위상 자료 분석(Topological Data Analysis)의 통계적 추정 및 기계 학습에의 응용

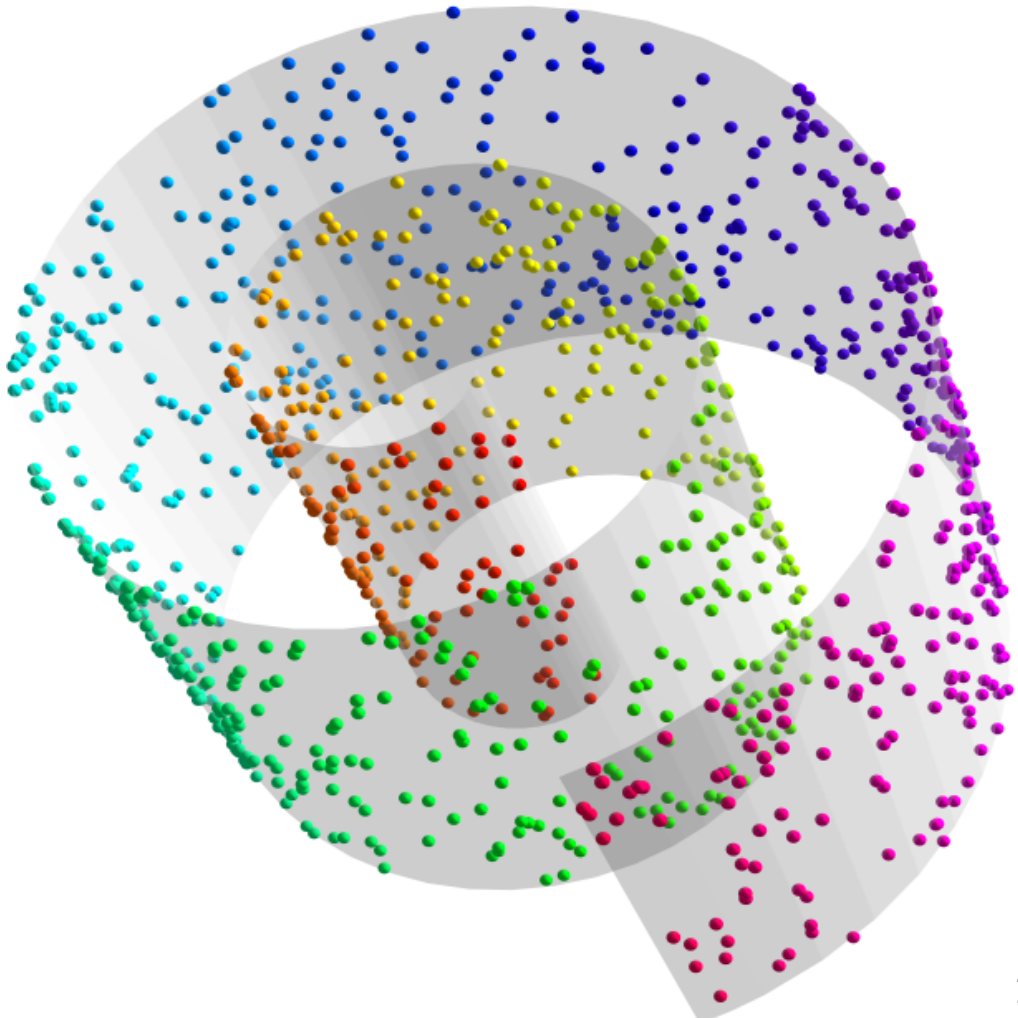
김지수(Jisu Kim)

한양대 ERICA

2025-05-16

위상 자료 분석(Topological Data Analysis) 소개

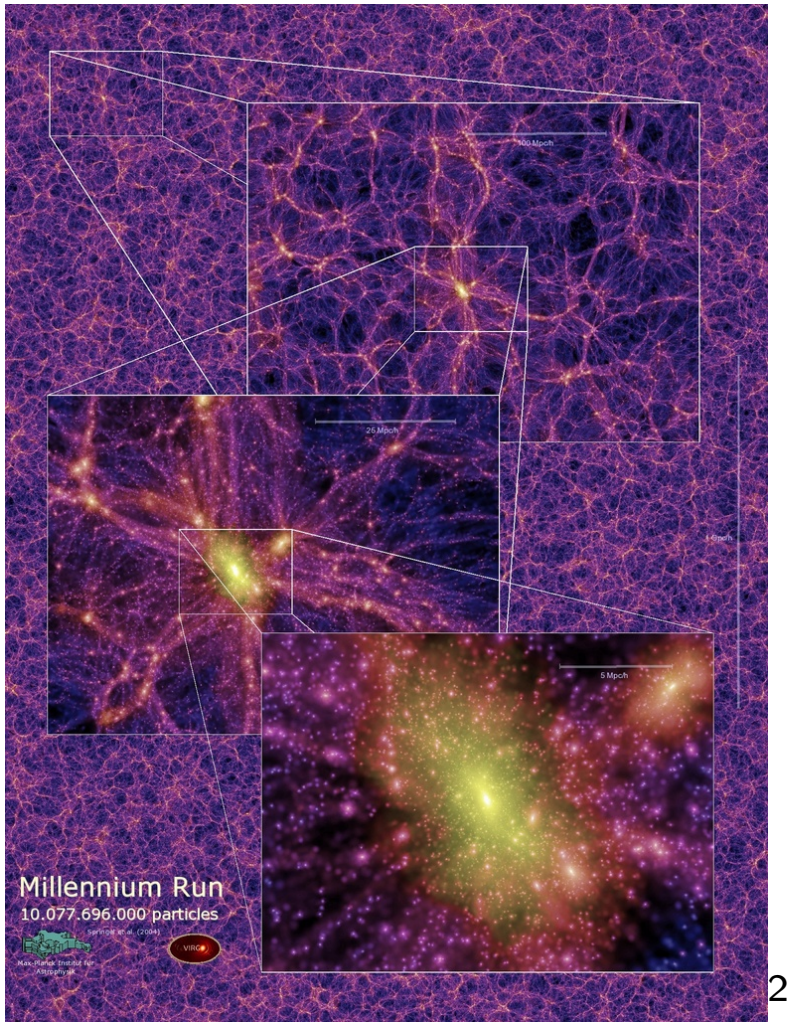
저차원 기하 구조를 가정함으로써 고차원 자료의 차원의 저주(curse of dimensionality)를 경감시킬 수 있습니다.



1

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

자료의 위상학적 구조로부터 정보를 얻을 수 있습니다.



[2]

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

위상 구조를 여러 해상도에서 바라봅니다.

위상 구조를 여러 해상도에서 바라봅니다.

위상 구조를 여러 해상도에서 바라봅니다.

위상 구조를 여러 해상도에서 바라봅니다.

- 조르주 쇠라 (Georges Seurat), 그랑드 자트 섬의 일요일 오후 (Un dimanche après-midi à l'Île de la Grande Jatte)

기하학적 및 위상학적 자료를 통계적으로 어떻게 추정하는지 알아봅니다.

- 다양체(manifold)의 기하학적 모수 추정의 미니맥스 위험(minimax risk)
  - Minimax Rates for Estimating the Dimension of a Manifold (Kim, Rinaldo, Wasserman, 2019)
  - The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory (Aamari, Kim, Chazal, Michel, Rinaldo, Wasserman, 2019)
- 위상 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)
- Persistent Homology를 통계적으로 추정하기
  - Confidence sets for persistence diagrams (Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan, Singh, 2014b)

위상 자료 분석(Topological Data Analysis)의 기계학습(Machine Learning)에의 응용을 소개합니다.

- 위상 자료 분석(Topological Data Analysis)을 기계학습(Machine Learning)에 응용
  - A Survey of Topological Machine Learning Methods (Hensel, Moor, Rieck, 2021)
- 위상 자료 분석을 이용하여 특성(Feature) 만들기
  - 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)
- 자료나 모형의 품질을 TDA로 평가
  - TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)
- R 패키지 TDA: 위상 자료 분석을 위한 통계 계산 도구
  - Introduction to the R package TDA (Fasy, Kim, Lecci, Maria, Millman, Rouvreau, 2014a)

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 호몰로지(Homology)에 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.

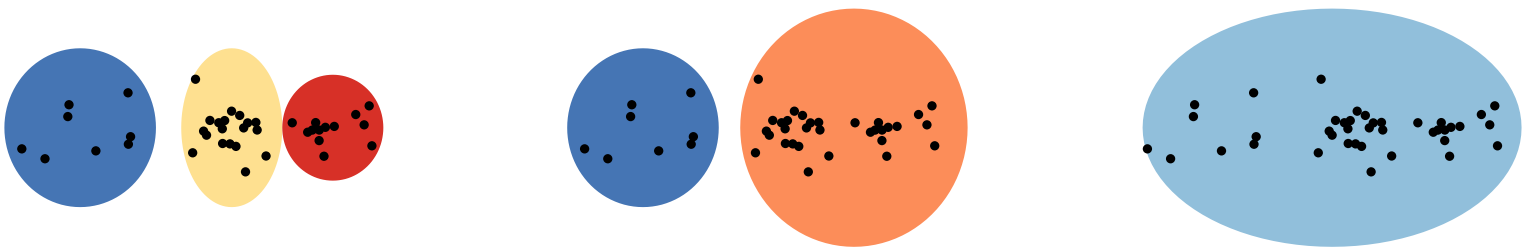
- Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)

자료를 군집으로 묶고자 합니다.

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

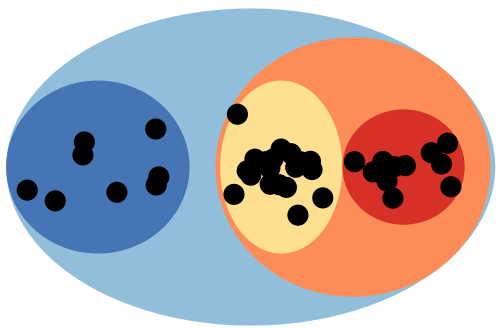
원하는 해상도에 따라 다른 군집이 생길 수 있습니다.

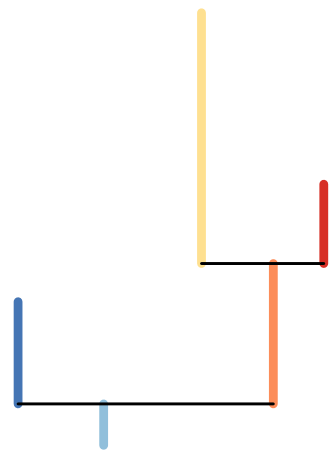
- Statistical Inference for Cluster Trees (Kim, Chen, Balakrishnan, Rinaldo, Wasserman, 2016)
- 국소적(local)이고 상세한 정보를 묘사하고 싶으면 (높은 해상도), 작은 규모의 많은 군집이 생깁니다.
- 대역적(global)이고 개략적인 정보를 묘사하고 싶으면 (낮은 해상도), 큰 규모의 적은 군집이 생깁니다.



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)

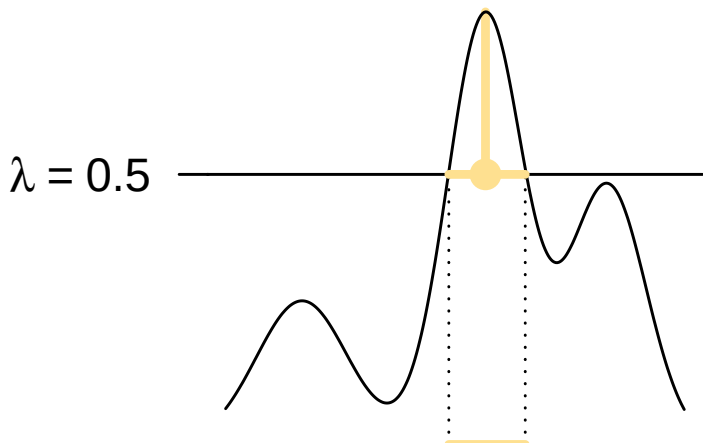




군집 나무(Cluster Tree)는 고밀도 군집들의 계층 구조입니다.

Definition

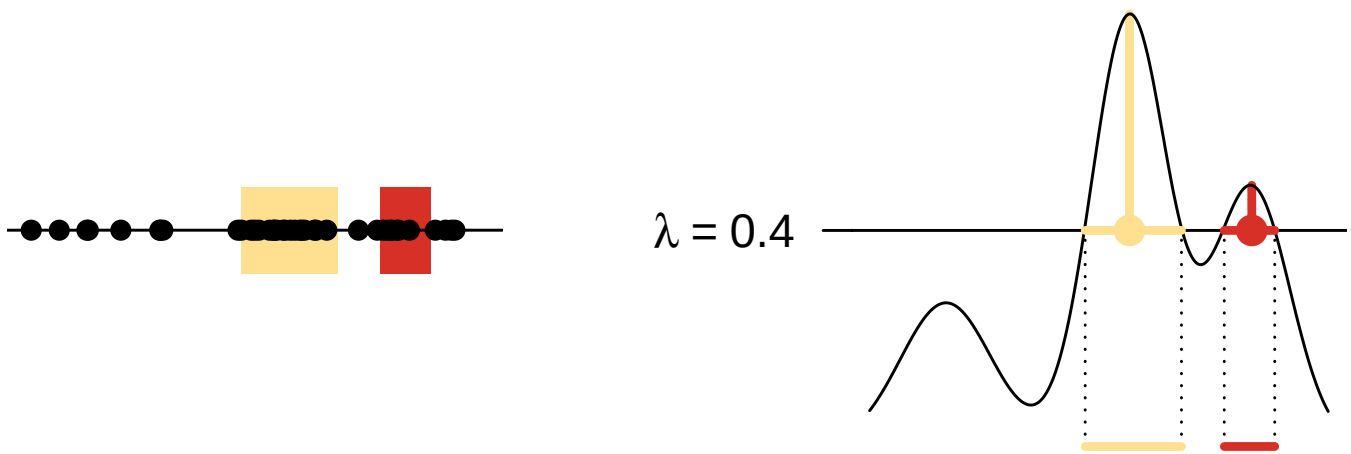
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



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

Definition

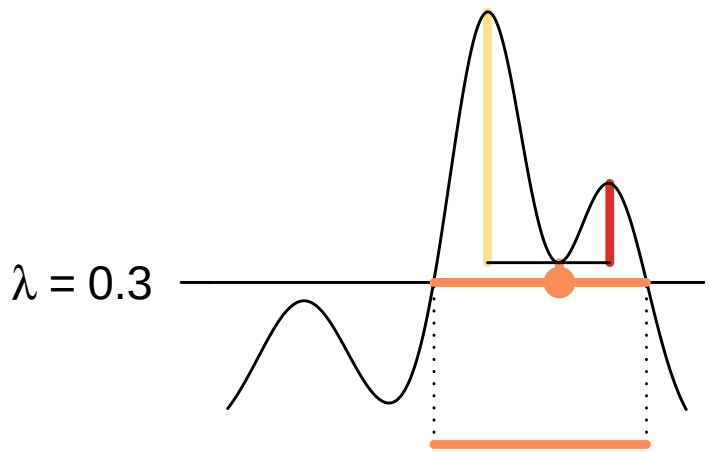
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



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

Definition

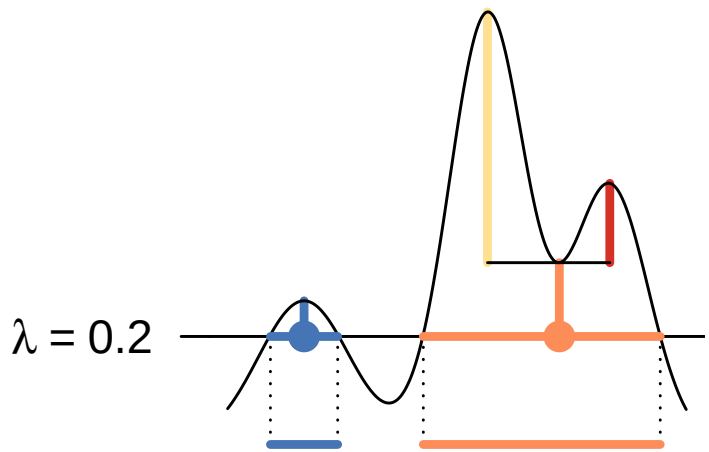
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



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

Definition

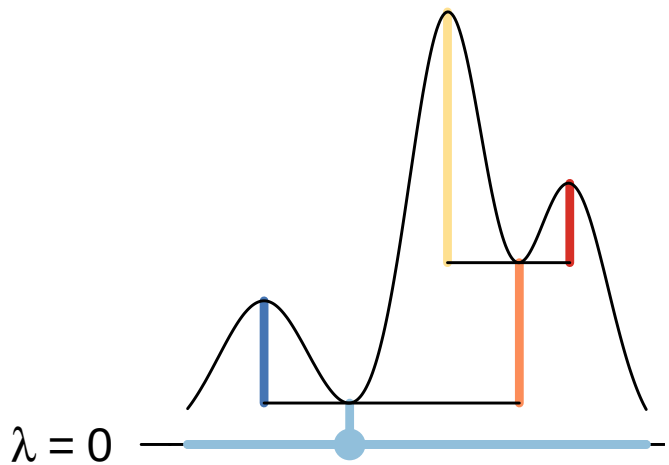
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



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

Definition

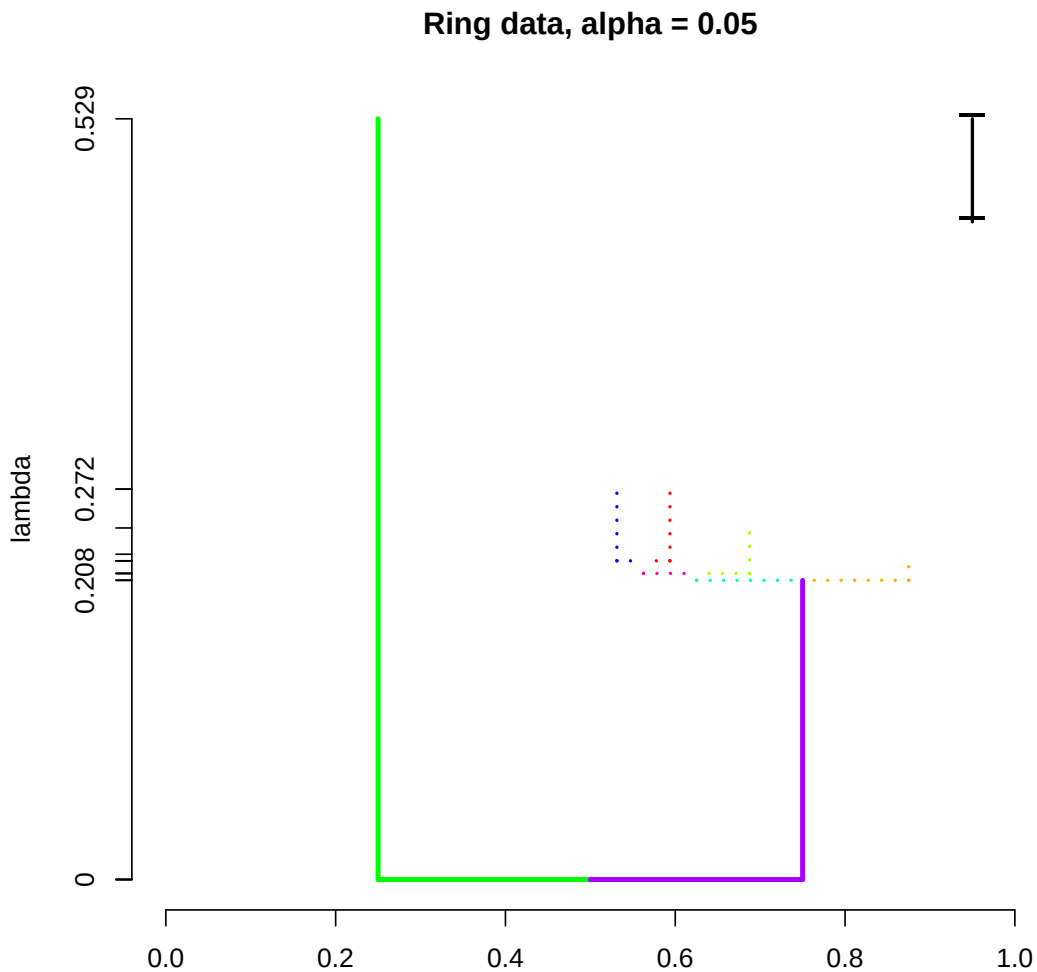
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



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,

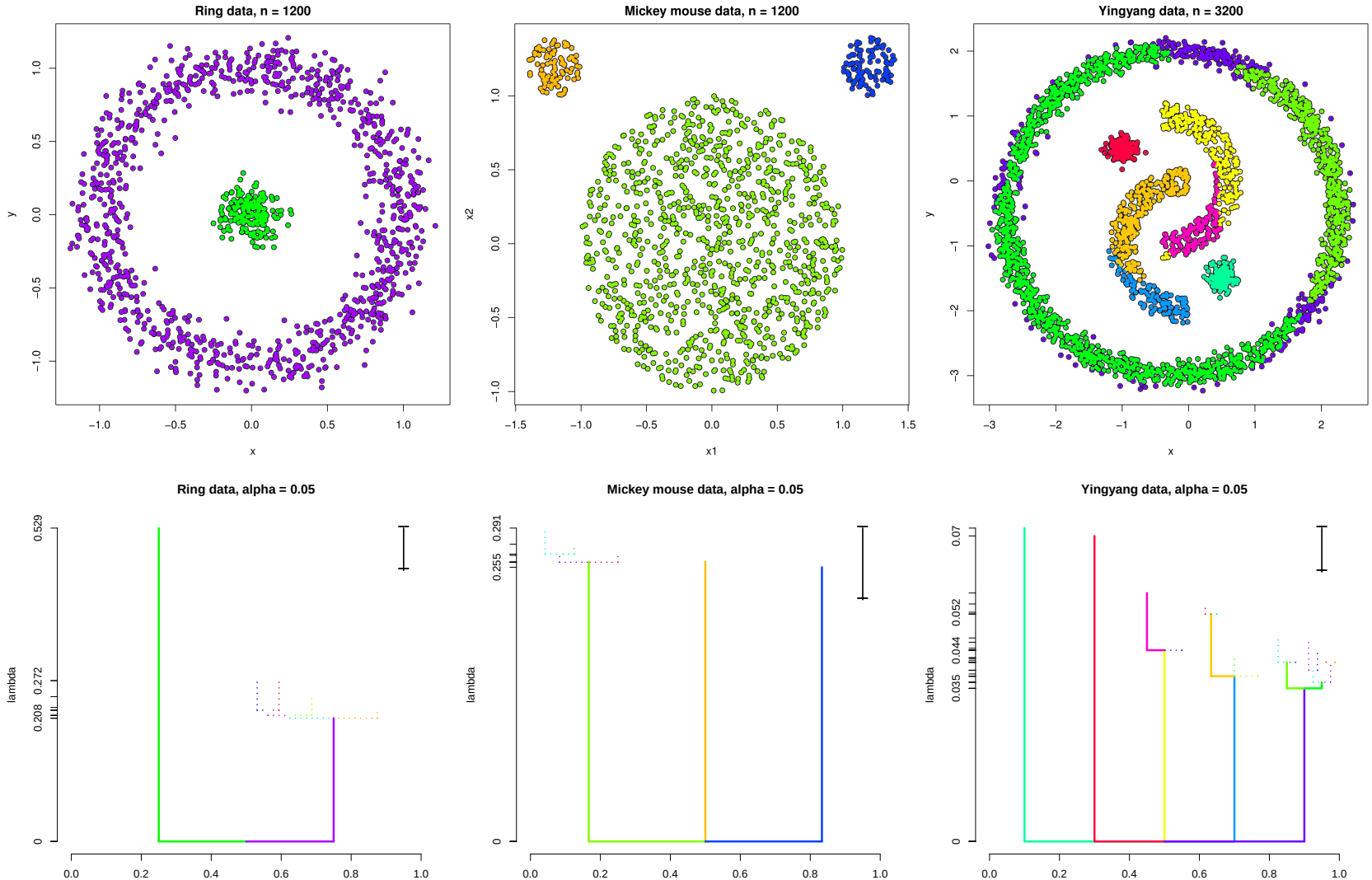
$$\hat{C}_\alpha = \{T : d_\infty(T, T_{\hat{p}_h}) \leq t_\alpha\}.$$

Theorem

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

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

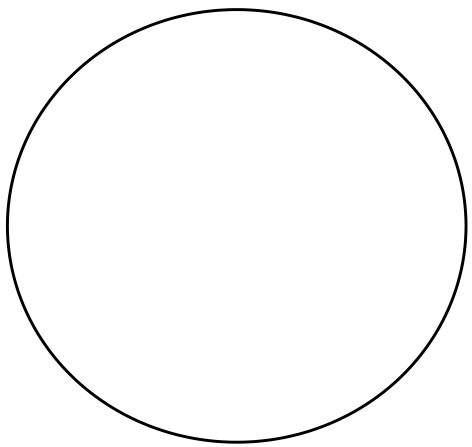
신뢰집합을 이용하여 가지치기한 군집나무로 실제 군집나무를 찾을 수 있습니다.



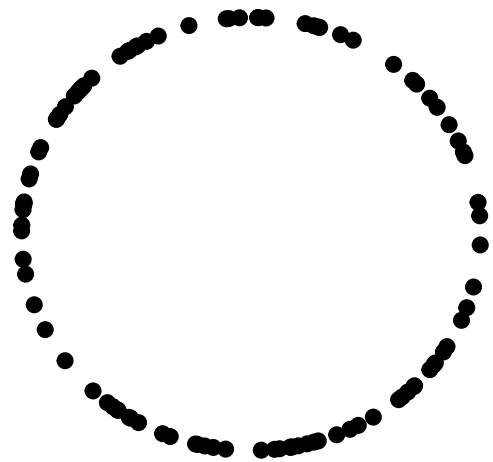
유한한 자료의 호몰로지는 기저 구조의 호몰로지와 다르기 때문에, 유한한 자료로 직접 기저 구조의 호몰로지를 추정할 수는 없습니다.

- 자료를 분석할 때, 기저 구조의 특성을 자료의 특성으로부터 추정할 수 있는 로버스트(robust)한 특성을 선호합니다.
- 호몰로지(Homology)는 로버스트하지 않습니다:

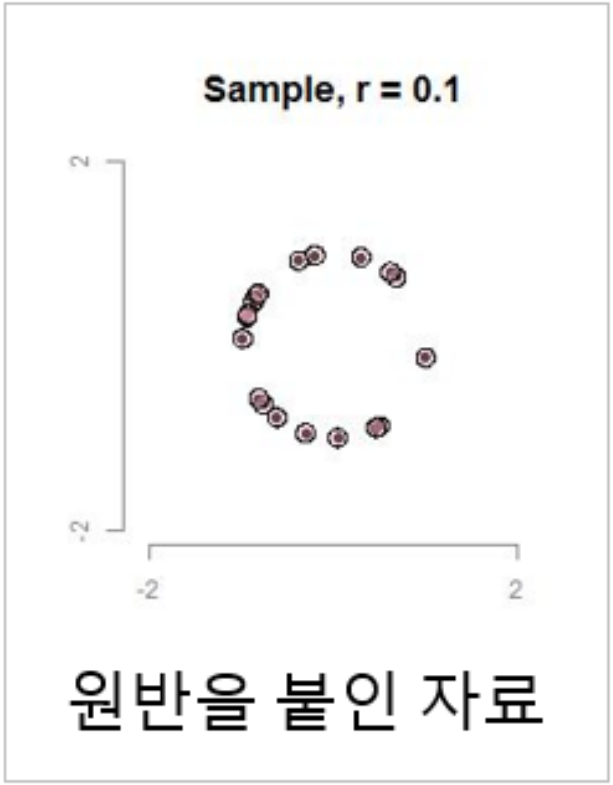
Underlying circle: $\beta_0 = 1$, $\beta_1 = 1$



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

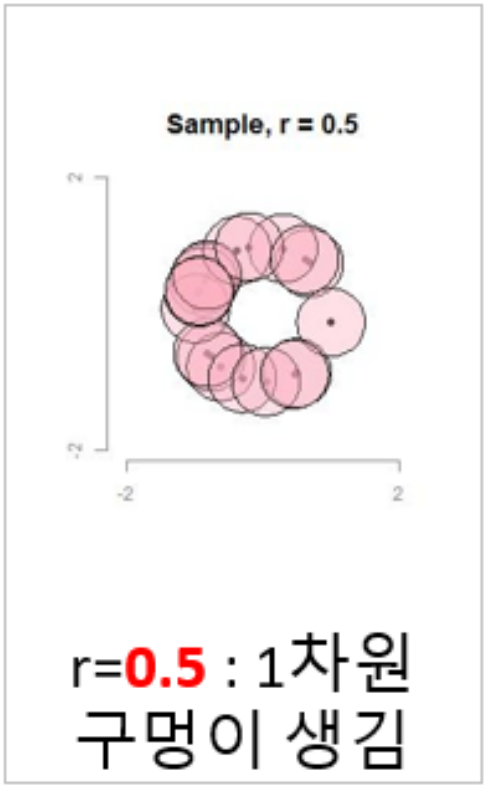


Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.

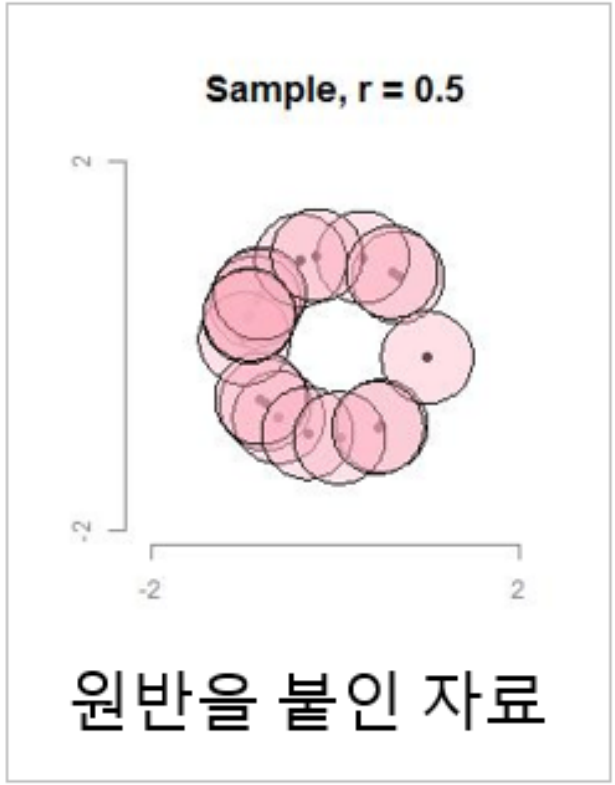


원반을 붙인 자료

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

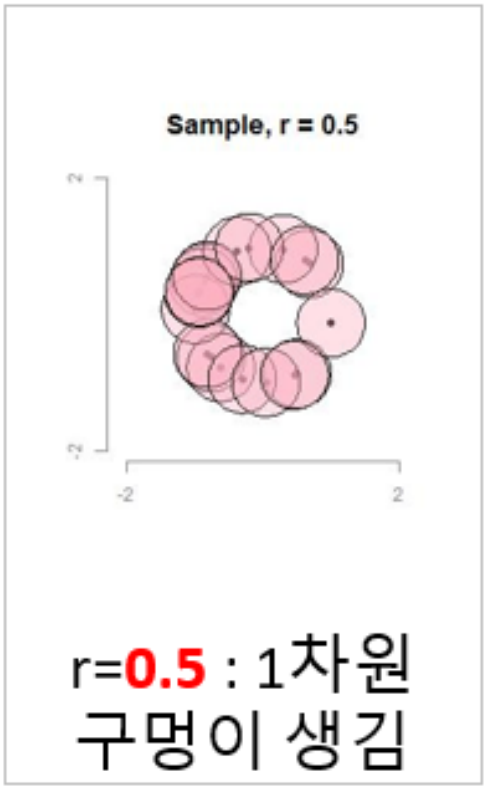


r=**0.5** : 1차원 구멍이 생김

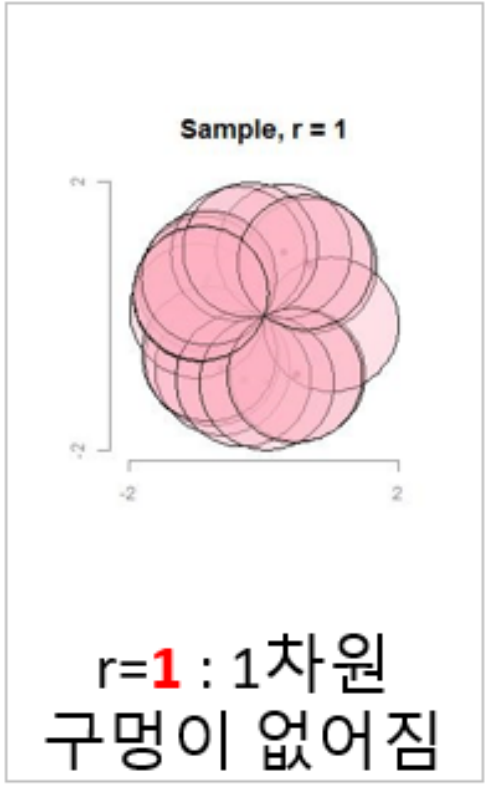


원반을 붙인 자료

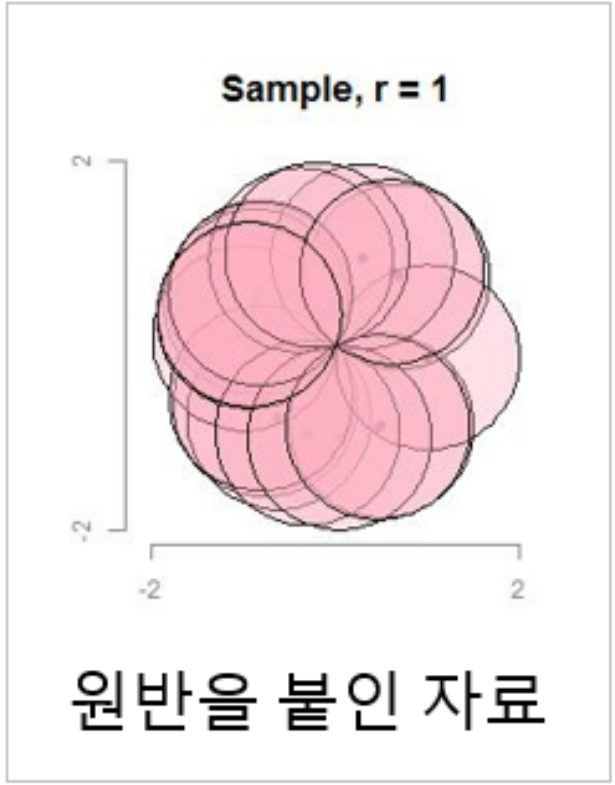
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



r=**0.5** : 1차원 구멍이 생김

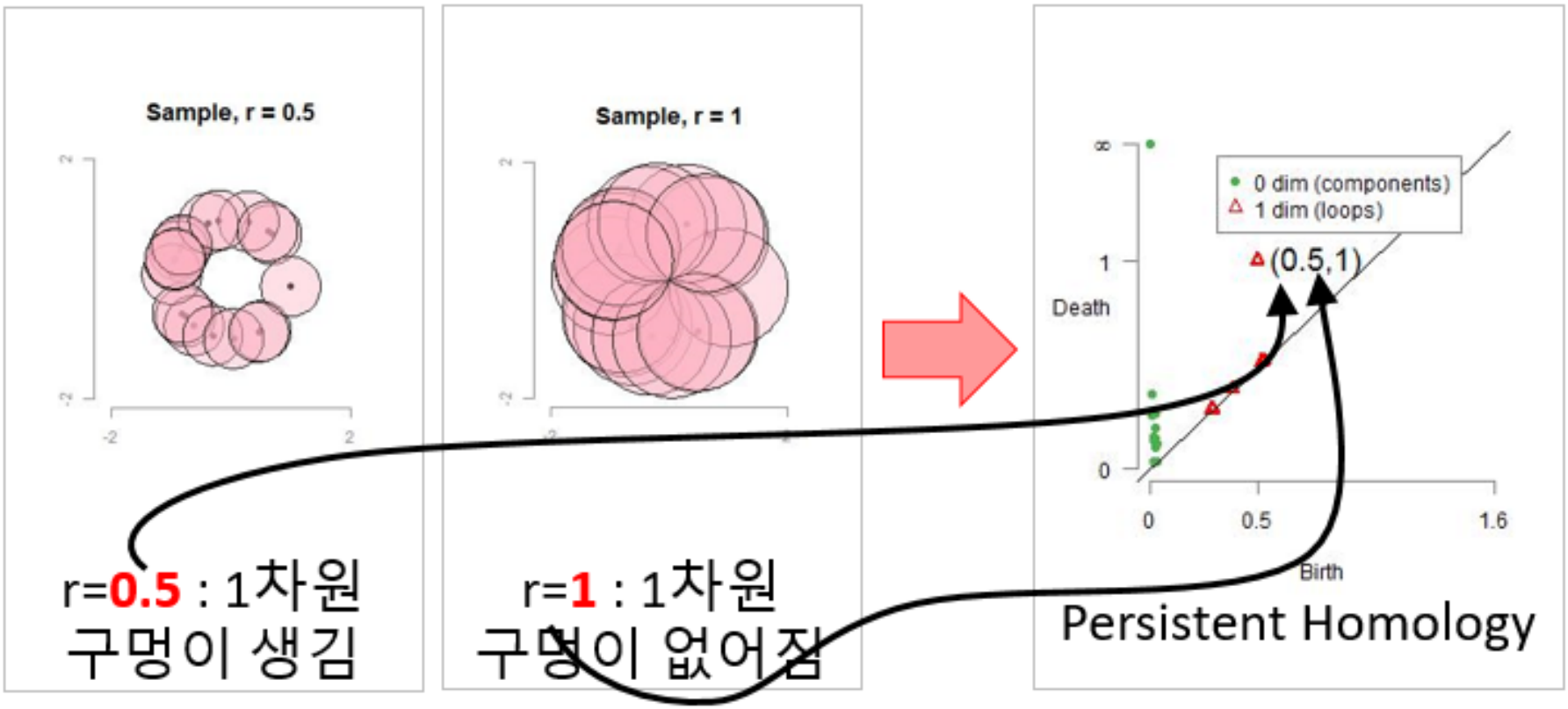


r=**1** : 1차원 구멍이 없어짐



원반을 붙인 자료

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

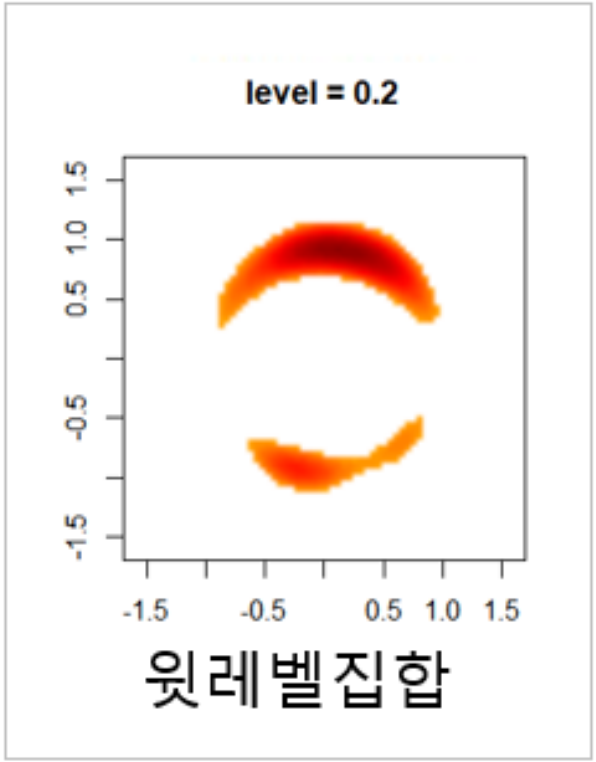


기저 구조의 위상학적 정보를 추출하는 데에 핵밀도추정(kernel density estimator)을 사용합니다.

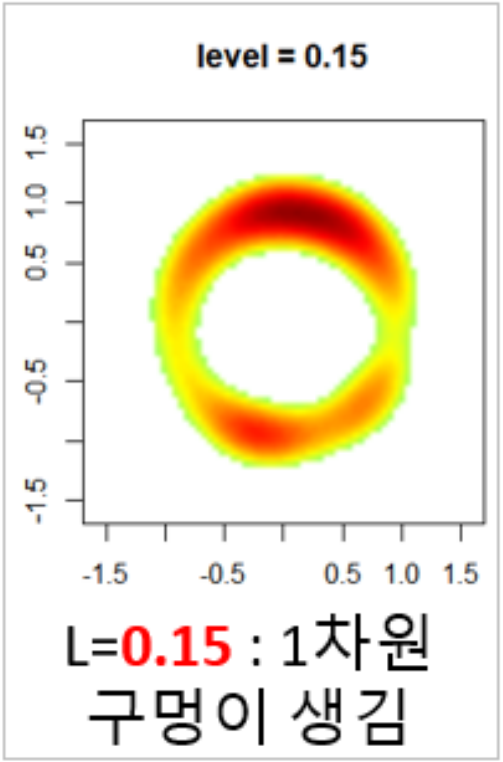
- 핵밀도추정(kernel density estimator)은 다음과 같습니다:

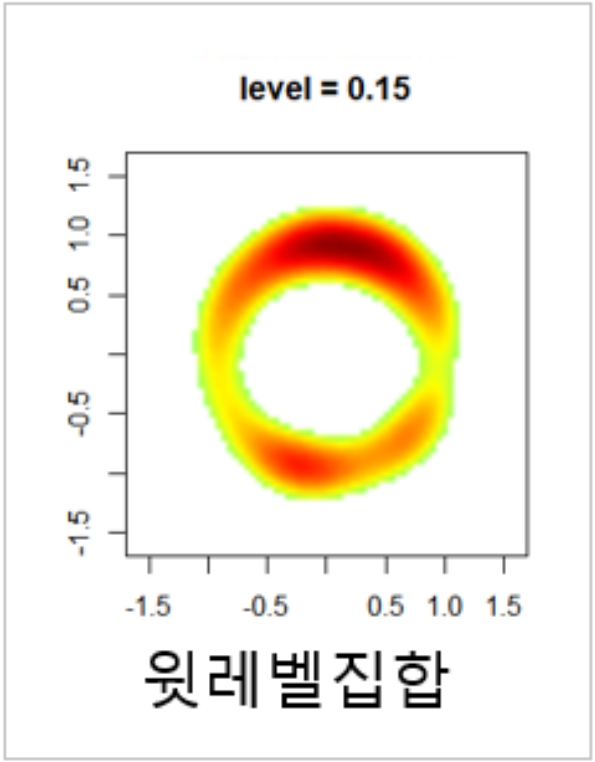
$$\hat{p}_h(x) = \frac{1}{nh^m} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

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

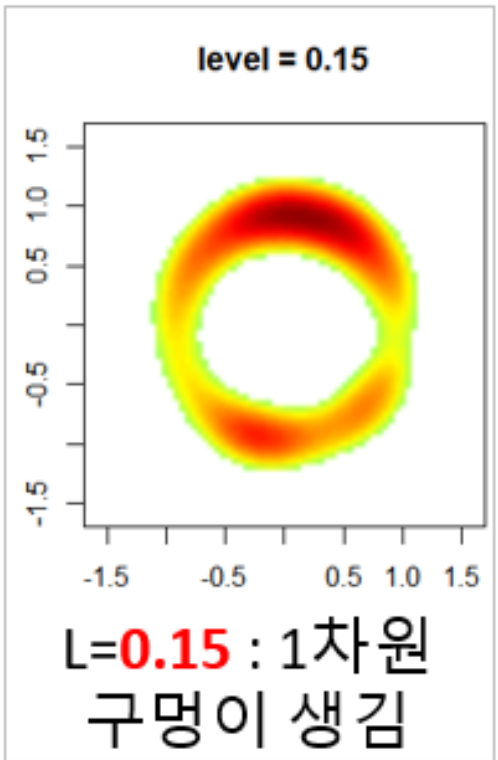


Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.

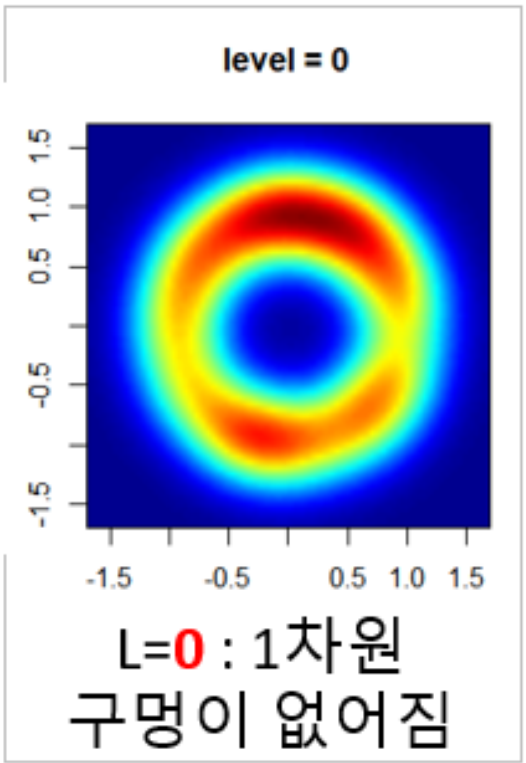




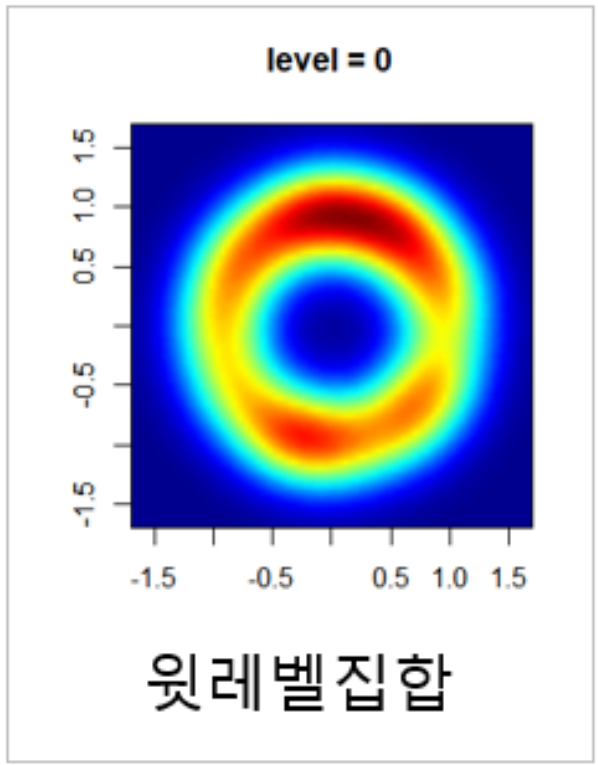
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



L=**0.15** : 1차원 구멍이 생김

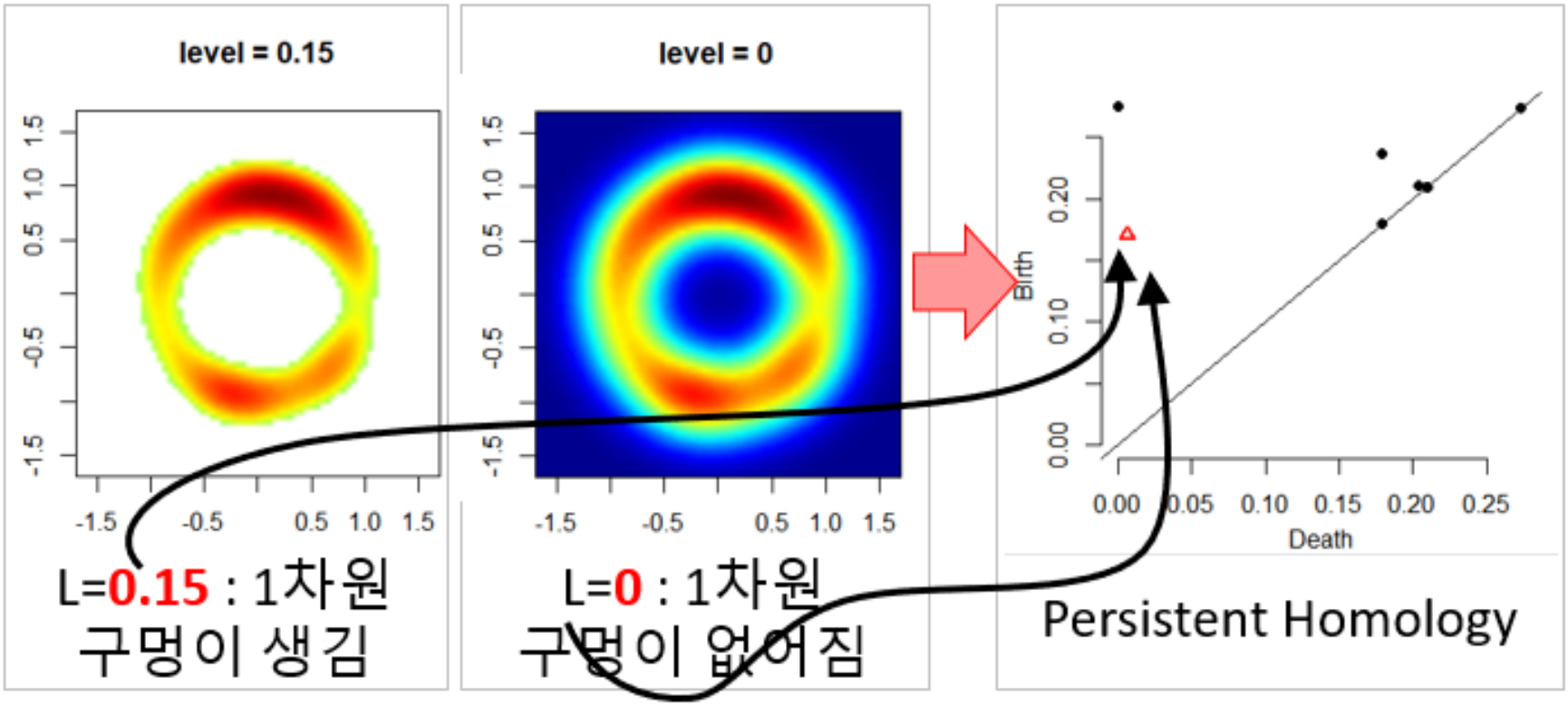


L=**0** : 1차원 구멍이 없어짐

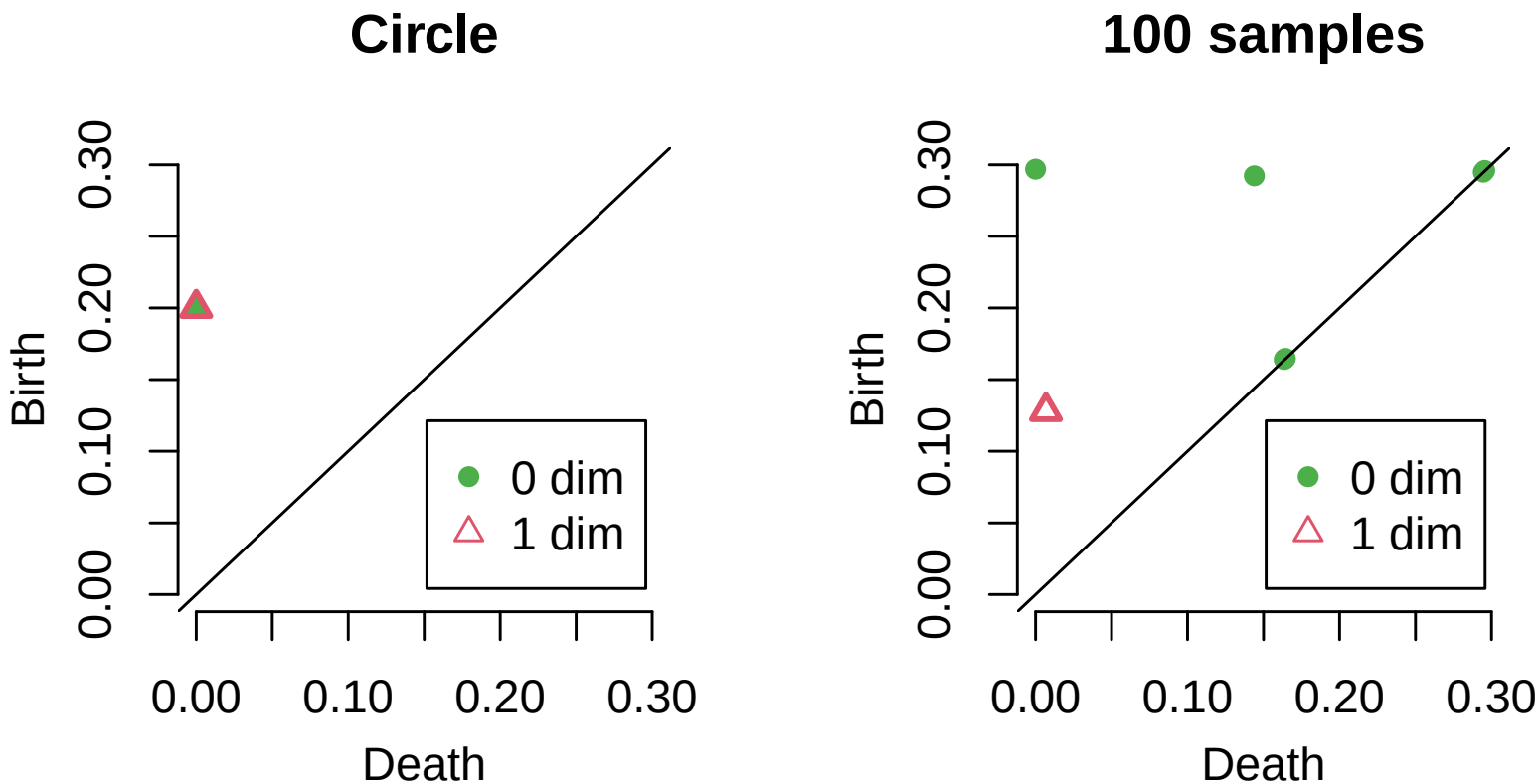


윗레벨집합

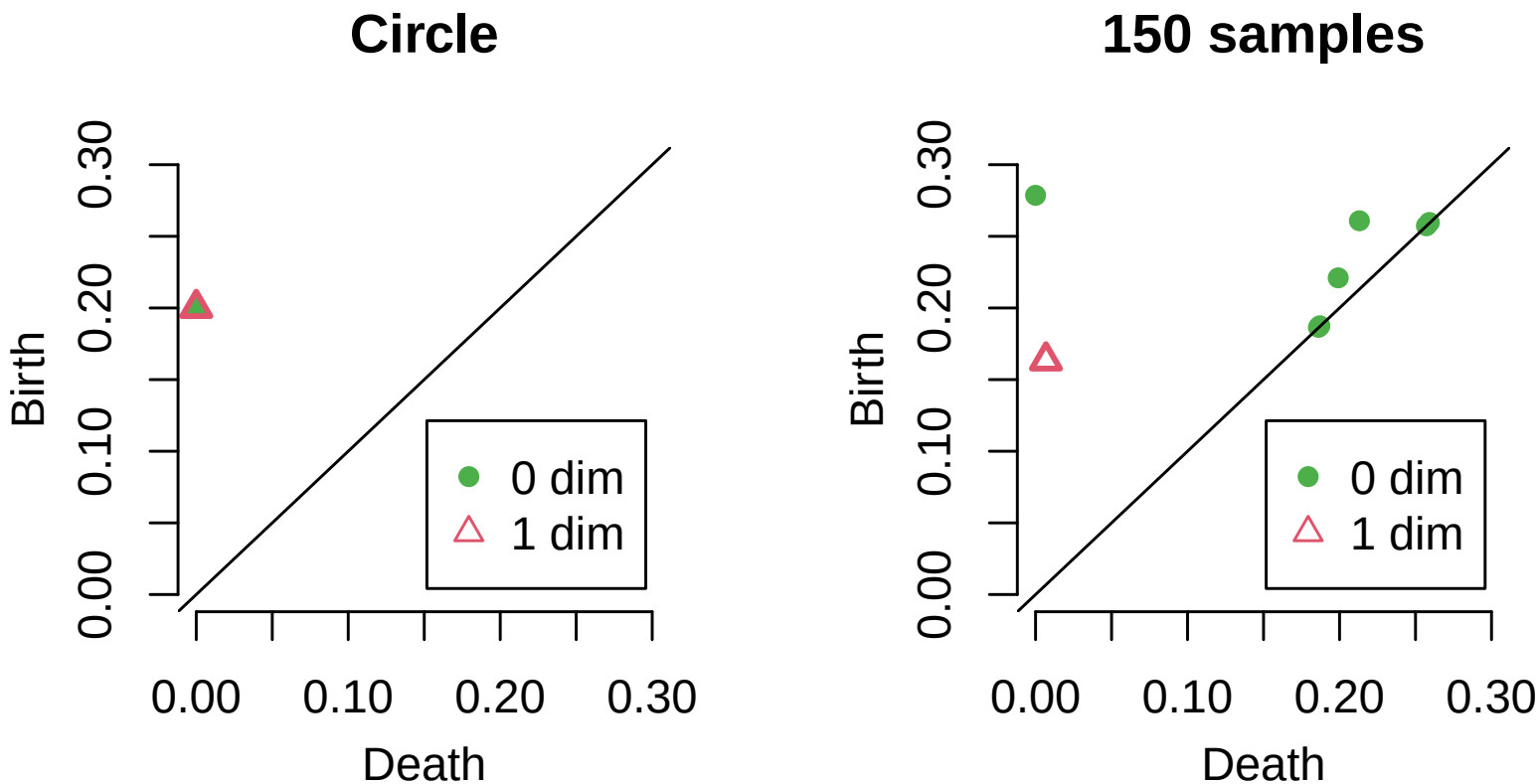
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



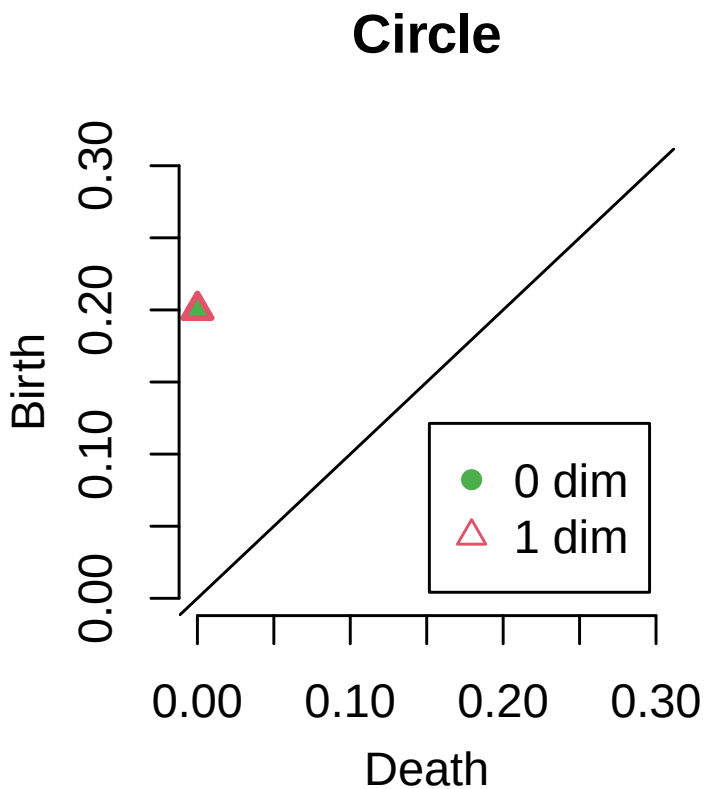
유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.

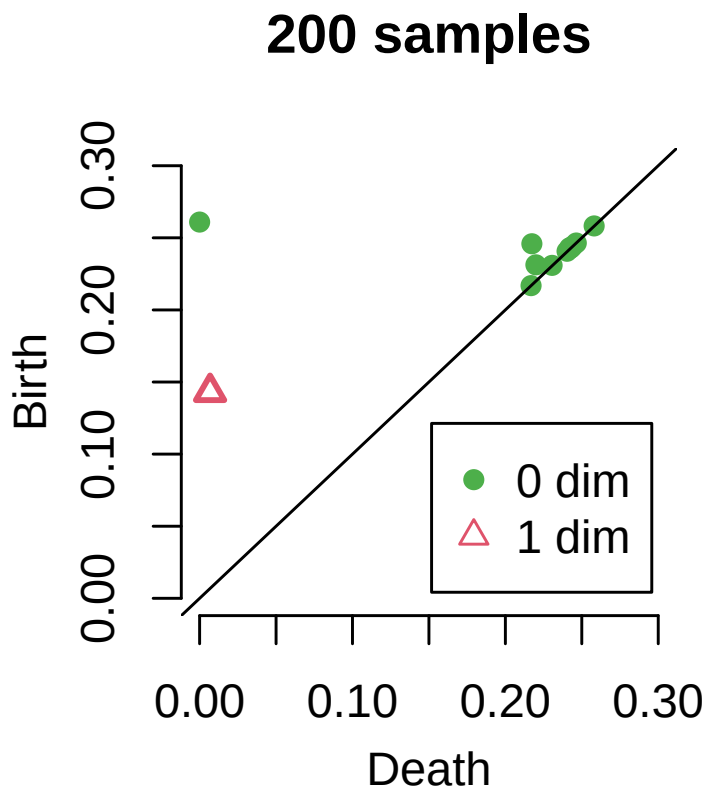


유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.

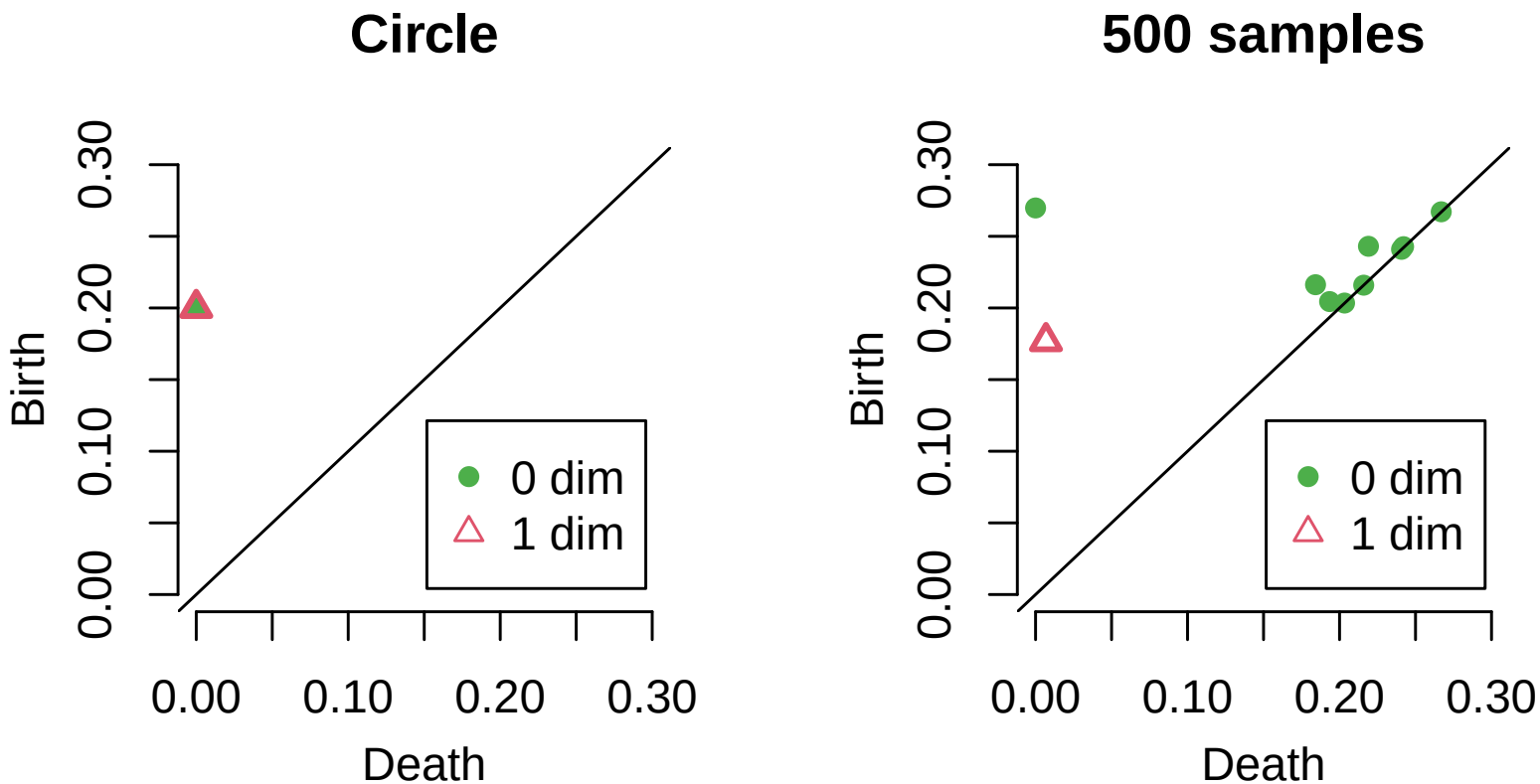


유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.

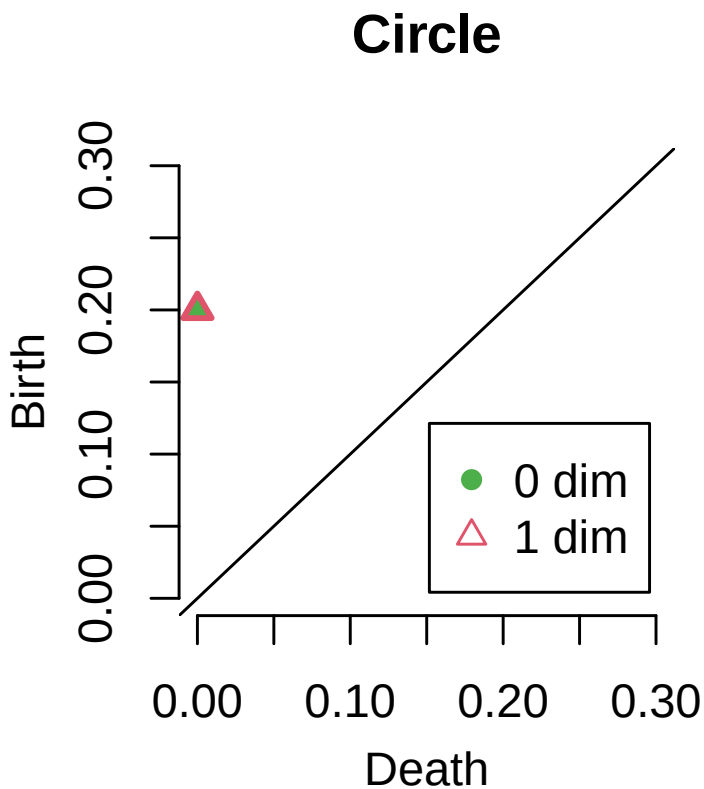


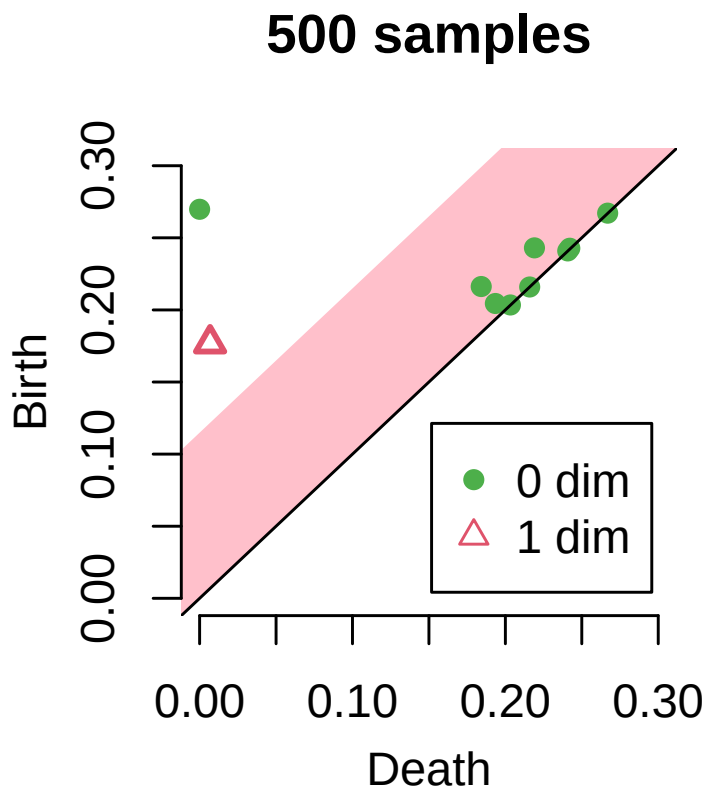


유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.



통계적으로 유의한 호몰로지 특성과 그렇지 않은 호몰로지 특성을 어떻게 구분할까요?





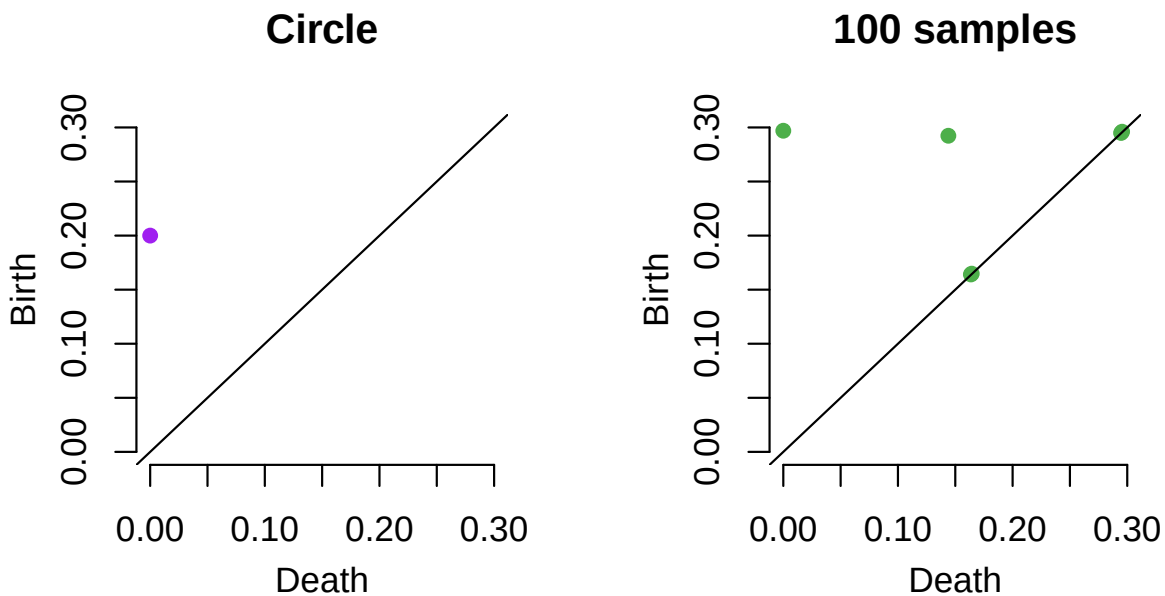
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

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.



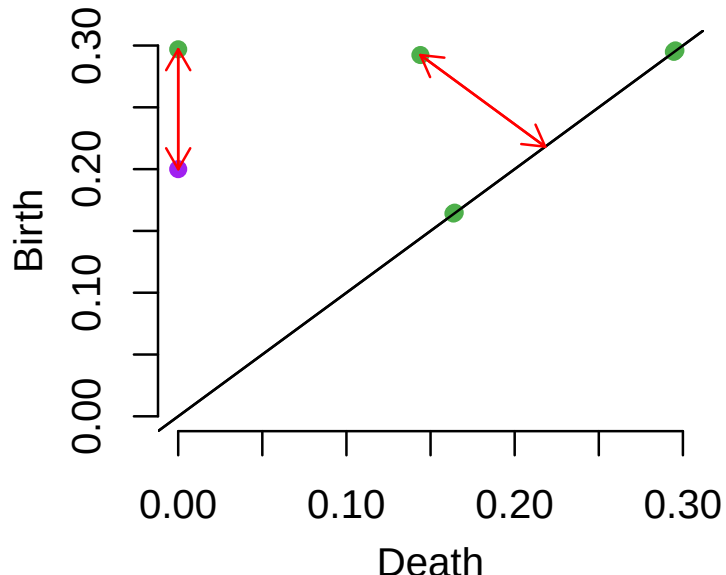
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

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.



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

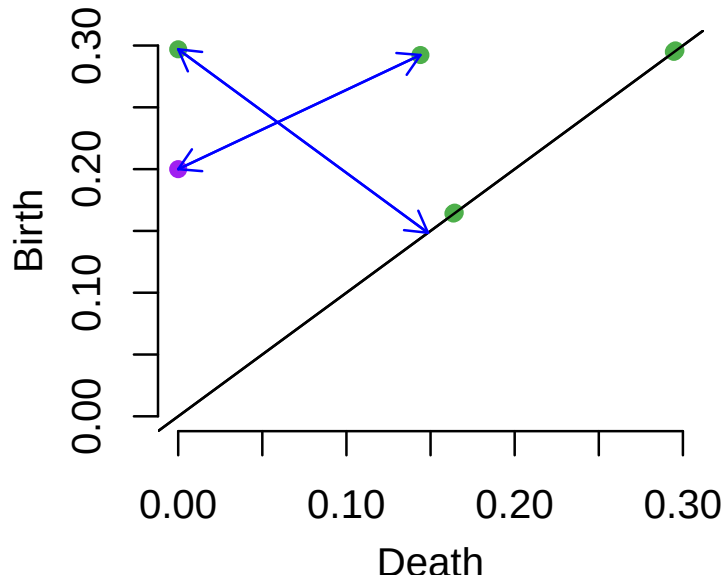
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

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.



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

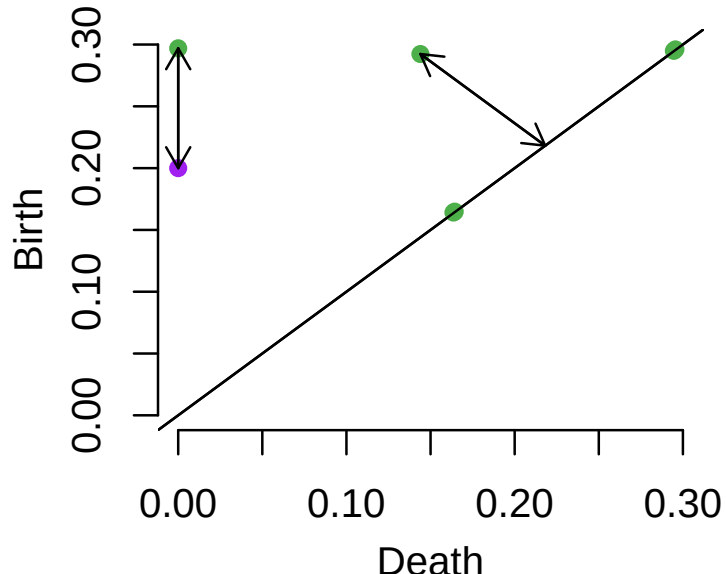
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

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.

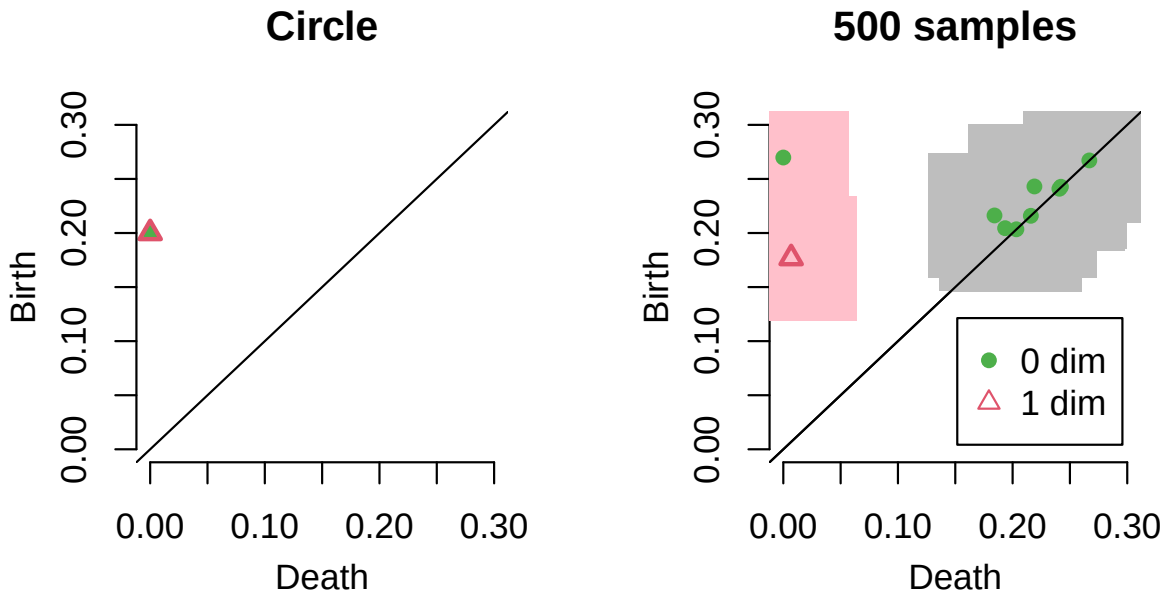


$$\inf_\gamma \sup_{x \in D_1} \|x - \gamma(x)\|_\infty = 0.1$$

Persistent homology의 신뢰집합(Confidence Set)은 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(Confidence Set) $\{D \in Dgm : \ W_\infty(Dgm(X), D) \leq c_n\}$은 다음을 만족하는 랜덤집합입니다:

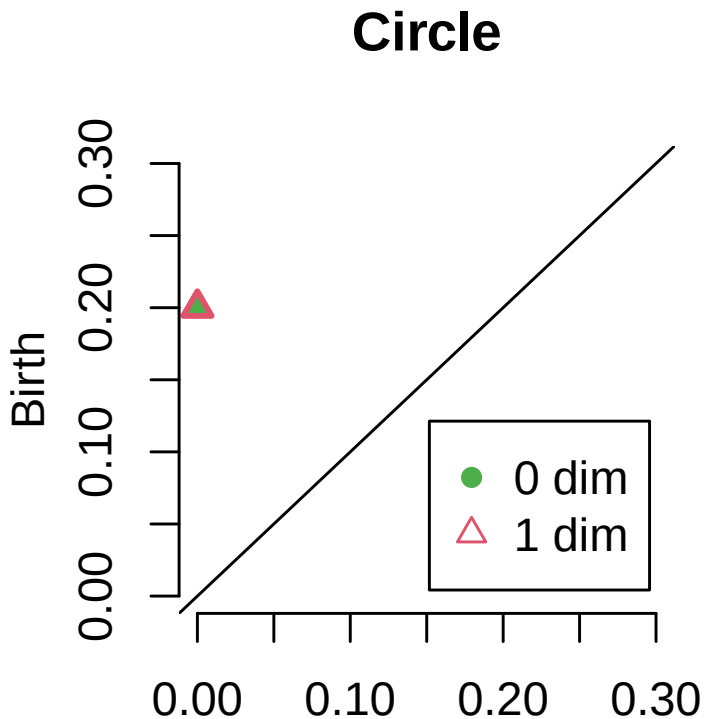
$$\mathbb{P}\left(Dgm(M) \in \{D \in Dgm : \ W_\infty(Dgm(X), D) \leq c_n\}\right) \geq 1-\alpha.$$

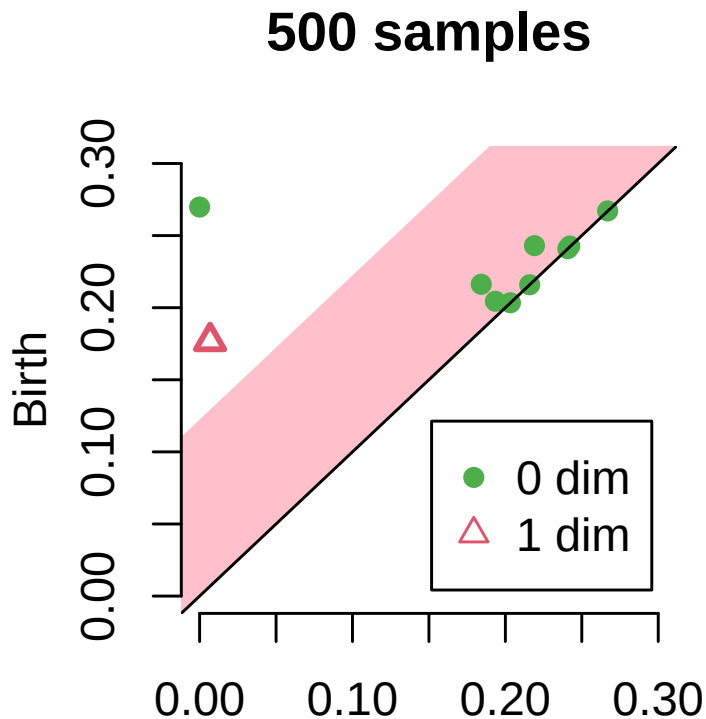


Persistent Homology의 신뢰띠(Confidence Band)를 이용하여 통계적으로 유의한 호몰로지 특성과 그렇지 않은 호몰로지 특성을 구분합니다.

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(Confidence Band) $c_n = c_n(X)$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, $\theta^* = \sqrt{nh^m}||\hat{p}_h^*(x) - \hat{p}_h(x)||_\infty$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 $\left[\hat{p}_h - \frac{\hat{z}_\alpha}{\sqrt{nh^m}}\ ,\ \hat{p}_h + \frac{\hat{z}_\alpha}{\sqrt{nh^m}}\right]$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: $\arg\min_\theta \mathcal{L}(f_\theta, X)$.
- 많은 경우, 최소해의 명시적 형태(explicit formula)를 구하는 것은 불가능하거나 너무 비쌉니다(e.g. 큰 역행렬을 계산). 따라서, $\nabla_\theta \mathcal{L}(f_\theta, X)$를 이용한 경사법(gradient descent)을 사용합니다:

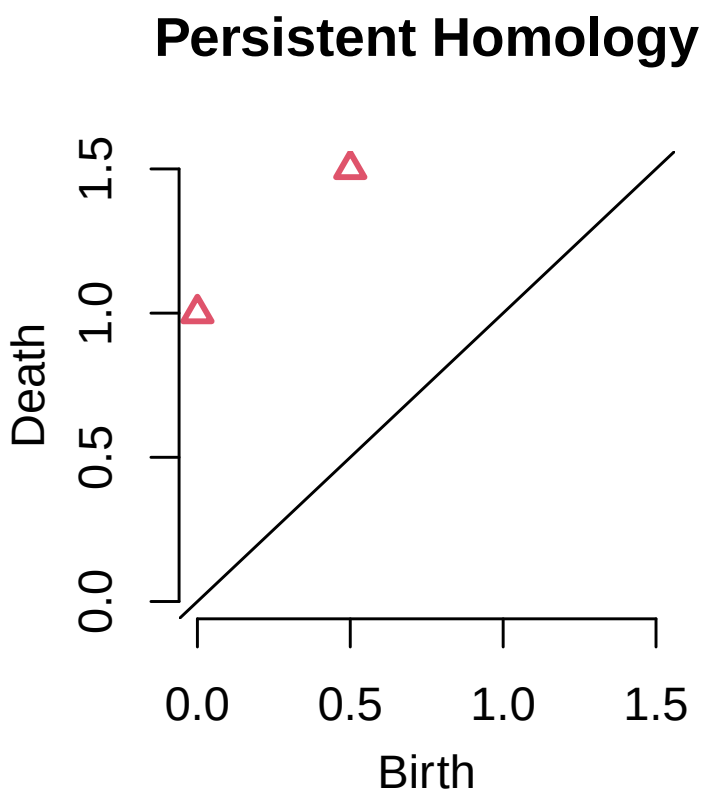
$$\theta_{n+1} = \theta_n - \lambda \nabla_\theta \mathcal{L}(f_\theta, X).$$

위상 자료 분석(Topological Data Analysis)을 기계학습(Machine Learning)에 응용합니다.

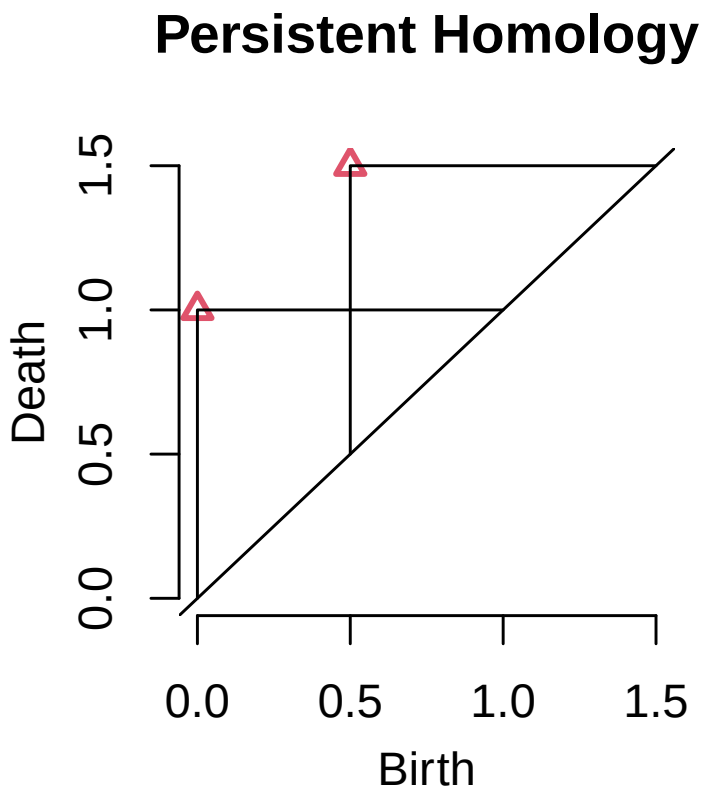
- A Survey of Topological Machine Learning Methods (Hensel, Moor, Rieck, 2021)
- 위상 자료 분석(Topological Data Analysis)을 기계학습(Machine Learning)에 응용하는 데에는 크게 두 가지 방향이 있습니다:
  - 위상 자료 분석을 이용하여 특성(feature)을 만들어, 자료 X에 위상학적 특성을 추가하기: 더 흔한 방식
    - PLLay: Efficient Topological Layer based on Persistence Landscapes (Kim, Kim, Zaheer, Kim, Chazal, Wasserman, 2020)
    - Generalized penalty for circular coordinate representation (Luo, Patania, Kim, Vejdemo-Johansson, 2021)
  - 자료 X나 모형 f_θ의 품질을 TDA로 평가: 최근 주목
    - TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)

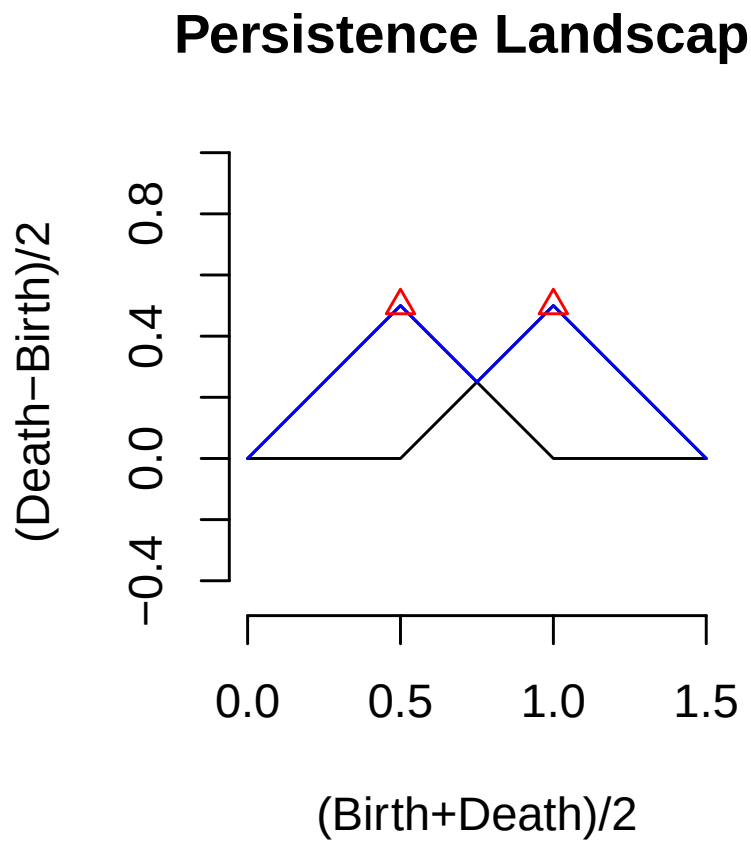
Persistent homology를 한 번 더 요약해서 유클리드 공간 또는 함수 공간에 넣습니다.

- Persistent homology의 공간은 구조적으로 복잡하여 기계학습 (machine learning) 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.



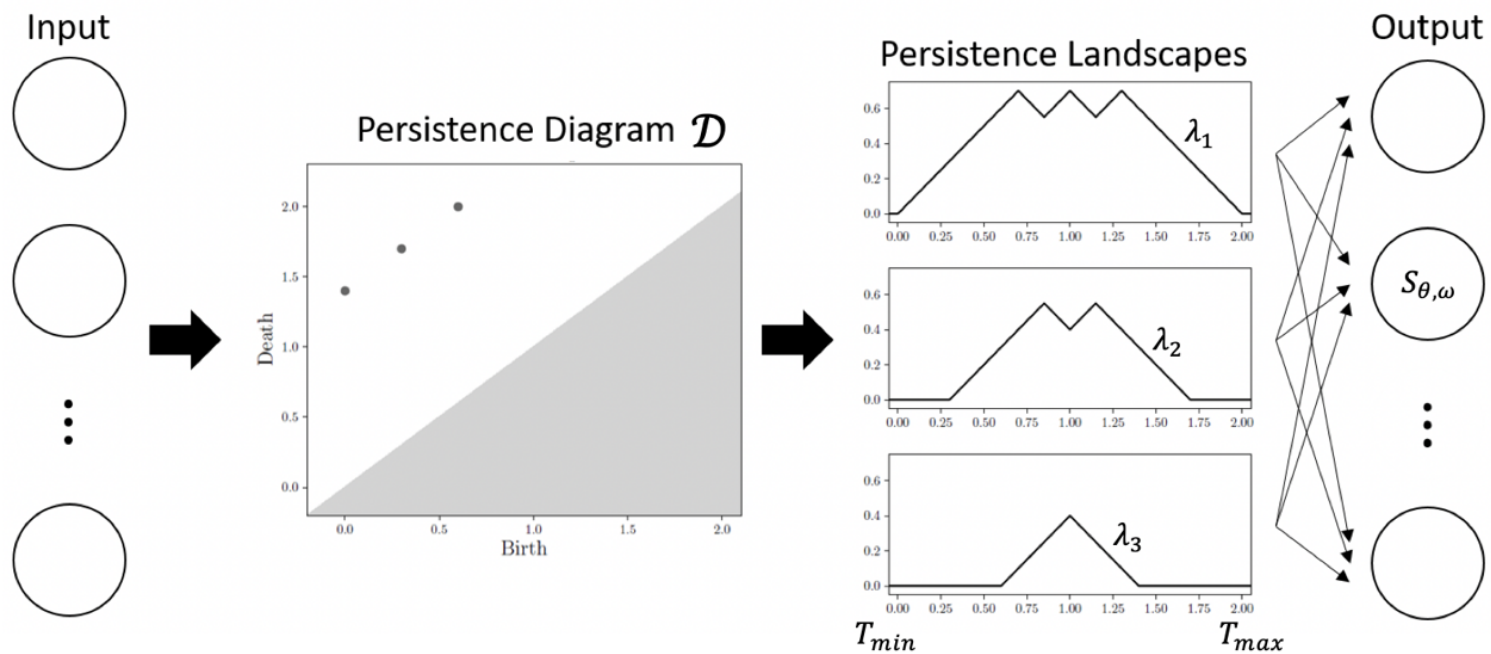
Persistence Landscape은 Persistent homology의 함수 요약입니다.





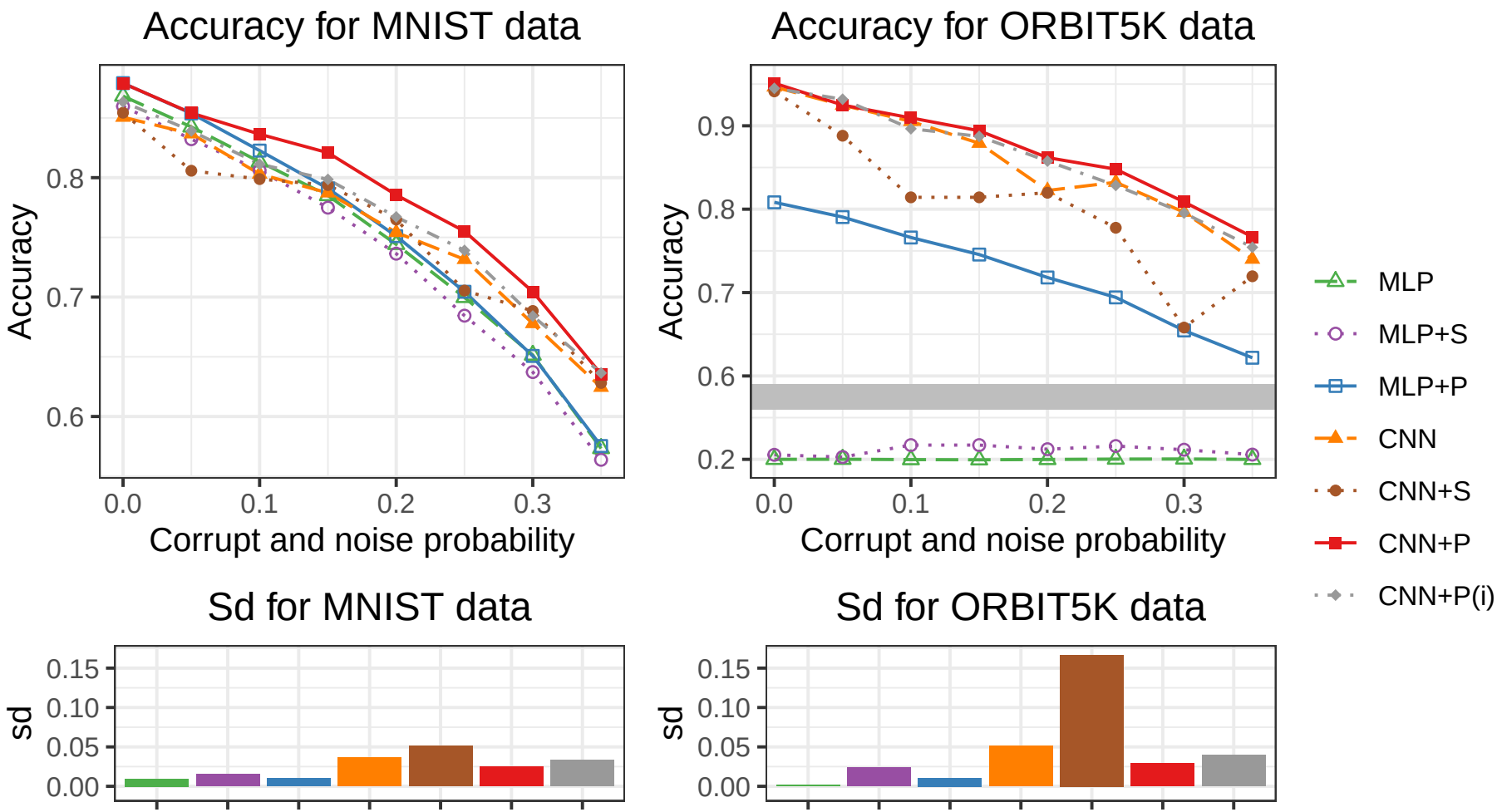
Persistence Landscape으로 위상학적 층(topological layer) 만들기

1. 자료 X의 Persistent Homology $\mathcal{D}$를 계산합니다.
2. $\mathcal{D}$로부터 Persistence 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 계산하고, 이를 벡터화하여 $\bar{\Lambda}_\omega \in \mathbb{R}^m$을 만듭니다.
4. 매개화된 미분가능한 함수 $g_\theta : \mathbb{R}^m \to \mathbb{R}$을 사용하여, $S_{\theta,\omega}(\mathcal{D}) := g_\theta(\bar{\Lambda}_\omega)$를 계산합니다.



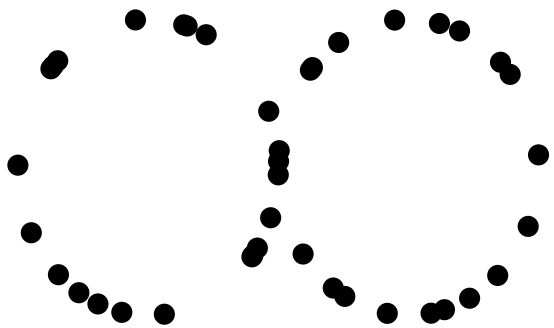
Persistence Landscape으로 위상학적 층(topological layer) 만들기

- PLLay: Efficient Topological Layer based on Persistence Landscapes (Kim, Kim, Zaheer, Kim, Chazal, Wasserman, 2020)

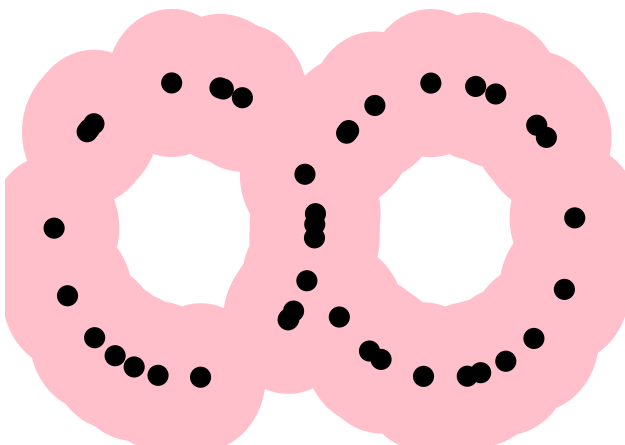


Circular coordinates 는 자료의 위상 구조를 반영하는 차원 축소 방법입니다.

data

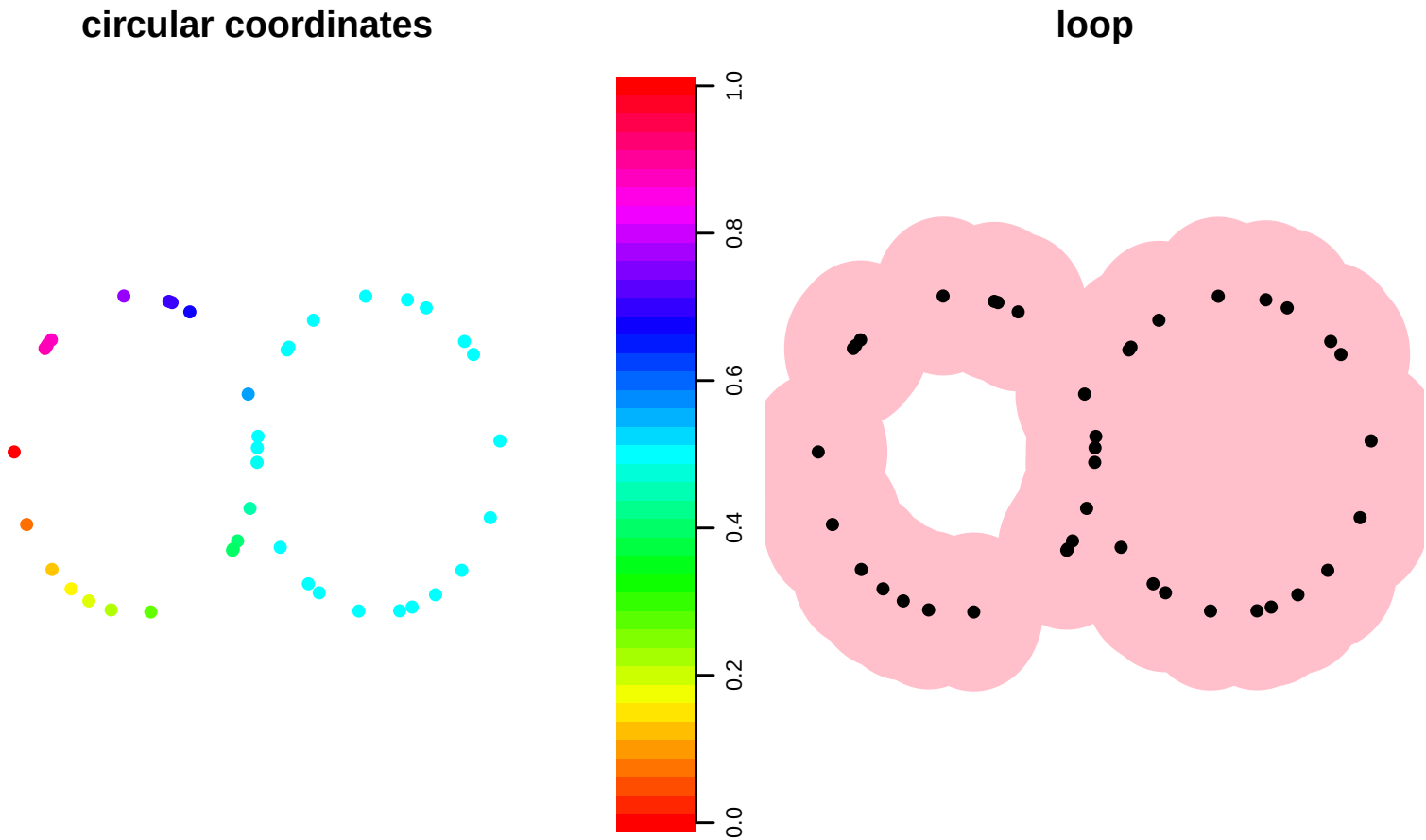


loop



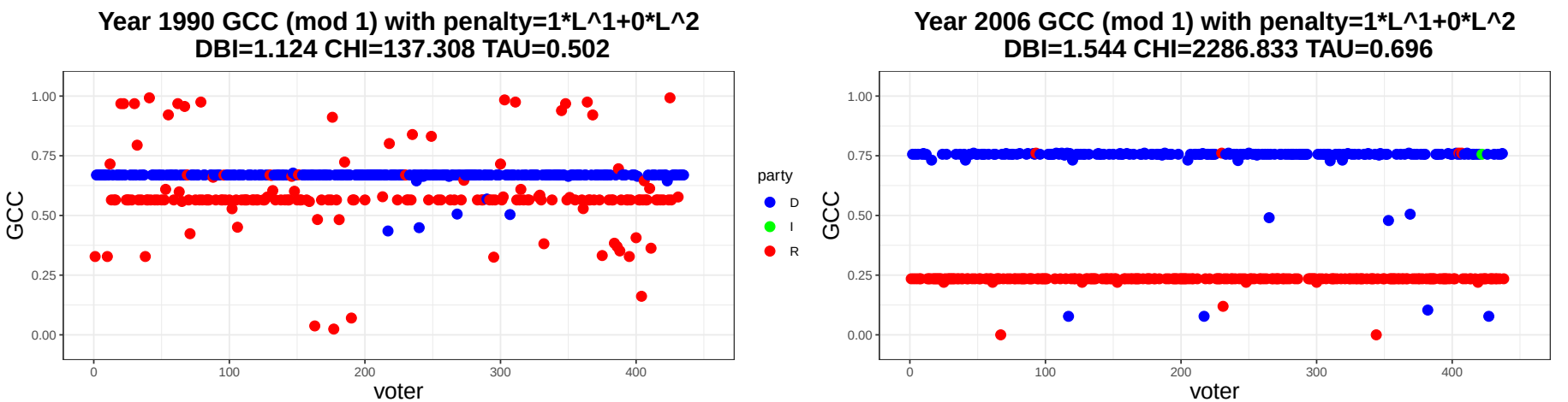
Circular coordinates 는 자료의 위상 구조를 반영하는 차원 축소 방법입니다.

- circuiar coordinate 는 자료 X 에서 원 S^1 으로 가는 함수입니다.



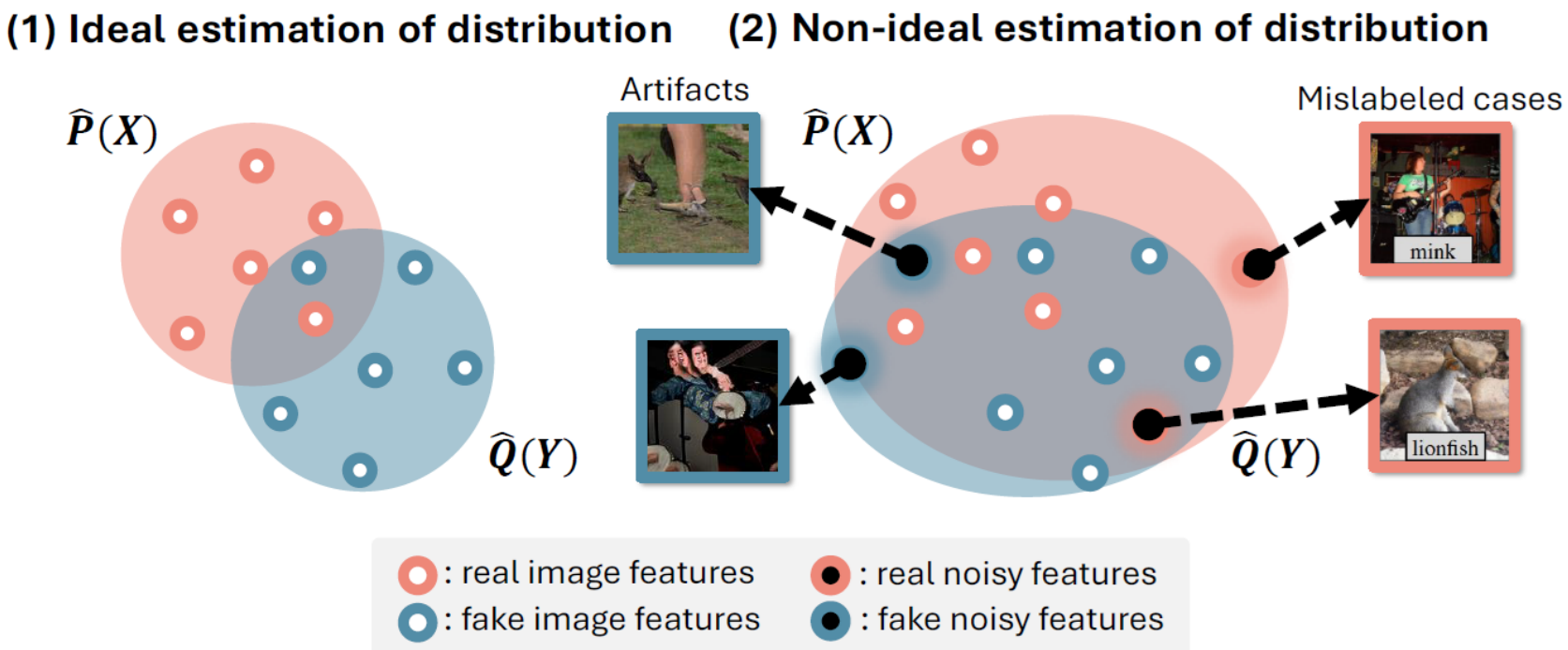
Circular coordinates 를 계산할 때 일반화된 규제 함수 (generalized penalty function)를 사용하면 자료의 위상적인 정보를 더 잘 시각화할 수 있습니다.

- Generalized penalty for circular coordinate representation (Luo, Patania, Kim, Vejdemo-Johansson, 2021)



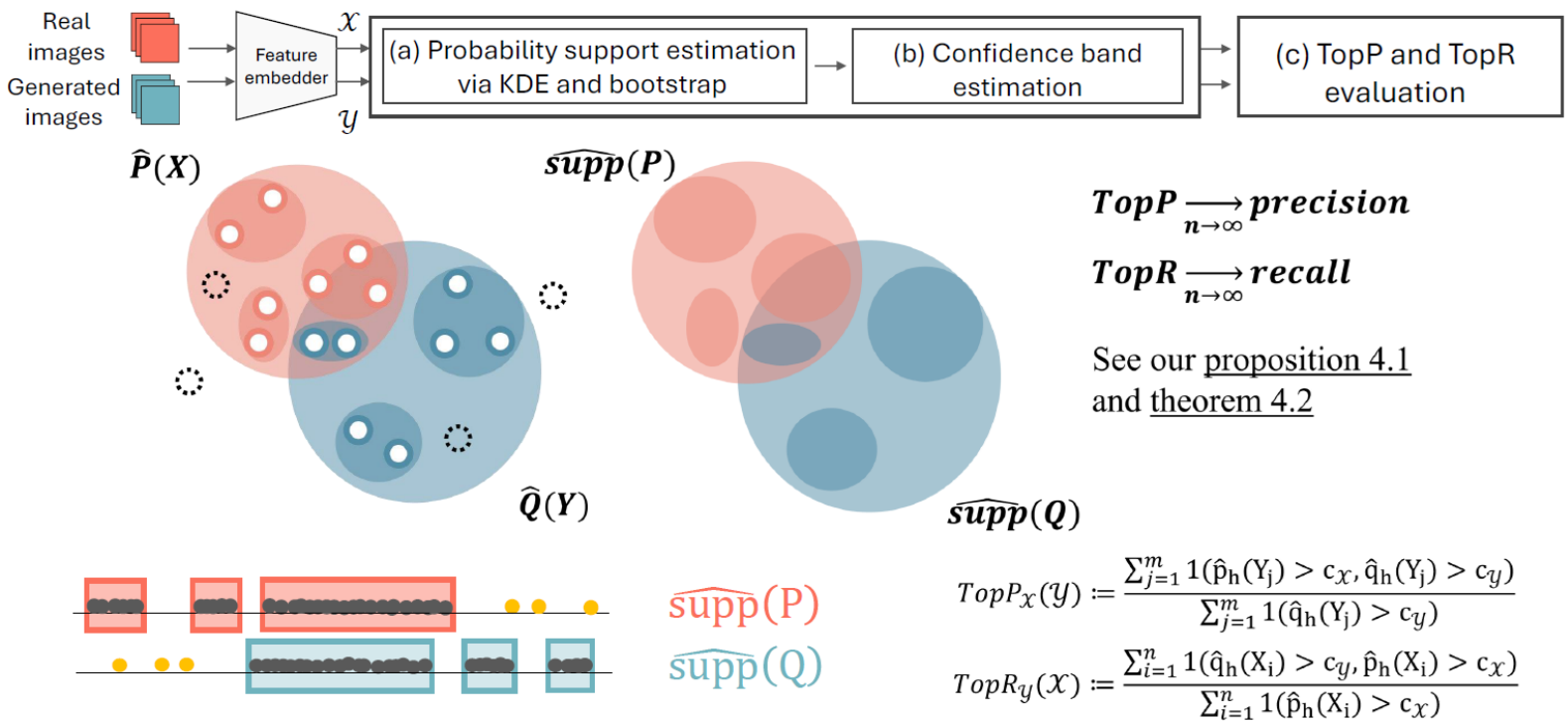
기존에 있는 생성 모형(generative model)의 평가 거리(evaluation metric)는 잡음(noise)에 취약합니다.

- TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)
- 생성 모형(generative model)을 평가(evaulate)할 때, 실제 화상(real image)의 분포(distribution)의 지지집합(support)과 가짜 화상(fake image)의 분포의 지지집합을 거리(metric)를 사용하여 비교합니다.
- 기존의 평가 거리(evaluation metric)는 자료 분포(data distribution)의 지지집합을 과대 추정합니다: 잡음(noise)에 취약합니다.



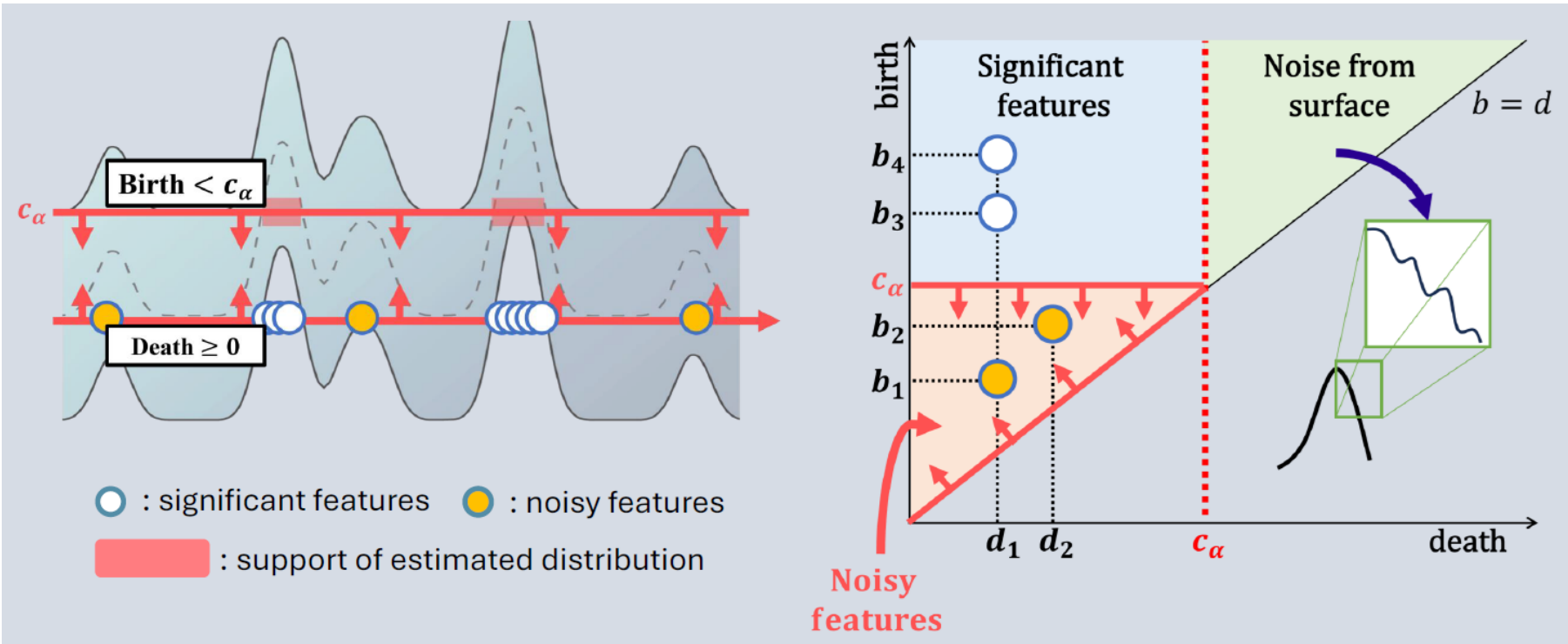
TopP&R은 위상적이고 통계적으로 유의미한 특성(feature)들만 골라냄으로써 로버스트(robust)하게 생성 모형(generative model)을 평가(evaluate)합니다.

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



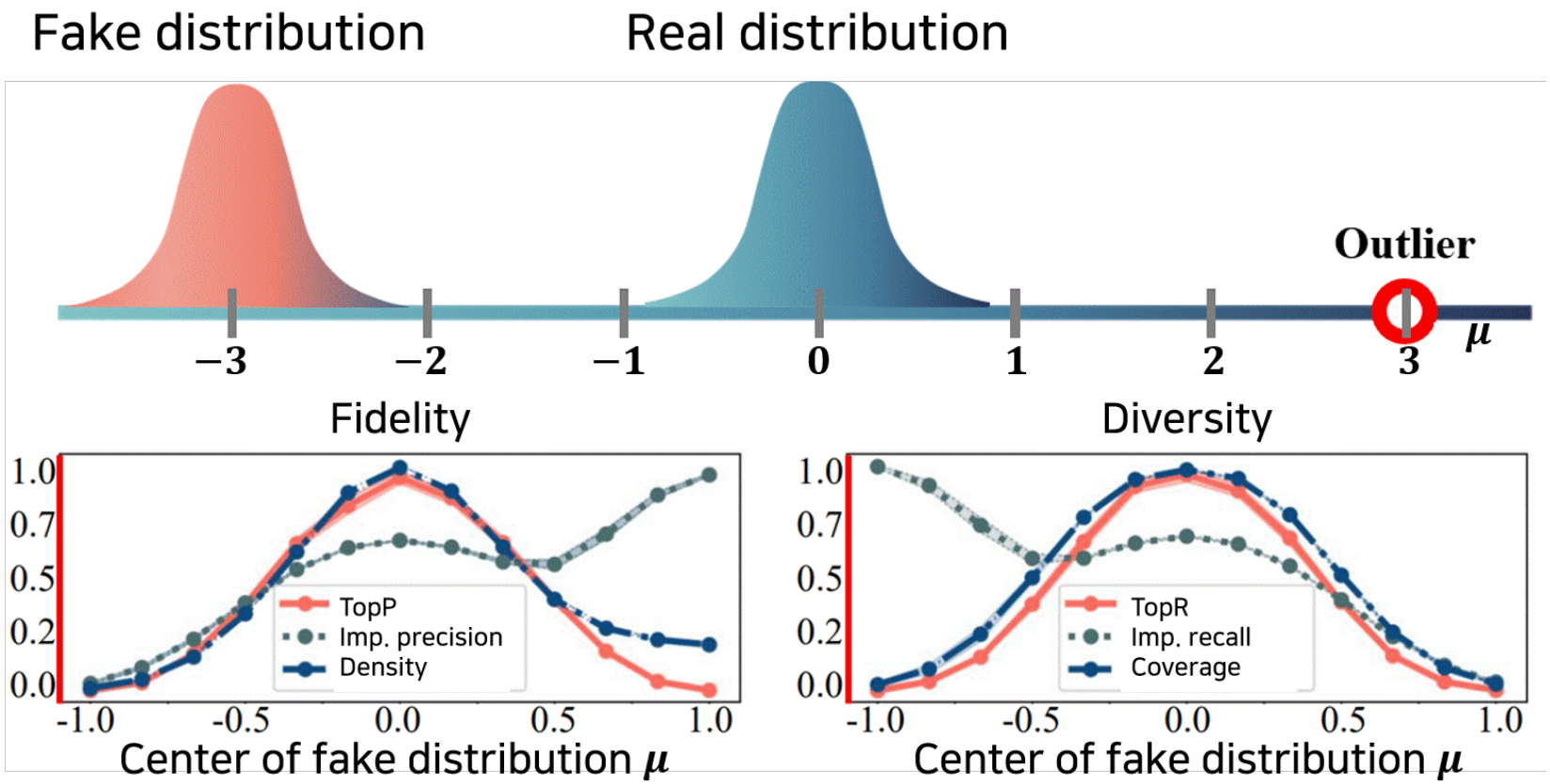
통계적 및 위상적으로 유의미한 특성들을 골라내는 문턱(threshold)을 찾아냅니다.

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



실험

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



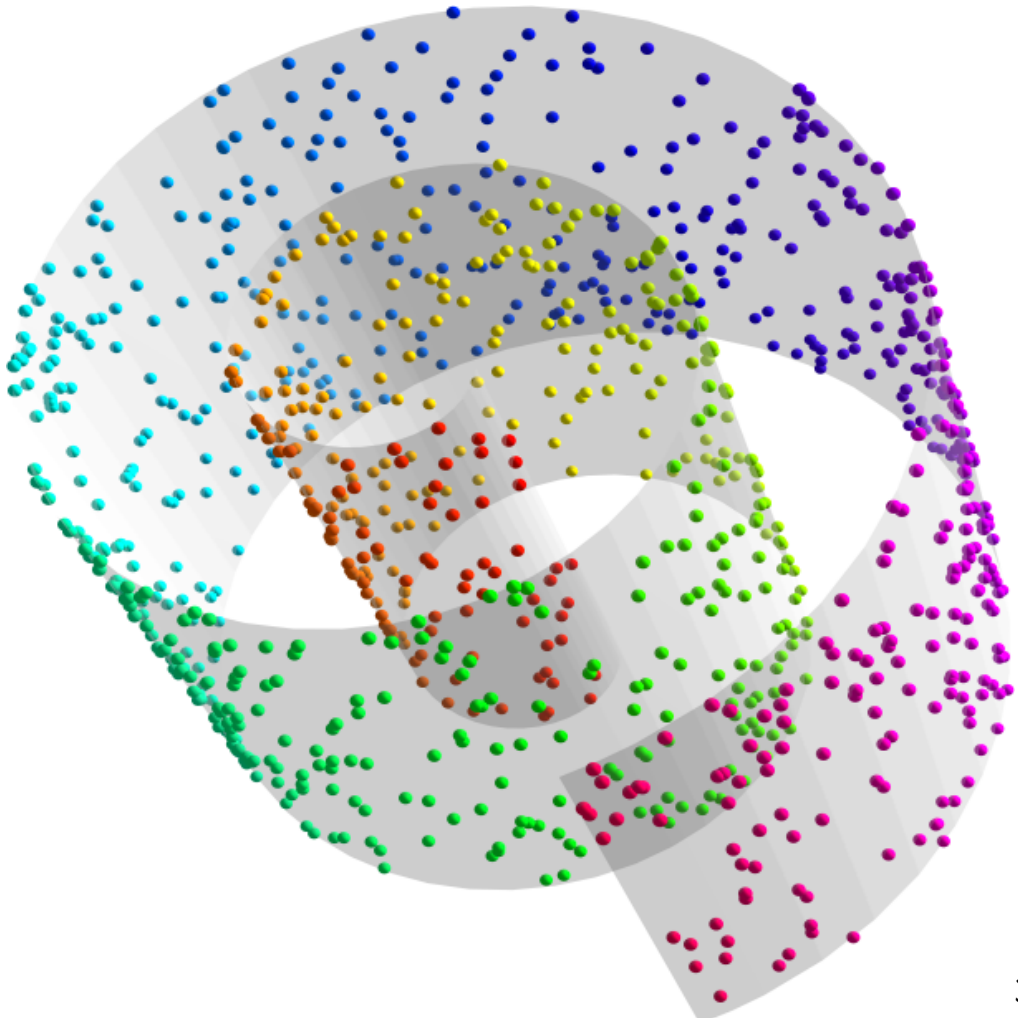
위상 자료 분석(Topological Data Analysis)를 해주는 많은 프로그램들이 있습니다.

- 위상 자료 분석을 해주는 프로그램들 예시: Dionysus, DIPHA, GUDHI, javaPlex, Perseus, PHAT, Ripser, TDA, TDAstats

R 패키지 TDA는 위상 자료 분석을 해주는 C++ 라이브러리의 R 인터페이스(interface)를 제공합니다.

- 웹사이트: `https://cran.r-project.org/web/packages/TDA/index.html`
- 저자: Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, Clément Maria, David Milman, and Vincent Rouvreau.
- R은 통계 계산과 시각화를 위한 프로그래밍 언어입니다.
- R은 개발시간이 짧고, C/C++는 실행시간이 짧습니다.
- R package TDA 는 위상 자료 분석을 해주는 C++ 라이브러리인 GUDHI/Dionysus/PHAT의 R 인터페이스(interface)를 제공합니다.

다양체(manifold)는 국소적으로 유클리드 공간을 닮은 저차원 기하 구조입니다.



3

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

추정량(estimator)의 최대위험(maximum risk)은 추정량의 최악의 경우에 오차의 기대값(expected error)입니다.

- 추정량(estimator) $\hat{\theta}_n$ 의 최대위험(maximum risk)은 최악의 경우에 추정량 $\hat{\theta}_n$ 이 만들어낼 수 있는 오차의 기대값(expected error)입니다.
- $$\sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[\ell \left(\hat{\theta}_n(X), \ \theta(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$는 고정된 분포 P에서 추출하고, P는 확률분포의 집합 $\mathcal{P}$에 속합니다.
  - 추정량 $\hat{\theta}_n$ 은 자료 X의 임의의 함수입니다.
  - 손실함수 $\ell(\cdot, \cdot)$는 추정량 $\hat{\theta}_n$의 오차를 잽니다.

미니맥스 위험(minimax risk)은 모수(parameter) 추정의 통계적 어려움을 묘사합니다.

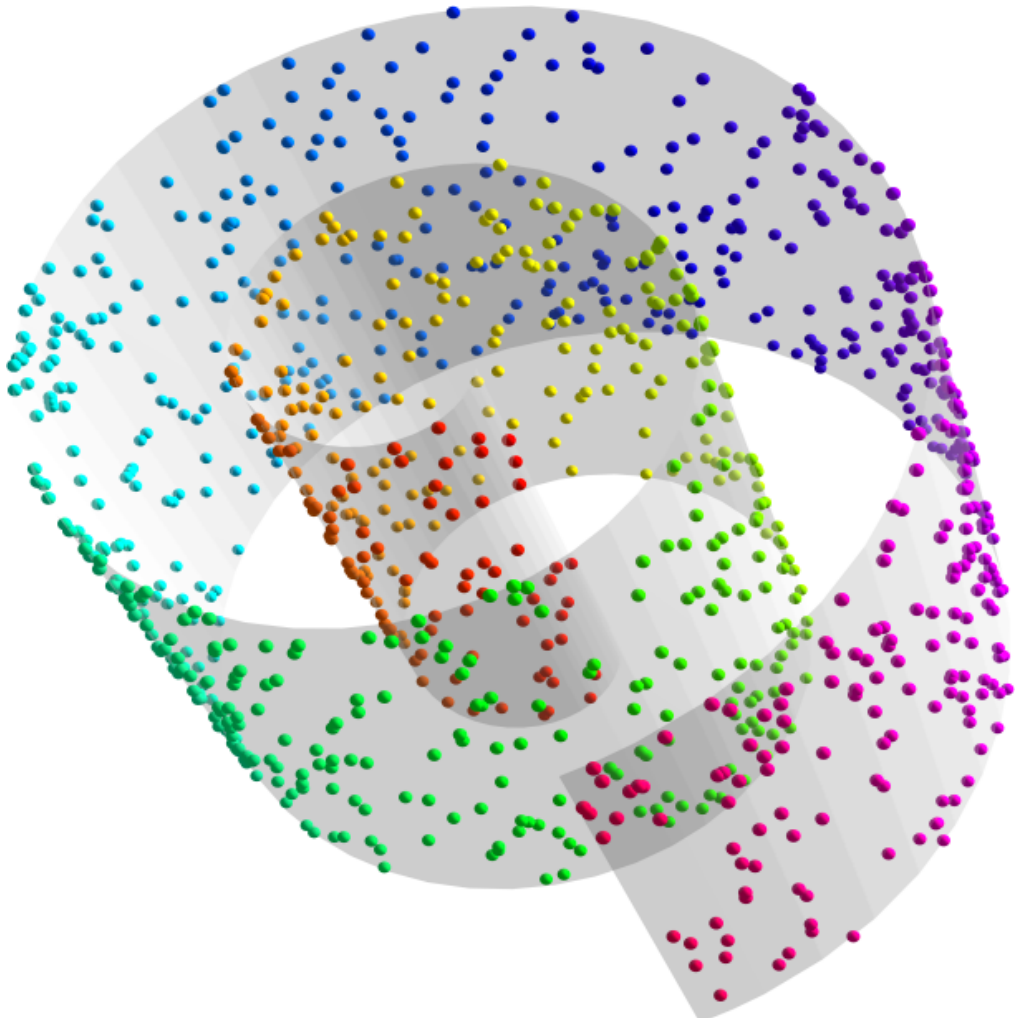
- 미니맥스 위험(minimax risk) R_n은 최악의 경우에도 잘 작동하는 추정량(estimator)의 위험(risk)입니다. 이를 표본크기(sample size)의 함수로 봅니다.
- $$R_n = \inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[\ell \left(\hat{\theta}_n(X), \ \theta(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$는 고정된 분포 P에서 추출하고, P는 확률분포의 집합 $\mathcal{P}$에 속합니다.
  - 추정량 $\hat{\theta}_n$ 은 자료 X의 임의의 함수입니다.
  - 손실함수 $\ell(\cdot, \cdot)$는 추정량 $\hat{\theta}_n$의 오차를 잽니다.

다양체(manifold)의 기하학적 모수(paramter) 추정의 통계적 어려움을 미니맥스 위험(minimax risk)으로 잽니다.

- Minimax Rates for Estimating the Dimension of a Manifold (Kim, Rinaldo, Wasserman, 2019)
- The Origin of the Reach: Better Understanding Regularity Through Minimax Estimation Theory (Aamari, Kim, Chazal, Michel, Rinaldo, Wasserman, 2019)

다양체를 배우기 전에 다양체의 내재적 차원을 추정해야 합니다.

- 대부분의 다양체 학습 알고리즘에서 다양체의 내재적 차원을 입력으로 넣어야 합니다.
- 내재적 차원이 미리 알려진 경우는 드물고, 따라서 학습해야 합니다.



차원 추정의 미니맥스 위험(minimax risk)

- $$R_n = \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[1 \left(\hat{\dim}_n(X) \neq \dim(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$는 고정된 분포 P에서 추출하고, P는 확률분포의 집합 $\mathcal{P}$에 속합니다.
  - 추정량 $\hat{\dim}_n$ 은 자료 X의 임의의 함수입니다.
  - $0-1$ 손실함수를 사용합니다, 따라서 모든 $x, y \in \mathbb{R}$에 대해, $\ell(x, y) = 1(x \neq y)$입니다.

차원 추정의 미니맥스 위험(minimax risk): 먼저 차원이 d_1 vs d_2일 때를 생각합니다.

- $$R_n = \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[1 \left(\hat{\dim}_n(X) \neq \dim(P) \right) \right]$$
  - $X = (X_1, \cdots, X_n)$는 고정된 분포 P에서 추출하고, P는 확률분포의 집합 $\mathcal{P} = \mathcal{P}^{d_1} \cup \mathcal{P}^{d_2}$에 속합니다, 여기서 $\mathcal{P}^d$는 d-차원 확률분포의 집합입니다.
  - 추정량 $\hat{\dim}_n$ 은 자료 X의 임의의 함수입니다.
  - $0-1$ 손실함수를 사용합니다, 따라서 모든 $x, y \in \mathbb{R}$에 대해, $\ell(x, y) = 1(x \neq y)$입니다.

차원 추정의 미니맥스 위험(minimax risk)

Theorem

(Proposition 16 and 17)

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

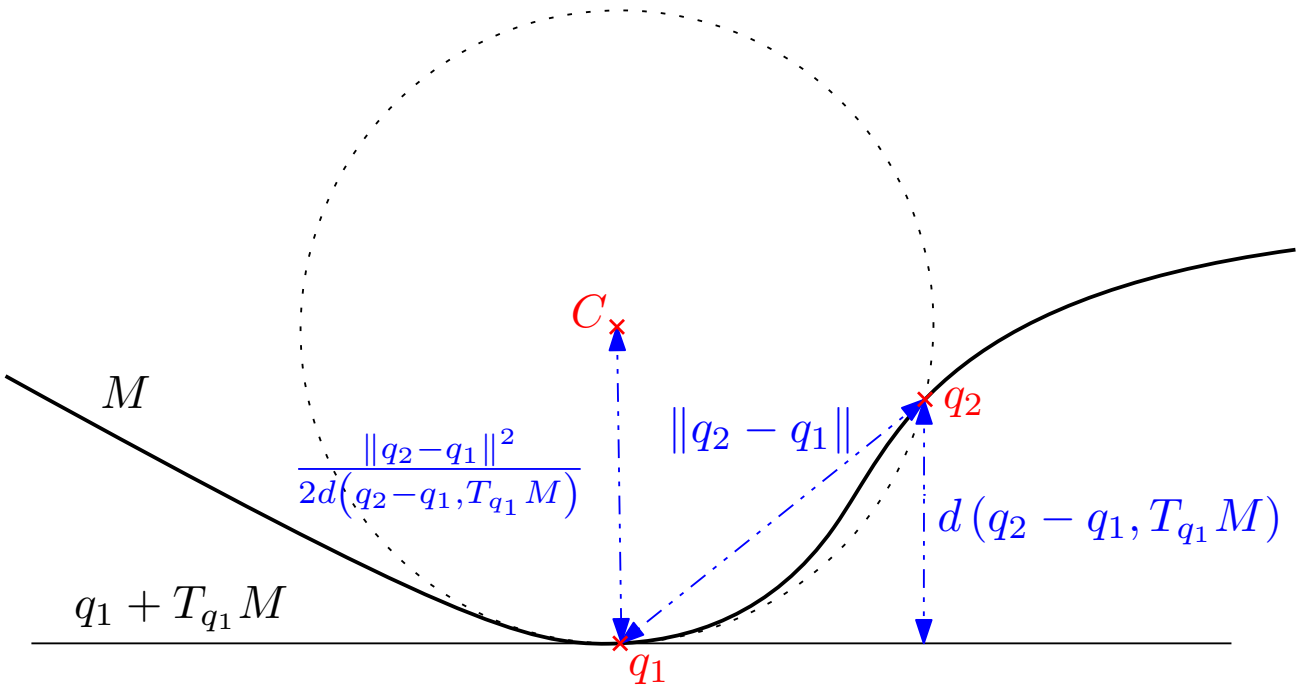
reach는 다양체(manifold) 위에서 구를 수 있는 공의 최대 반지름입니다.

Definition

$M \subset \mathbb{R}^m$ 이 다양체(manifold)일 때, M의 reach($\tau(M)$으로 표기)는 다음과 같이 정의합니다:

$$\tau(M) = \inf_{q_2 \neq q_1 \in M} \frac{\|q_2 - q_1\|_2^2}{2d(q_2 - q_1,\ T_{q_1}M)},$$

여기서 T_aM는 M의 a에서의 접공간(tangent space)입니다.



reach는 많은 기하 추정 문제에서 정칙성(regularity)를 나타내는 모수(parameter)입니다.

- reach는 다음과 같은 문제에서 중요한 모수(parameter)입니다:
  - 차원 추정
  - 호몰로지(homology) 추정
  - 부피(volume) 추정
  - 다양체(manifold) 군집화
  - 확산 사상(diffusion map) 추정

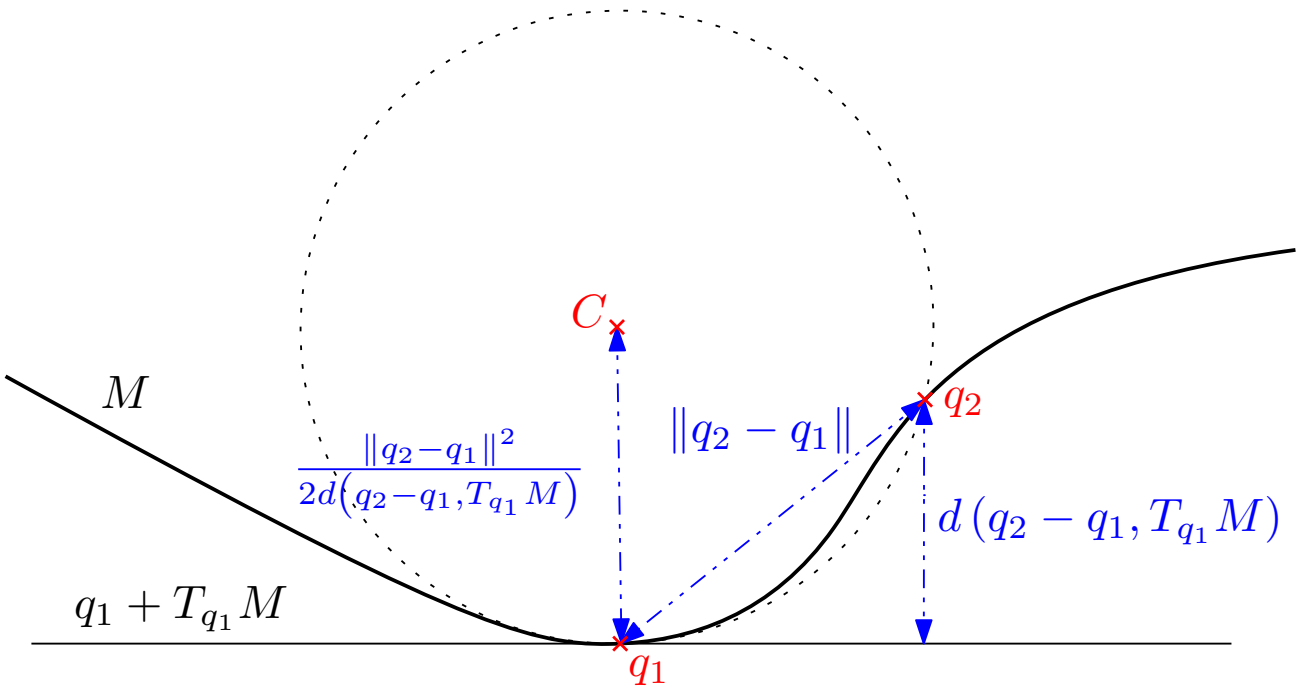
reach 추정의 미니맥스 위험(minimax risk)

- $$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[\left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}_n(X)} \right|^q \right]$$
  - $X = (X_1, \cdots, X_n)$는 고정된 분포 P에서 추출하고, P는 확률분포의 집합 $\mathcal{P}$에 속합니다.
  - 추정량 $\hat{\tau}_n$은 자료 X의 임의의 함수입니다.
  - 역-l_q 손실함수를 사용합니다, 따라서 모든 $x, y \in \mathbb{R}$에 대해, $\ell(x, y) = \left| \frac{1}{x} - \frac{1}{y} \right|^q$입니다.

reach 추정량 $\hat{\tau}_n$을 점구름 자료(point cloud)에서 구를 수 있는 공의 최대 반지름으로 정의합니다.

- 자료 $X = (X_1, \ldots, X_n)$가 주어졌을 때, reach 추정량 $\hat{\tau}_n$은 다음과 같은 삽입(plugin) 추정량입니다:

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



reach 추정의 미니맥스 위험(minimax risk)

Theorem

(Theorem 5.1 and Proposition 5.6)

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

참조문헌 I

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the Reach of a Manifold. *ArXiv e-prints*, May 2019.

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: $10.3389/\mathrm{frai}.2021.667963$. 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: $10.1090/\mathrm{mbk}/069$. URL `https://doi.org/10.1090/mbk/069`. An introduction.

참조문헌 II

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.

참조문헌 III

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`.

Jisu Kim, Alessandro Rinaldo, and Larry Wasserman. Minimax Rates for Estimating the Dimension of a Manifold. *ArXiv e-prints*, May 2019.

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.

참조문헌 IV

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.

감사합니다!

호몰로지(Homology)와 Persistent Homology

Persistent Homology를 통계적으로 추정하기

위상 자료 분석(Topological Data Analysis)을 기계학습에 응용
- Persistence Landscape를 이용하여 특성(Feature) 만들기
- Circular Coordinates를 이용하여 특성(Feature) 만들기

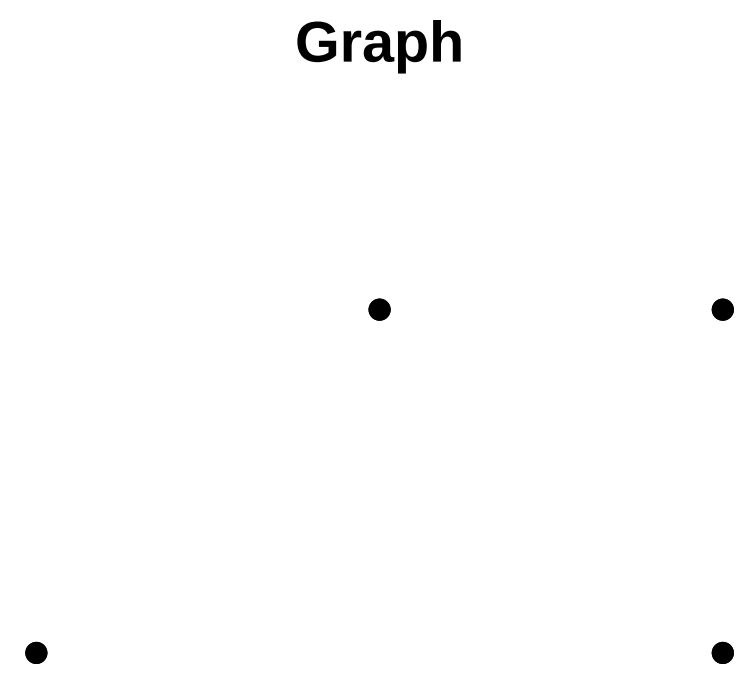
R 패키지 TDA: 위상 자료 분석을 위한 통계 계산 도구
- 다양체(manifold)에서의 표본 추출, 거리 함수, 밀도 함수
- Persistent Homology와 Persistence Landscape
- Persistence Homology와 Persistence Landscape의 통계적 추정

다양체(manifold)의 차원 추정의 미니맥스 위험(minimax risk)
- 정칙성(regularity) 조건
- 상한(upper bound)
- 하한(lower bound)
- 일반적인 경우의 상한과 하한

다양체(manifold)의 reach 추정의 미니맥스 위험(minimax risk)
- reach와 기하 구조
- reach 추정량과 분석
- 미니맥스 추정량

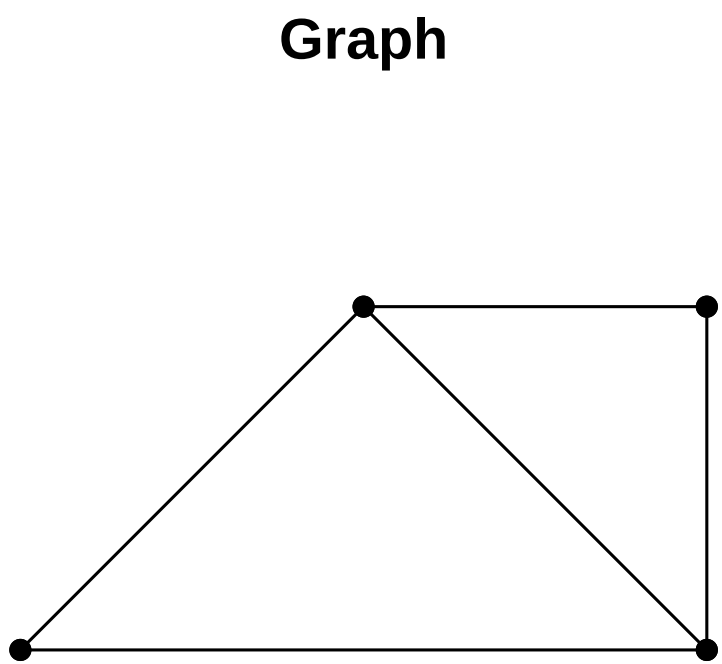
그래프(graph)는 꼭지점(vertex)과 변(edge)로 이루어진 이산 구조입니다.

- 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.



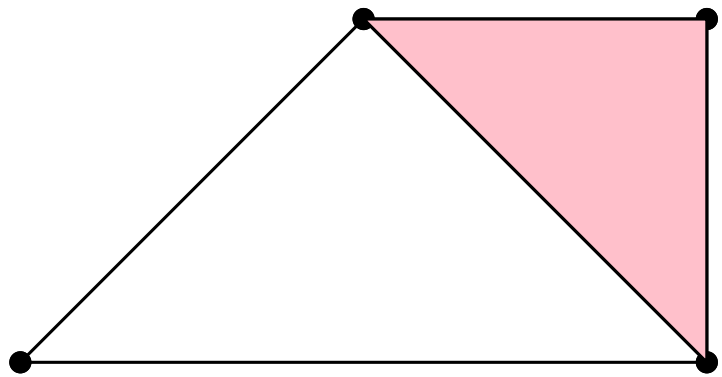
단체 복합체(Simplicial complex)는 고차원으로 일반화한 그래프입니다.

- 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.

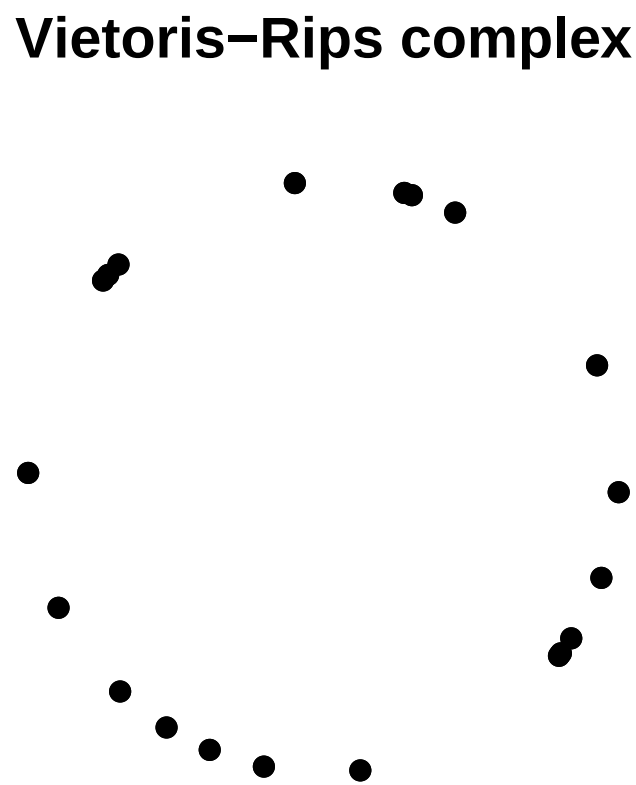
Simplicial complex



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:

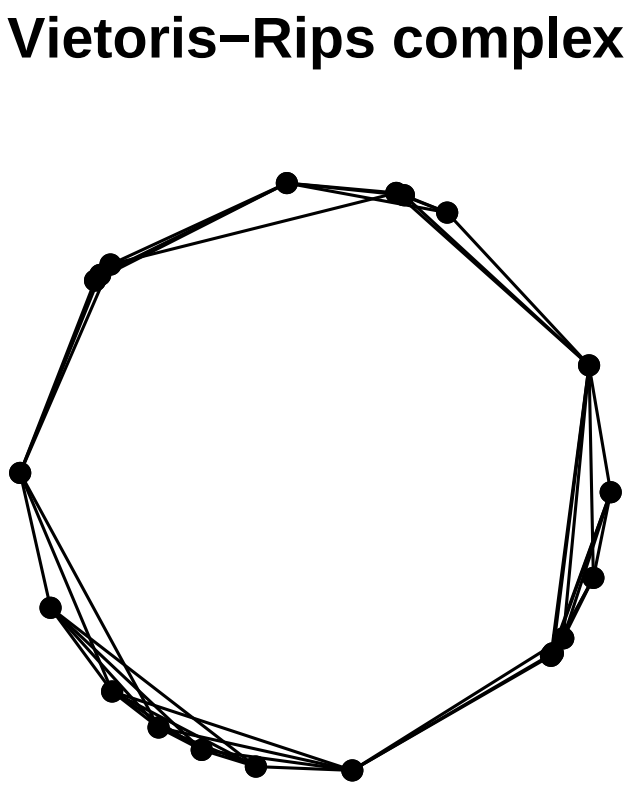
$$\mathrm{Rips}(\mathcal{X}, r) = \{\{x_1, \ldots, x_k\} \subset \mathcal{X} : \, d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k\}.$$



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:

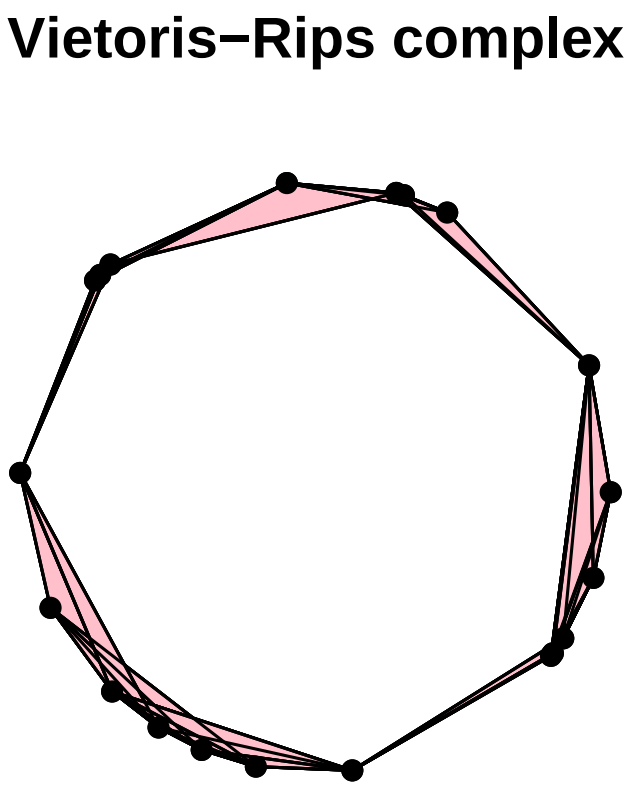
$$\mathrm{Rips}(\mathcal{X}, r) = \{\{x_1, \ldots, x_k\} \subset \mathcal{X} : \, d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k\}.$$



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:

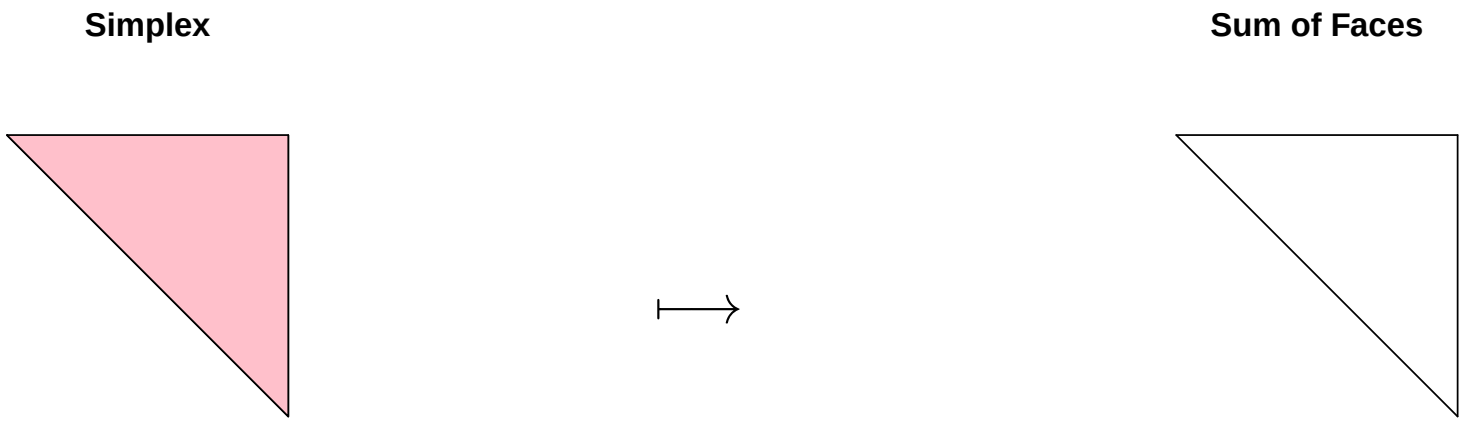
$$c + c' = \sum_{i=1}^{p} (a_i + a_i')\sigma_i, \qquad \lambda \cdot c = \sum_{i=1}^{p} (\lambda a_i)\sigma_i.$$

- 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:

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

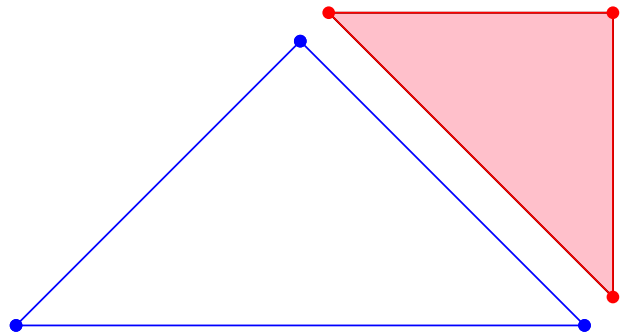
호몰로지(homology)는 cycle을 boundary로 자른 몫공간(quotient space)입니다.

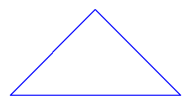
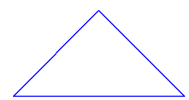
- K의 k-th 호몰로지 $H_k(K)$는 k-cycle $Z_k(K)$를 k-boundary $B_k(K)$로 자른 몫공간(quotient space)입니다:

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

- 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.

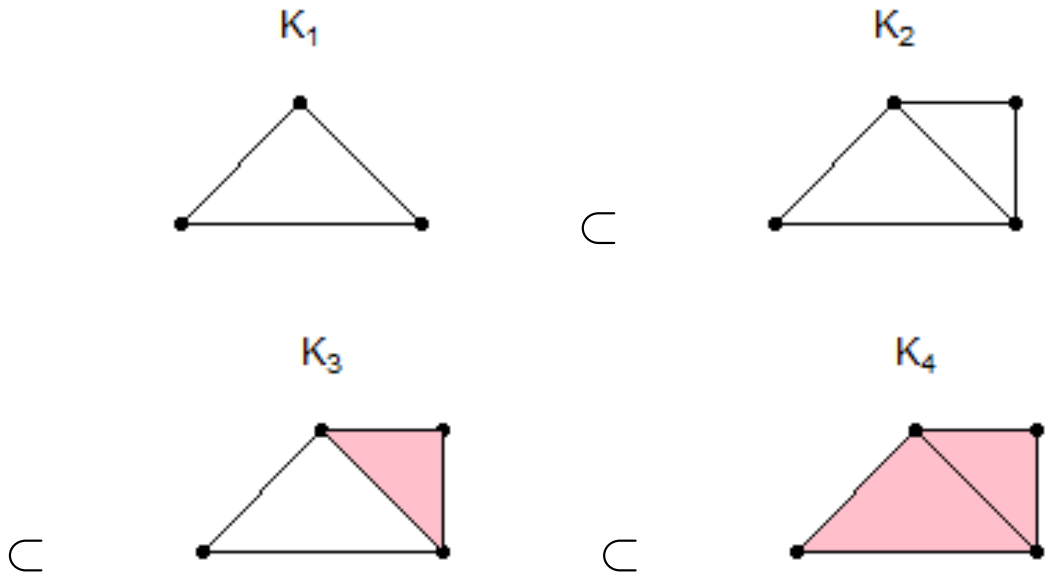


- $Z_1(K) = \ker \partial_1 = (\mathbb{Z}/2\mathbb{Z})^2 =<$  ,  $>$
- $B_1(K) = \mathrm{im}\partial_2 = \mathbb{Z}/2\mathbb{Z} =<$  $>$
- $H_1(K) = Z_1(K)/B_1(K) = \mathbb{Z}/2\mathbb{Z} =<$  $>$, $\beta_1(K) = 1$

filtration은 증가하는 단체 복합체들의 모임입니다.

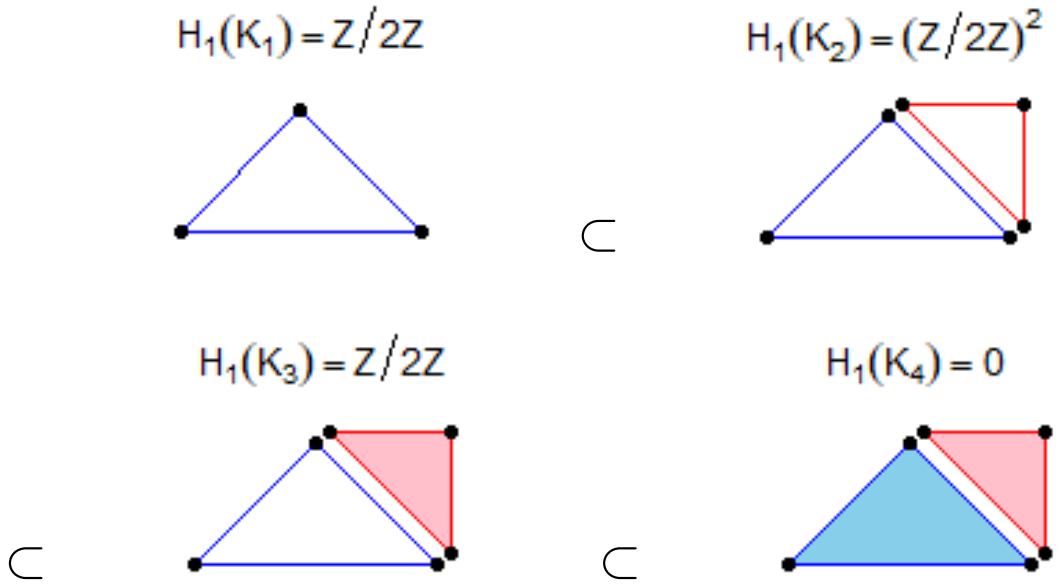
- 단체 복합체 K가 있을 때, filtration $\mathcal{F} = \{K_a\}_{a \in \mathbb{R}}$는 다음을 만족하는 K의 부분 복합체(subcomplex) K_a들의 모임입니다:

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



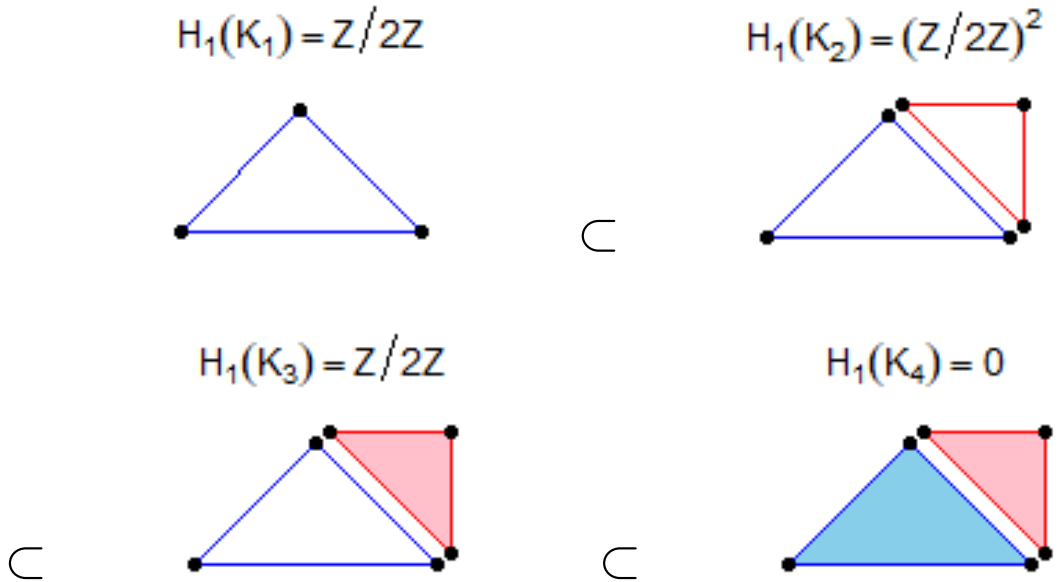
Persistent Homology는 filtration에서 호몰로지가 어떻게 변화하는지 추적합니다.

- 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:

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

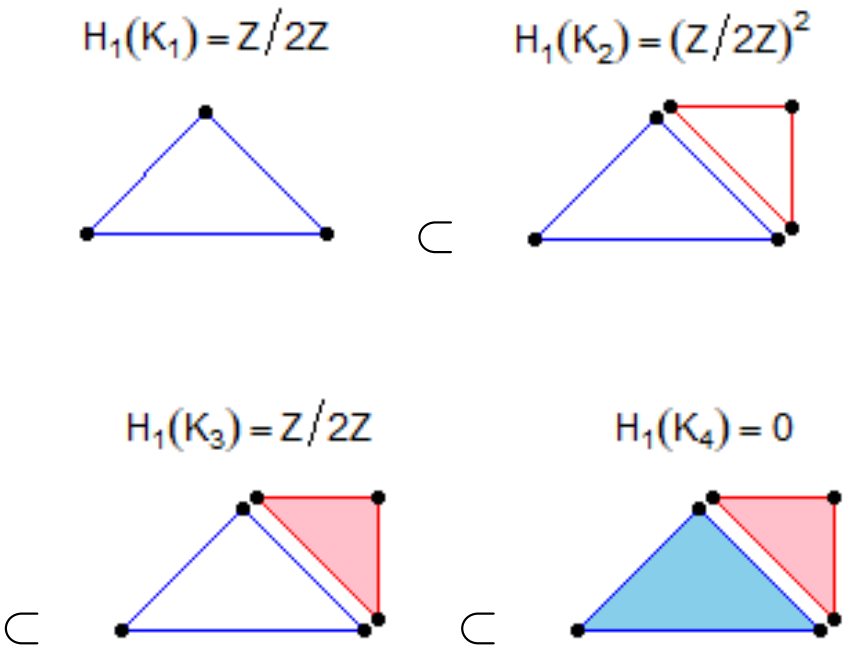
- $(\mathbb{R} \cup \{\infty\})^2$ 위에 점 (a_i, a_j) 를 multiplicity $\mu_k^{a_i,a_j}$로 찍으면 persistence diagram 이 됩니다.

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.



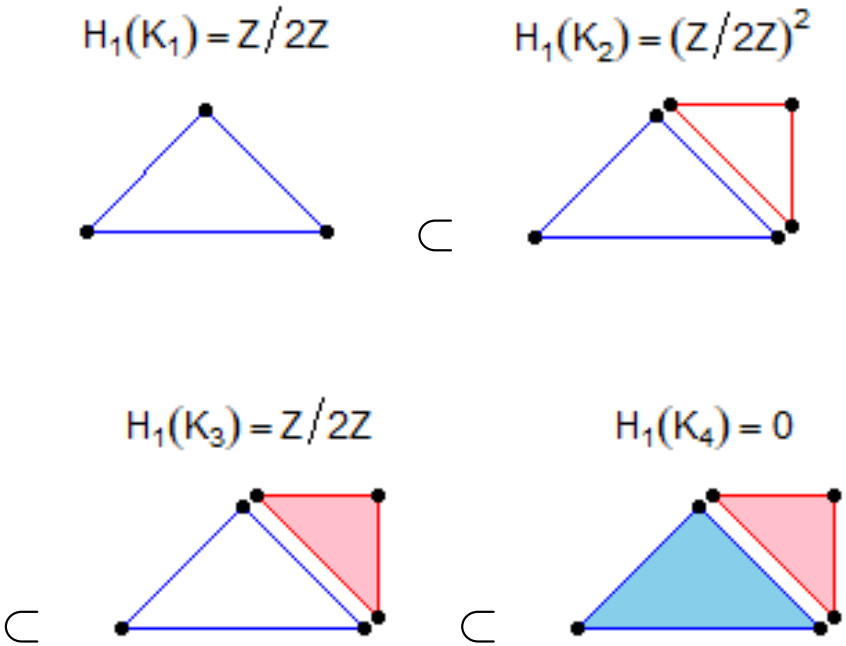
$\mu_1^{i,j}$	$i=1$	$i=2$	$i=3$	$i=4$
$j=4$	0	0	0	0
$j=3$	1	1	1	
$j=2$	1	2		
$j=1$	1			

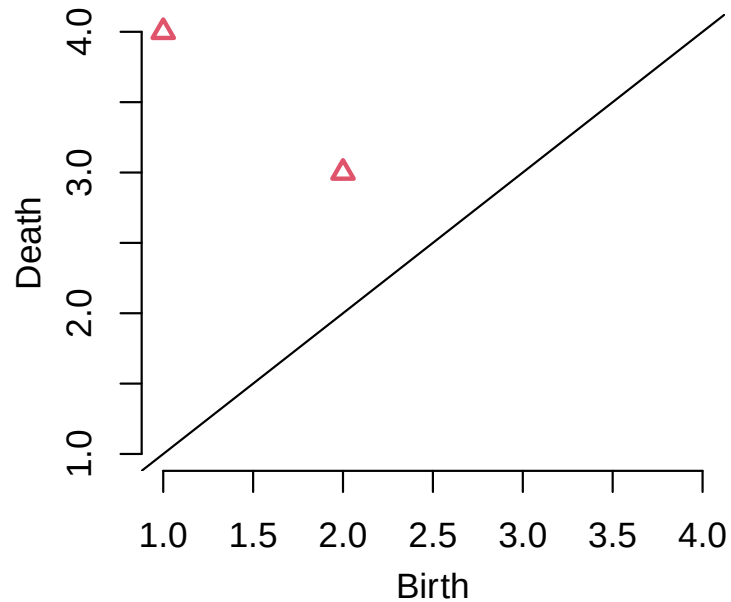
Persistence Diagram 은 Persistent Homology 를 평면 위의 점들로 나타냅니다.

- 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.

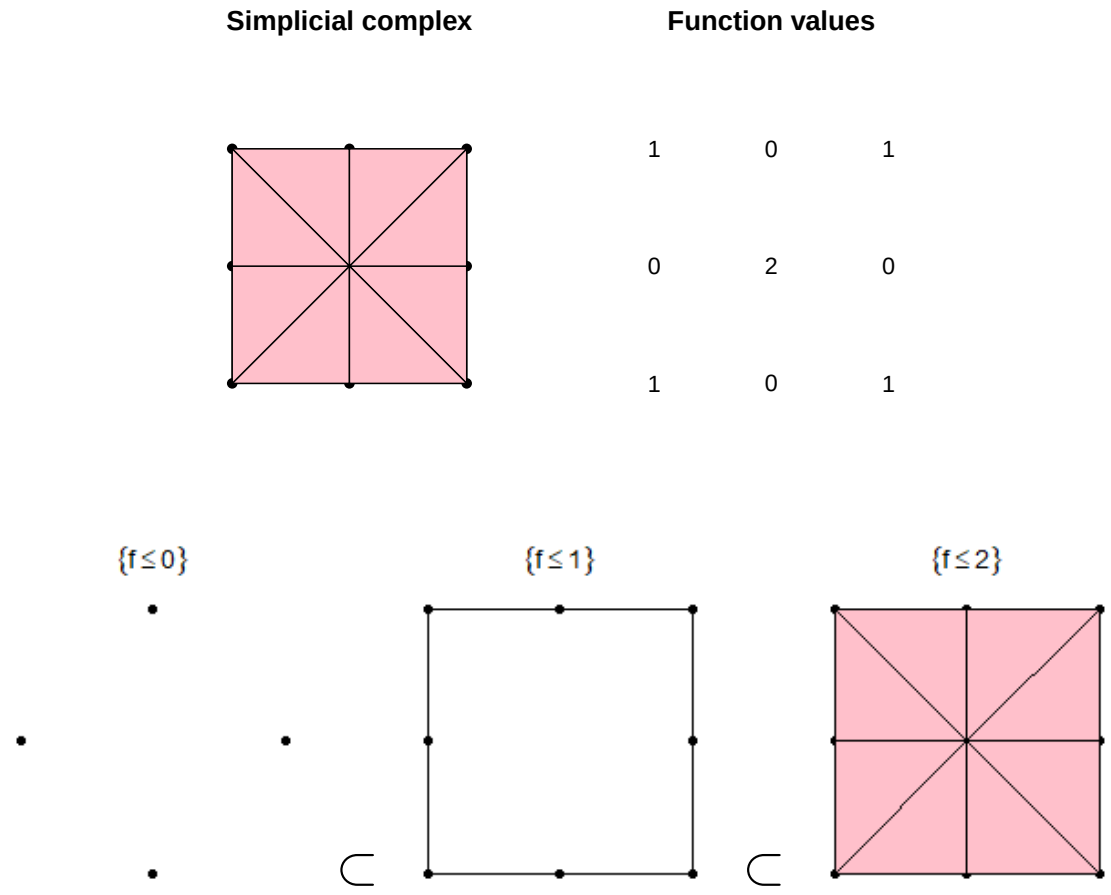




함수의 레벨집합으로부터 filtration을 만들 수 있습니다.

- 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:

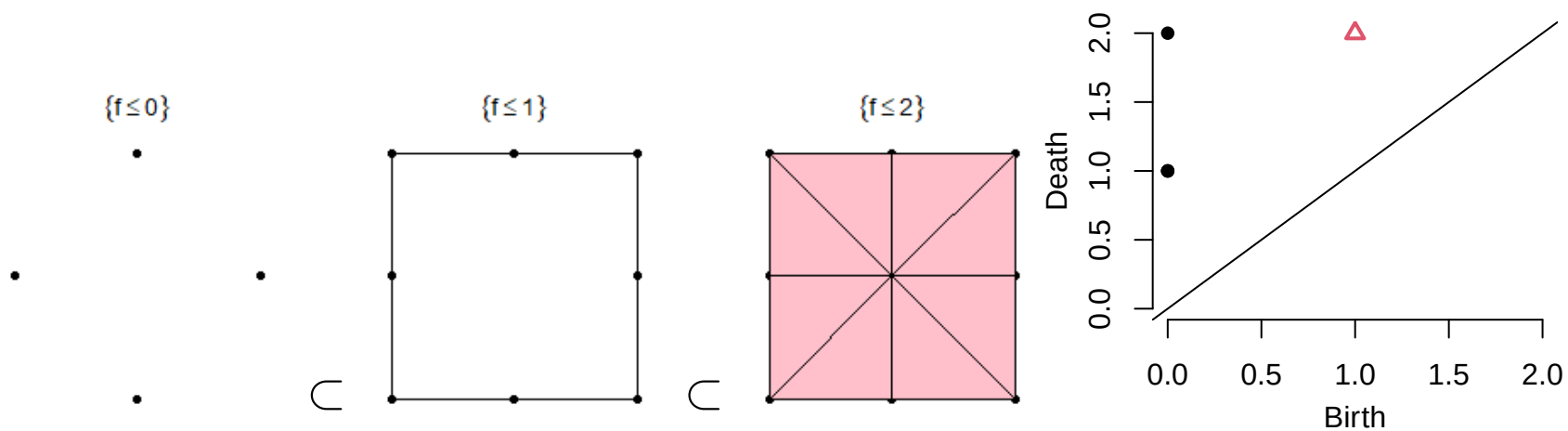
$$\mathrm{super}(f) := \{\{\sigma \in K : f(\sigma) \geq L\}\}_{L \in \mathbb{R}} .$$

함수의 레벨집합으로부터 persistent homology를 계산할 수 있습니다.

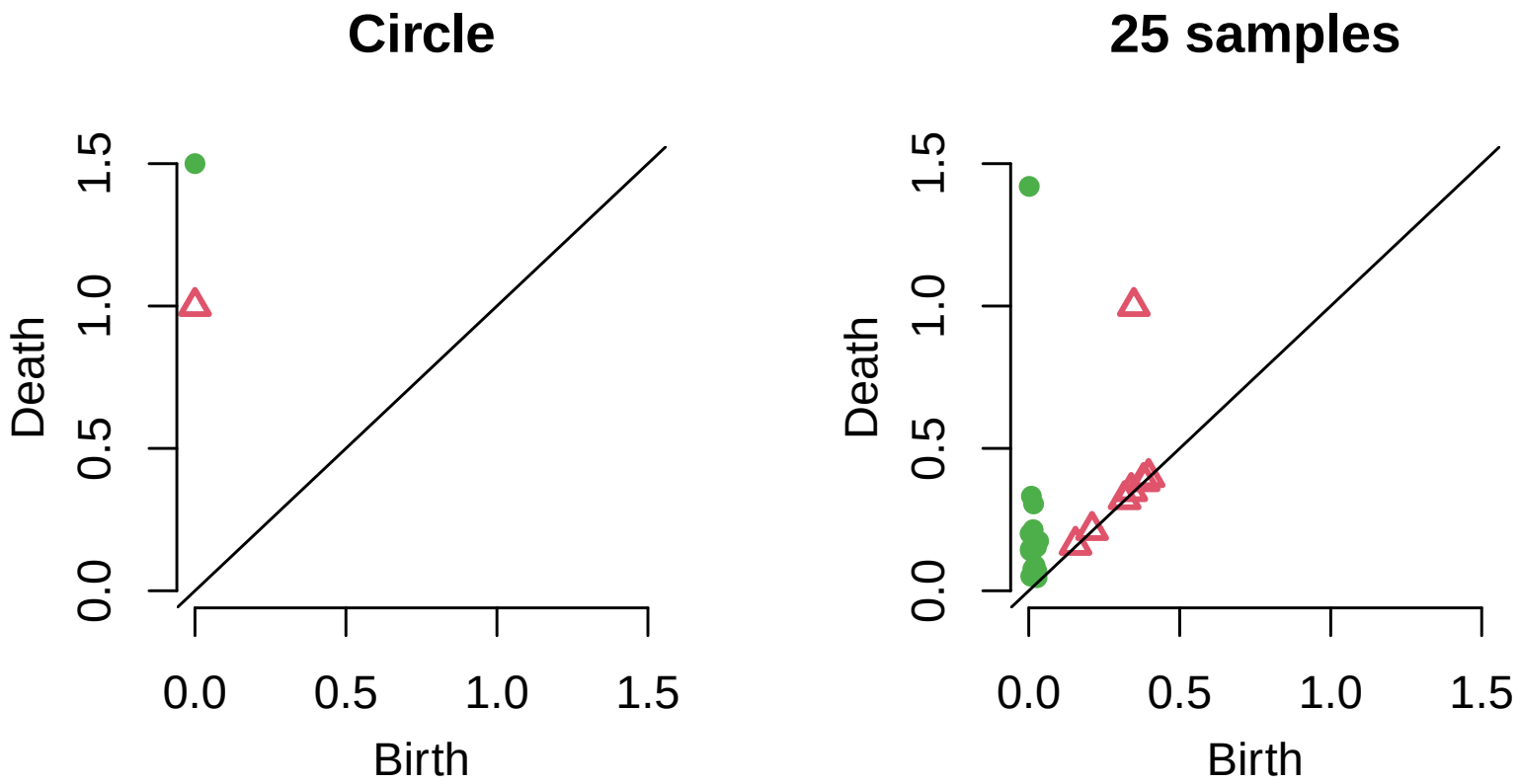
- 단체 복합체 K와 그 위에서 정의된 함수 $f : K \to \mathbb{R}$가 있을 때, f의 아랫레벨(sub-level) filtration $\mathrm{sub}(f)$를 다음과 같이 정의합니다:

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

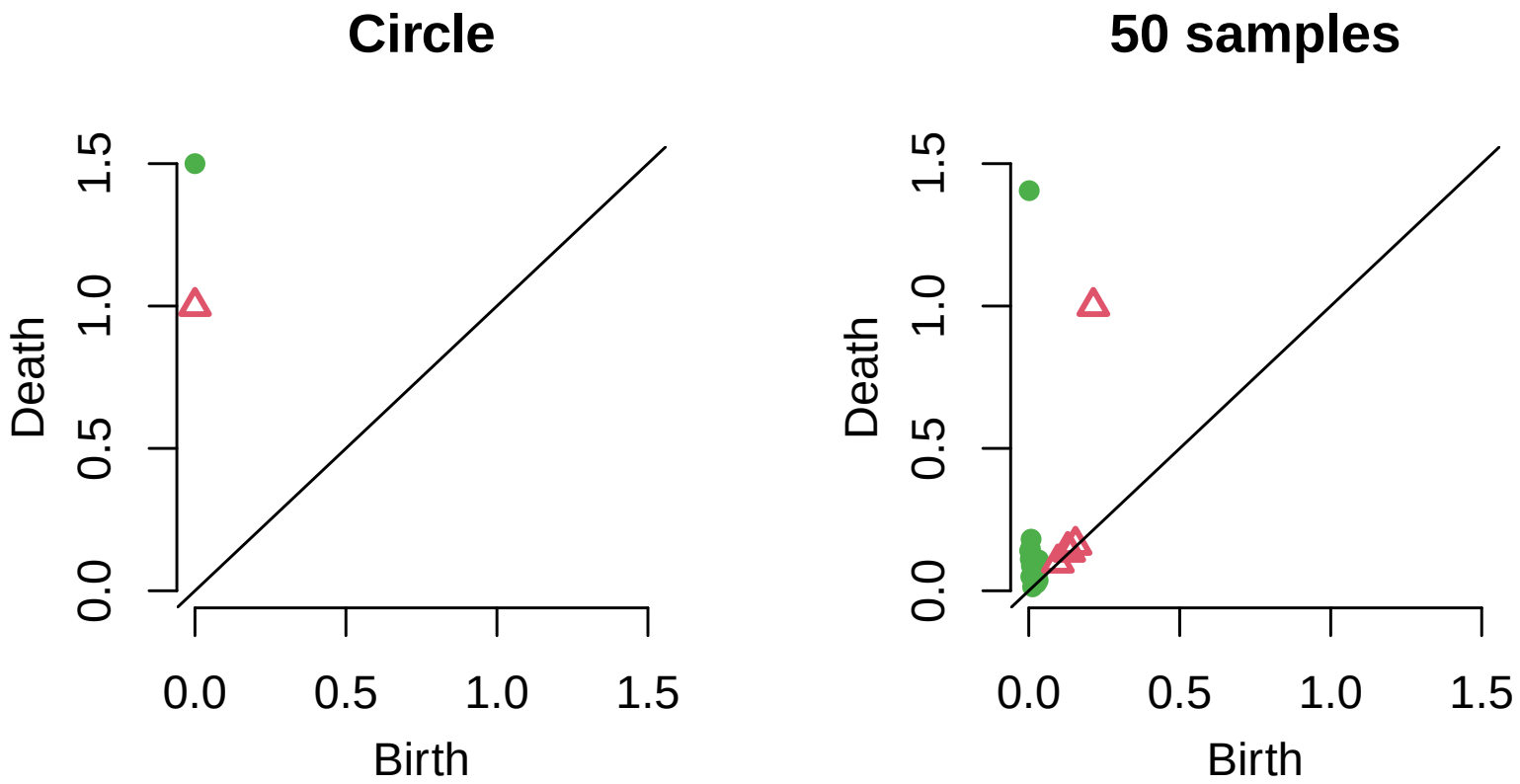
- 그로부터 계산한 persistent homology 또는 persistence diagram을 $Dgm(f)$로 씁니다.



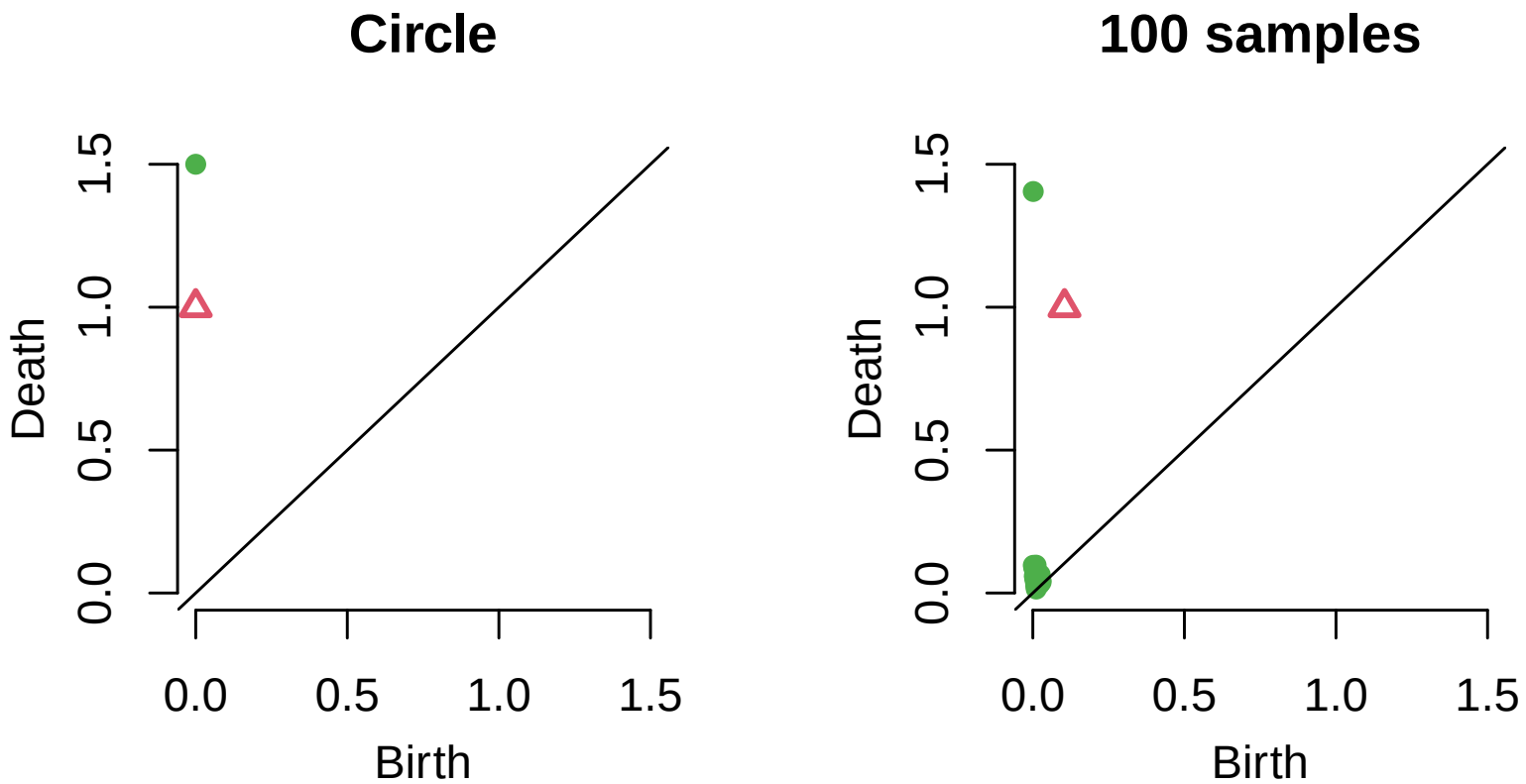
유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.



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



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



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 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:

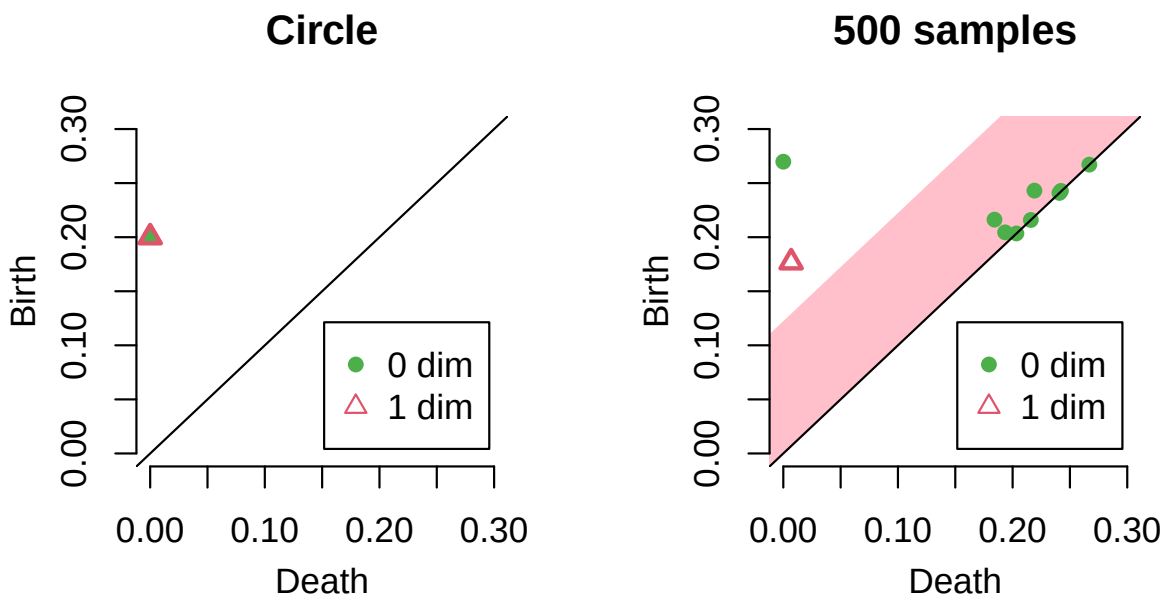
$$\mathbb{P}(W_\infty(Dgm(f_M), Dgm(f_X)) \leq c_n) \geq \mathbb{P}(||f_M - f_X||_\infty \leq c_n) \geq 1 - \alpha,$$

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

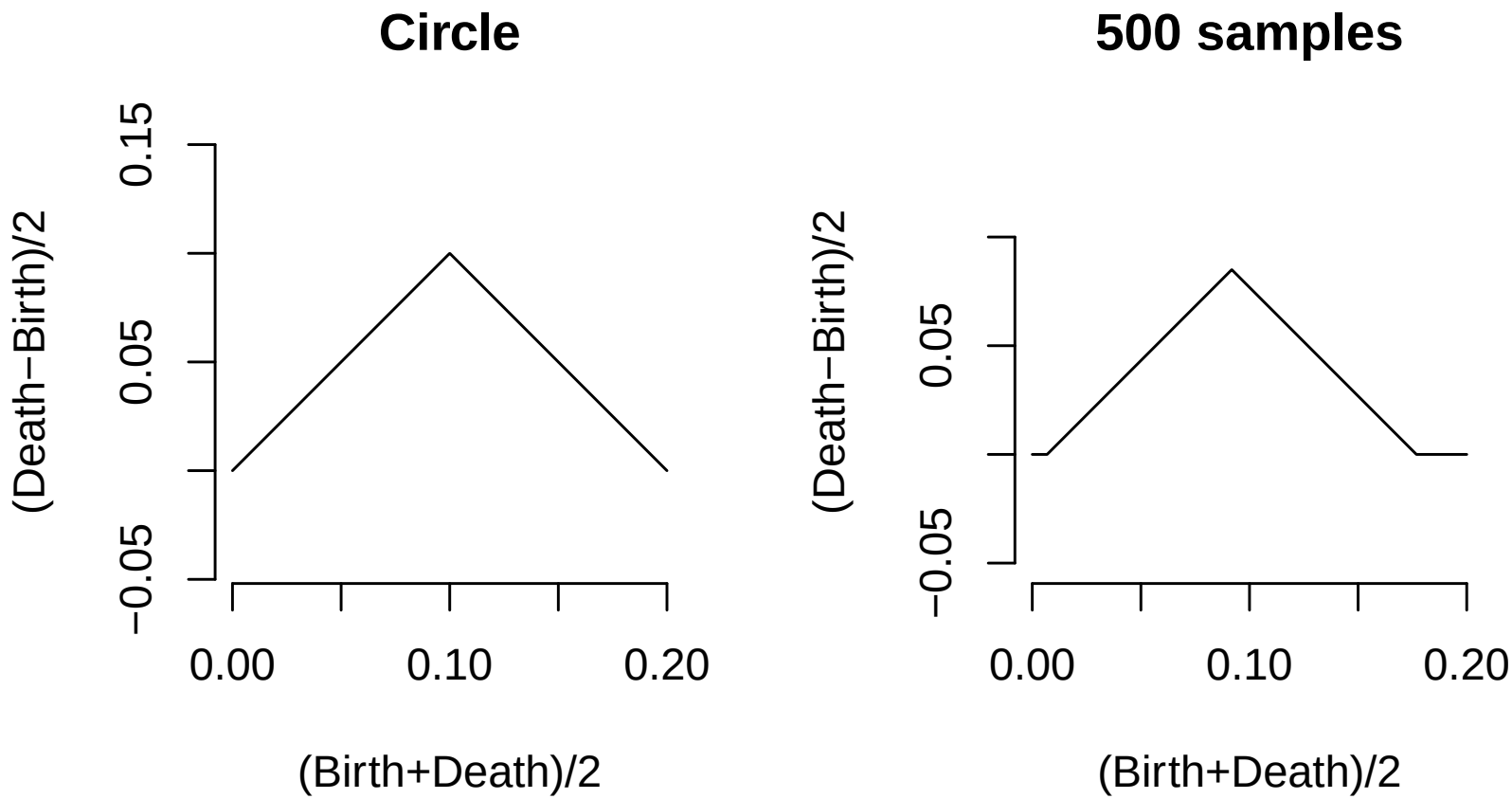
Persistent homology의 신뢰띠는 붓스트랩으로 계산할 수 있습니다.

붓스트랩 알고리즘을 persistent homology에 적용할 수 있다는 것이 증명되었습니다.

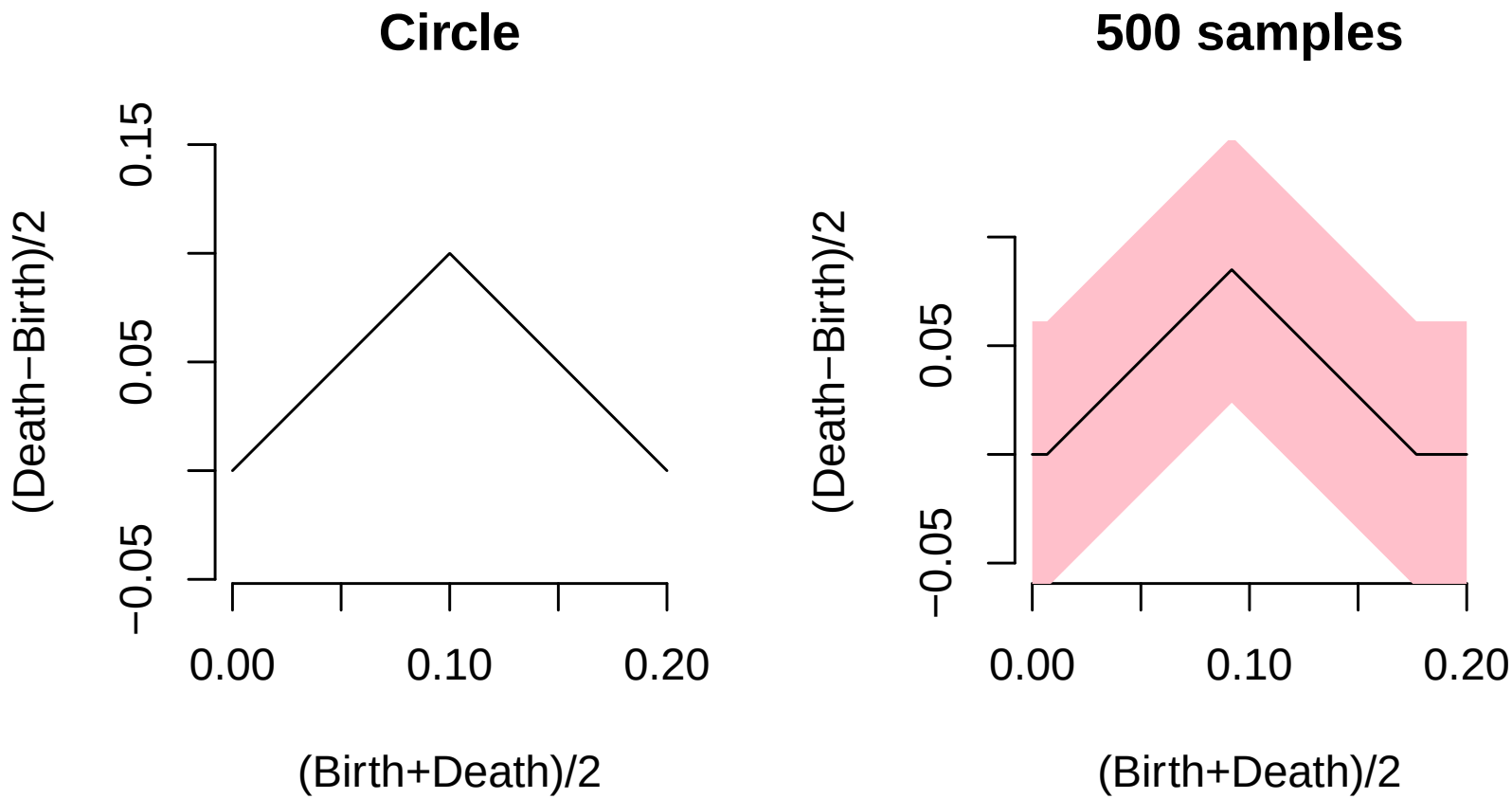
- Fasy et al. [2014b] 이 핵밀도추정(kernel density estimator)에서 보였고,
- Chazal et al. [2014] 이 distance to measure와 kernel distance에서 보였습니다.



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



Persistent homology의 신뢰띠로 Persistence Landscape의 랜덤성을 정량화할 수 있습니다.

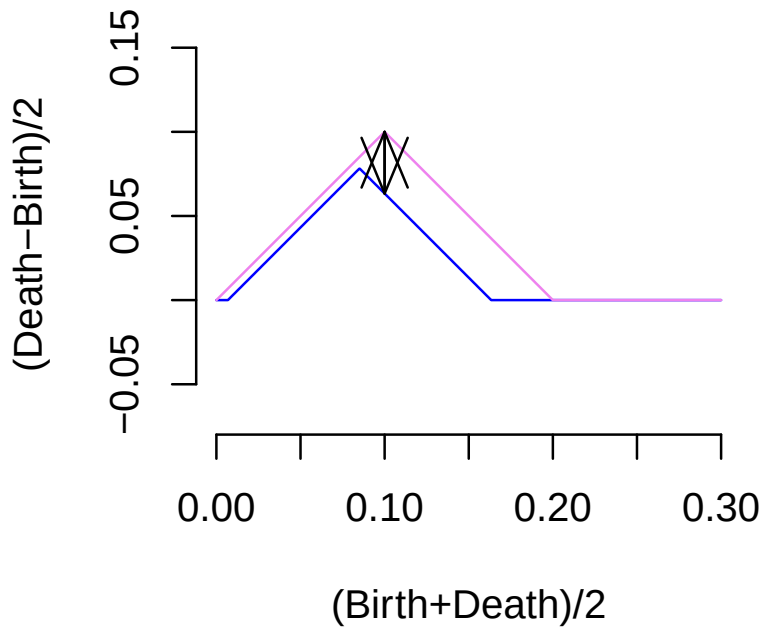


∞-landscape 거리는 persistence landscape 공간에 거리를 줍니다.

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,

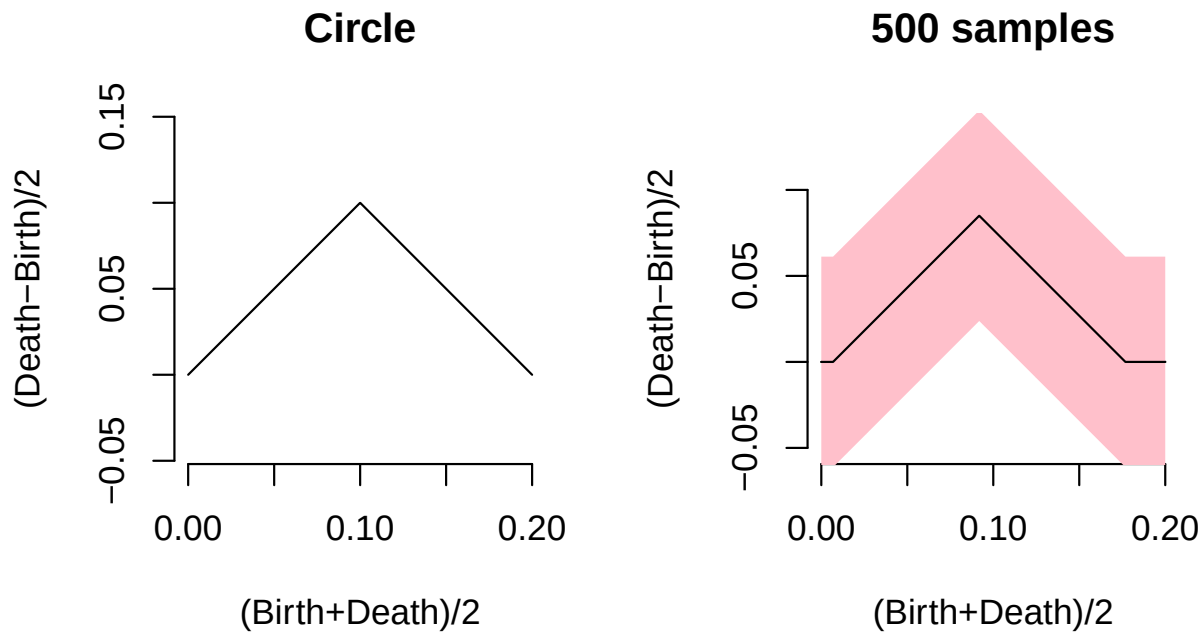
$$\Lambda_\infty(\lambda(f), \lambda(g)) \leq \|f - g\|_\infty.$$

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].

PLLay는 미분 가능(differentiable)합니다.

- 심층학습(deep learning) 모형은 매개변수(parameter)를 역전파(back propagation)으로 배우는데, 이는 경사법(gradient descent)을 층(layer)마다 적용하는 것입니다.
- 심층학습 층이 학습 가능하려면, 층이 미분 가능(differentiable)해야 합니다.

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

PLLay 함수 $S_{\theta,\omega}$는 입력 X에 대해 미분 가능(differentiable)합니다.

PLLay는 안정적(stable)입니다.

- PLLay는 persistence diagram 의 변화에 대해 안정적(stable)입니다:

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

두 persistence diagrams $\mathcal{D}, \mathcal{D}'$에 대해,

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

여기서 W_∞는 bottleneck distance입니다.

PLLay는 안정적(stable)입니다.

- PLLay는 입력 X의 변화에 대해 안정적(stable)입니다:

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

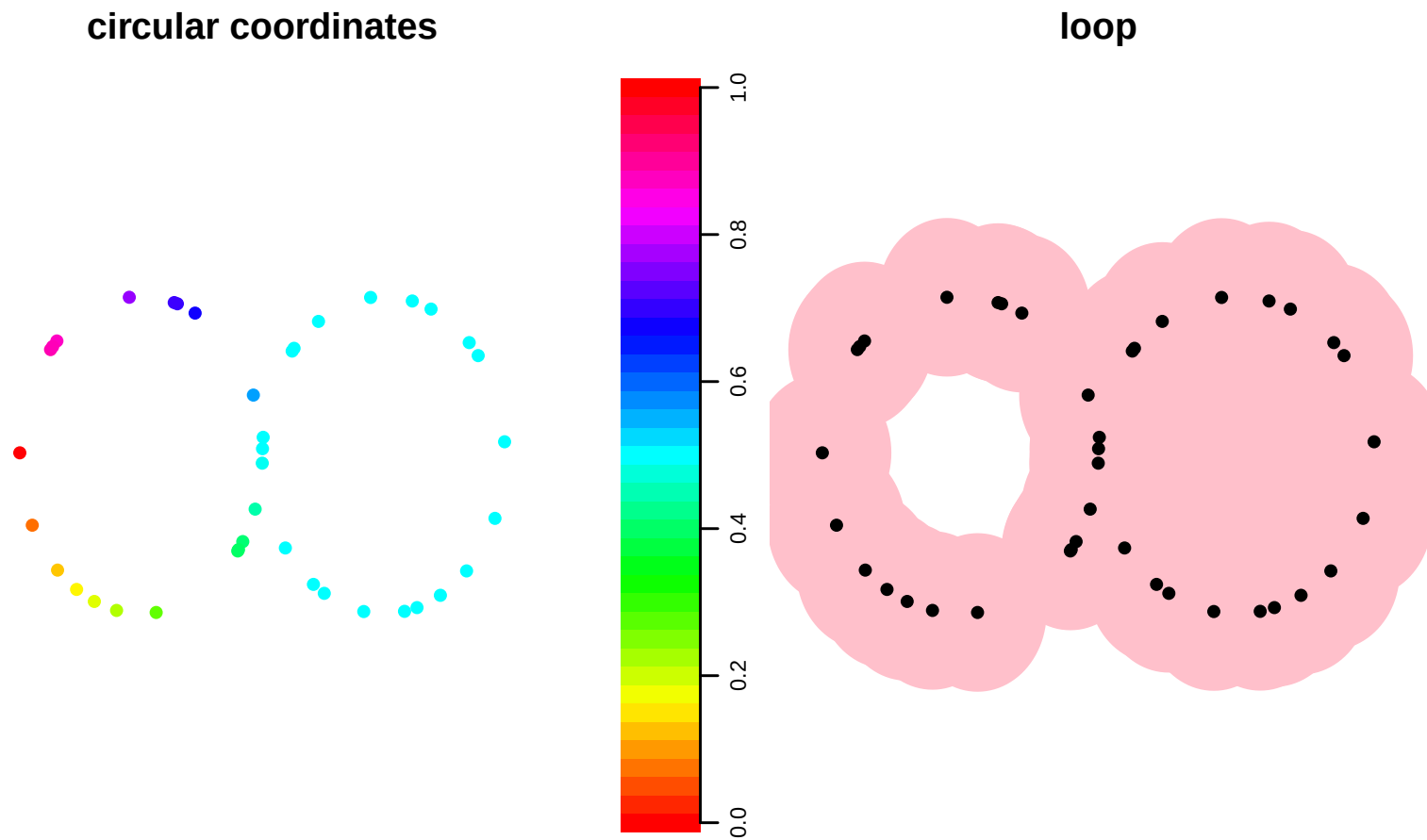
$X \sim P$이고 P_n을 경험적 분포(empirical distribution)으로 놓습니다. $\mathcal{D}_P, \mathcal{D}_X$를 각각 P, X의 persistence diagram으로 놓습니다. 그러면

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

여기서 W_2는 2-Wasserstein distance입니다.

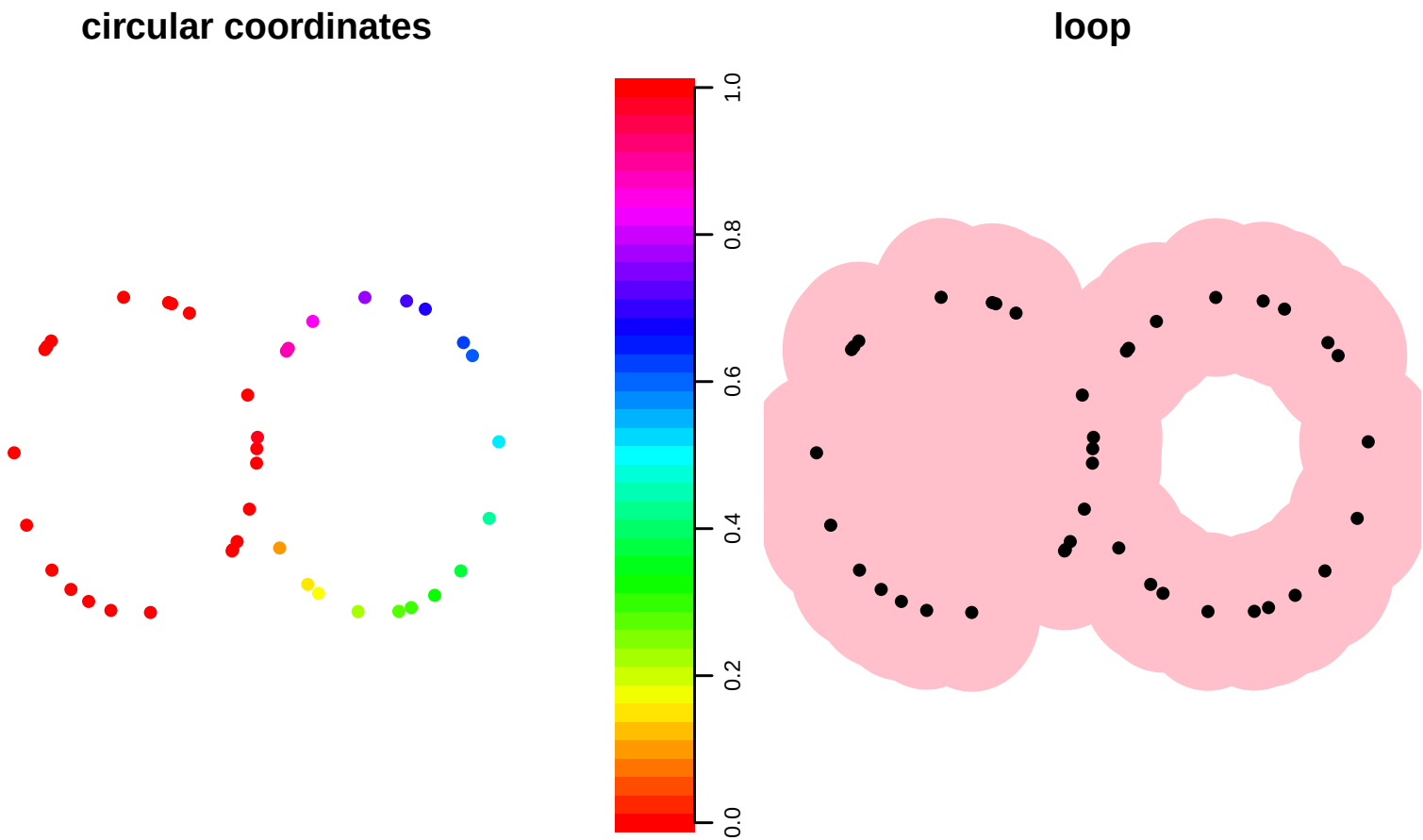
Circular Coordinates 는 자료의 위상 구조를 반영하는 차원 축소 방법입니다.

- circuiar coordinate 는 자료 X 에서 원 S^1 으로 가는 함수입니다.



Circular Coordinates 는 자료의 위상 구조를 반영하는 차원 축소 방법입니다.

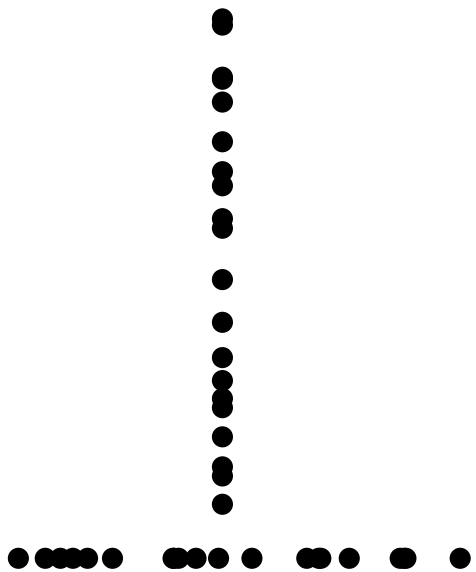
- circuiar coordinate 는 자료 X 에서 원 S^1 으로 가는 함수입니다.



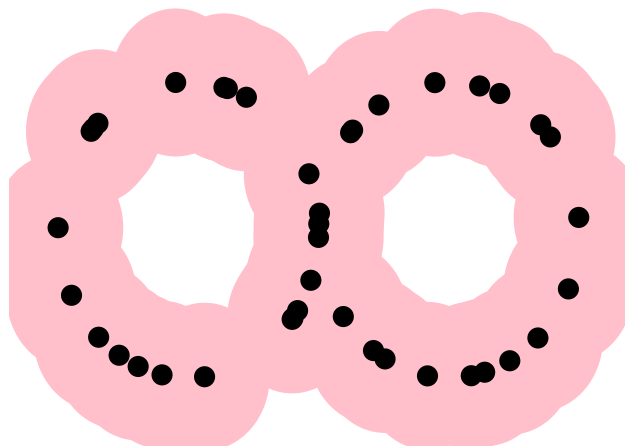
Circular Coordinates 는 자료의 위상 구조를 반영하는 차원 축소 방법입니다.

- circuiar coordinate 는 자료 X 에서 원환면 $\mathbb{T}^k = (S^1)^k$ 으로 가는 함수입니다.

circular coordinates

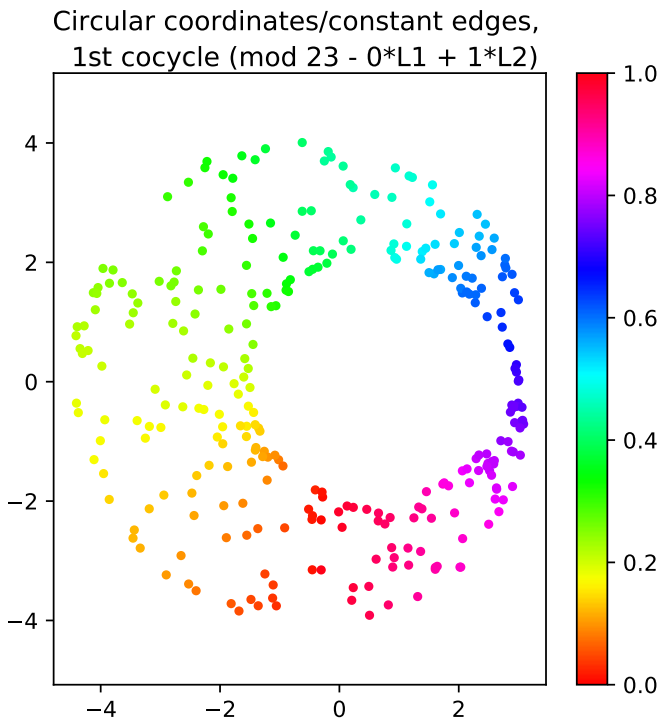


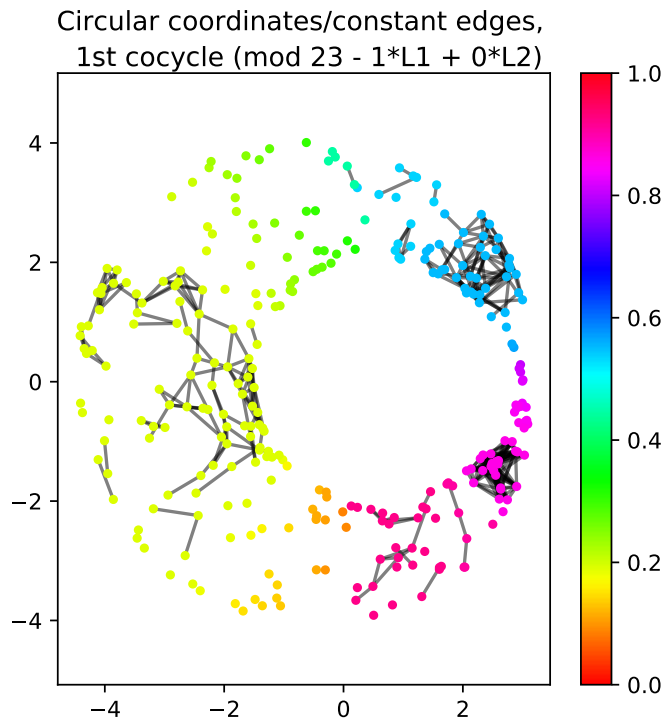
loop



Circular coordinates 를 계산할 때 일반화된 규제 함수 (generalized penalty function)를 사용하면 자료의 위상적인 정보를 더 잘 시각화할 수 있습니다.

- circular coordinates 를 계산할 때, 최적화 문제(optimization problem)를 풉니다.
- L_2 손실(loss)을 L_1 손싱로 바꿈으로써 circuiar coordinate 값이 더 급격하게 바뀌게 할 수 있습니다: 자료의 위상적인 정보를 더 잘 시각화합니다.

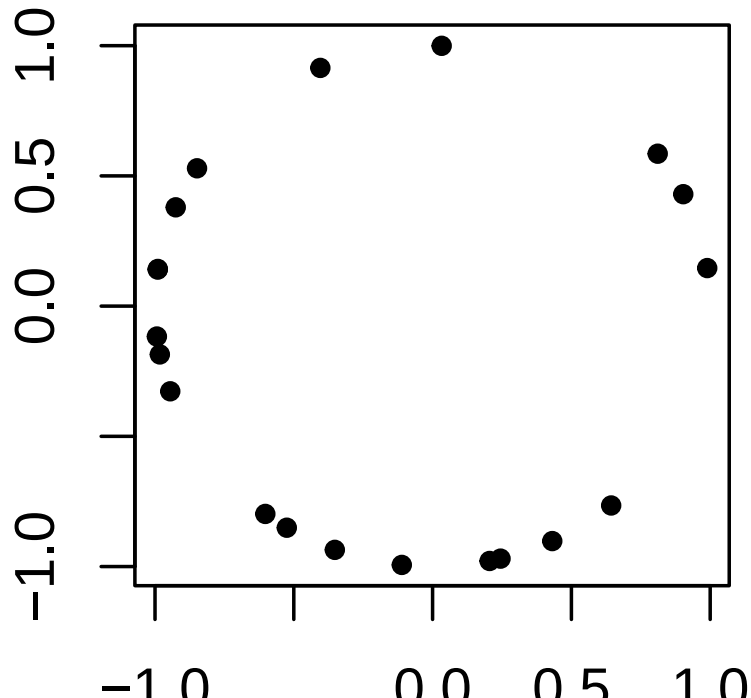




R 패키지 TDA는 원 위에서 표본 추출할 수 있는 함수를 제공합니다.

함수 circleUnif()는 $\mathbb{R}^2$ 상 에 있는 반지름이 r인 원 위의 균등분포에서 n 개의 자료를 생성합니다.

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



R 패키지 TDA는 격자 위에서의 거리 함수와 밀도 함수를 제공합니다.

단위원으로부터 $n = 400$개의 자료가 생성되었고, 격자점들이 있다고 가정합니다.

```
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 패키지 TDA는 격자 위에서의 핵밀도추정(KDE)을 제공합니다.

가우스 핵밀도추정 (Kernel Density Estimator, KDE) $\hat{p}_h : \mathbb{R}^d \to [0, \infty)$는 다음과 같이 정의됩니다:

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

여기서 h는 평활매개변수(smoothing parameter)입니다.
함수 kde()는 격자 위의 점에서 핵밀도추정(KDE) $\hat{p}_h$를 계산합니다.

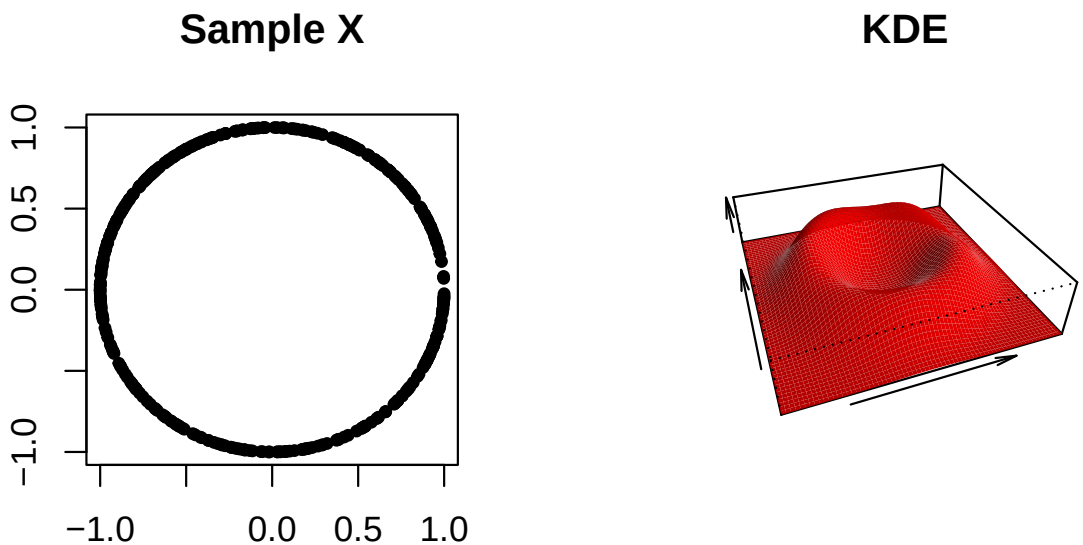
```
h <- 0.3
KDE <- kde(X = X, Grid = Grid, h = h)

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 패키지 TDA는 격자 위에서의 핵밀도추정(KDE)을 제공합니다.

가우스 핵밀도추정 (Kernel Density Estimator, KDE) $\hat{p}_h : \mathbb{R}^d \to [0, \infty)$는 다음과 같이 정의됩니다:

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

여기서 h는 평활매개변수(smoothing parameter)입니다.
함수 kde()는 격자 위의 점에서 핵밀도추정(KDE) $\hat{p}_h$를 계산합니다.



R 패키지 TDA는 격자 위에서의 Persistent Homology를 계산합니다.

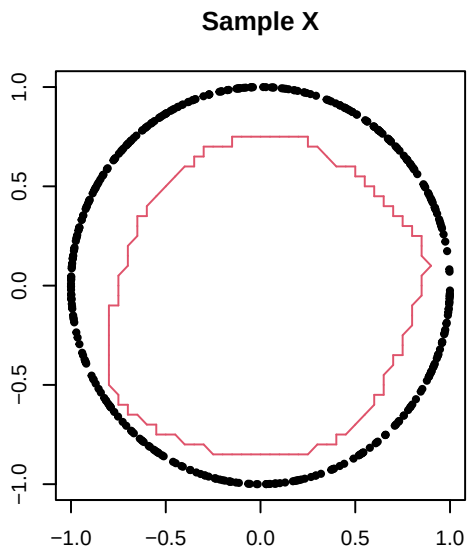
- 함수 gridDiag()는 입력함수의 아랫레벨(sublevel) 및 윗레벨(superlevel) 집합들의 persistence diagram을 계산합니다.
  - gridDiag()는 격자 위에서 실수값 입력함수를 계산합니다.
  - gridDiag()는 입력함수의 값으로 단체(simplex)들의 filtration을 만듭니다.
  - gridDiag()는 filtration의 persistent homology를 계산합니다.
- 사용자는 persistent homology를 계산하는 데에 C++ 라이브러리 GUDHI, Dionysus, 또는 PHAT을 선택할 수 있습니다.

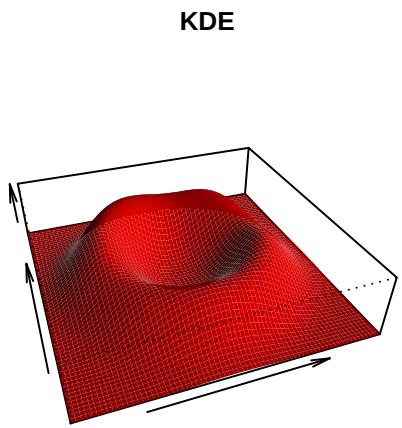
R 패키지 TDA는 격자 위에서의 Persistent Homology를 계산합니다.

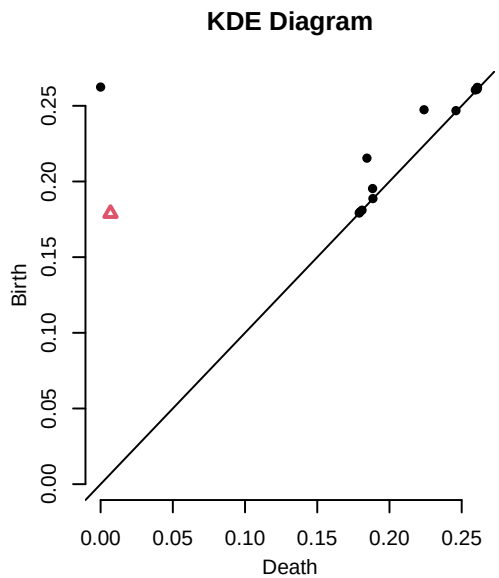
```
DiagGrid <- gridDiag(X = X, FUN = kde, lim = c(lim, lim), by = by,
    sublevel = FALSE, library = "Dionysus", location = TRUE,
    printProgress = FALSE, h = h)

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 패키지 TDA는 격자 위에서의 Persistent Homology를 계산합니다.

- 함수 gridDiag()는 입력함수의 아랫레벨(sublevel) 및 윗레벨(superlevel) 집합들의 persistence diagram을 계산합니다.
  - gridDiag()는 격자 위에서 실수값 입력함수를 계산합니다.
  - gridDiag()는 입력함수의 값으로 단체(simplex)들의 filtration을 만듭니다.
  - gridDiag()는 filtration의 persistent homology를 계산합니다.
- 사용자는 persistent homology를 계산하는 데에 C++ 라이브러리 GUDHI, Dionysus, 또는 PHAT을 선택할 수 있습니다.



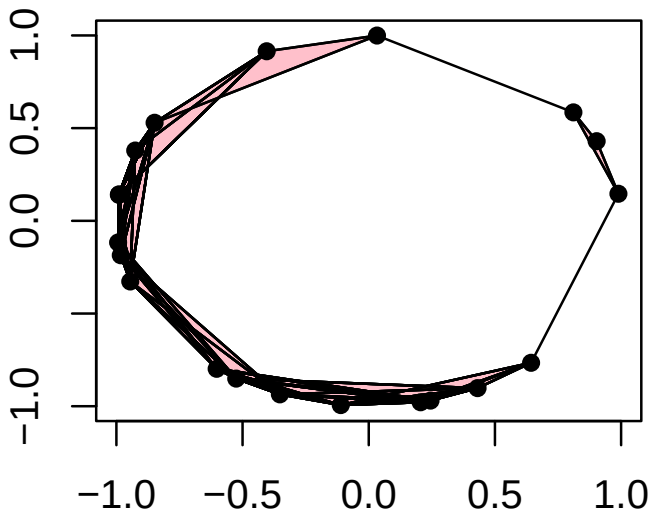




R 패키지 TDA는 Vietoris-Rips Persistent Homology를 계산합니다.

- Vietoris-Rips 복합체(complex)는 사이의 거리가 최대 $2r$ 이내인 꼭지점들로 이루어진 단체(simplex)들의 모임입니다. 즉,

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



- VIetoris-Rips filtration은 Vietoris-Rips 복합체에서 r을 서서히 증가시키면서 만들어집니다.

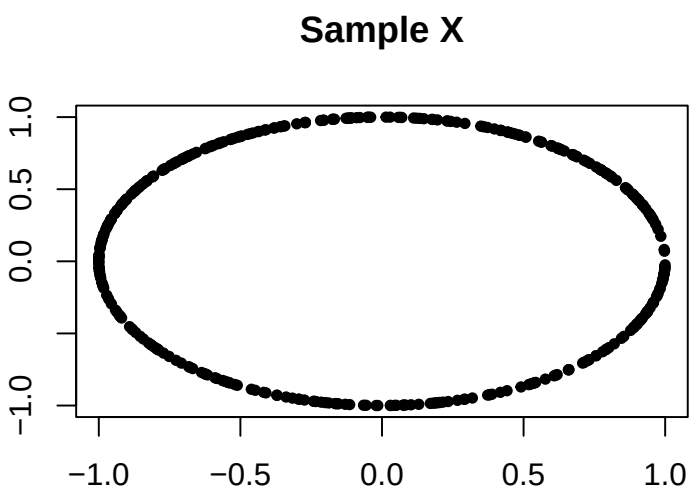
R 패키지 TDA는 Vietoris-Rips Persistent Homology를 계산합니다.

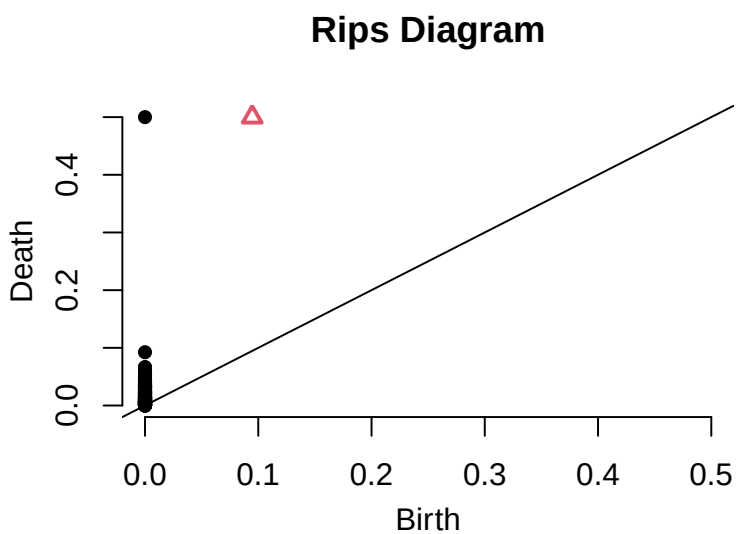
- 함수 ripsDiag()는 점집합 위에서 만들어진 Vietoris-Rips filtration의 persistence diagram을 계산합니다.
  - ripsDiag()는 자료로부터 Vietoris-Rips filtration을 만듭니다.
  - ripsDiag()는 Vietoris-Rips filtration으로부터 persistent homology를 계산합니다.
- 사용자는 persistent homology를 계산하는 데에 C++ 라이브러리 GUDHI, Dionysus, 또는 PHAT을 선택할 수 있습니다.

```
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 패키지 TDA는 Vietoris-Rips Persistent Homology를 계산합니다.

- 함수 ripsDiag()는 점집합 위에서 만들어진 Vietoris-Rips filtration의 persistence diagram을 계산합니다.
  - ripsDiag()는 자료로부터 Vietoris-Rips filtration을 만듭니다.
  - ripsDiag()는 Vietoris-Rips filtration으로부터 persistent homology를 계산합니다.
- 사용자는 persistent homology를 계산하는 데에 C++ 라이브러리 GUDHI, Dionysus, 또는 PHAT을 선택할 수 있습니다.





R 패키지 TDA는 Persistence Landscape를 계산합니다.

- persistence diagram D의 birth-death 쌍 (b, d)로부터 점 $p = (x, y) = \left(\frac{b+d}{2}, \frac{d-b}{2}\right)$를 생각하고, 이 p를 꼭지점으로 한 텐트 모양의 함수 Λ_p를 생각합니다.
- D의 persistence landscape는 다음과 같음 함수들의 모임입니다:

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

  여기서 kmax는 집합에서 k번째로 큰 값을 줍니다.
- 함수 landscape()는 persistence landscape 함수 $\lambda_k(t)$를 계산합니다.

```
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)
```

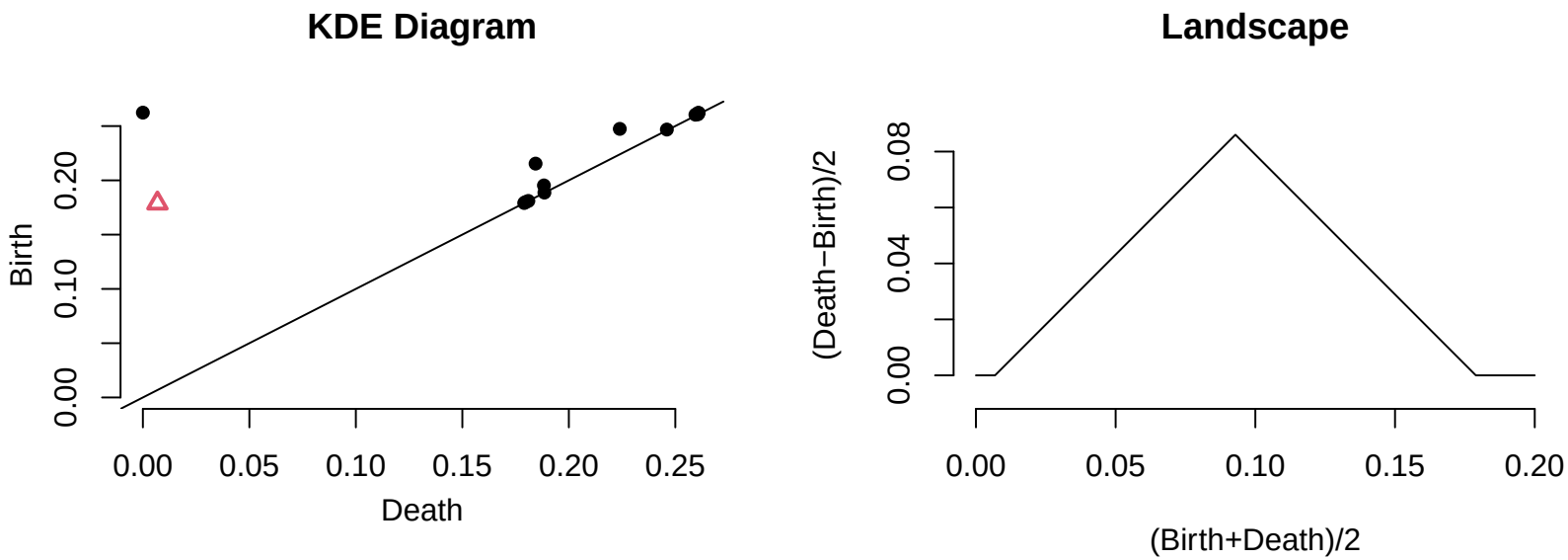
R 패키지 TDA는 Persistence Landscape를 계산합니다.

- persistence diagram D의 birth-death 쌍 (b, d)로부터 점 $p = (x, y) = \left(\frac{b+d}{2}, \frac{d-b}{2}\right)$를 생각하고, 이 p를 꼭지점으로 한 텐트 모양의 함수 Λ_p를 생각합니다.
- D의 persistence landscape는 다음과 같음 함수들의 모임입니다:

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

여기서 kmax는 집합에서 k번째로 큰 값을 줍니다.

- 함수 landscape()는 persistence landscape 함수 $\lambda_k(t)$를 계산합니다.



R 패키지 TDA는 함수의 붓스트랩 신뢰띠를 계산합니다.

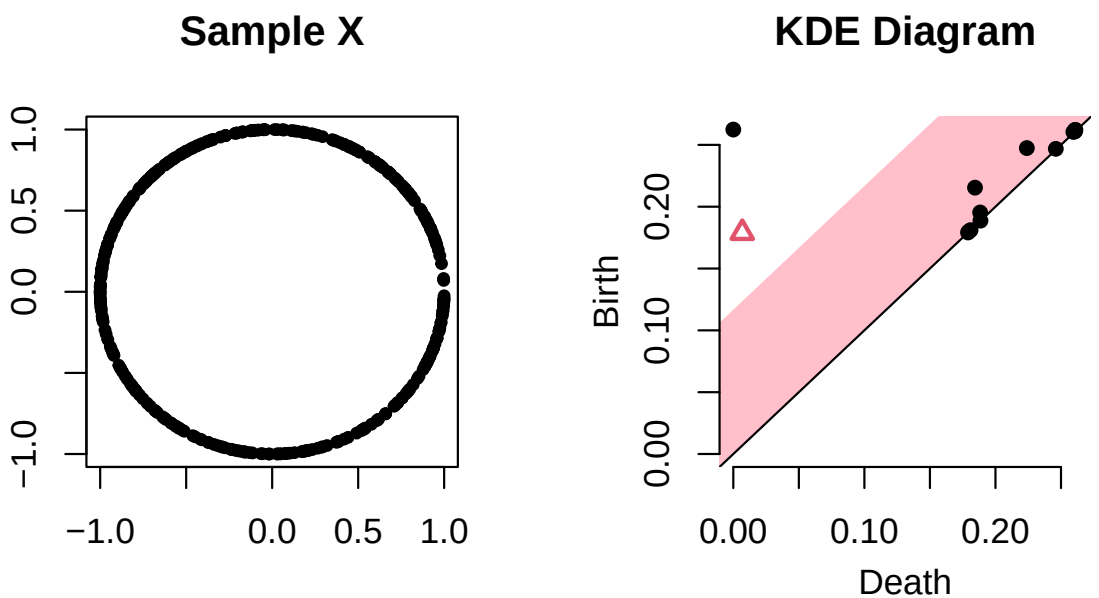
함수 bootstrapBand()는 $\mathbb{E}[\hat{p}_h]$의 $(1-\alpha)$ 붓스트랩 신뢰띠(bootstrap confidence band)를 계산합니다.

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

##         90%
## 0.05836494
```

R 패키지 TDA는 persistent homology의 붓스트랩 신뢰띠를 계산합니다.

$\mathbb{E}[\hat{p}_h]$의 $(1-\alpha)$ 붓스트랩 신뢰띠(bootstrap confidence band)가 persistent homology의 붓스트랩 신뢰띠로 사용됩니다.

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



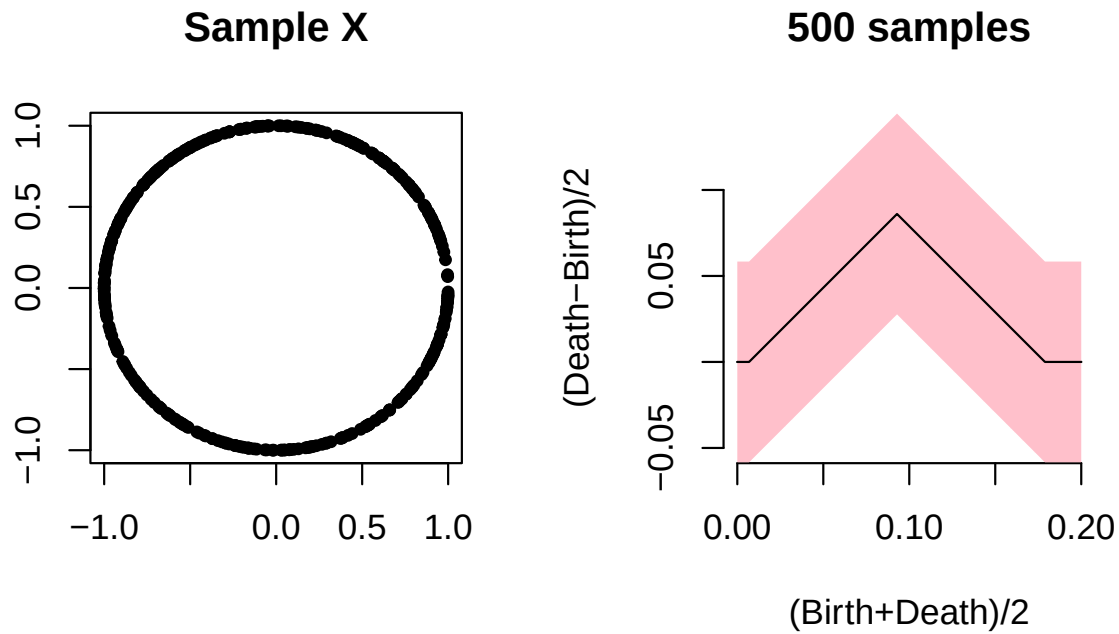
R 패키지 TDA는 persistence landscape의 붓스트랩 신뢰띠를 계산합니다.

$\mathbb{E}[\hat{p}_h]$의 $(1-\alpha)$ 붓스트랩 신뢰띠(bootstrap confidence band)가 persistent homology의 붓스트랩 신뢰띠로 사용됩니다.

```
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 패키지 TDA는 persistence landscape의 붓스트랩 신뢰띠를 계산합니다.

$\mathbb{E}[\hat{p}_h]$의 $(1-\alpha)$ 붓스트랩 신뢰띠(bootstrap confidence band)가 persistent homology의 붓스트랩 신뢰띠로 사용됩니다.

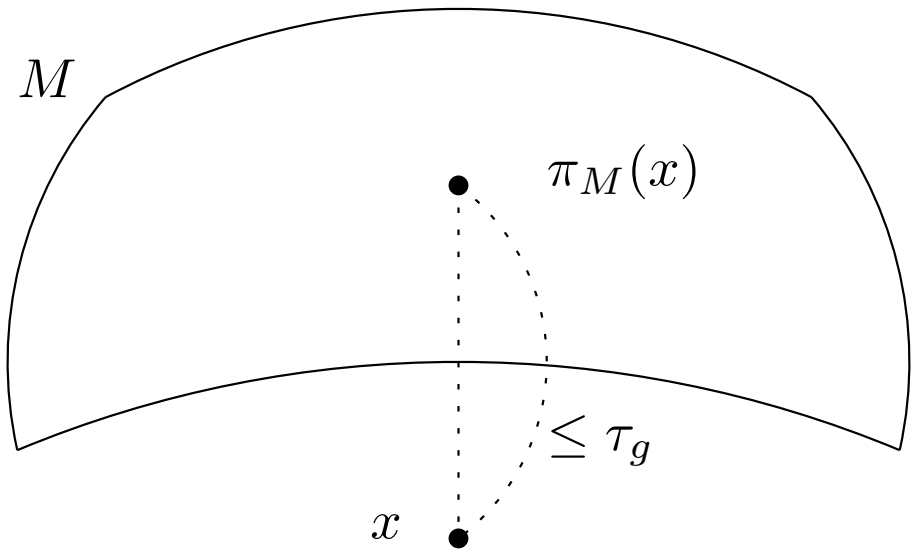


받침 다양체(supporting manifold) M은 유계입니다 (bounded).

$$M \subset I := [-K_I, K_I]^m \subset \mathbb{R}^m \text{ with } K_I \in (0, \infty)$$

임의로 복잡한 다양체(manifold)를 피하기 위해, reach에 하한이 있다고 가정합니다.

- $\mathcal{P}$ is a set of distributions P that is supported on a bounded manifold M, with its reach $\tau(M) \geq \tau_g$, and with other regularity assumptions.



임의로 복잡한 다양체(manifold)를 피하기 위해, reach에 하한이 있다고 가정합니다.

- M is of local reach $\geq \tau_\ell$, if for all points $p \in M$, there exists a neighborhood $U_p \subset M$ such that U_p is of reach $\geq \tau_\ell$.

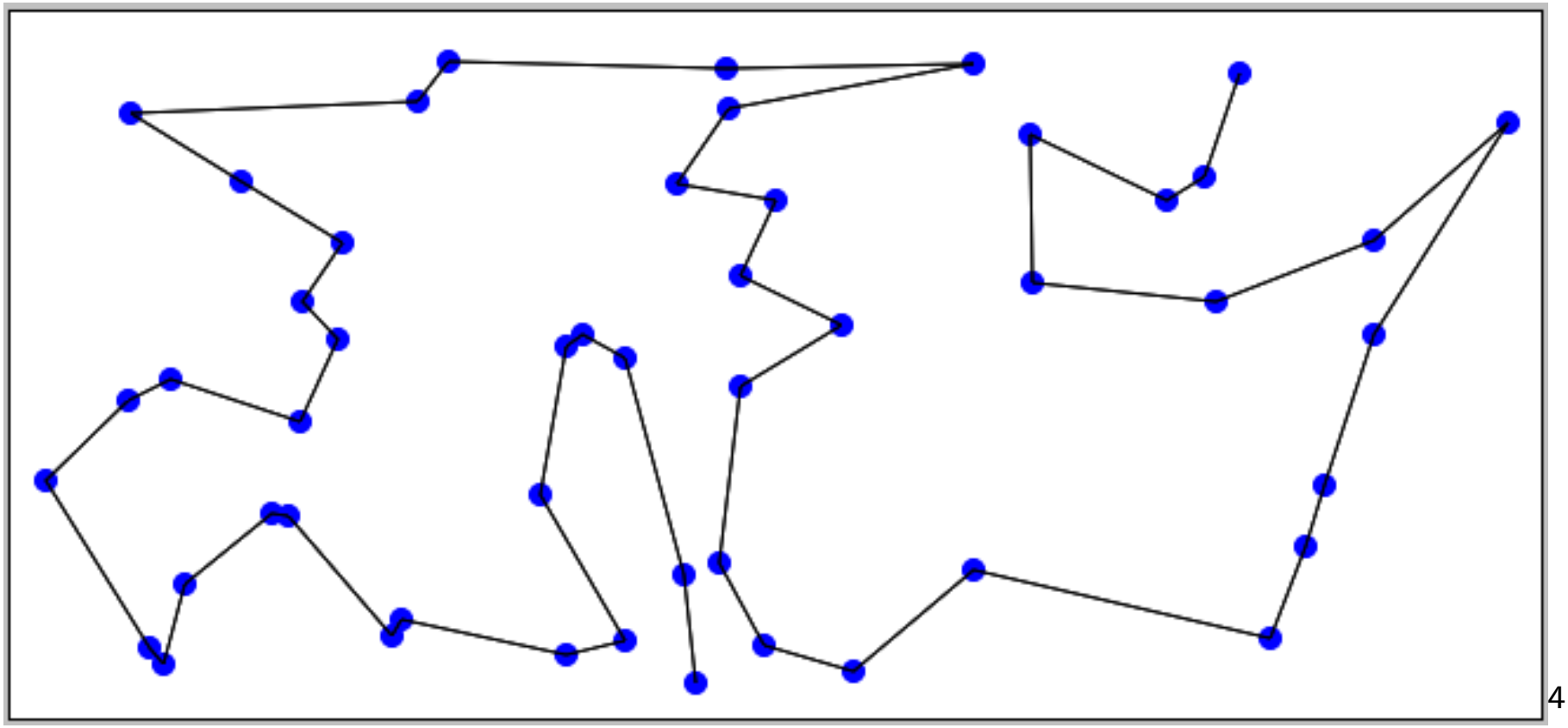
균등분포에 대한 밀도함수의 상한이 존재합니다.

- Distribution P is absolutely continuous to induced Lebesgue measure vol_M, and $\frac{dP}{dvol_M} \leq K_p$ for fixed K_p.
- This implies that the distribution on the manifold is of essential dimension d.
- $\mathcal{P}^d_{\kappa_l, \kappa_g, K_p}$ denotes set of distributions P that is supported on d-dimensional manifold of (global) reach $\geq \tau_g$, local reach $\geq \tau_\ell$, and density is bounded by K_p.

임의의 추정량의 최대위험(maximum risk)이 미니맥스 위험(minimax risk)의 상한(upper bound)이 됩니다.

$$
\begin{aligned}
R_n &= \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[1 \left(\hat{\dim}_n(X) \neq \dim(P) \right) \right] \\
&\leq \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[1 \left(\hat{\dim}_n(X) \neq \dim(P) \right) \right]}_{\text{임의의 추정량의 최대위험(maximum risk)}}
\end{aligned}
$$

외판원 문제(Traveling Salesman Problem, TSP) 경로는 각 점을 정확히 한 번씩만 지나는 가장 짧은 경로를 찾습니다.



[4]

[4] http://www.heatonresearch.com/fun/tsp/anneal

우리의 추정량(estimator)은 자료로부터 생성된 TSP 경로의 d_1-제곱한 길이가 길 때 차원을 d_2로 추정합니다.

- 내제적 차원이 높을 때, TSP 경로의 길이는 더 커지는 경향이 있습니다.
- $$\hat{\dim}_n(X) = d_1 \iff \min_{\sigma \in S_n} \sum_{i=1}^{n-1} \|X_{\sigma(i+1)} - X_{\sigma(i)}\|_{\mathbb{R}^m}^{\boldsymbol{d_1}} \leq C,$$

  여기에서 C는 상수입니다.

우리의 추정량은 최대위험(maximum risk)이 $O\left(n^{-\left(\frac{d_2}{d_1}-1\right)n}\right)$입니다.

- Our estimator makes error with probability at most $O\left(n^{-\left(\frac{d_2}{d_1}-1\right)n}\right)$ if intrinsic dimension is d_2.
- Our estimator is always correct when the intrinsic dimension is d_1.

우리의 추정량은 내재적 차원이 d_2일 때 확률 $O\left(n^{-\left(\frac{d_2}{d_1}-1\right)n}\right)$로 오류를 냅니다.

- Based on the following lemma:

Lemma

(Lemma 6) Let $X_1, \cdots, X_n \sim P \in \mathcal{P}^{d_2}_{\kappa_l, \kappa_g, K_p}$, then

$$P^{(n)}\left[\sum_{i=1}^{n-1} \|X_{i+1} - X_i\|^{d_1} \leq L\right] \lesssim n^{-\frac{d_2}{d_1} n}.$$

우리의 추정량은 내재적 차원이 d_1일 때는 언제나 정확합니다.

- Based on following lemma:

Lemma

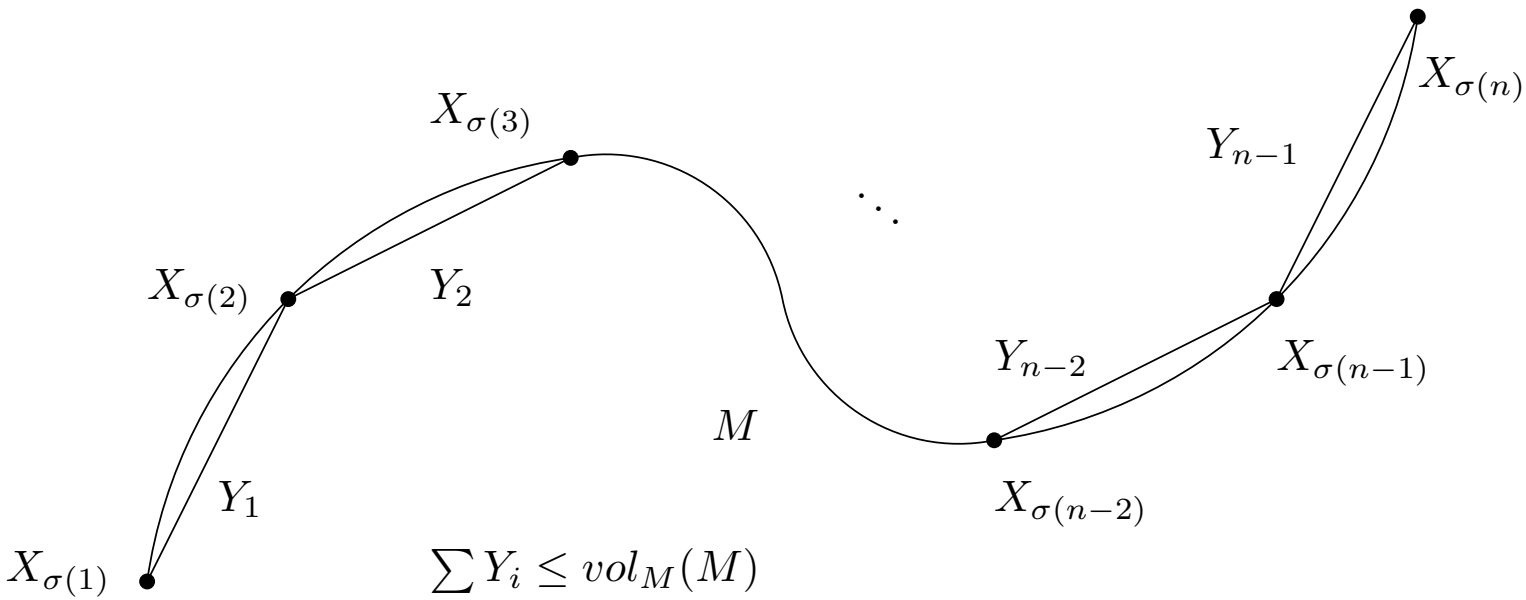
(Lemma 7) Let M be a d_1-dimensional manifold with global reach $\geq \tau_g$ and local reach $\geq \tau_\ell$, and $X_1, \cdots, X_n \in M$. Then there exists C which depends only on m, d_1 and K_I, and there exists $\sigma \in S_n$ such that

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

우리의 추정량은 내재적 차원이 d_1일 때는 언제나 정확합니다.

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

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

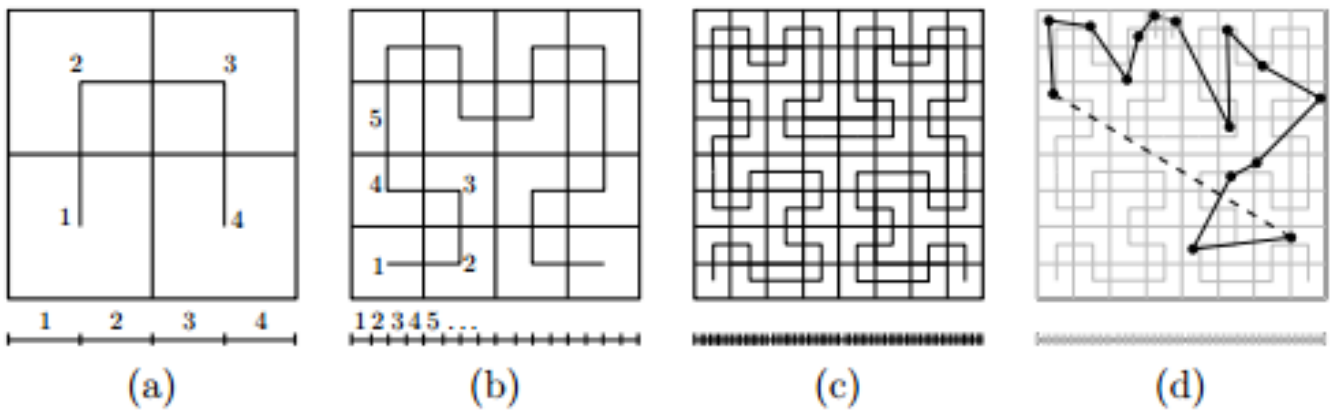


- Global reach$\geq \tau_g$ implies $vol_M(M)$ is bounded.

우리의 추정량은 내재적 차원이 d_1일 때는 언제나 정확합니다.

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

- When $d_1 > 1$, Several conditions implied by regularity conditions combined with Hölder continuity of d_1-dimensional space-filling curve is used.



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

우리의 추정량은 내재적 차원이 d_1일 때는 언제나 정확합니다.

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

- When $d_1 > 1$, Several conditions implied by regularity conditions combined with Hölder continuity of d_1-dimensional space-filling curve is used.

Lemma

(Lemma 22, Space-filling curve) There exists surjective map $\psi_d : \mathbb{R} \to \mathbb{R}^d$ which is Hölder continuous of order $1/d$, i.e.

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

미니맥스 위험(minimax risk)는 $O\left(n^{-\left(\frac{d_2}{d_1}-1\right)n}\right)$로 상한이 됩니다.

Proposition

(Proposition 9) $1 \leq d_1 < d_2 \leq m$일 때, 다음이 성립합니다:

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

Le Cam의 보조정리가 차원의 다른 확률분포들의 통계적 차이에 기반하여 차원을 추정하는 데에 하한을 제공합니다.

Lemma

(Lemma 10, Le Cam's Lemma) $\mathcal{P}$*를 확률측도의 집합이라 놓고,* $\mathcal{P}^{d_1}, \mathcal{P}^{d_2} \subset \mathcal{P}$*를 모든* $P \in \mathcal{P}^{d_i}$*에 대해,* $i = 1, 2$*일 때* $\theta(P) = \theta_i$*를 만족한다고 가정합니다. 임의의* $Q_i \in co(\mathcal{P}_i)$*에 대해,* q_i *를 측도* ν*에 대한* Q_i*의 밀도함수로 놓습니다. 그러면,*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[1 \left(\hat{\dim}_n(X) \neq \dim(P) \right) \right]$$
$$\geq \frac{1(\theta_1 \neq \theta_2)}{4} \sup_{Q_i \in co(\mathcal{P}^{d_i})} \int [q_1(x) \wedge q_2(x)] d\nu(x).$$

$X = (X_1, \cdots, X_n) \in T$을 만족할 때 두 모델을 구분하기 힘들도록 부분집합 $T \subset [-K_l, K_l]^n$과 확률분포의 집합 $\mathcal{P}_1^{d_1}$, $\mathcal{P}_2^{d_2}$를 찾습니다.

- The lower bound measures how hard it is to tell whether the data come from a d_1 or d_2 -dimensional manifold.
- T, $\mathcal{P}_1^{d_1}$ and $\mathcal{P}_2^{d_2}$ are linked to the lower bound by using Le Cam's lemma.

Le Cam의 보조정리는 두 밀도함수의 최소 $q_1 \wedge q_2$에 기반하여 하한을 찾는데, 이 때 q_1, q_2은 각각 $\mathcal{P}_1^{d_1}$의 최소볼록집합(convex hull)과 $\mathcal{P}_2^{d_2}$의 최소볼록집합에 있습니다.

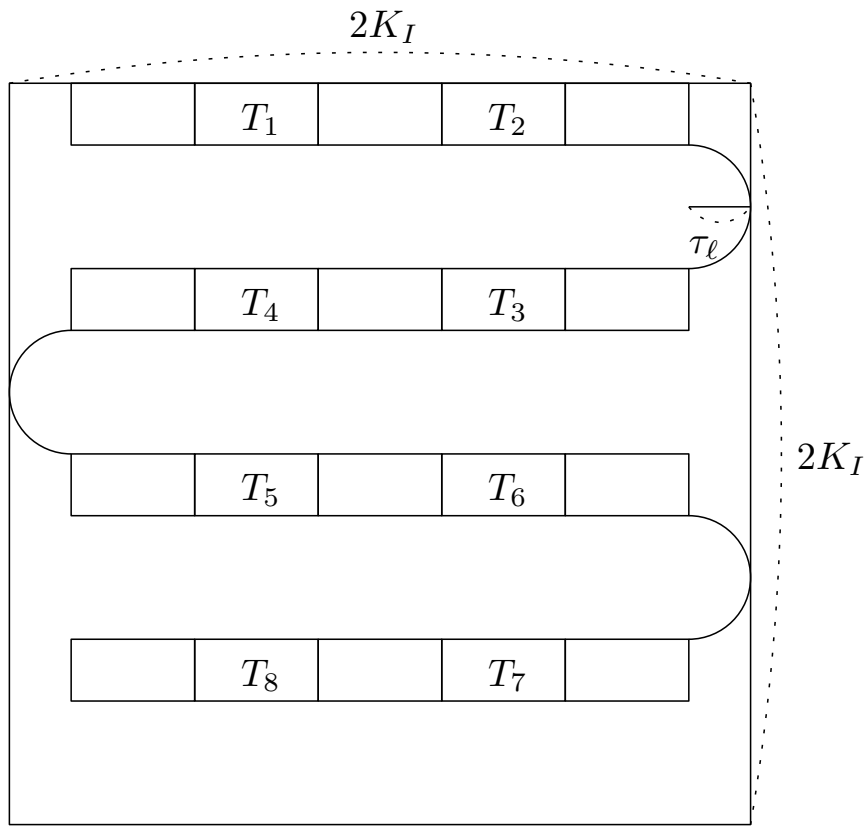
Lemma

(Lemma 10, Le Cam's Lemma) Let $\mathcal{P}$ be a set of probability measures, and $\mathcal{P}^{d_1}, \mathcal{P}^{d_2} \subset \mathcal{P}$ be such that for all $P \in \mathcal{P}^{d_i}$, $\theta(P) = \theta_i$ for $i = 1, 2$. For any $Q_i \in \mathrm{conv}(\mathcal{P}_i)$, let q_i be density of Q_i with respect to measure ν. Then

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[1 \left(\hat{\dim}_n(X) \neq \dim(P) \right) \right]$$
$$\geq \frac{1(\theta_1 \neq \theta_2)}{4} \sup_{Q_i \in \mathrm{conv}(\mathcal{P}^{d_i})} \int [q_1(x) \wedge q_2(x)] d\nu(x).$$

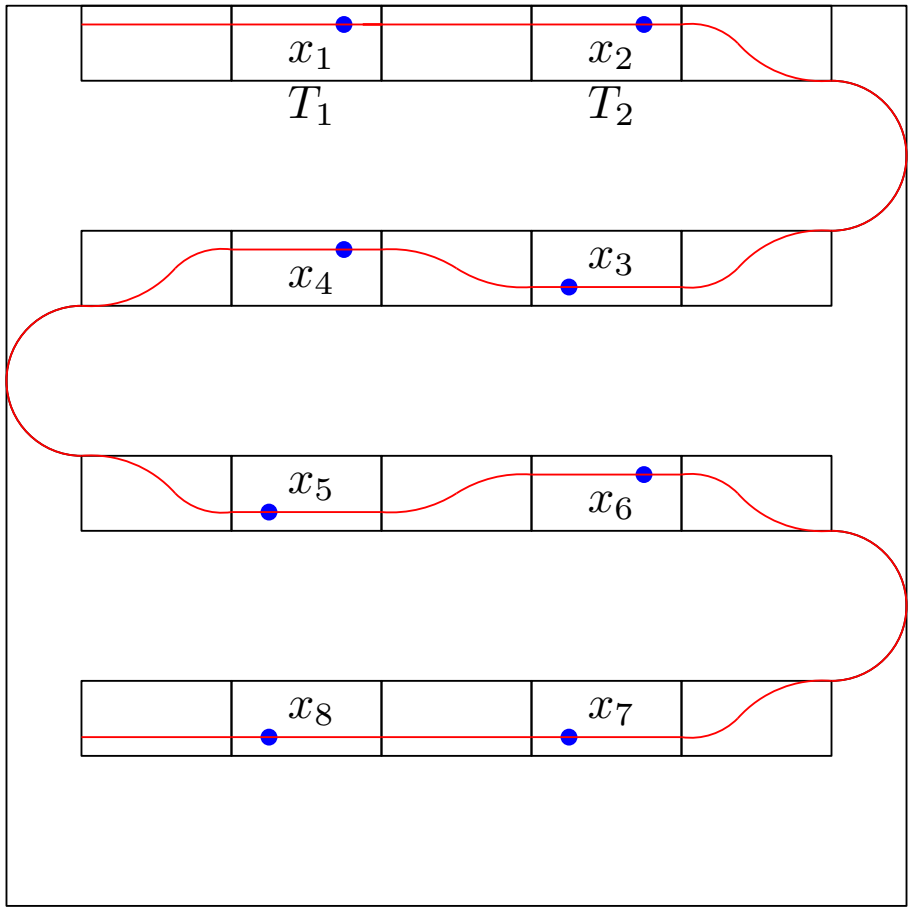
임의의 $x = (x_1, \cdots, x_n) \in T$에 대해, 정칙성(regularity) 조건을 만족하고 $x_1, \cdots, x_n$을 지나는 d_1-차원 다양체(manifold)가 존재하도록 T를 구성합니다.

- T_i's are cylinder sets in $[-K_I, K_I]^{d_2}$, and then T is constructed as $T = S_n \prod_{i=1}^{n} T_i$, where the permutation group S_n acts on $\prod_{i=1}^{n} T_i$ as a coordinate change.



임의의 $x = (x_1, \cdots, x_n) \in T$에 대해, 정칙성(regularity) 조건을 만족하고 $x_1, \cdots, x_n$을 지나는 d_1-차원 다양체(manifold)가 존재하도록 T를 구성합니다.

- Given $x_1, \cdots, x_n \in T$ (blue points), manifold of global reach $\geq \tau_g$ and local reach $\geq \tau_\ell$ (red line) passes through $x_1, \cdots, x_n$.



$\mathcal{P}_1^{d_1}$ 은 $x = (x_1, \cdots, x_n) \in T$일 때 $x_1, \cdots, x_n$을 지나는 다양체를 받침으로 하는확률분포의 집합이고, $\mathcal{P}_2^{d_2}$는 is a $[-K_l, K_l]^{d_2}$ 위의 균등분포 하나로 구성된 집합입니다.

$X \in T$이면, X의 확률분포가 $\mathcal{P}_1^{d_1}$아니면 $\mathcal{P}_2^{d_2}$에 속해있는지 판단하기 힘듭니다.

- There exists $Q_1 \in co(\mathcal{P}_1^{d_1})$ and $Q_2 \in co(\mathcal{P}_2^{d_2})$ such that $q_1(x) \geq Cq_2(x)$ for every $x \in T$ with $C < 1$.
- Then $q_1(x) \wedge q_2(x) \geq Cq_2(x)$ if $x \in T$, so $C \int_T q_2(x)dx$ can serve as lower bound of minimax rate.
- Based on following claim:

Claim

(Claim 25) Let $T = S_n \prod_{i=1}^{n} T_i$. Then for all $x \in \operatorname{int} T$, there exists $C > 0$ that depends only on κ_l, K_l, and $r_x > 0$ such that for all $r < r_x$,

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

미니맥스 위험(minimax risk)는 $\Omega\left(n^{-2(d_2-d_1)n}\right)$로 하한이 됩니다.

Proposition

(Proposition 14)

$$\inf_{\hat{\dim} P \in \mathcal{P}^{d_1} \cup \mathcal{P}^{d_2}} \sup \mathbb{E}_{P^{(n)}} \left[1 \left(\hat{\dim}_n(X) \neq \dim(P) \right) \right] \gtrsim n^{-2(d_2-d_1)n}.$$

다중 분류문제와 $0-1$ 손실함수를 고려합니다.

- $$R_n = \inf_{\hat{\dim}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[1 \left(\hat{\dim}_n(X) \neq \dim(P) \right) \right]$$
- Now the manifolds are of any dimensions between 1 and m, so considered distribution set is $\mathcal{P} = \bigcup_{d=1}^{m} \mathcal{P}^d$.
- $0-1$ loss function is considered, so for all $x, y \in \mathbb{R}$, $\ell(x, y) = I(x = y)$.

미니맥스 위험(Minimax risk)은 $O\left(n^{-\frac{1}{m-1}n}\right)$으로 상한이 되고, $\Omega\left(n^{-2n}\right)$ 으로 하한이 됩니다.

Proposition

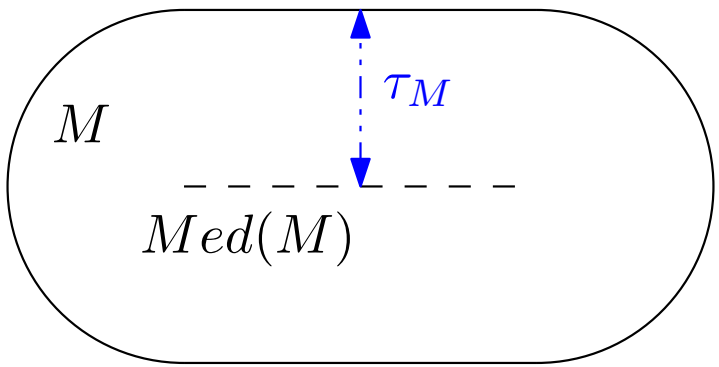
(Proposition 16 and 17)

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

집합 M의 medial axis는 M 상에서 가장 가까운 지점이 최소한 두 개 이상인 점들의 집합입니다.

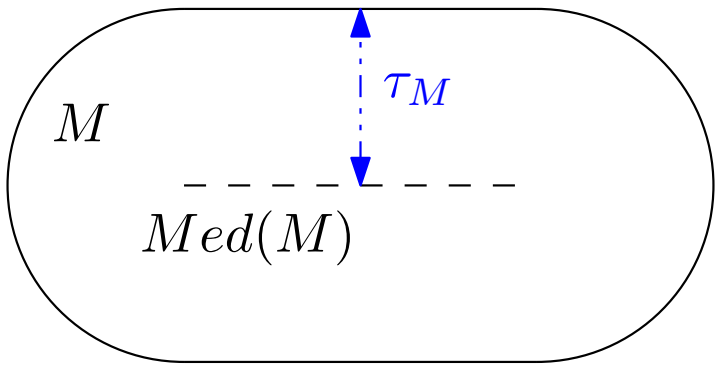
▶

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



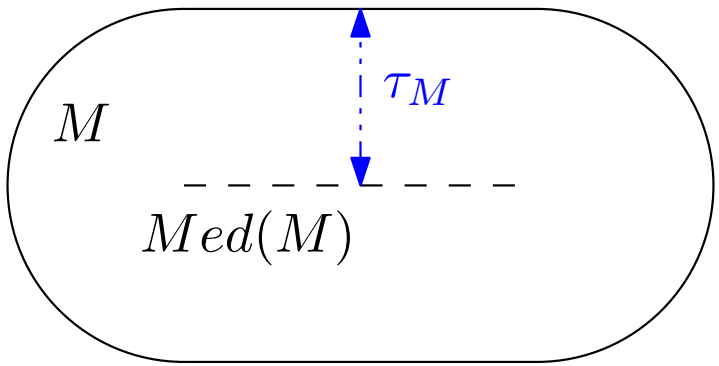
M의 reach는 (τ_M로 표기) M의 medial axis $Med(M)$으로부터 M까지의 거리입니다.

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



reach τ_M은 M으로의 사영(projection)이 잘 정의되는 최대의 오프셋(offset) 크기입니다.

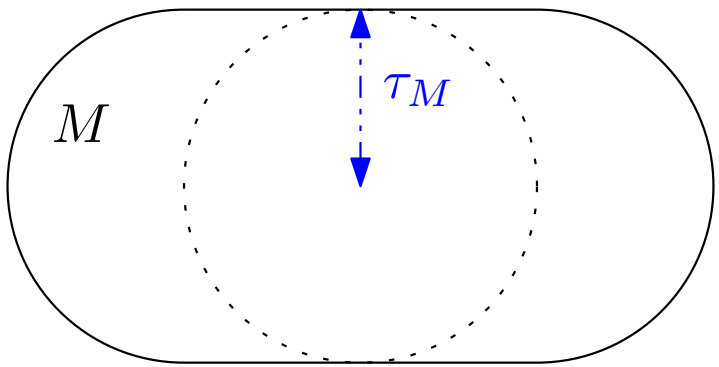
- $$\tau_M = \inf_{x \in Med(M), y \in M} \|x - y\| .$$



reach τ_M은 M 위에서 굴러갈 수 있는 공(ball)의 반지름의 최대 크기입니다.

- When $M \subset \mathbb{R}^m$ is a manifold,

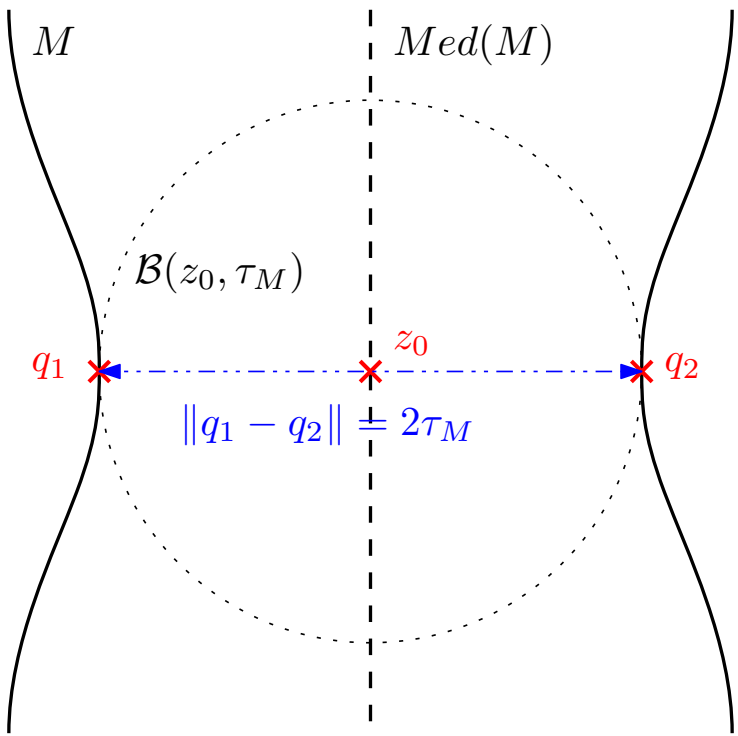
$$\tau_M = \inf_{q_2 \neq q_1 \in M} \frac{\|q_2 - q_1\|^2}{2d(q_2 - q_1, T_{q_1} M)}.$$



병목(bottleneck)은 다양체에서 자기교차(self-intersecting)에 가깝게 생긴 기하학적 구조입니다.

Definition

(Definition 3.1) A pair of points (q_1, q_2) in M is said to be a bottleneck of M if there exists $z_0 \in Med(M)$ such that $q_1, q_2 \in \mathcal{B}(z_0, \tau_M)$ and $\|q_1 - q_2\| = 2\tau_M$.

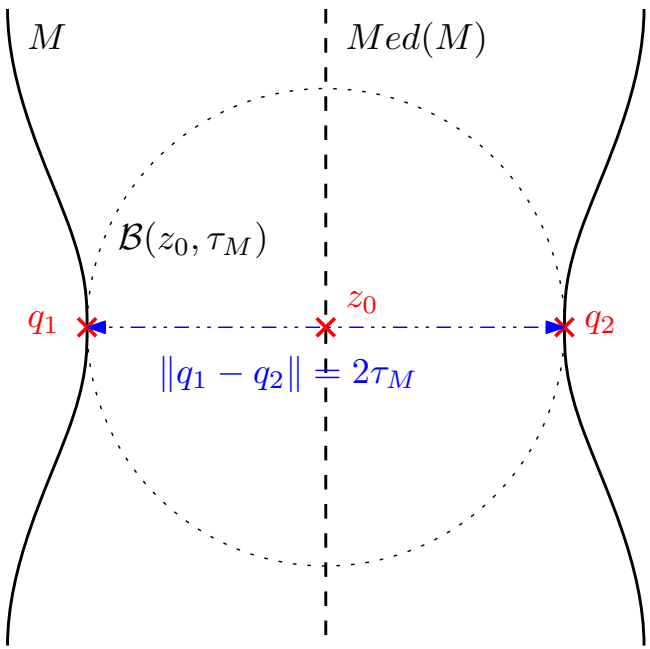


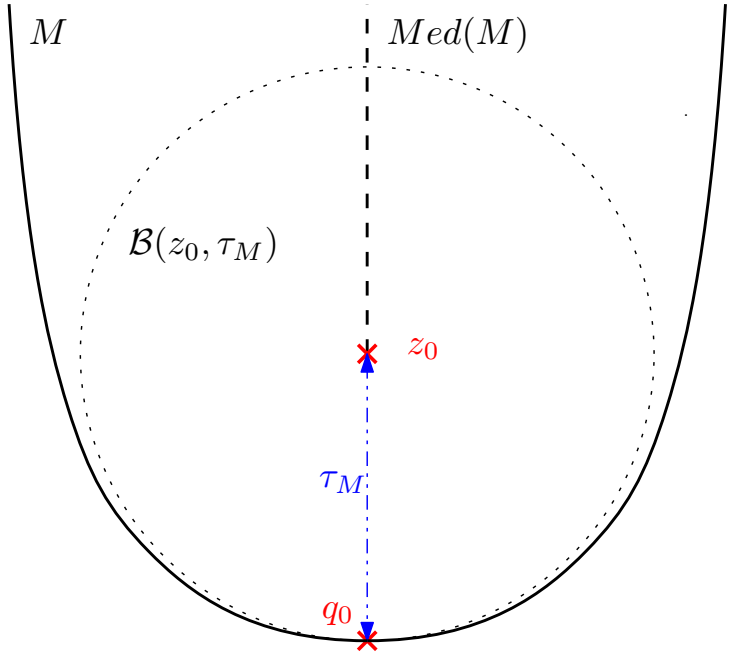
reach는 병목(bottleneck)에서 생기거나 (대역적, global) 곡률이 큰 부분에서 생깁니다 (국소적, local).

Theorem

(Theorem 3.4) At least one of the following two assertions holds:

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





reach 추정량 $\hat{\tau}$의 통계적 효율성을 위험(risk)로 분석합니다.

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

$$\mathbb{E}_{P^{(n)}}\left[\ell\left(\hat{\tau}(\mathcal{X}),\ \tau_M\right)\right].$$

  - $\mathcal{X} = \{X_1, \ldots, X_n\}$ is drawn from a fixed distribution P with its support M.
  - The loss function used is $\ell(\tau, \tau') = \left|\frac{1}{\tau} - \frac{1}{\tau'}\right|^p$, $p \geq 1$.

reach 추정량 $\hat{\tau}$의 위험을 분석합니다.

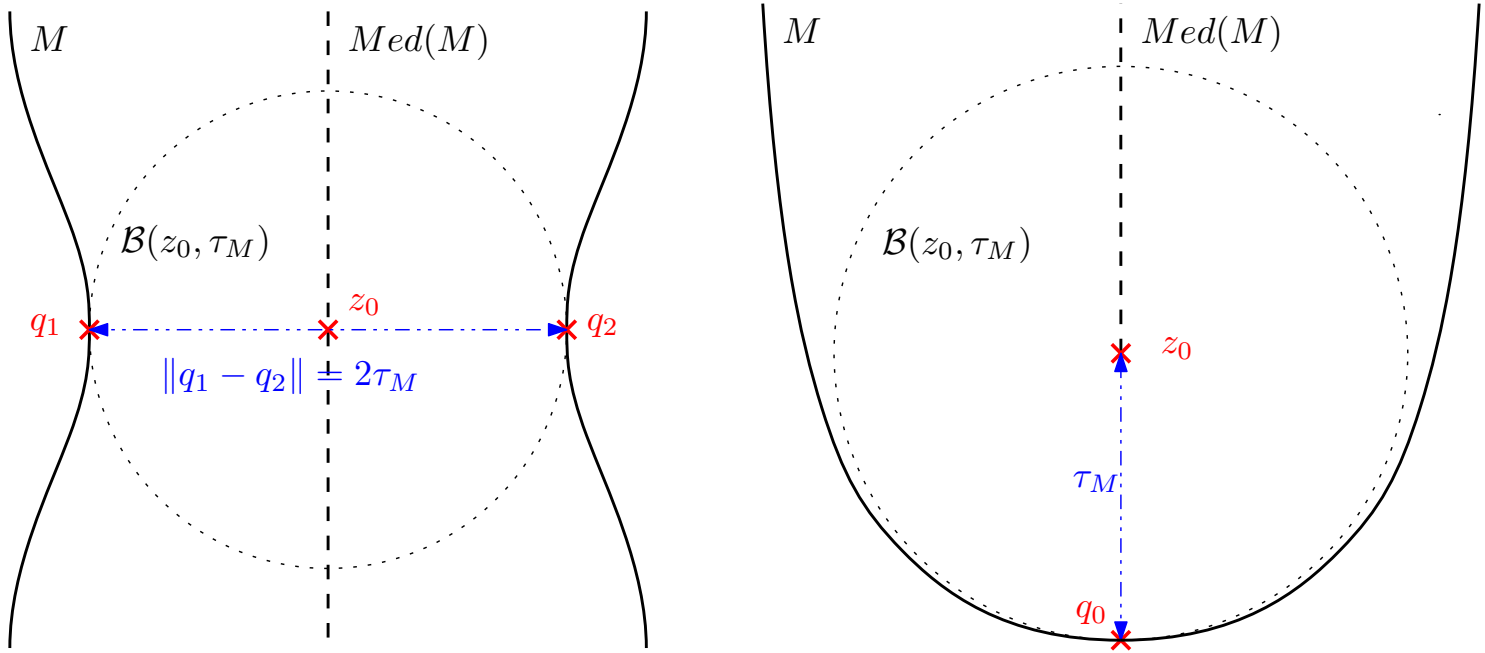
- The risk of the estimator $\hat{\tau}$ is the expected loss the estimator

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

  - $\mathcal{X}=\{X_1,\ldots,X_n\}$ is drawn from a fixed distribution P with its support M.
  - The loss function used is $\ell(\tau,\tau')=\left|\frac{1}{\tau}-\frac{1}{\tau'}\right|^q$, $q\geq 1$.

reach 추정량은 $O\left(n^{-\frac{2q}{3d-1}}\right)$의 위험을 지닙니다.

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

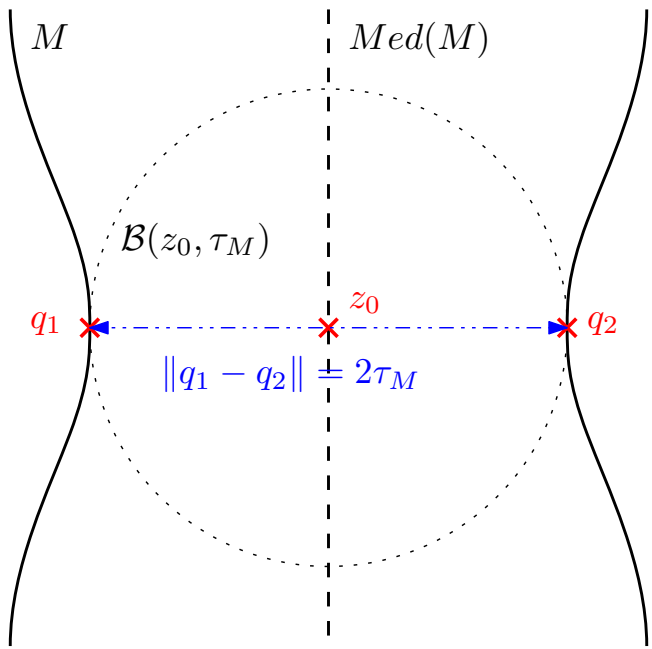


reach 추정량은 대역적 경우에 $O\left(n^{-\frac{q}{d}}\right)$의 최대위험 (maximum risk)를 지닙니다.

Proposition

(Proposition 4.3) Assume that the support M has a bottleneck. Then,

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

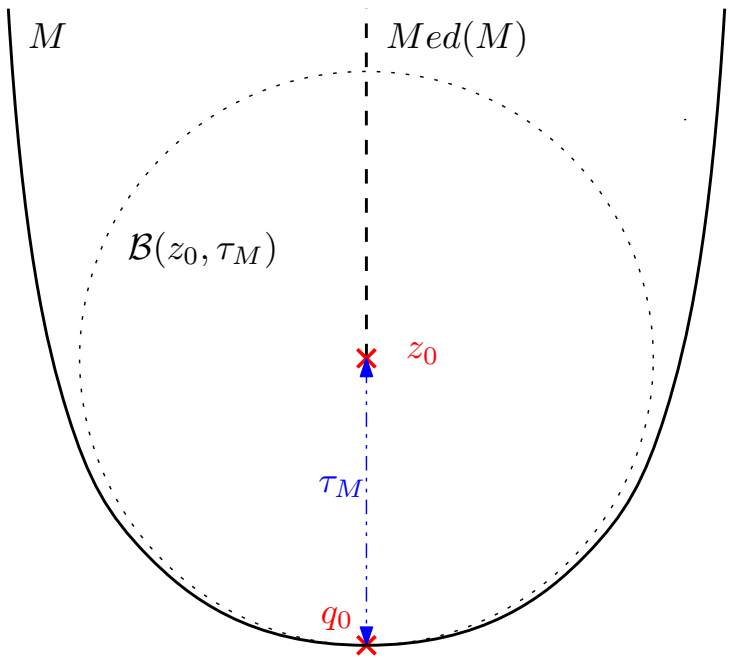


reach 추정량은 국소적 경우에 $O\left(n^{-\frac{2q}{3d-1}}\right)$의 최대 위험 (maximum risk)를 지닙니다.

Proposition

(Proposition 4.7) Suppose there exists $q_0 \in M$ and a geodesic γ_0 with $\gamma_0(0) = q_0$ and $\|\gamma_0''(0)\| = \frac{1}{\tau_M}$. Then,

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



reach 추정 문제의 통계적인 어려움을 미니맥스 위험 (minimax risk)으로 분석합니다.

- Minimax rate is the risk of an estimator that performs best in the worst case, as a function of sample size.
- $$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\ell \left(\hat{\tau}_n(\mathcal{X}),\ \tau_M \right) \right].$$
  - $\mathcal{X} = \{X_1, \ldots, X_n\}$ is drawn from a fixed distribution P with its support M, where P is contained in set of distributions $\mathcal{P}$.
  - An estimator $\hat{\tau}_n$ is any function of data $\mathcal{X}$.
  - The loss function used is $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^q$, $q \geq 1$.

reach 추정 문제의 통계적인 어려움을 미니맥스 위험 (minimax risk)으로 분석합니다.

- Minimax rate is the risk of an estimator that performs best in the worst case, as a function of sample size.
- $$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n(\mathcal{X})} \right|^q \right].$$
  - $\mathcal{X} = \{X_1, \ldots, X_n\}$ is drawn from a fixed distribution P with its support M, where P is contained in set of distributions $\mathcal{P}$.
  - An estimator $\hat{\tau}_n$ is any function of data $\mathcal{X}$.
  - The loss function used is $\ell(\tau, \tau') = \left| \frac{1}{\tau} - \frac{1}{\tau'} \right|^q$, $q \geq 1$.

우리의 추정량의 최대위험(maximum risk)이 미니맥스 위험(minimax risk)의 상한(upper bound)이 됩니다.

$$
\begin{aligned}
R_n &= \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}_n(X)} \right|^q \right] \\
&\leq \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}(X)} \right|^q \right]}_{\text{우리의 추정량의 최대위험(maximum risk)}}
\end{aligned}
$$

미니맥스 위험(minimax risk)는 $O\left(n^{-\frac{2q}{3d-1}}\right)$로 상한이 됩니다.

Theorem
(Theorem 5.1)

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

Le Cam의 보조정리는 두 확률분포의 reach 차이와 통계적 차이에 기반하여 reach를 추정하는 데에 하한을 제공합니다.

- 두 확률분포의 전변동거리(total variance distance)는 다음과 같이 정의됩니다:

$$TV(P, P') = \sup_{A \in \mathcal{B}(\mathbb{R}^D)} |P(A) - P'(A)| .$$

Lemma

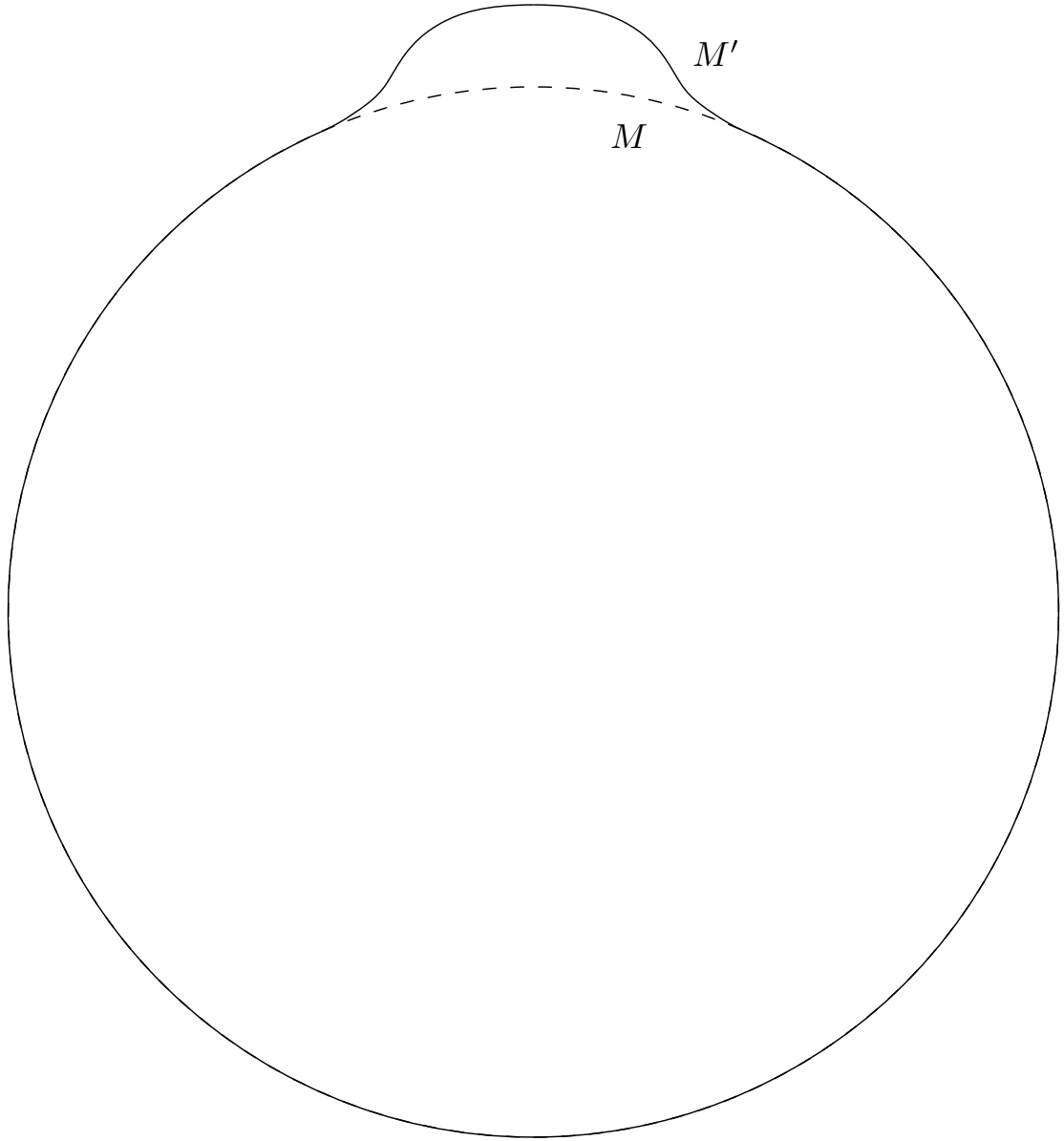
(Lemma 5.2) $P, P' \in \mathcal{P}$ 각각의 받침(support)을 M과 M'이라 놓습니다. 그러면

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

두 개의 확률분포 P, P'를 찾는데, 그들의 reach는 다르지만 확률분포 자체는 통계적으로 구분하기 어렵게 되도록 찾습니다.

- $$\inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| \frac{1}{\tau_M} - \frac{1}{\hat{\tau}_n} \right|^q \right] \gtrsim \left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^q \left(1 - TV(P, P')\right)^{2n}.$$
- The lower bound measures how hard it is to tell whether the data is from distributions with different reaches.
- P and P' are found so that $\left| \frac{1}{\tau_M} - \frac{1}{\tau_{M'}} \right|^q$ is large while $\left(1 - TV(P, P')\right)^{2n}$ is small.

P는 구(sphere) 위에 놓인 확률분포이고, P'는 혹이 달린 구(bumped sphere) 위에 높인 확률분포입니다.



미니맥스 위험(minimax risk)는 $\Omega\left(n^{-\frac{p}{d}}\right)$로 하한이 됩니다.

Proposition

(Proposition 5.6)

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