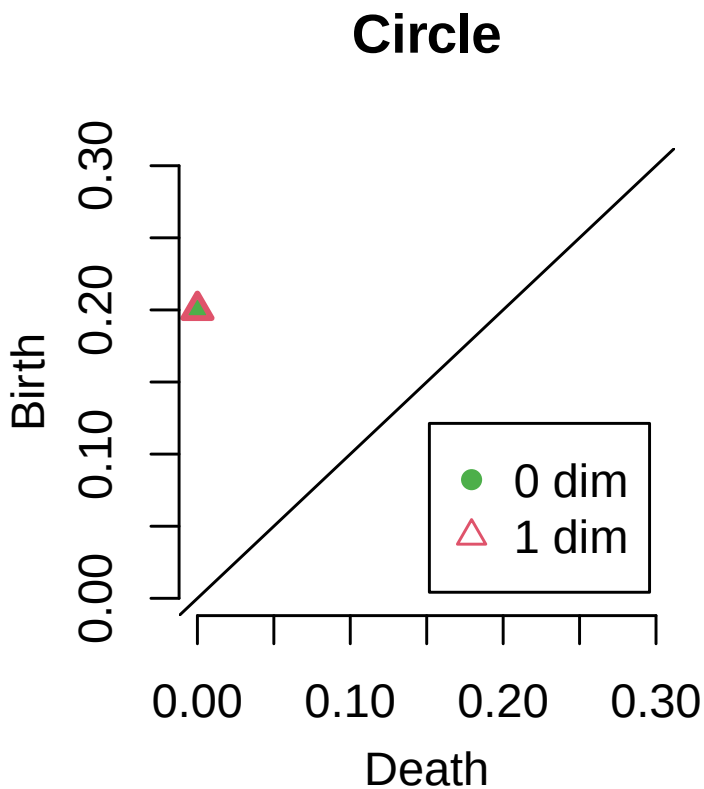
유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.

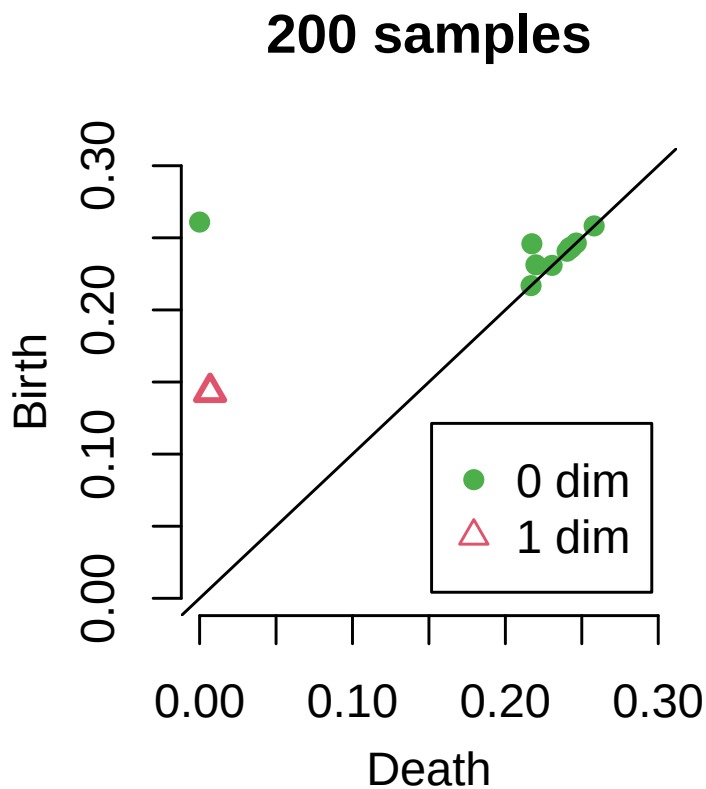


유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.

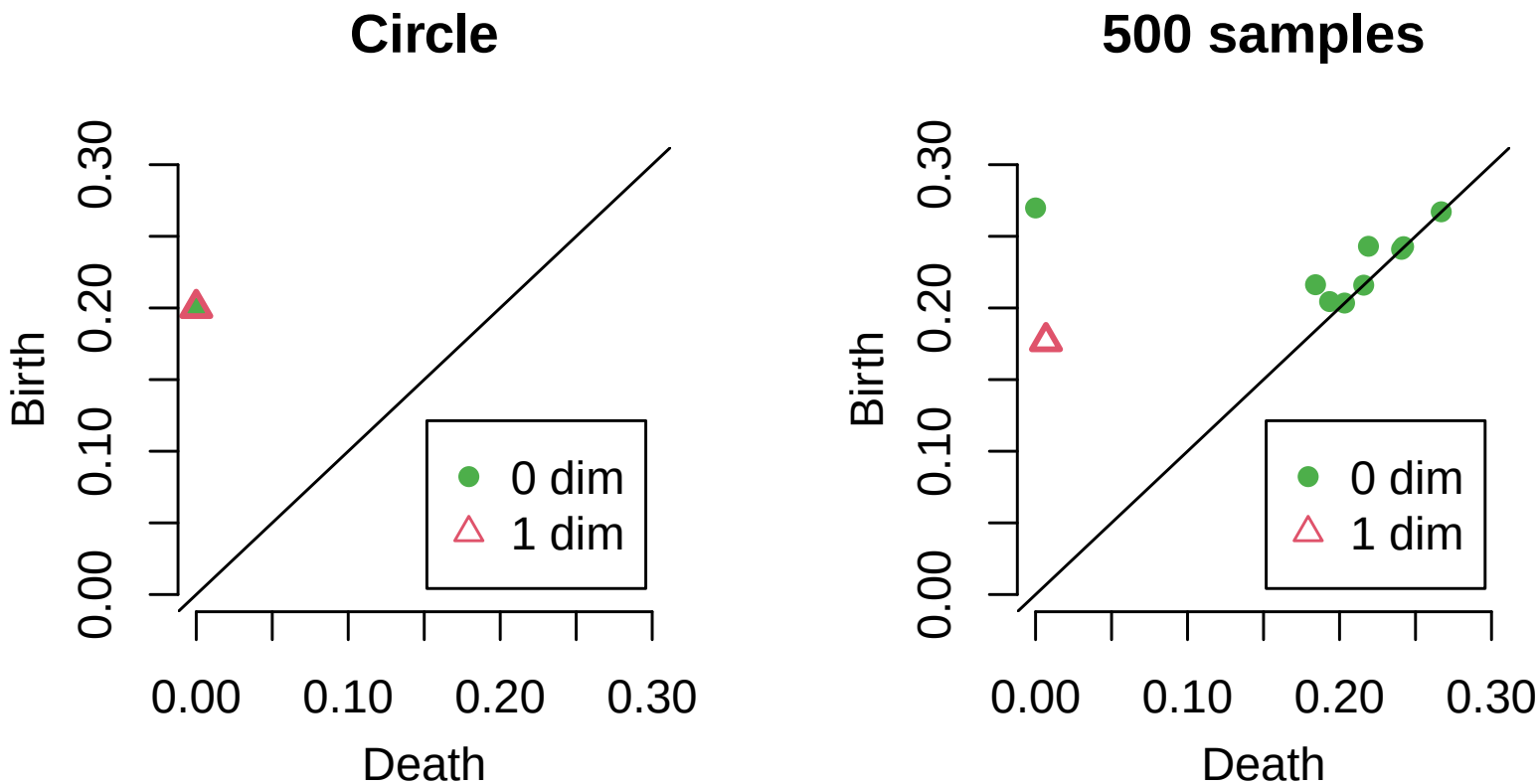


유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.

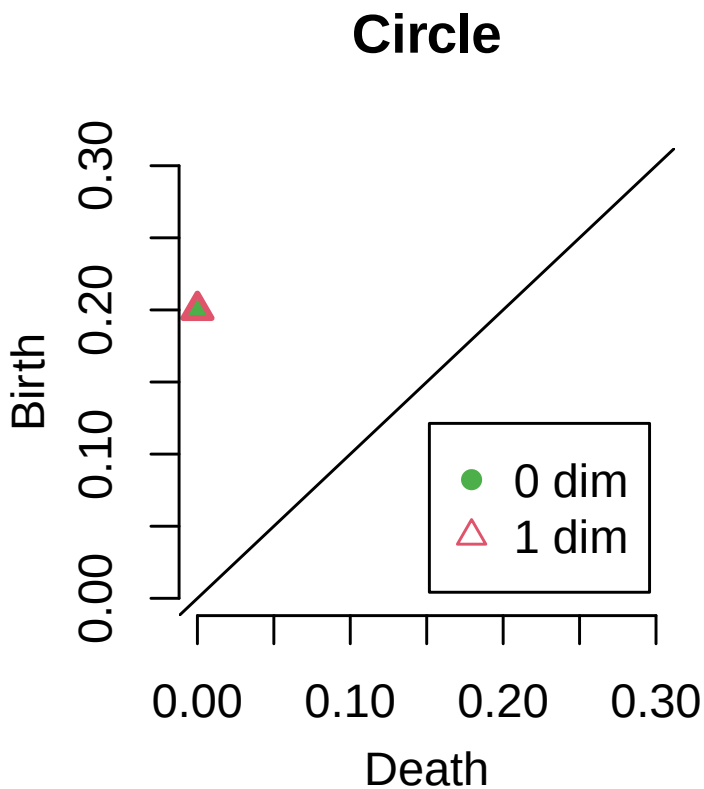


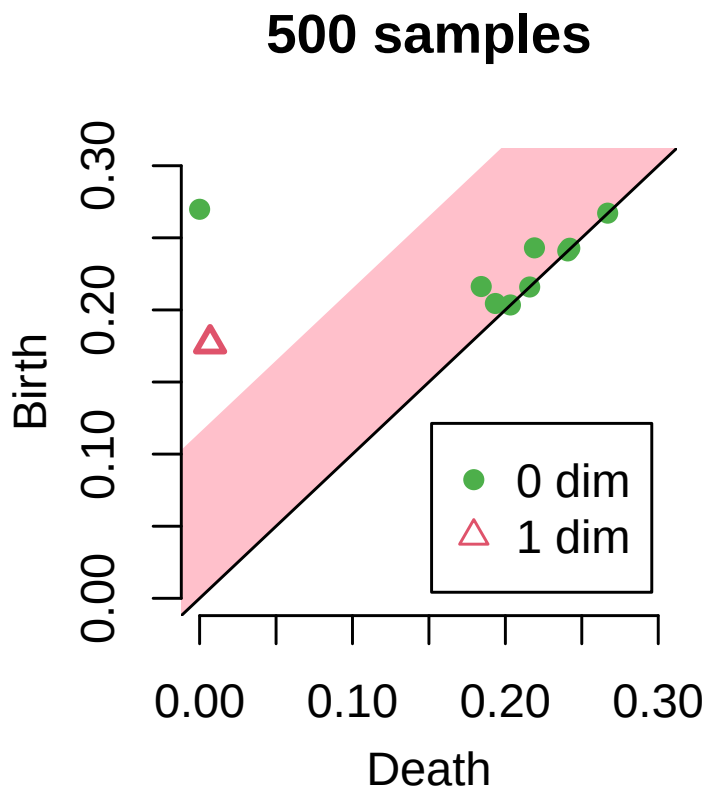


유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.



통계적으로 유의한 호몰로지 특성과 그렇지 않은 호몰로지 특성을 어떻게 구분할까요?





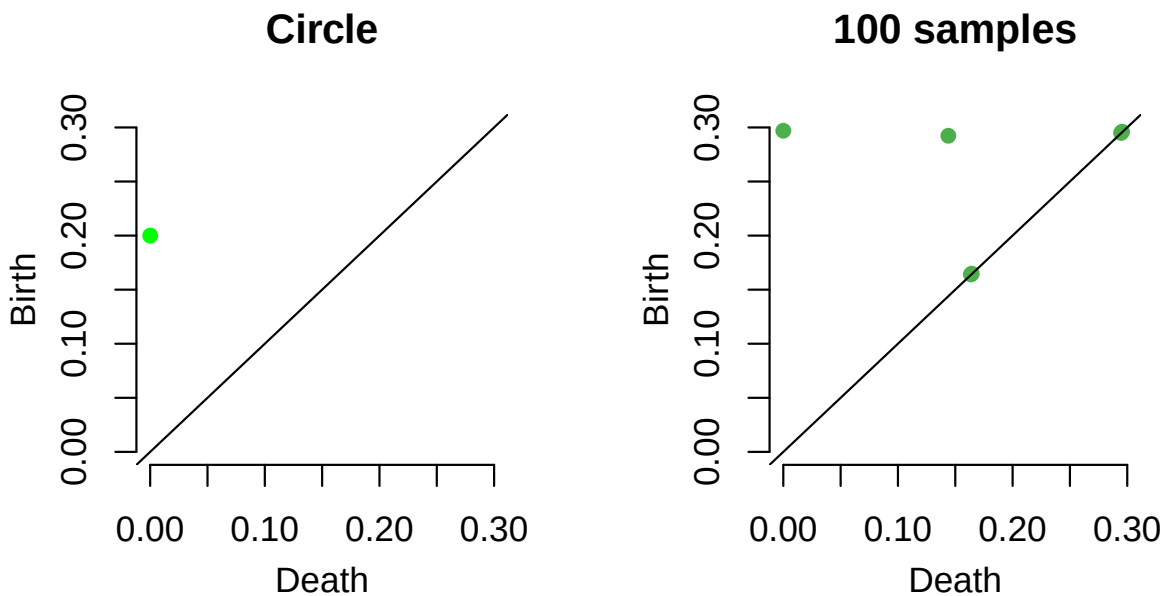
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

Definition

D_1, D_2를 두 Persistent homology라고 하면, Bottleneck distance는 다음과 같이 정의됩니다:

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

이 때, γ는 D_1에서 D_2로 가는 모든 일대일대응이 될 수 있습니다.



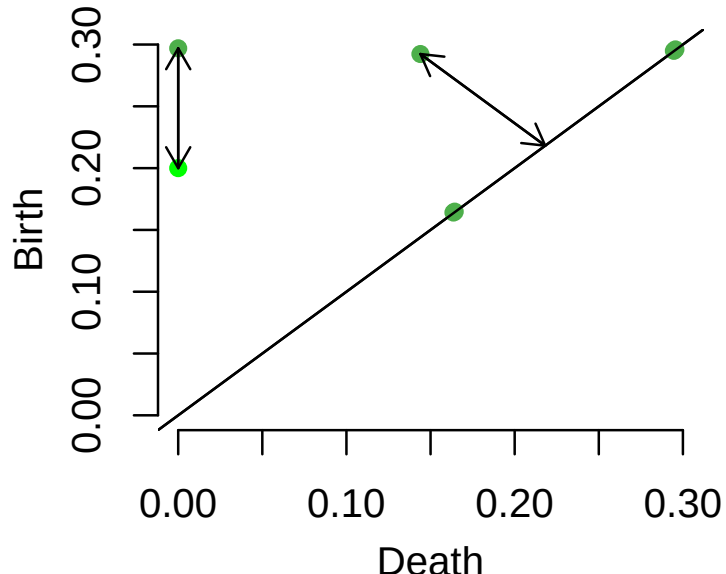
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

Definition

D_1, D_2를 두 Persistent homology라고 하면, Bottleneck distance는 다음과 같이 정의됩니다:

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

이 때, γ는 D_1에서 D_2로 가는 모든 일대일대응이 될 수 있습니다.

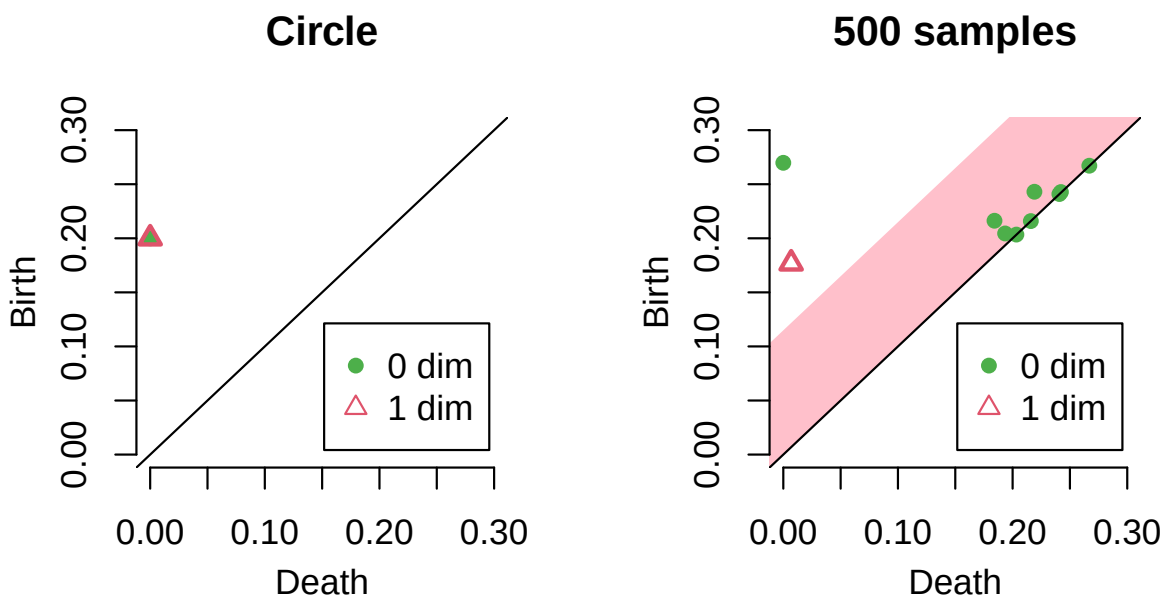


$$\inf_\gamma \sup_{x \in D_1} \|x - \gamma(x)\|_\infty = 0.1$$

Persistent homology의 신뢰띠는 Persistent homology를 높을 확률로 포함하는 확률변수입니다.

기저 M과 자료 X의 Persistent homology를 각각 $Dgm(M)$과 $Dgm(X)$라고 놓습니다. 유의수준 $\alpha \in (0,1)$가 주어졌을 때, $(1-\alpha)$ 신뢰띠 $c_n = c_n(X)$는 다음을 만족하는 확률변수입니다:

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



Persistent homology의 신뢰띠는 붓스트랩으로 계산할 수 있습니다.

1. 주어진 자료 $X = \{x_1, \ldots, x_n\}$에서 핵밀도추정(kernel density estimator) $\hat{p}_h$를 계산합니다.
2. $X = \{x_1, \ldots, x_n\}$로부터 $X^* = \{x_1^*, \ldots, x_n^*\}$를 복원추출하고, X^*의 핵밀도추정 $\hat{p}_h^*$을 계산한 후, $\theta^* = \sqrt{nh^d}||\hat{p}_h^*(x) - \hat{p}_h(x)||_\infty$를 계산합니다.
3. 전단계를 B번 반복하여 $\theta_1^*, \ldots, \theta_B^*$를 얻습니다.
4. 분위수 $\hat{z}_\alpha = \inf\left\{q : \frac{1}{B}\sum_{j=1}^{B} I(\theta_j^* \geq q) \leq \alpha\right\}$를 계산합니다.
5. $\mathbb{E}[\hat{p}_h]$의 $(1-\alpha)$ 신뢰띠는 $\left[\hat{p}_h - \frac{\hat{z}_\alpha}{\sqrt{nh^d}}\ ,\ \hat{p}_h + \frac{\hat{z}_\alpha}{\sqrt{nh^d}}\right]$이 됩니다.

기계학습(machine learning) (아주) 대충 보기

- 주어진 문제와 자료에서, 기계학습(machine learning) / 심층학습(deep learning)은 매개화된 모형(parametrized model)을 학습합니다.
  - 주어진 자료 X,
  - 매개화된 모형(parametrized model) f_θ,
  - 문제에 맞춰진 손실함수(loss function) $\mathcal{L}$,
  - 기계학습은 손실함수를 최소화하는 해를 계산합니다: $\arg\min_\theta \mathcal{L}(f_\theta, \mathcal{X})$.
- 많은 경우, 최소해의 명시적 형태(explicit formula)를 구하는 것은 불가능하거나 너무 비쌉니다(e.g. 큰 역행렬을 계산). 따라서, $\nabla_\theta \mathcal{L}(f_\theta, \mathcal{X})$를 이용한 경사법(gradient descent)을 사용합니다:

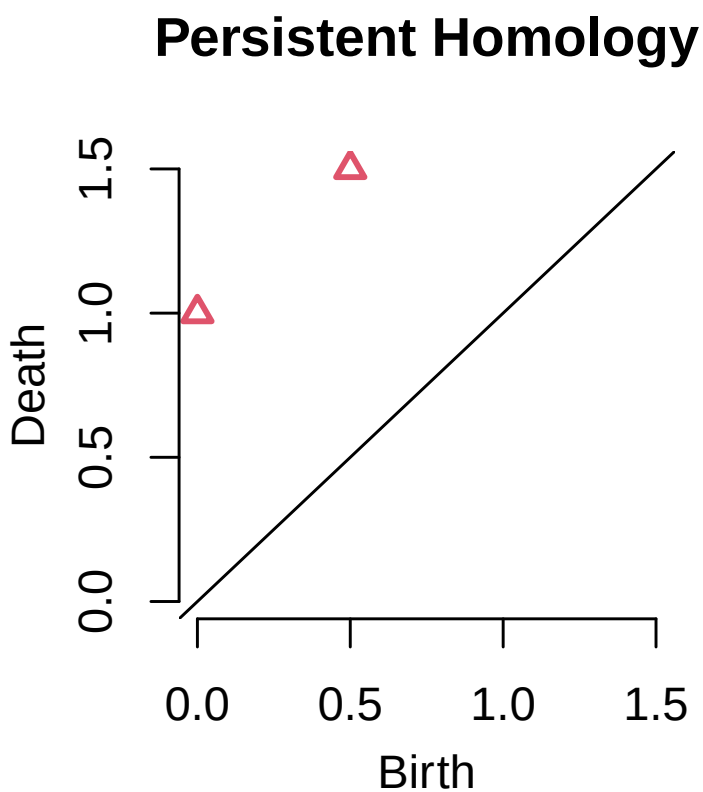
$$\theta_{n+1} = \theta_n - \lambda \nabla_\theta \mathcal{L}(f_\theta, \mathcal{X}).$$

위상학적 자료 분석(Topological Data Analysis)을 기계학습에 응용합니다.

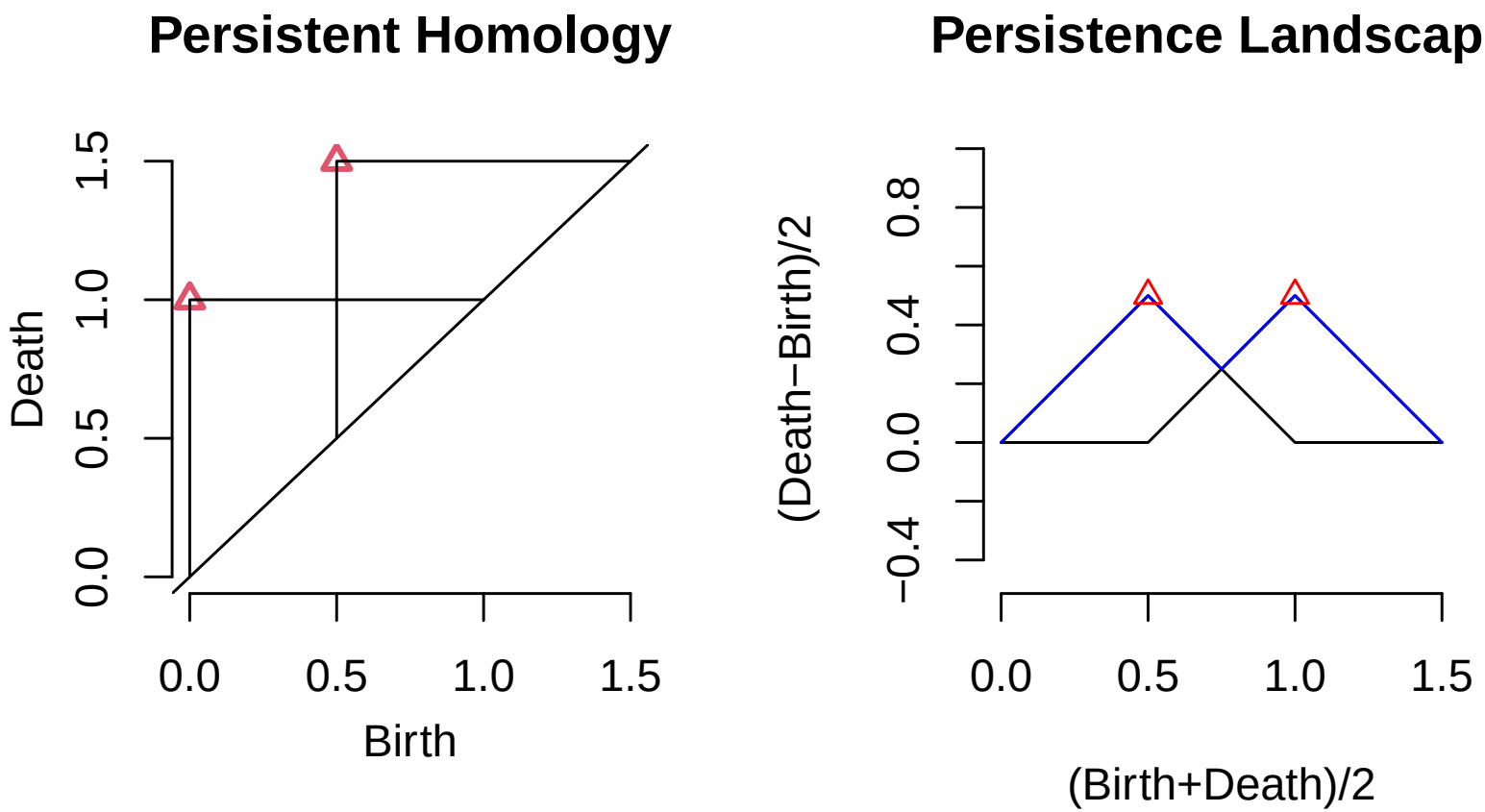
- A Survey of Topological Machine Learning Methods (Hensel, Moor, Rieck, 2021)
- 위상학적 자료 분석을 기계학습에 응용하는 데에는 크게 두 가지 방향이 있습니다:
  - 위상학적 자료 분석을 이용하여 특성(feature)을 만들어, 자료 X에 위상학적 특성을 추가하기: 더 흔한 방식
    - PLLay: Efficient Topological Layer based on Persistence Landscapes (Kim, Kim, Zaheer, Kim, Chazal, Wasserman, 2020)
  - 손실함수(loss function) $\mathcal{L}$에 위상학적 손실 고려하기: 최근 주목

Persistent homology를 한 번 더 요약해서 유클리드 공간 또는 함수 공간에 넣습니다.

- Persistent homology의 공간은 구조적으로 복잡하여 기계학습 알고리즘과 같이 사용하기는 힘듭니다.
- Persistent homology를 한 번 더 요약해서 유클리드 공간 또는 함수 공간에 넣으면 기계학습의 알고리즘에 사용하기 편합니다.
  - Persistence Landscape, Persistence Silhouette, Persistence Image 등 여러 방법이 있습니다.

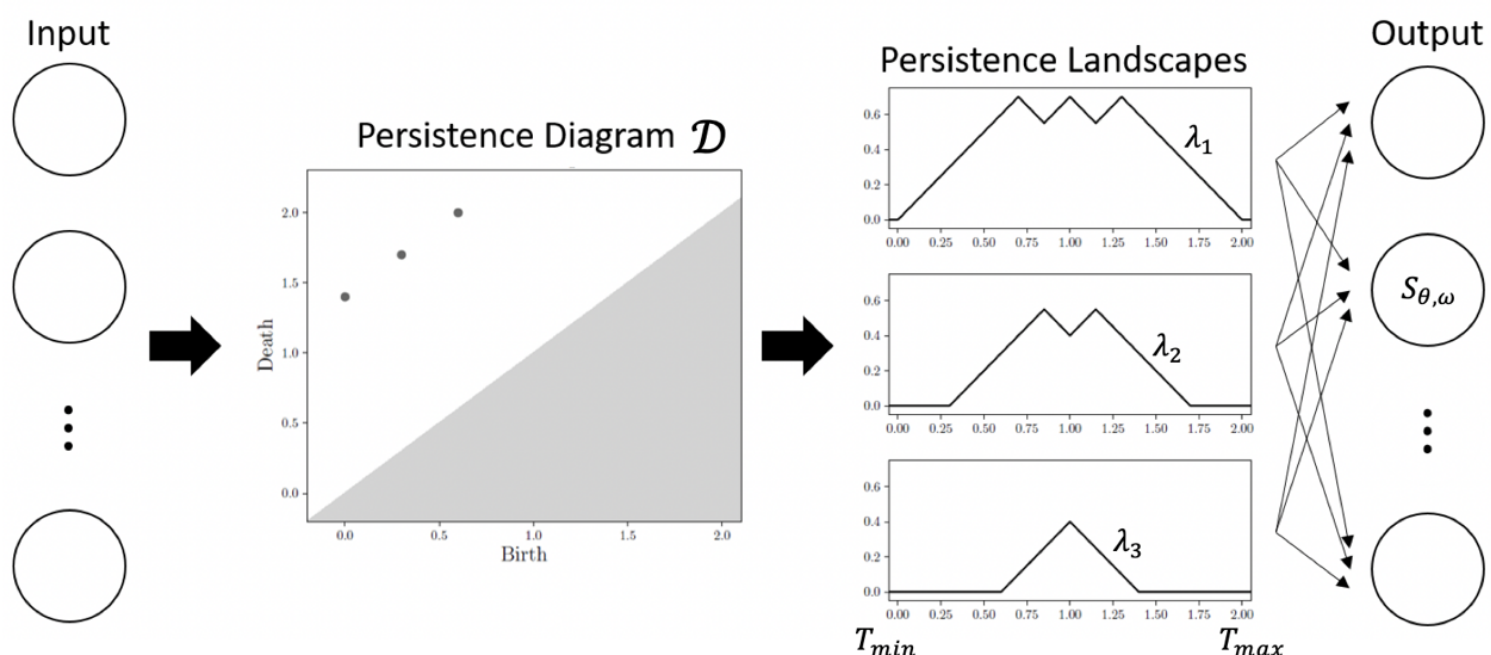


Persistence Landscape은 Persistent homology의 함수 요약입니다.

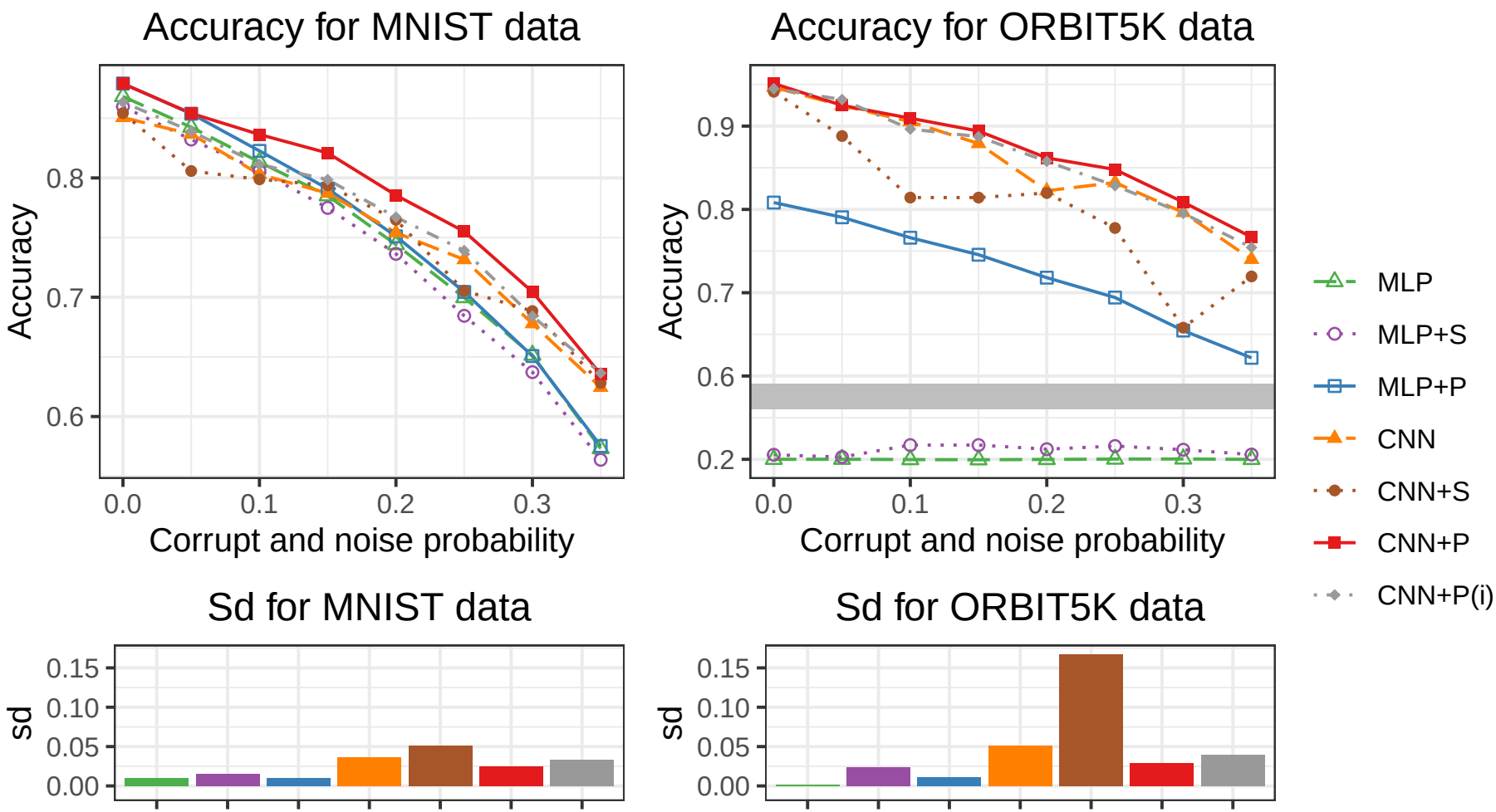


Persistence Landscape으로 위상학적 층(topological layer) 만들기

1. 자료 X에 적당한 simplicial complex K와 함수 f를 선택하여 Persistent Homology $\mathcal{D}$를 계산합니다.
2. $\mathcal{D}$로부터 Landscape $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$을 계산합니다.
3. 매개변수 $\omega \in \mathbb{R}^{K_{\max}}$를 이용하여 가중평균함수 $\bar{\lambda}_\omega(t) := \sum_{k=1}^{K_{\max}} \omega_k \lambda_k(t)$를 계산합니다.
4. $\bar{\lambda}_\omega$를 벡터화하여 $\bar{\Lambda}_\omega \in \mathbb{R}^m$을 만듭니다.
5. 매개화된 미분가능함 함수 $g_\theta : \mathbb{R}^m \to \mathbb{R}$을 사용하여, $S_{\theta,\omega}(\mathcal{D}) := g_\theta(\bar{\Lambda}_\omega)$를 계산합니다.



Persistence Landscape으로 위상학적 층(topological layer) 만들기



위상학적 자료 분석(Topological Data Analysis) 소개

호몰로지(homology)를 통계적으로 추정하기
- 군집 나무(Cluster Tree)와 이를 통계적으로 추정하기

Persistent Homology를 통계적으로 추정하기

위상학적 자료 분석(Topological Data Analysis)을 기계학습에 응용
- 위상학적 자료 분석을 이용하여 특성(Feature) 만들기

참조문헌

참조문헌 I

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.

참조문헌 II

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, 2014. 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.

Jisu KIM, Yen-Chi Chen, Sivaraman Balakrishnan, Alessandro Rinaldo, and Larry Wasserman. Statistical inference for cluster trees. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1839–1847. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6508-statistical-inference-for-cluster-trees.pdf.

참조문헌 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.

Larry Wasserman. Topological data analysis, 2016.

감사합니다!

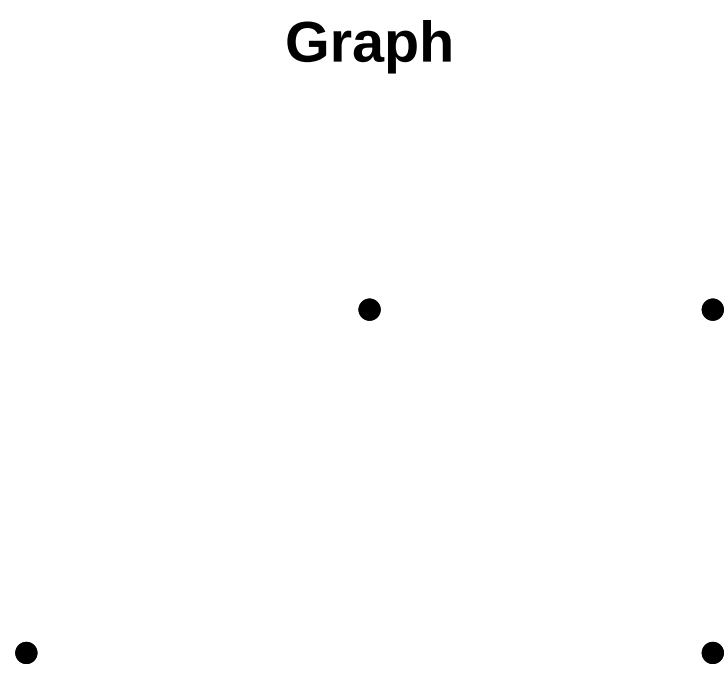
호몰로지(Homology)와 Persistent Homology

Persistent Homology를 통계적으로 추정하기

Persistent Homology를 이용하여 특성(Feature) 만들기

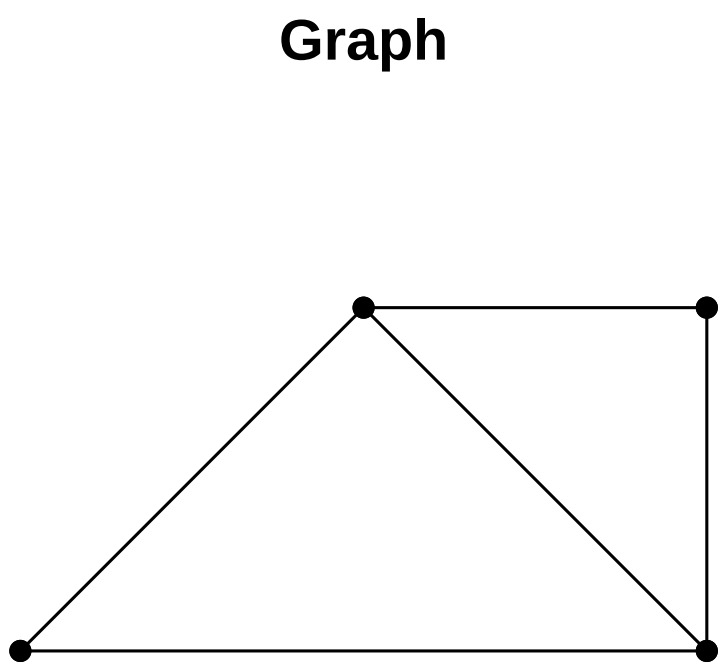
그래프(graph)는 꼭지점(vertex)과 변(edge)로 이루어진 이산 구조입니다.

- 주어진 거리공간 $\mathbb{X}$의 부분집합 $\mathcal{X} \subset \mathbb{X}$ 에 대해, 그래프(graph) $G = (\mathcal{X}, E)$는 꼭지점(vertex) 집합 $\mathcal{X}$와 변(edge)의 집합 E로 이루어져 있으면서 $E \subset \{\{x, y\} | \, x, y \in \mathcal{X}, x \neq y\}$를 만족합니다.



그래프(graph)는 꼭지점(vertex)과 변(edge)로 이루어진 이산 구조입니다.

- 주어진 거리공간 $\mathbb{X}$의 부분집합 $\mathcal{X} \subset \mathbb{X}$ 에 대해, 그래프(graph) $G = (\mathcal{X}, E)$는 꼭지점(vertex) 집합 $\mathcal{X}$와 변(edge)의 집합 E로 이루어져 있으면서 $E \subset \{\{x, y\} | \, x, y \in \mathcal{X}, x \neq y\}$를 만족합니다.



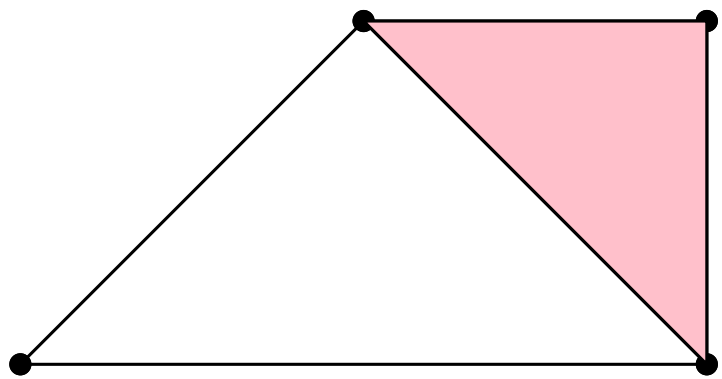
단체 복합체(Simplicial complex)는 고차원으로 일반화한 그래프입니다.

- 주어진 거리공간 $\mathbb{X}$의 부분집합 $\mathcal{X} \subset \mathbb{X}$ 에 대해, 단체 복합체 (Simplicial complex) K는 $\mathcal{X}$의 유한집합들의 집합이면서 다음을 만족합니다:

$$\alpha \in K,\ \beta \subset \alpha \Longrightarrow \beta \in K.$$

이 때, 각 단체 α의 차원은 $\dim \alpha := |\alpha| - 1$로 정의합니다.

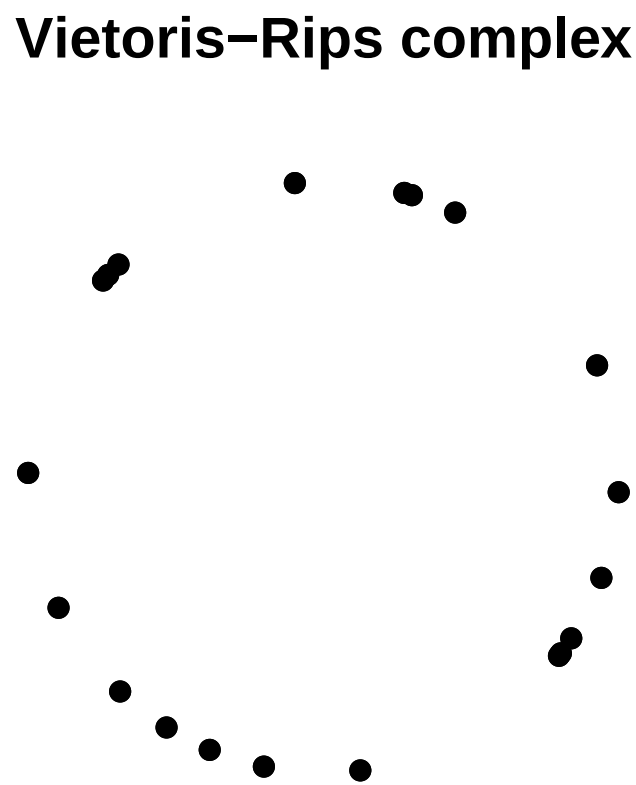
Simplicial complex



Vietoris-Rips 복합체(Vietoris-Rips complex)는 서로 가까운 꼭지점들을 모아 놓은 복합체입니다.

- 주어진 거리공간 $\mathbb{X}$의 부분집합 $\mathcal{X} \subset \mathbb{X}$ 와 $r > 0$에 대해, Vietoris-Rips 복합체(Vietoris-Rips complex) $\mathrm{Rips}(\mathcal{X}, r)$는 다음과 같이 정의됩니다:

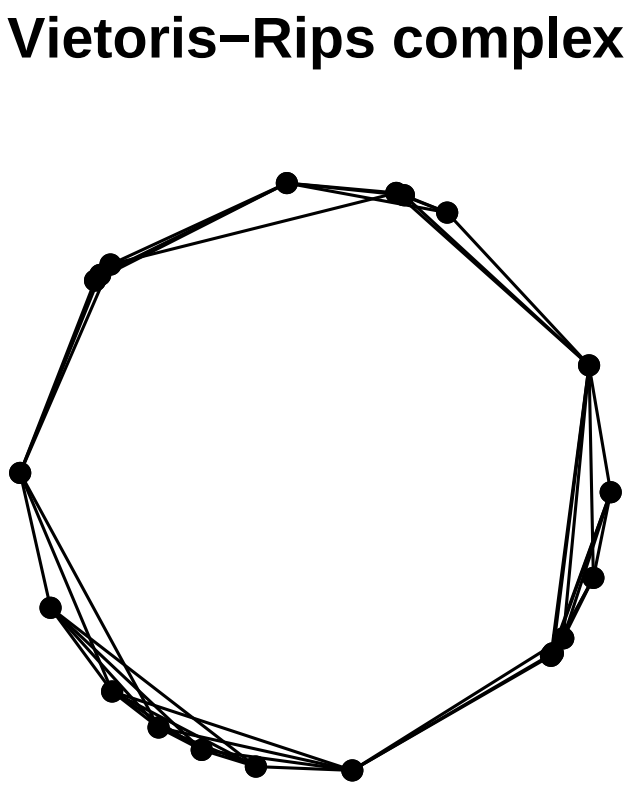
$$\mathrm{Rips}(\mathcal{X}, r) = \{\{x_1, \ldots, x_k\} \subset \mathcal{X} : \, d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k\}.$$



Vietoris-Rips 복합체(Vietoris-Rips complex)는 서로 가까운 꼭지점들을 모아 놓은 복합체입니다.

- 주어진 거리공간 $\mathbb{X}$의 부분집합 $\mathcal{X} \subset \mathbb{X}$ 와 $r > 0$에 대해, Vietoris-Rips 복합체(Vietoris-Rips complex) $\mathrm{Rips}(\mathcal{X}, r)$는 다음과 같이 정의됩니다:

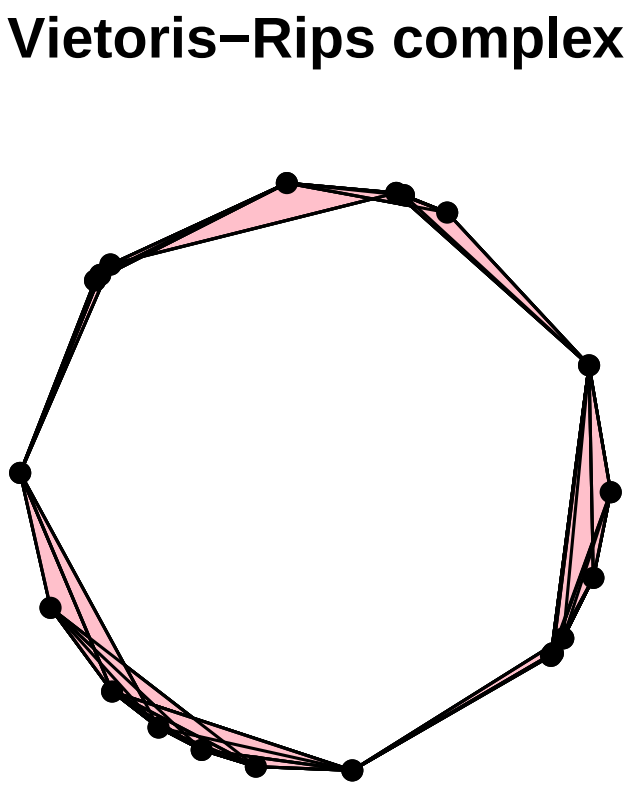
$$\mathrm{Rips}(\mathcal{X}, r) = \{\{x_1, \ldots, x_k\} \subset \mathcal{X} : \, d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k\}.$$



Vietoris-Rips 복합체(Vietoris-Rips complex)는 서로 가까운 꼭지점들을 모아 놓은 복합체입니다.

- 주어진 거리공간 $\mathbb{X}$의 부분집합 $\mathcal{X} \subset \mathbb{X}$ 와 $r > 0$에 대해, Vietoris-Rips 복합체(Vietoris-Rips complex) $\mathrm{Rips}(\mathcal{X}, r)$는 다음과 같이 정의됩니다:

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



단체 복합체의 k-연쇄(k-chain)는 단체들로 생성된 선형 공간입니다.

- 주어진 단체 복합체 K와 차원 $k \geq 0$에 대해, K의 k-연쇄(k-chain)는 K의 k-차원 단체들의 형식적 합(formal sum)입니다:

$$c = \sum_{i=1}^{p} a_i \sigma_i,\ \sigma_i \in K,\ a_i \in \mathbb{Z}/2\mathbb{Z} = \{0, 1\}.$$

- $\mathbb{Z}/2\mathbb{Z}$의 연산: $0 + 0 = 1 + 1 = 0$, $0 + 1 = 1 + 0 = 1$, $0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0$, $1 \cdot 1 = 1$.
- k-연쇄의 합과 스칼라곱:

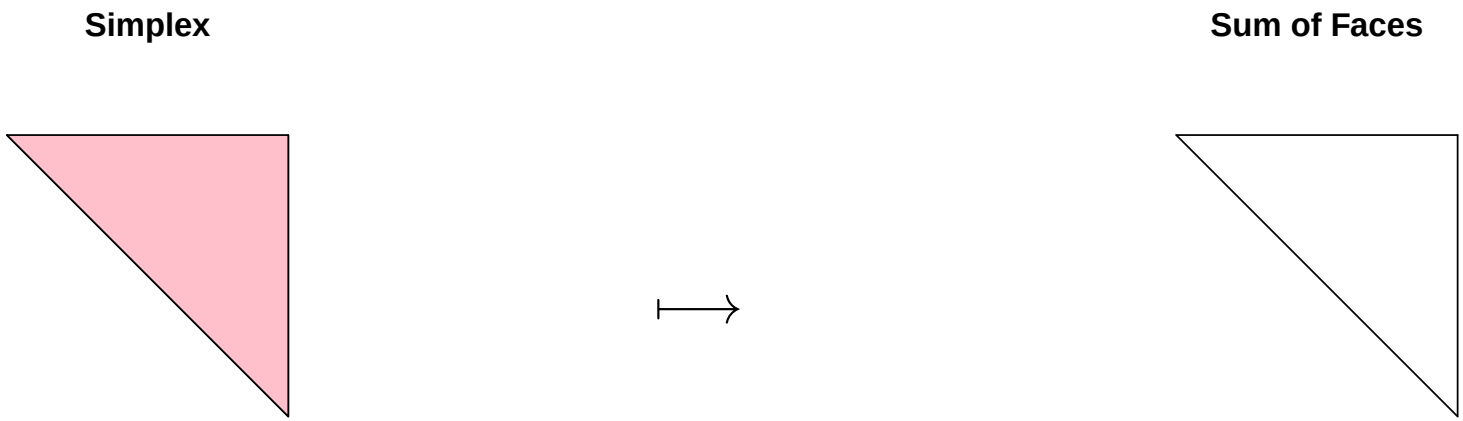
$$c + c' = \sum_{i=1}^{p} (a_i + a_i')\sigma_i, \qquad \lambda \cdot c = \sum_{i=1}^{p} (\lambda a_i)\sigma_i.$$

- K의 k-연쇄를 모은 집합 $C_k(K)$는 선형 공간이 됩니다.

경계 사상(boundary map)은 단체 복합체의 k-연쇄(k-chain) 간의 사상입니다.

- 경계 사상(boundary map) ∂_k은 각 k-차원 단체 σ를 그의 k-1차원 면(face)들의 합으로 보냅니다:

$$\sigma = \{v_0, \ldots, v_k\} \longmapsto \partial_k \sigma = \sum_{i=0}^{k} \{v_0, \ldots, v_k\} \backslash \{v_i\}.$$



경계 사상(boundary map)은 단체 복합체의 k-연쇄 (k-chain) 간의 사상입니다.

- 경계 사상(boundary map) ∂_k은 각 k-차원 단체 σ를 그의 k-1차원 면(face)들의 합으로 보냅니다:

$$\sigma = \{v_0, \ldots, v_k\} \longmapsto \partial_k \sigma = \sum_{i=0}^{k} \{v_0, \ldots, v_k\} \backslash \{v_i\}.$$

- 경계 사상을 $\partial_k : C_k(K) \to C_{k-1}(K)$로 자연스럽게 확장합니다:

$$\begin{array}{ccc} C_k(K) & \longmapsto & C_{k-1}(K) \\ c = \sum a_i \sigma_i & \longmapsto & \partial_k c = \sum a_i \partial_k \sigma_i \end{array}.$$

- $$\partial_k \circ \partial_{k+1} = 0.$$

호몰로지(homology)는 cycle을 boundary로 자른 몫공간(quotient space)입니다.

- K의 k-cycle $Z_k(K)$는 경계 사상에 의해 0으로 가는 k-연쇄의 집합입니다:

$$Z_k(K) := \ker \partial_k = \{c \in C_k : \partial_k c = 0\}.$$

- K의 k-boundary $B_k(K)$는 경계 사상에 의한 $k+1$-연쇄의 상(image)입니다:

$$B_k(K) := \mathrm{im} \partial_{k+1} = \{c \in C_k : \exists c' \in C_{k+1}, \partial_{k+1} c' = c\}.$$

- $\partial_k \circ \partial_{k+1} = 0$에 의해, k-boundary $B_k(K)$는 k-cycle $Z_k(K)$의 선형부분공간(linear subspace)입니다:

$$B_k(K) \subset Z_k(K) \subset C_k(K).$$

- k-th 호몰로지 $H_k(K)$는 k-cycle $Z_k(K)$를 k-boundary $B_k(K)$로 자른 몫공간(quotient space)입니다:

$$H_k(K) := Z_k(K)/B_k(K).$$

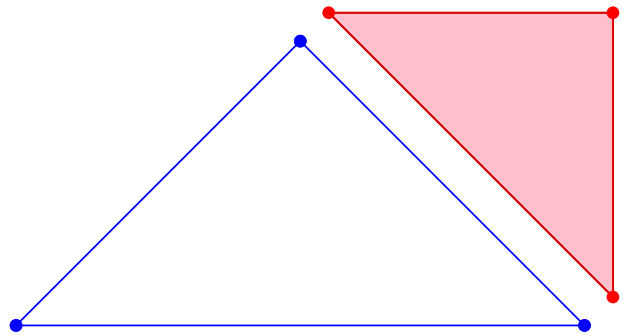
호몰로지(homology)는 cycle을 boundary로 자른 몫공간(quotient space)입니다.

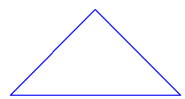
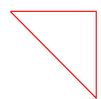
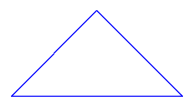
- K의 k-th 호몰로지 $H_k(K)$는 k-cycle $Z_k(K)$를 k-boundary $B_k(K)$로 자른 몫공간(quotient space)입니다:

$$H_k(K) := Z_k(K)/B_k(K).$$

- K의 k-th Betti number $\beta_k(K)$는 선형공간 $H_k(K)$의 랭크입니다: $\beta_k(K) = \mathrm{rank}(H_k(K))$.

호몰로지(homology)는 cycle을 boundary로 자른 몫공간(quotient space)입니다.

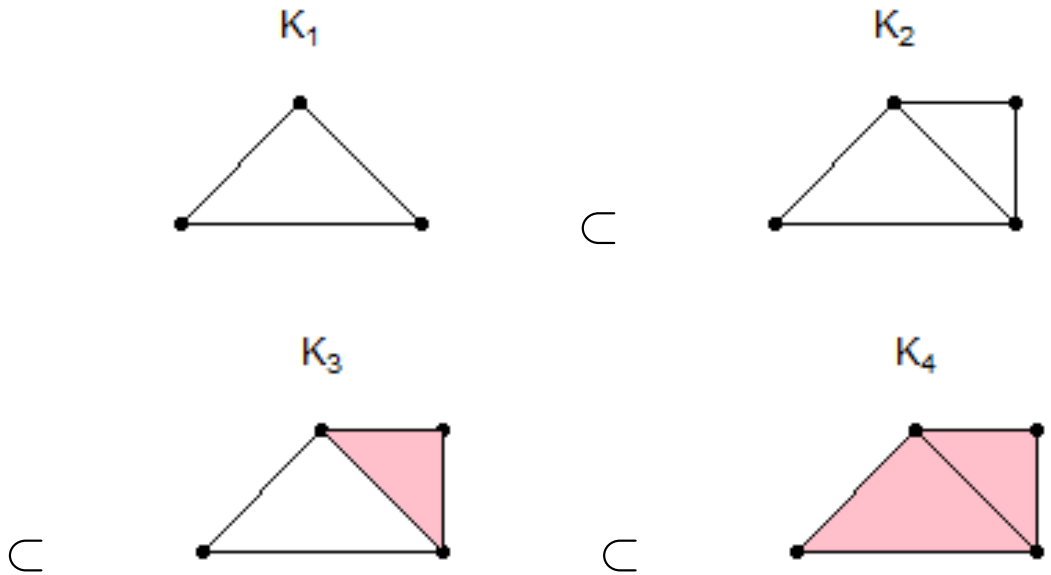


- $Z_1(K) = \ker \partial_1 = (\mathbb{Z}/2\mathbb{Z})^2 =<$  ,  $>$
- $B_1(K) = \mathrm{im}\partial_2 = \mathbb{Z}/2\mathbb{Z} =<$  $>$
- $H_1(K) = Z_1(K)/B_1(K) = \mathbb{Z}/2\mathbb{Z} =<$  $>$, $\beta_1(K) = 1$

filtration은 증가하는 단체 복합체들의 모임입니다.

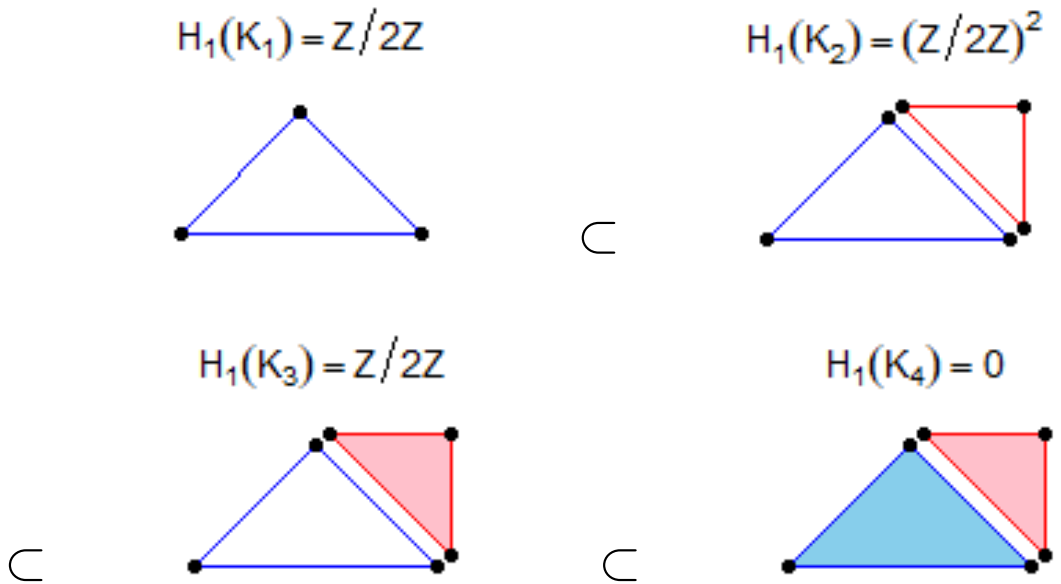
- 단체 복합체 K가 있을 때, filtration $\mathcal{F} = \{K_a\}_{a \in \mathbb{R}}$는 다음을 만족하는 K의 부분 복합체(subcomplex) K_a들의 모임입니다:

$$a \leq b \Longrightarrow K_a \subset K_b.$$



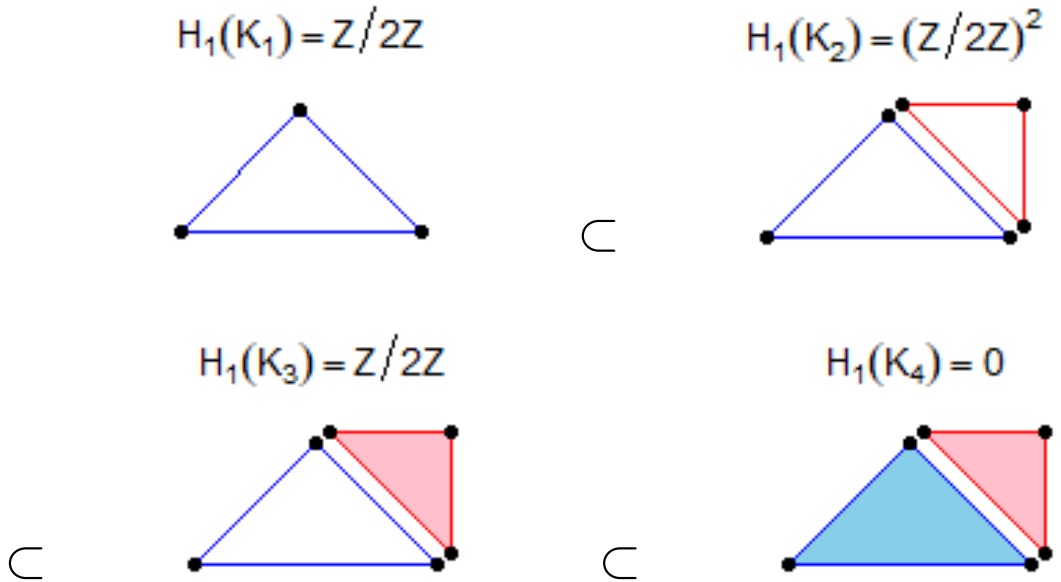
Persistent Homology는 filtration에서 호몰로지가 어떻게 변화하는지 추적합니다.

- 단체 복합체 K 위의 filtration $\mathcal{F} = \{K_a \subset K : a \in \mathbb{R}\}$가 있을 때, k-th persistent homology $\mathrm{PH}_k\mathcal{F}$는 호몰로지들 $\{H_k(K_a) : a \in \mathbb{R}\}$과 선형사상들 $\{\imath_k^{a,b} : a \leq b\}$의 모임인데, 이 때 선형사상 $\imath_k^{a,b}$는 포함관계 $K_a \subset K_b$로부터 유도됩니다.
- Persistence betti number 는 $\beta_k^{a,b} := \mathrm{rank}(\mathrm{im}\imath_k^{a,b})$ 입니다.



Persistent Homology는 filtration에서 호몰로지가 어떻게 변화하는지 추적합니다.

- 단체 복합체 K 위의 filtration $\mathcal{F} = \{K_a \subset K : a \in \mathbb{R}\}$가 있을 때, k-th persistent homology $\mathrm{PH}_k\mathcal{F}$는 호몰로지들 $\{H_k(K_a) : a \in \mathbb{R}\}$과 선형사상들 $\{i_k^{a,b} : a \leq b\}$의 모임인데, 이 때 선형사상 $i_k^{a,b}$는 포함관계 $K_a \subset K_b$로부터 유도됩니다.
- 각 homology class γ는 K_a에서 생기고 K_b에서 $\gamma = 0$이 됩니다. 이 때, a를 γ의 birth time이라 하고, b를 γ의 death time이라고 합니다.



Persistence Diagram 은 Persistent Homology 를 평면 위의 점들로 나타냅니다.

- 편의상 filtration $\mathcal{F} = \{K_a \subset K : a \in \mathbb{R}\}$의 부분단체 K_a들이 유한 번 바뀐다고 가정합니다:

$$K_{a_1} \subset \cdots \subset K_{a_n}.$$

- 각 filtration 값들의 쌍 (a_i, a_j)에 대해, K_{a_i}에서 생기고 K_{a_j}에서 없어지는 homology class 의 개수를 셉니다:

$$\mu_k^{a_i,a_j} = (\beta_k^{a_i,a_{j-1}} - \beta_k^{a_i,a_j}) - (\beta_k^{a_{i-1},a_{j-1}} - \beta_k^{a_{i-1},a_j}).$$

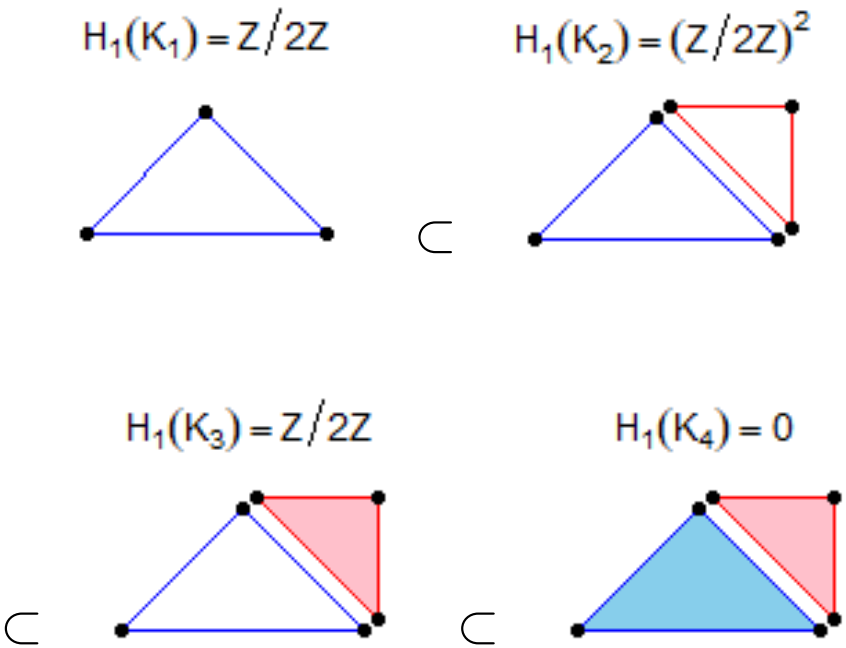
- $(\mathbb{R} \cup \{\infty\})^2$ 위에 점 (a_i, a_j) 를 multiplicity $\mu_k^{a_i,a_j}$로 찍으면 persistence diagram 이 됩니다.

Persistence Diagram 은 Persistent Homology 를 평면 위의 점들로 나타냅니다.

- 각 filtration 값들의 쌍 (a_i, a_j)에 대해, K_{a_i}에서 생기고 K_{a_j}에서 없어지는 homology class 의 개수를 셉니다:

$$\mu_k^{a_i,a_j} = (\beta_k^{a_i,a_{j-1}} - \beta_k^{a_i,a_j}) - (\beta_k^{a_{i-1},a_{j-1}} - \beta_k^{a_{i-1},a_j}).$$

- $(\mathbb{R} \cup \{\infty\})^2$ 위에 점 (a_i, a_j) 를 multiplicity $\mu_k^{a_i,a_j}$로 찍으면 persistence diagram 이 됩니다.



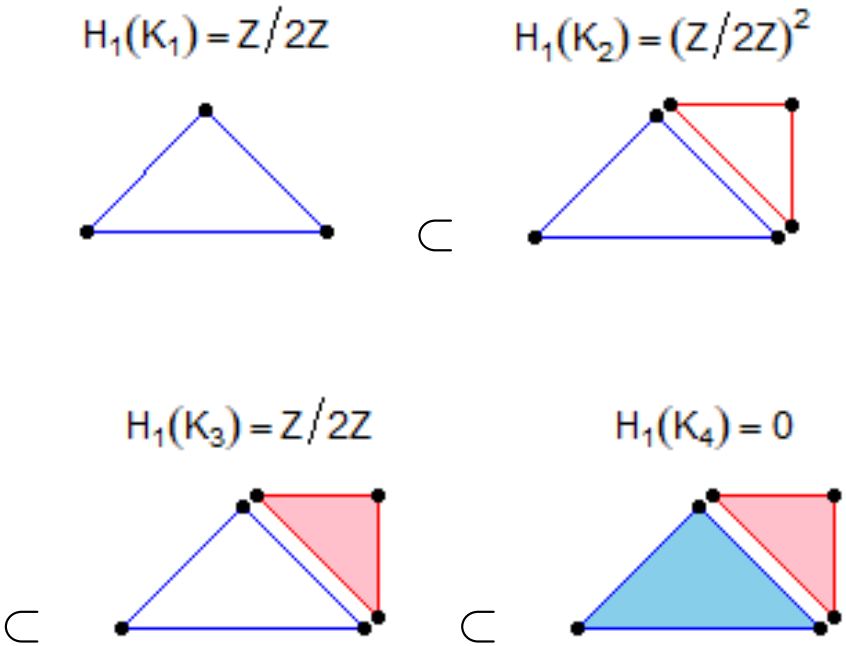
$\mu_1^{i,j}$	$i=1$	$i=2$	$i=3$	$i=4$
$j=4$	0	0	0	0
$j=3$	1	1	1	
$j=2$	1	2		
$j=1$	1			

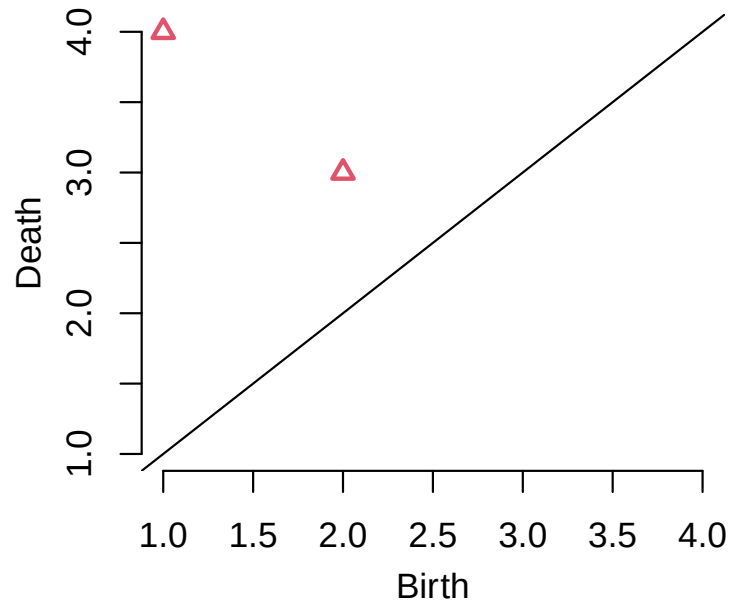
Persistence Diagram 은 Persistent Homology 를 평면 위의 점들로 나타냅니다.

- 각 filtration 값들의 쌍 (a_i, a_j)에 대해, K_{a_i}에서 생기고 K_{a_j}에서 없어지는 homology class 의 개수를 셉니다:

$$\mu_k^{a_i,a_j} = (\beta_k^{a_i,a_{j-1}} - \beta_k^{a_i,a_j}) - (\beta_k^{a_{i-1},a_{j-1}} - \beta_k^{a_{i-1},a_j}).$$

- $(\mathbb{R} \cup \{\infty\})^2$ 위에 점 (a_i, a_j) 를 multiplicity $\mu_k^{a_i,a_j}$로 찍으면 persistence diagram 이 됩니다.





호몰로지(Homology)와 Persistent Homology

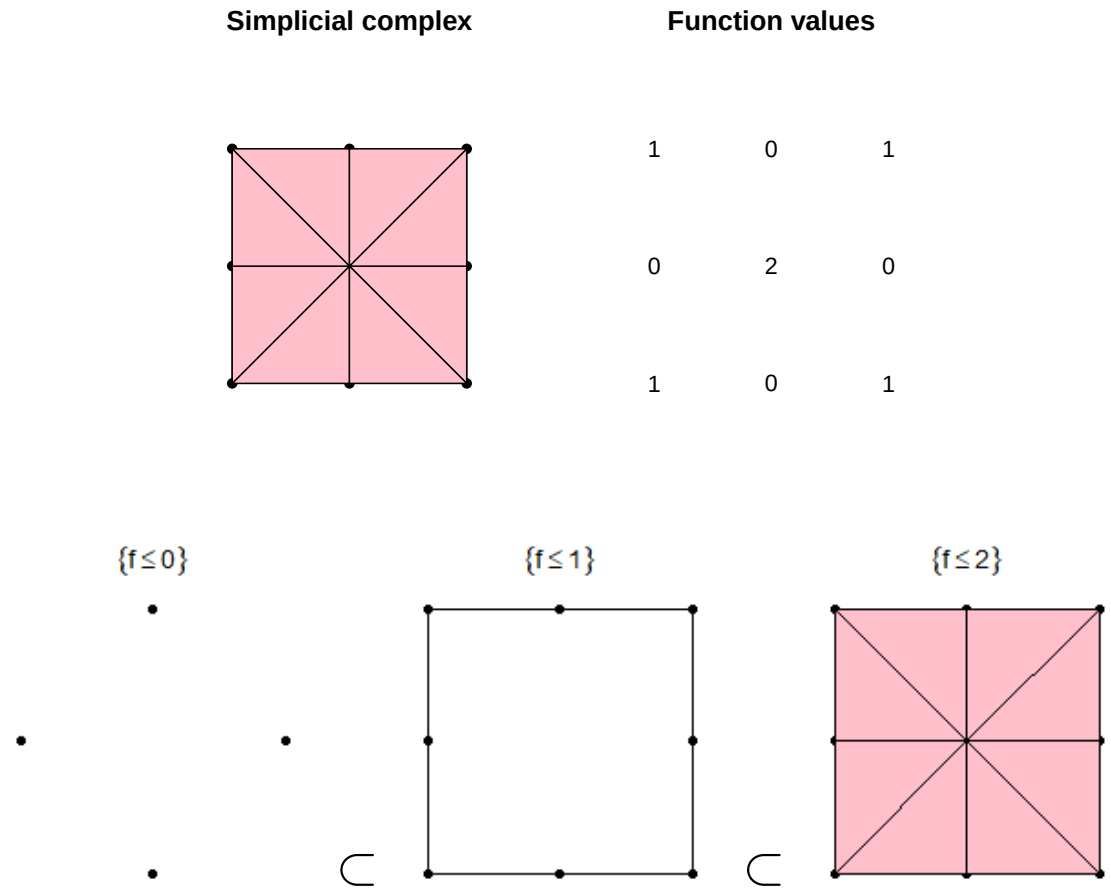
Persistent Homology를 통계적으로 추정하기

Persistent Homology를 이용하여 특성(Feature) 만들기

함수의 레벨집합으로부터 filtration을 만들 수 있습니다.

- 단체 복합체 K와 그 위에서 정의된 함수 $f : K \to \mathbb{R}$가 있을 때, f의 sub-level filtration $\mathrm{sub}(f)$를 다음과 같이 정의합니다:

$$\mathrm{sub}(f) := \{\{\sigma \in K : f(\sigma) \leq L\}\}_{L \in \mathbb{R}} .$$



함수의 레벨집합으로부터 filtration을 만들 수 있습니다.

- 단체 복합체 K와 그 위에서 정의된 함수 $f : K \to \mathbb{R}$가 있을 때, f의 아랫레벨(sub-level) filtration $\mathrm{sub}(f)$를 다음과 같이 정의합니다:

$$\mathrm{sub}(f) := \{\{\sigma \in K : f(\sigma) \leq L\}\}_{L \in \mathbb{R}} .$$

- 마찬가지로 f의 윗레벨(super-level) filtration $\mathrm{super}(f)$를 다음과 같이 정의합니다:

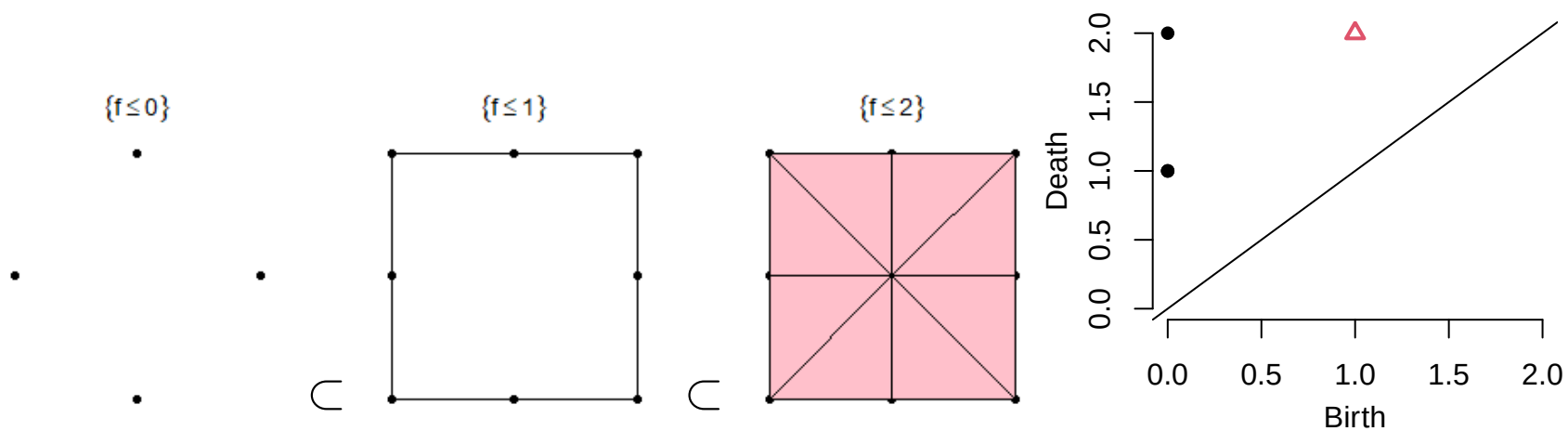
$$\mathrm{super}(f) := \{\{\sigma \in K : f(\sigma) \geq L\}\}_{L \in \mathbb{R}} .$$

함수의 레벨집합으로부터 persistent homology를 계산할 수 있습니다.

- 단체 복합체 K와 그 위에서 정의된 함수 $f : K \to \mathbb{R}$가 있을 때, f의 아랫레벨(sub-level) filtration $\mathrm{sub}(f)$를 다음과 같이 정의합니다:

$$\mathrm{sub}(f) := \{\{\sigma \in K : f(\sigma) \leq L\}\}_{L \in \mathbb{R}} .$$

- 그로부터 계산한 persistent homology 또는 persistence diagram을 $Dgm(f)$로 씁니다.



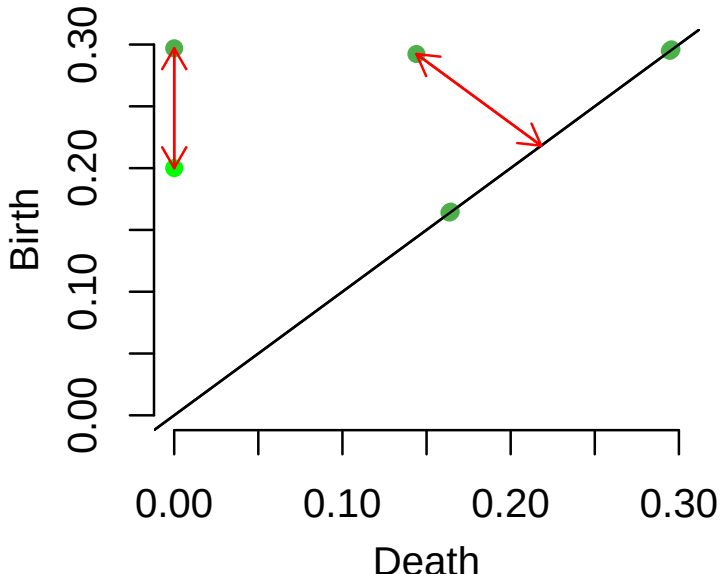
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

Definition

D_1, D_2를 두 Persistent homology라고 하면, Bottleneck distance는 다음과 같이 정의됩니다:

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

이 때, γ는 D_1에서 D_2로 가는 모든 일대일대응이 될 수 있습니다.



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

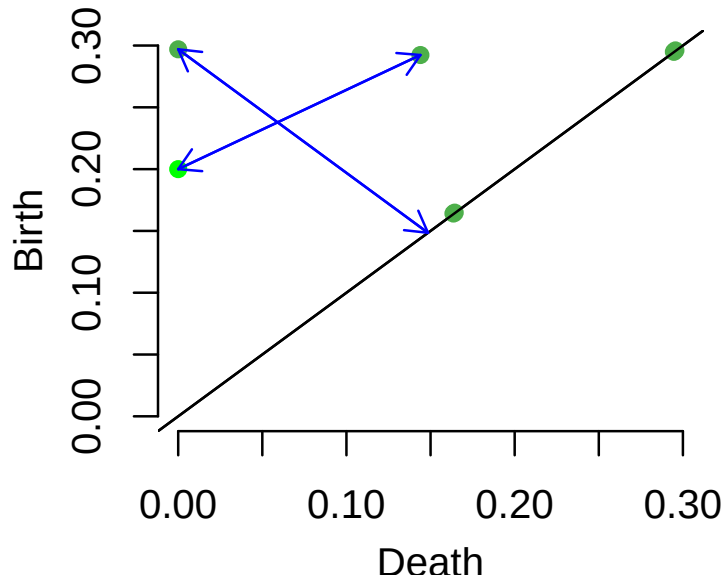
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

Definition

D_1, D_2를 두 Persistent homology라고 하면, Bottleneck distance는 다음과 같이 정의됩니다:

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

이 때, γ는 D_1에서 D_2로 가는 모든 일대일대응이 될 수 있습니다.



$$\sup_{x \in D_1} \|x - \gamma_2(x)\|_\infty = 0.15$$

Bottleneck distance는 그에 상응하는 함수간의 거리로 조정할 수 있습니다: 안정성 정리

Theorem

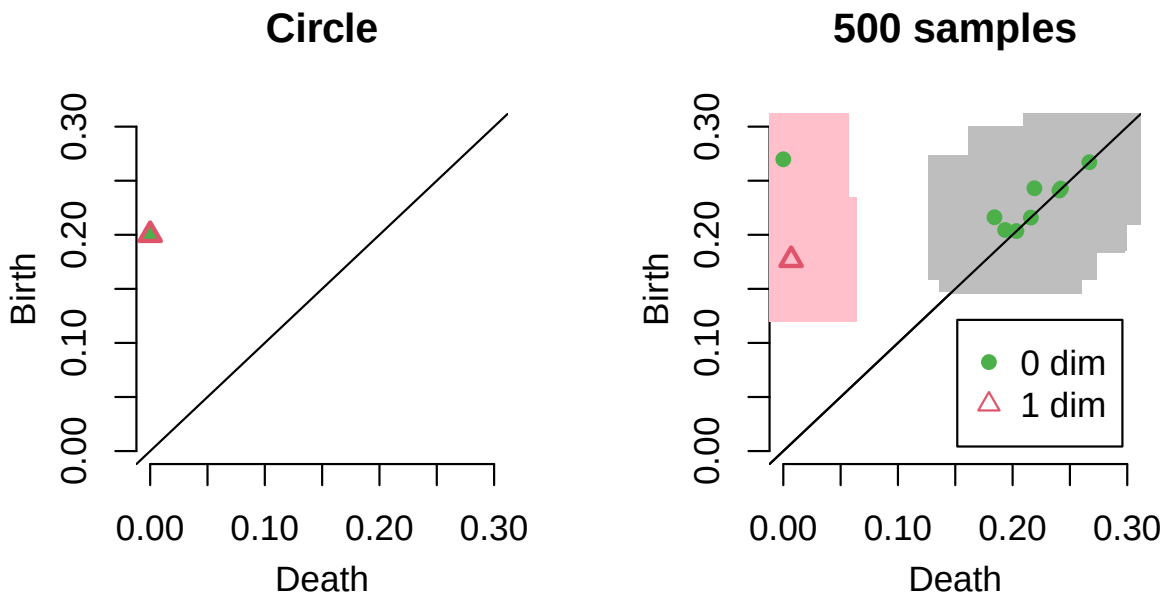
[Edelsbrunner and Harer, 2010][Chazal, de Silva, Glisse, and Oudot, 2012] K를 단체 복합체(simplicial complex)라 하고 $f, g : K \to \mathbb{R}$를 두 함수라 합니다. $Dgm(f)$와 $Dgm(g)$를 그에 상응하는 persistent homology라고 할 때, 다음이 성립합니다:

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

Persistent homology의 신뢰집합은 Persistent homology를 높을 확률로 포함하는 랜덤집합입니다.

기저 M과 자료 X의 Persistent homology를 각각 $Dgm(M)$과 $Dgm(X)$라고 놓습니다. 유의수준 $\alpha \in (0,1)$가 주어졌을 때, $(1-\alpha)$ 신뢰집합 $\{D \in Dgm : \ W_\infty(Dgm(X), D) \leq c_n\}$은 다음을 만족하는 랜덤집합입니다:

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



Persistent homology의 신뢰띠는 그에 상응하는 함수의 신뢰띠로 계산할 수 있습니다.

안정성 정리로부터, $\mathbb{P}(||f_M - f_X|| \leq c_n) \geq 1 - \alpha$는 다음을 유도합니다:

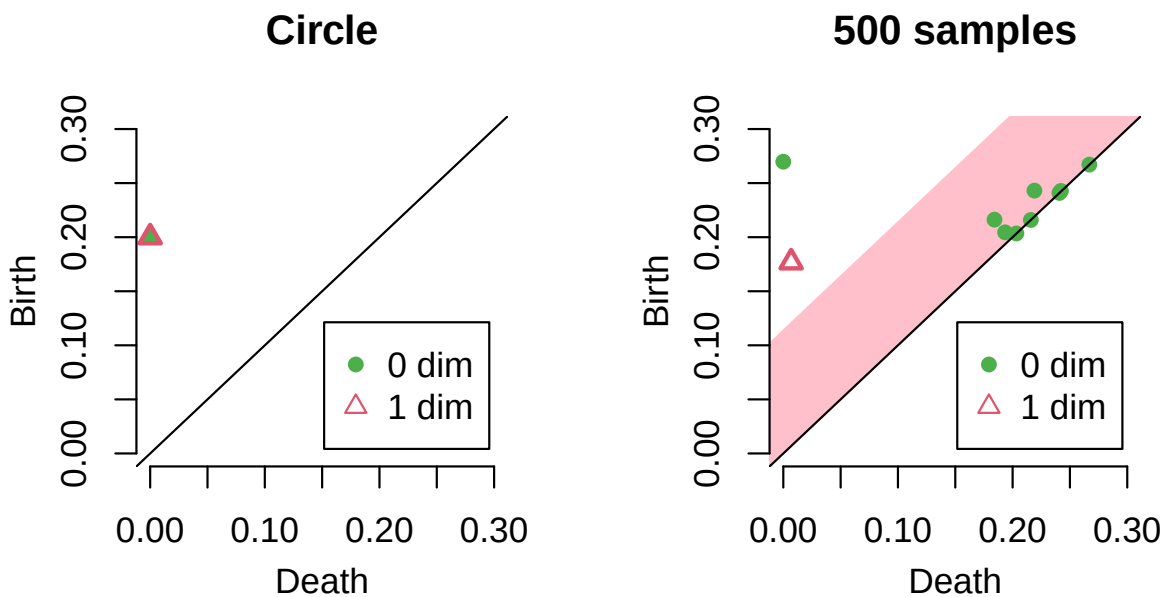
$$\mathbb{P}(W_\infty(Dgm(f_M), Dgm(f_X)) \leq c_n) \geq \mathbb{P}(||f_M - f_X||_\infty \leq c_n) \geq 1 - \alpha,$$

따라서 f_M의 신뢰띠를 persistent homology $Dgm(f_M)$의 신뢰띠로 이용할 수 있습니다.

Persistent homology의 신뢰띠는 붓스트랩으로 계산할 수 있습니다.

붓스트랩 알고리즘을 persistent homology에 적용할 수 있다는 것이 증명되었습니다.

- Fasy et al. [2014] 이 핵밀도추정(kernel density estimator)에서 보였고,
- Chazal et al. [2014] 이 distance to measure와 kernel distance에서 보였습니다.

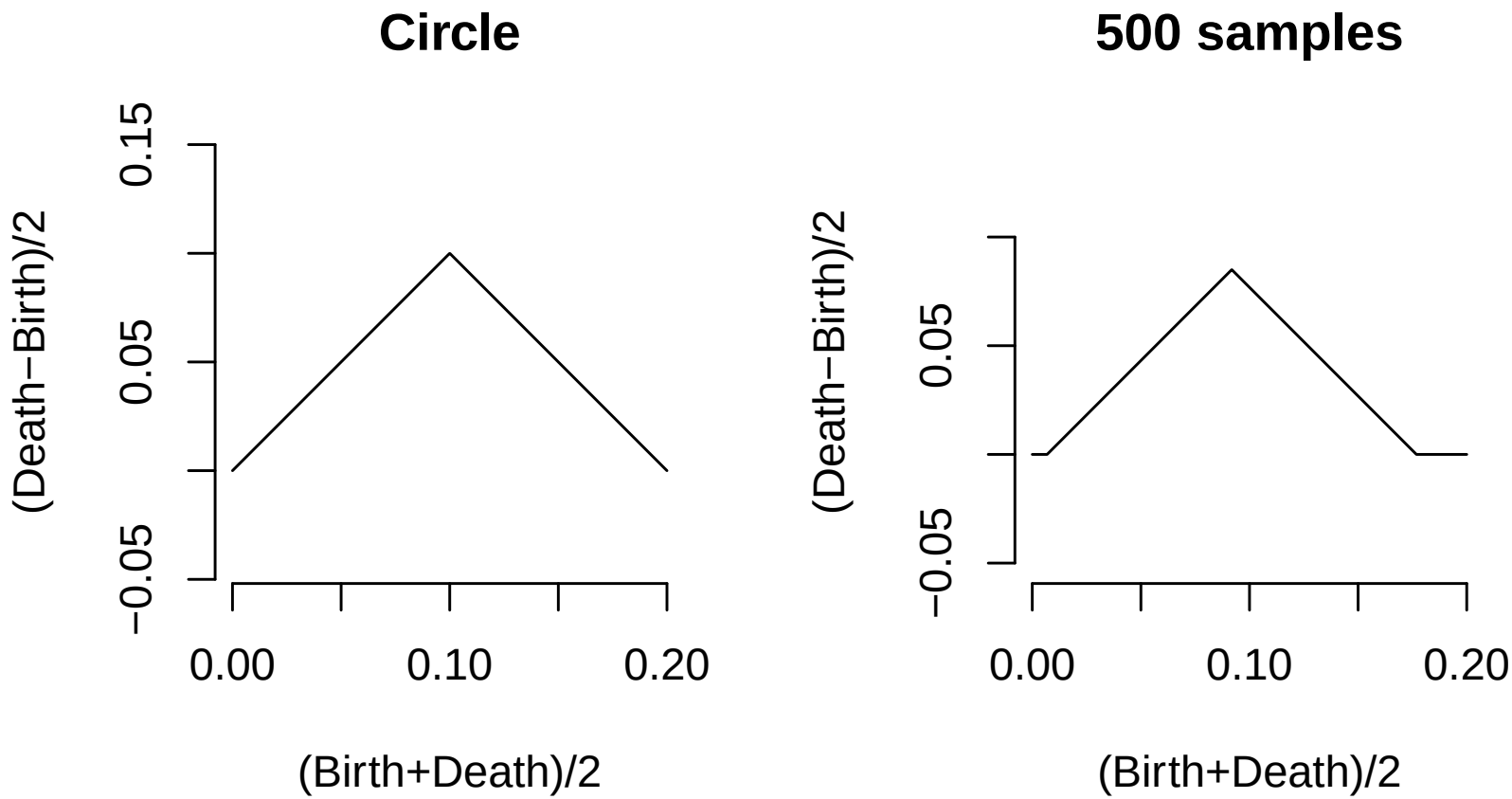


호몰로지(Homology)와 Persistent Homology

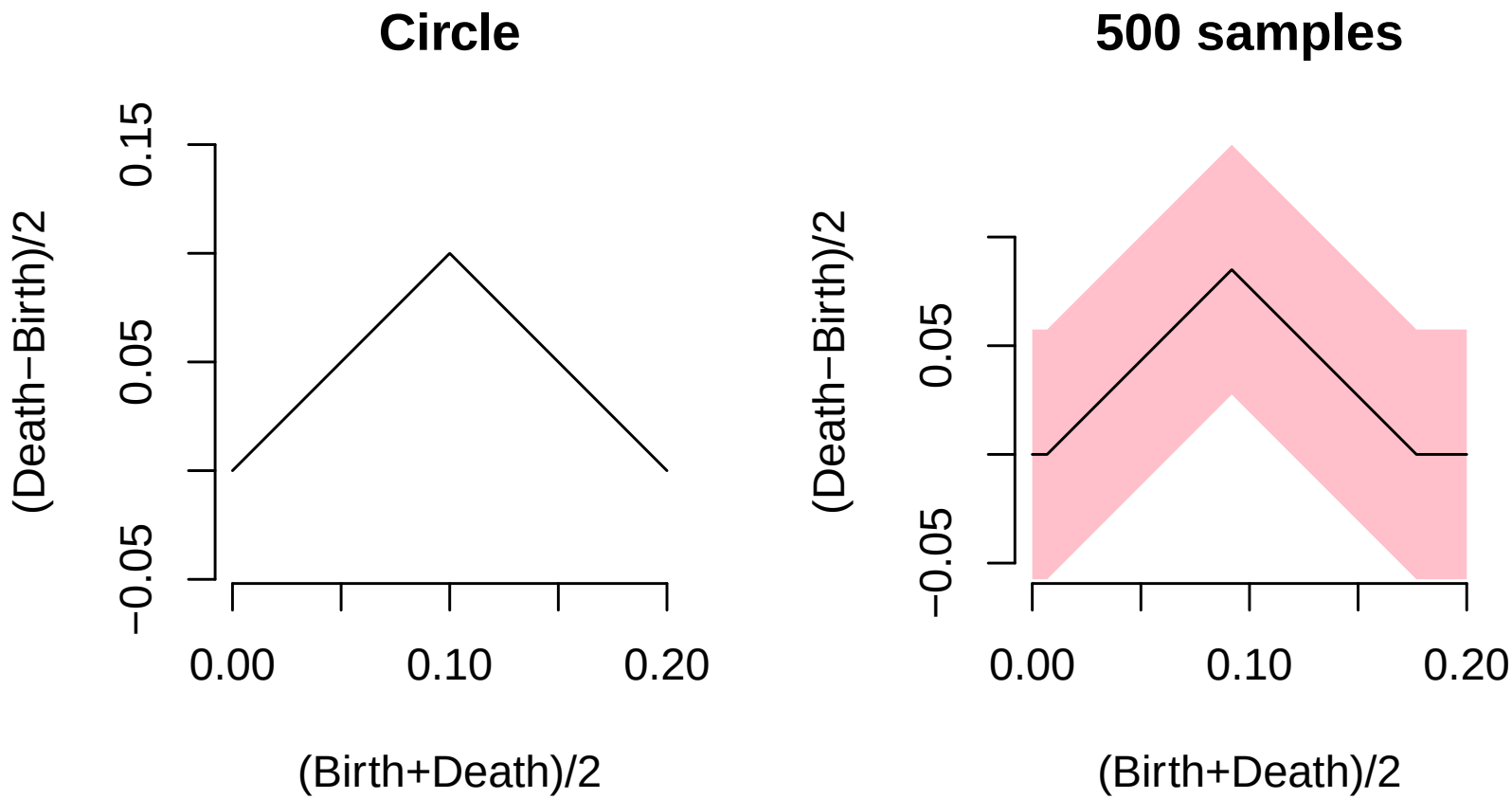
Persistent Homology를 통계적으로 추정하기

Persistent Homology를 이용하여 특성(Feature) 만들기

유한한 자료의 Persistence Landscape으로부터 기저 구조의 Persistence Landscape을 추정할 수 있습니다.



Persistent homology의 신뢰띠로 Persistence Landscape의 랜덤성을 정량화할 수 있습니다.

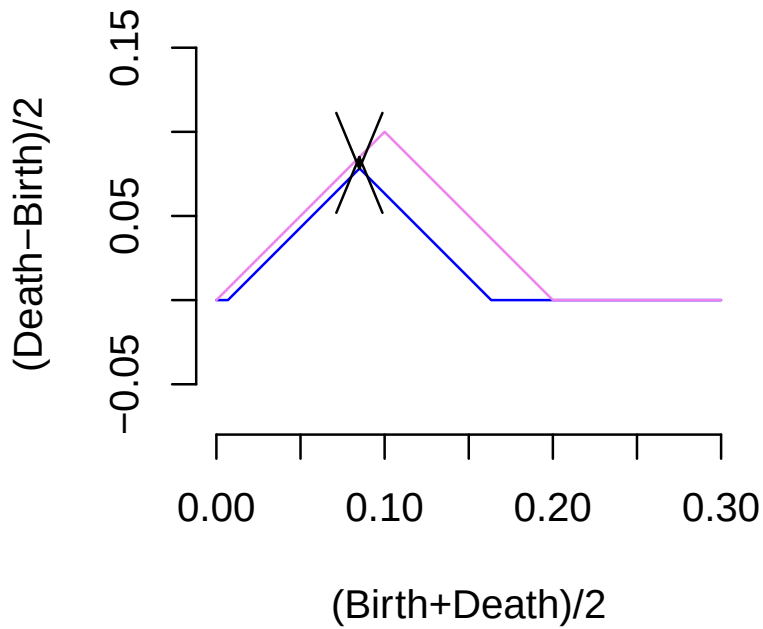


∞-landscape 거리는 persistence landscape 공간에 거리를 줍니다.

Definition

[?] D_1, D_2을 점들의 multiset이라 하고, 그에 해당하는 persistence landscape를 λ_1 , λ_2라고 놓습니다. ∞-landscape 거리는 다음과 같이 정의합니다:

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



∞-landscape 거리는 그에 대응되는 함수 간의 거리로 조정할 수 있습니다: 안정성 정리(stability theorem).

Theorem

$f, g : \mathbb{X} \to \mathbb{R}$를 두 함수로 놓고, 그에 해당하는 persistence landscape를 $\lambda(f)$과 $\lambda(g)$로 놓습니다. 그러면,

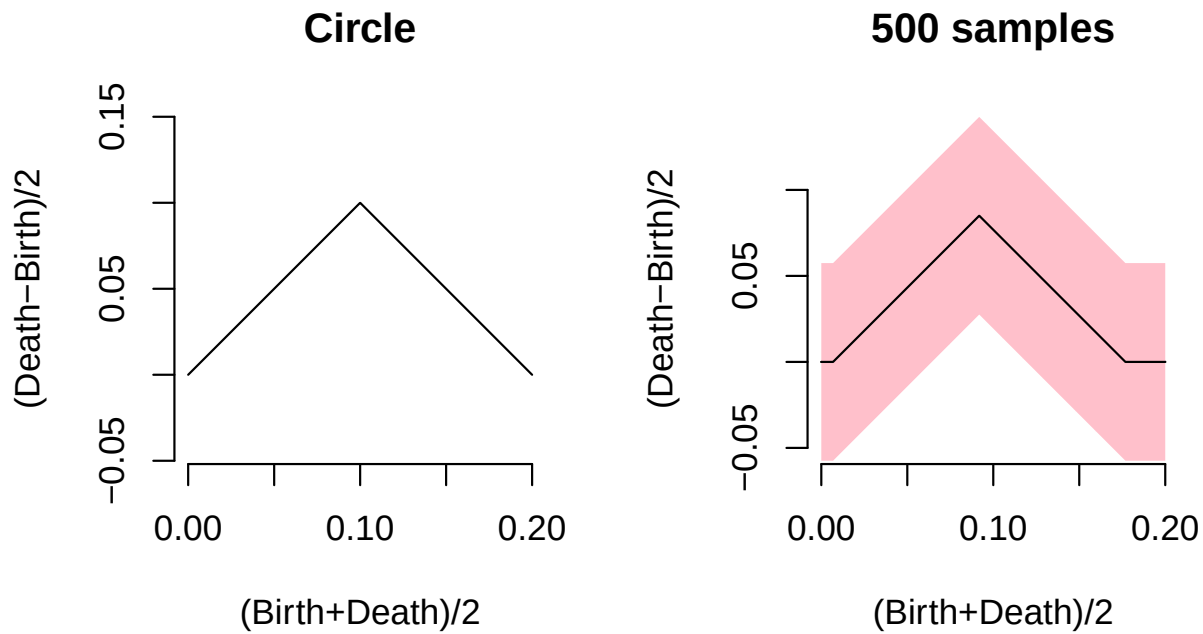
$$\Lambda_\infty(\lambda(f), \lambda(g)) \leq \|f - g\|_\infty.$$

persistence landscape의 신뢰띠는 붓스트랩으로 계산할 수 있습니다.

- 기저 M과 표본 X의 persistence landscape를 각각 λ_M과 λ_X로 놓습니다. 안정성 정리(stability theorem)으로부터, $\mathbb{P}(||f_M - f_X|| \leq c_n) \geq 1 - \alpha$는 다음을 유도합니다:

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

따라서 대응되는 함수인 f_M의 신뢰띠를 persistence landscape λ_M의 신뢰띠로 사용할 수 있습니다.



persistence landscape의 신뢰띠는 붓스트랩으로 계산할 수 있습니다.

- persistence landscape의 신뢰띠는 multiplier bootstrap으로도 계산할 수 있습니다; [Chazal, Fasy, Lecci, Michel, Rinaldo, and Wasserman, 2014].