

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

김지수(Jisu Kim)



숙명여자대학교 통계학과  
2025-10-29

## 위상 자료 분석(Topological Data Analysis) 소개

Persistent Homology를 통계적으로 추정하기

위상 자료 분석(Topological Data Analysis)을 자료분석 및 기계학습에  
응용

Persistence Landscape를 이용하여 특성(Feature) 만들기

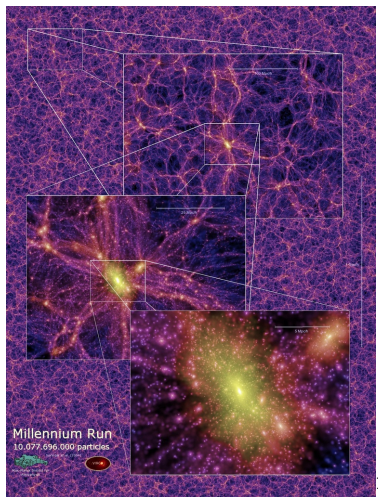
Circular Coordinates를 이용하여 특성(Feature) 만들기

위상 자료 분석(Topological Data Analysis)을 이용한 평가

R 패키지 TDA: 위상 자료 분석을 위한 통계 계산 도구

참조문헌

자료의 위상학적 구조로부터 정보를 얻을 수 있습니다.



<sup>1</sup>[http://www.mpa-garching.mpg.de/galform/virgo/millennium/poster\\_half.jpg](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)



# 위상학적 자료를 통계적으로 어떻게 추정하는지 알아봅니다.

- ▶ 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)

## 위상 자료 분석(Topological Data Analysis) 소개

### Persistent Homology를 통계적으로 추정하기

## 위상 자료 분석(Topological Data Analysis)을 자료분석 및 기계학습에 응용

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

## R 패키지 TDA: 위상 자료 분석을 위한 통계 계산 도구

## 참조문헌

# 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

- ▶ 숙, 명, 여, 대

- ▶ 여러 차원에서 구멍들의 개수들을 각각 고려합니다.

1.  $\beta_0$  = 연결된 성분의 개수



2.  $\beta_1$  = 고리(1차원 구의 구멍)의 개수



3.  $\beta_2$  = 2차원 구의 구멍의 개수



예제: 대상들을 호몰로지(Homology)에 따라 분류합니다.

1.  $\beta_0$  = 연결된 성분의 개수 ●

2.  $\beta_1$  = 고리의 개수 ○

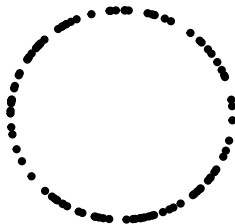
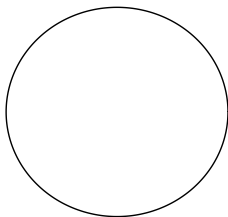
| $\beta_0 \setminus \beta_1$ | 0                         | 1                | 2    | 3 |
|-----------------------------|---------------------------|------------------|------|---|
| 1                           | ㄱ, ㄴ, ㄷ, ㄹ,<br>ㅅ, ㅈ, ㅊ, ㅌ | ㅁ, ㅅ, ㅈ,<br>ㅊ, ㅌ | ㅊ, ㅌ | ㅌ |
| 2                           | ㅊ, ㅌ                      |                  |      |   |
| 3                           | ㅌ                         | ㅌ                |      |   |

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

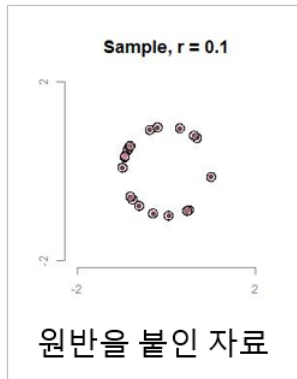
- ▶ 자료를 분석할 때, 기저 구조의 특성을 자료의 특성으로부터 추정할 수 있는 로버스트(robust)한 특성을 선호합니다.
- ▶ 호몰로지(Homology)는 로버스트하지 않습니다:

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

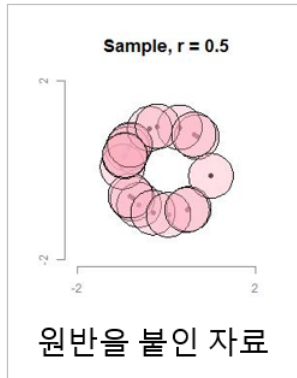
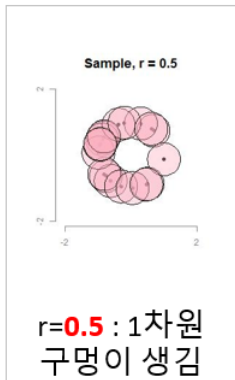
100 samples:  $\beta_0 = 100, \beta_1 = 0$



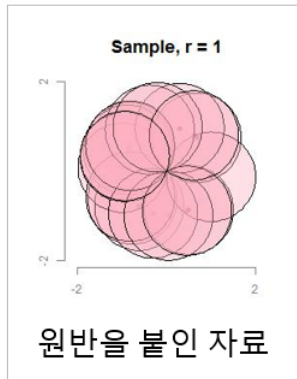
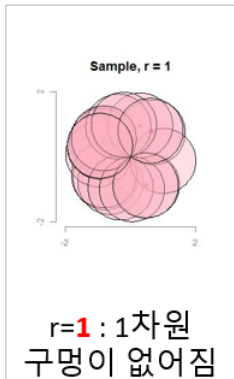
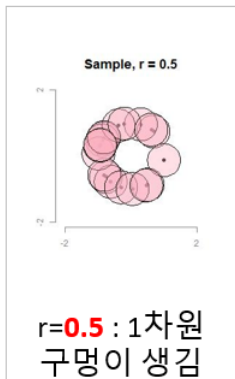
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



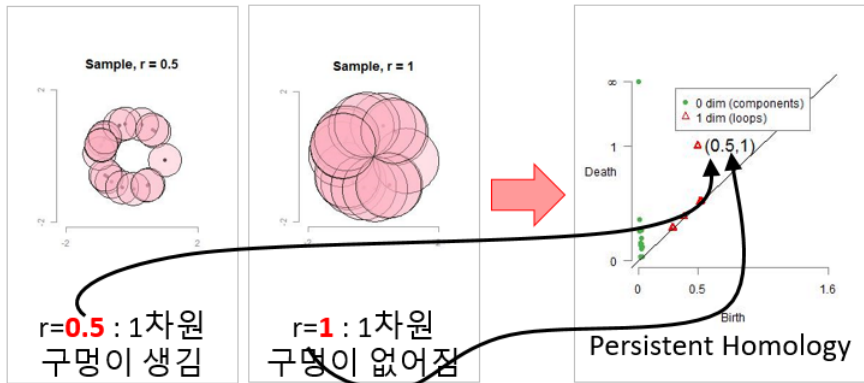
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



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



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

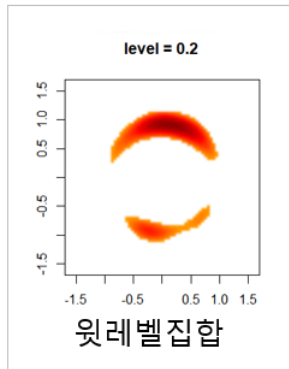


기저 구조의 위상학적 정보를 추출하는 데에 핵밀도추정(kernel density estimator)을 사용합니다.

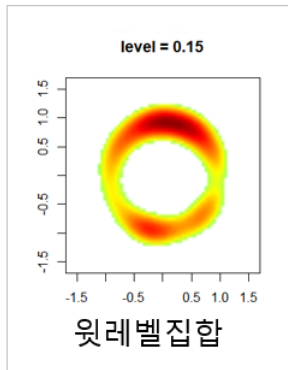
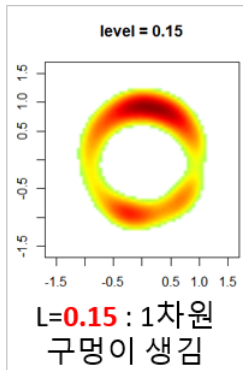
- ▶ 핵밀도추정(kernel density estimator)은 다음과 같습니다:

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

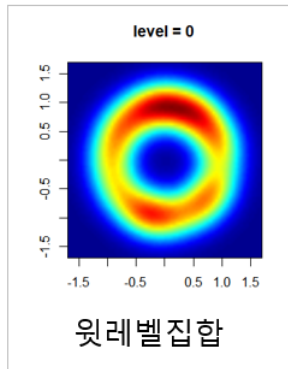
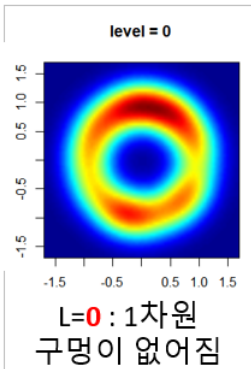
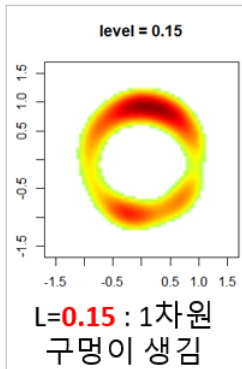
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



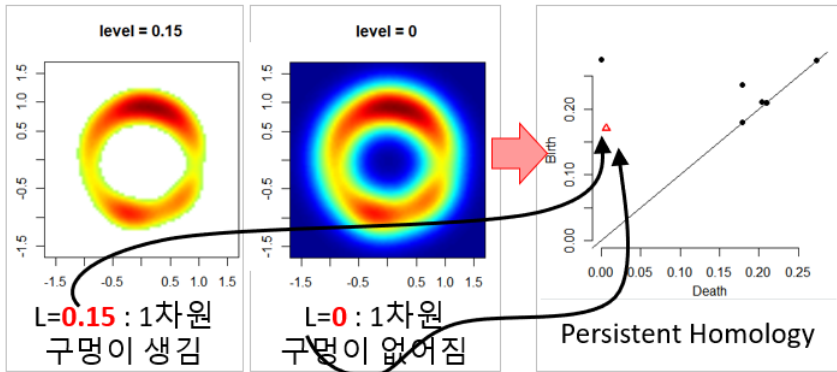
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



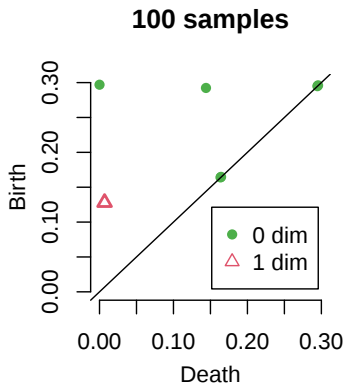
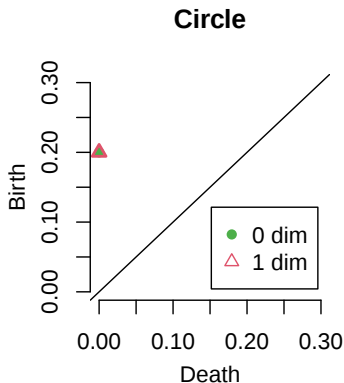
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



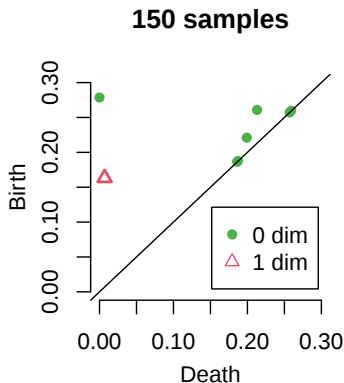
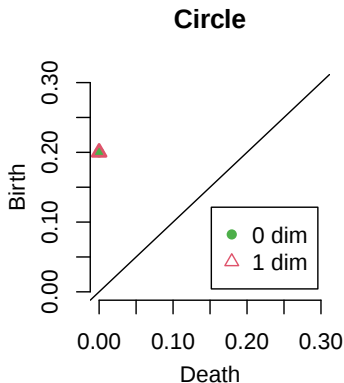
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



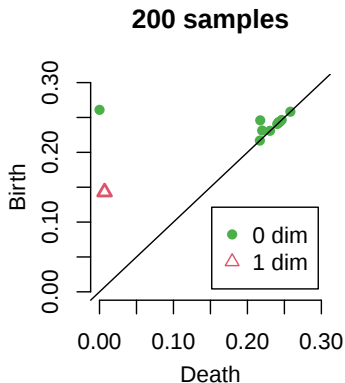
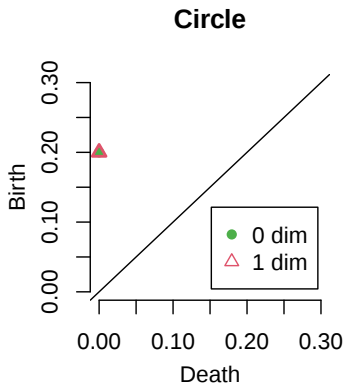
유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.



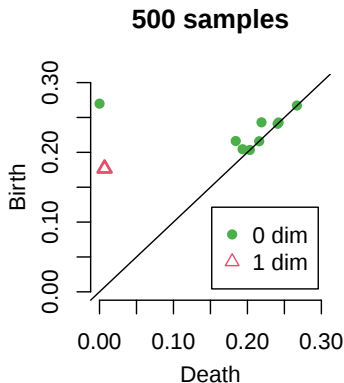
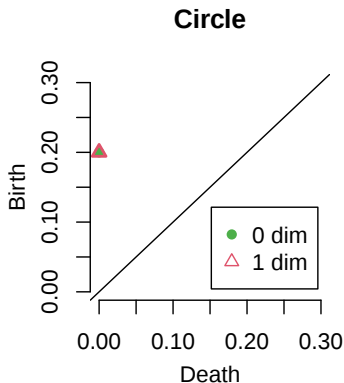
유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.



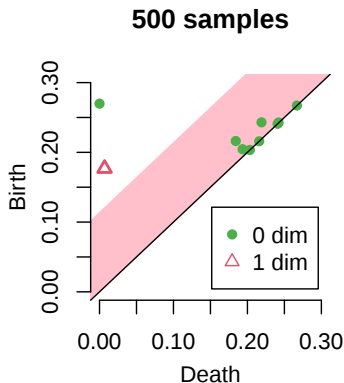
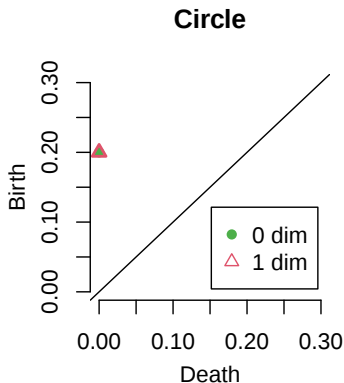
유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.



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



통계적으로 유의한 호몰로지 특성과 그렇지 않은 호몰로지 특성을 어떻게 구분할까요?



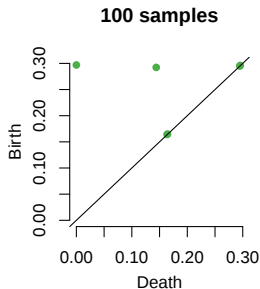
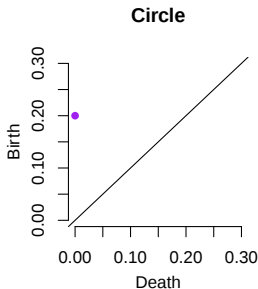
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

### Definition

$D_1, D_2$ 를 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,  $\gamma$ 는  $D_1$ 에서  $D_2$ 로 가는 모든 일대일대응이 될 수 있습니다.



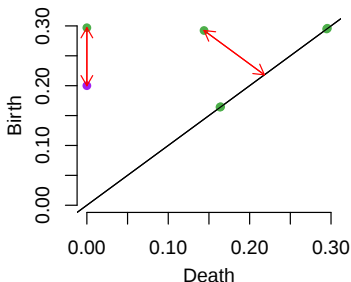
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

### Definition

$D_1, D_2$ 를 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,  $\gamma$ 는  $D_1$ 에서  $D_2$ 로 가는 모든 일대일대응이 될 수 있습니다.



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

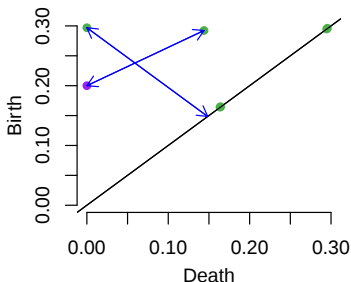
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

### Definition

$D_1, D_2$ 를 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,  $\gamma$ 는  $D_1$ 에서  $D_2$ 로 가는 모든 일대일대응이 될 수 있습니다.



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

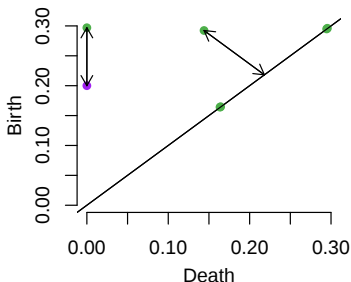
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

### Definition

$D_1, D_2$ 를 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,  $\gamma$ 는  $D_1$ 에서  $D_2$ 로 가는 모든 일대일대응이 될 수 있습니다.

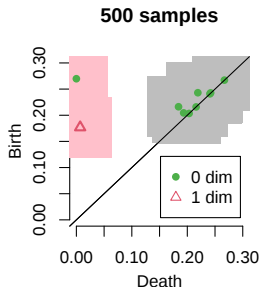
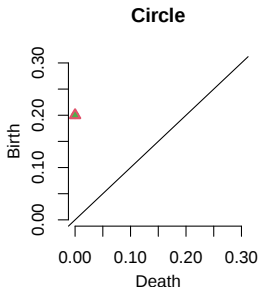


$$\inf_{\gamma} \sup_{x \in D_1} \|x - \gamma(x)\|_\infty = 0.1$$

# Persistent homology의 신뢰집합(Confidence Set)은 Persistent homology를 높을 확률로 포함하는 랜덤집합입니다.

기저  $M$ 과 자료  $X$ 의 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:

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

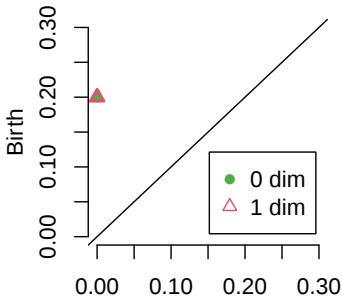


Persistent Homology의 신뢰띠(Confidence Band)를 이용하여 통계적으로 유의한 호몰로지 특성과 그렇지 않은 호몰로지 특성을 구분합니다.

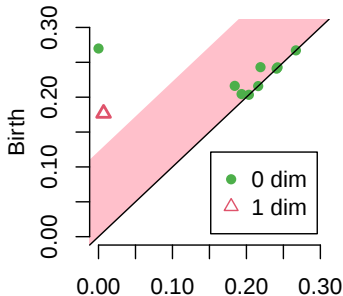
기저  $M$ 과 자료  $X$ 의 Persistent homology를 각각  $Dgm(M)$ 과  $Dgm(X)$ 라고 놓습니다. 유의수준  $\alpha \in (0, 1)$ 가 주어졌을 때,  $(1 - \alpha)$  신뢰띠 (Confidence Band)  $c_n = c_n(X)$ 는 다음을 만족하는 확률변수입니다:

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

**Circle**



**500 samples**



Persistent homology의 신뢰띠는 بوت스트랩으로 계산할 수 있습니다.

1. 주어진 자료  $X = \{x_1, \dots, x_n\}$ 에서 핵밀도추정(kernel density estimator)  $\hat{p}_h$ 를 계산합니다.
2.  $X = \{x_1, \dots, x_n\}$ 로부터  $X^* = \{x_1^*, \dots, x_n^*\}$ 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  $B$ 번 반복하여  $\theta_1^*, \dots, \theta_B^*$ 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.

## 위상 자료 분석(Topological Data Analysis) 소개

Persistent Homology를 통계적으로 추정하기

## 위상 자료 분석(Topological Data Analysis)을 자료분석 및 기계학습에 응용

Persistence Landscape를 이용하여 특성(Feature) 만들기

Circular Coordinates를 이용하여 특성(Feature) 만들기

위상 자료 분석(Topological Data Analysis)을 이용한 평가

R 패키지 TDA: 위상 자료 분석을 위한 통계 계산 도구

참조문헌

# 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)  $f_\theta$ ,
  - ▶ 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  $X$ 에 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  $X$ 나 모형  $f_\theta$ 의 품질을 TDA로 평가: 최근 주목
    - ▶ 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) 소개

Persistent Homology를 통계적으로 추정하기

위상 자료 분석(Topological Data Analysis)을 자료분석 및 기계학습에  
응용

Persistence Landscape를 이용하여 특성(Feature) 만들기

Circular Coordinates를 이용하여 특성(Feature) 만들기

위상 자료 분석(Topological Data Analysis)을 이용한 평가

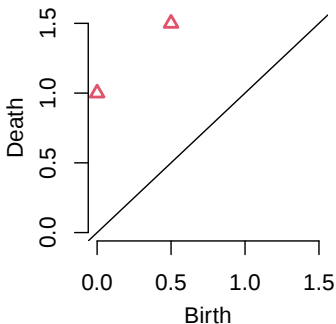
R 패키지 TDA: 위상 자료 분석을 위한 통계 계산 도구

참조문헌

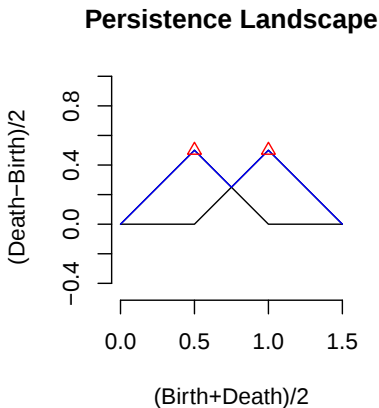
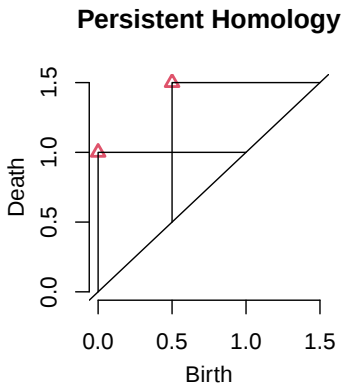
Persistent homology를 한 번 더 요약해서 유클리드 공간 또는 함수 공간에 넣습니다.

- ▶ 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.

### Persistent Homology

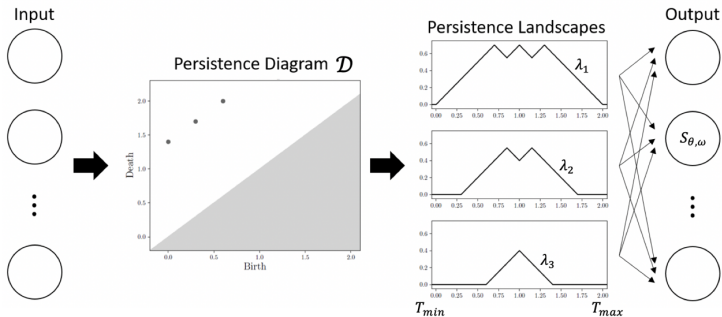


Persistence Landscape은 Persistent homology의 함수 요약입니다.



# Persistence Landscape으로 위상학적 층(topological layer) 만들기

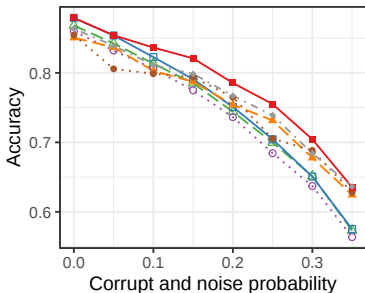
1. 자료  $X$ 의 Persistent Homology  $\mathcal{D}$ 를 계산합니다.
2.  $\mathcal{D}$ 로부터 Persistence Landscape  $\lambda : \mathbb{N} \times \mathbb{R} \rightarrow \mathbb{R}$ 을 계산합니다.
3. 매개변수  $\omega \in \mathbb{R}^{K_{\max}}$ 를 이용하여 가중평균함수  $\bar{\lambda}_{\omega}(t) := \sum_{k=1}^{K_{\max}} \omega_k \lambda_k(t)$ 를 계산하고, 이를 벡터화하여  $\bar{\Lambda}_{\omega} \in \mathbb{R}^m$ 을 만듭니다.
4. 매개화된 미분가능한 함수  $g_{\theta} : \mathbb{R}^m \rightarrow \mathbb{R}$ 을 사용하여,  $S_{\theta, \omega}(\mathcal{D}) := g_{\theta}(\bar{\Lambda}_{\omega})$ 를 계산합니다.



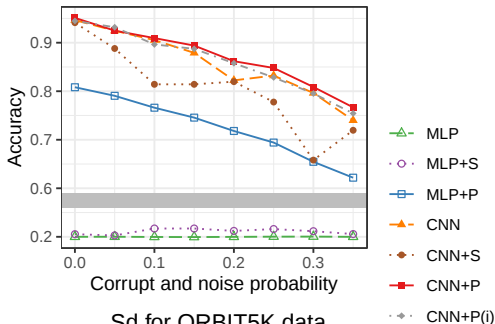
# Persistence Landscape으로 위상학적 층(topological layer) 만들기

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

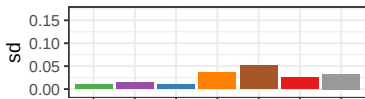
Accuracy for MNIST data



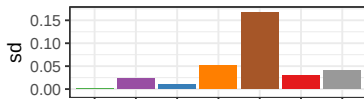
Accuracy for ORBIT5K data



Sd for MNIST data



Sd for ORBIT5K data



위상 자료 분석(Topological Data Analysis) 소개

Persistent Homology를 통계적으로 추정하기

위상 자료 분석(Topological Data Analysis)을 자료분석 및 기계학습에  
응용

Persistence Landscape를 이용하여 특성(Feature) 만들기

Circular Coordinates를 이용하여 특성(Feature) 만들기

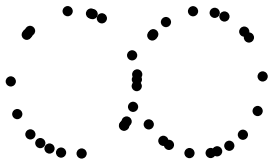
위상 자료 분석(Topological Data Analysis)을 이용한 평가

R 패키지 TDA: 위상 자료 분석을 위한 통계 계산 도구

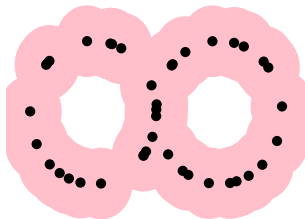
참조문헌

Circular coordinates 는 자료의 위상 구조를 반영하는 차원 축소 방법입니다.

data



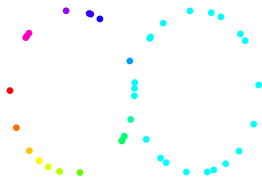
loop



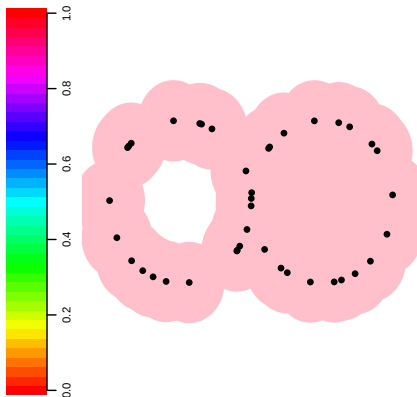
Circular coordinates 는 자료의 위상 구조를 반영하는 차원 축소 방법입니다.

- ▶ circular coordinate 는 자료  $X$  에서 원  $S^1$  으로 가는 함수입니다.

circular coordinates

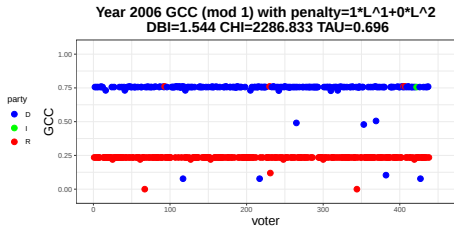
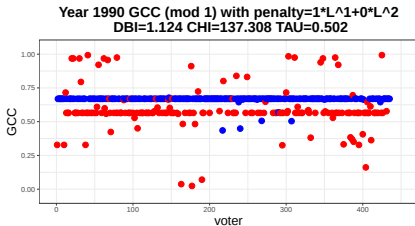


loop



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

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



위상 자료 분석(Topological Data Analysis) 소개

Persistent Homology를 통계적으로 추정하기

위상 자료 분석(Topological Data Analysis)을 자료분석 및 기계학습에  
응용

Persistence Landscape를 이용하여 특성(Feature) 만들기

Circular Coordinates를 이용하여 특성(Feature) 만들기

위상 자료 분석(Topological Data Analysis)을 이용한 평가

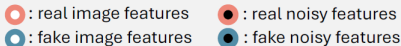
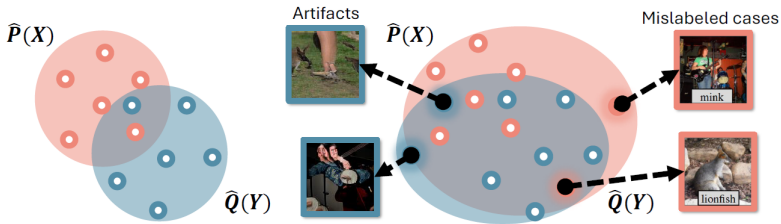
R 패키지 TDA: 위상 자료 분석을 위한 통계 계산 도구

참조문헌

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

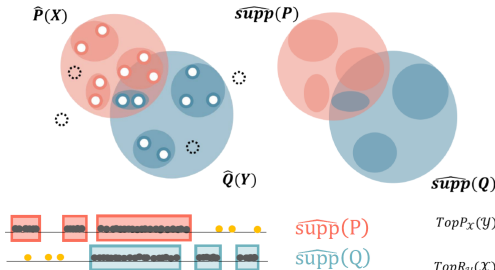
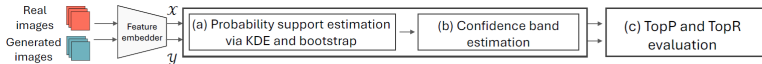
- ▶ TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models (Kim, Jang, Kim, Yoo, 2024)
- ▶ 생성 모형(generative model)을 평가(evaluate)할 때, 실제 화상(real image)의 분포(distribution)의 지지집합(support)과 가짜 화상(fake image)의 분포의 지지집합을 거리(metric)를 사용하여 비교합니다.
- ▶ 기존의 평가 거리(evaluation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

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



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)



$TopP \xrightarrow{n \rightarrow \infty} \text{precision}$

$TopR \xrightarrow{n \rightarrow \infty} \text{recall}$

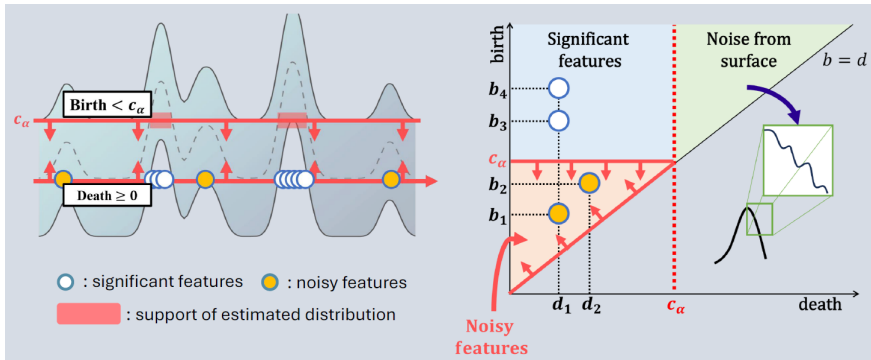
See our [proposition 4.1](#)  
and [theorem 4.2](#)

$$TopP_X(Y) = \frac{\sum_{j=1}^m 1(\hat{p}_h(Y_j) > c_X, \hat{q}_h(Y_j) > c_Y)}{\sum_{j=1}^m 1(\hat{q}_h(Y_j) > c_Y)}$$

$$TopR_Y(X) = \frac{\sum_{i=1}^n 1(\hat{q}_h(X_i) > c_Y, \hat{p}_h(X_i) > c_X)}{\sum_{i=1}^n 1(\hat{p}_h(X_i) > c_X)}$$

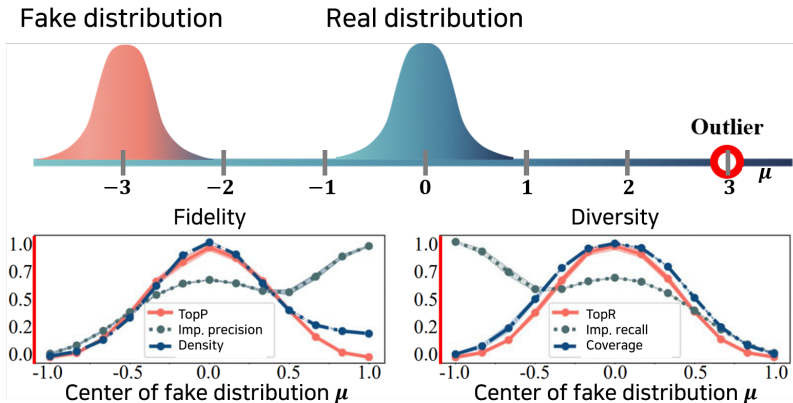
통계적 및 위상적으로 유의미한 특성들을 골라내는 문턱 (threshold)을 찾아냅니다.

- ▶ 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) 소개

Persistent Homology를 통계적으로 추정하기

위상 자료 분석(Topological Data Analysis)을 자료분석 및 기계학습에  
응용

Persistence Landscape를 이용하여 특성(Feature) 만들기

Circular Coordinates를 이용하여 특성(Feature) 만들기

위상 자료 분석(Topological Data Analysis)을 이용한 평가

R 패키지 TDA: 위상 자료 분석을 위한 통계 계산 도구

참조문헌

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.

## 위상 자료 분석(Topological Data Analysis) 소개

Persistent Homology를 통계적으로 추정하기

위상 자료 분석(Topological Data Analysis)을 자료분석 및 기계학습에  
응용

Persistence Landscape를 이용하여 특성(Feature) 만들기  
Circular Coordinates를 이용하여 특성(Feature) 만들기  
위상 자료 분석(Topological Data Analysis)을 이용한 평가

R 패키지 TDA: 위상 자료 분석을 위한 통계 계산 도구

참조문헌

- Peter Bubenik. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16:77–102, 2015. ISSN 1532-4435.
- 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 T. Fasy, Jisu Kim, Fabrizio Lecci, Clément Maria, David L. Millman, and Vincent Rouvreau. Introduction to the R package TDA. *CoRR*, abs/1411.1830, 2014a. URL <http://arxiv.org/abs/1411.1830>.
- Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *Ann. Statist.*, 42(6):2301–2339, 2014b. ISSN 0090-5364. doi: 10.1214/14-AOS1252. URL <https://doi.org/10.1214/14-AOS1252>.
- Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers Artif. Intell.*, 4:681108, 2021. doi: 10.3389/frai.2021.681108. URL <https://doi.org/10.3389/frai.2021.681108>.
- Kwangho Kim, Jisu Kim, 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.

## 참조문헌 III

Pum Jun Kim, Yoojin Jang, Jisu Kim, and Jaejun Yoo. TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models. *arXiv e-prints*, art. arXiv:2306.08013, June 2024. doi: 10.48550/arXiv.2306.08013.

Hengrui Luo, Alice Patania, Jisu Kim, and Mikael Vejdemo-Johansson. Generalized penalty for circular coordinate representation. *Foundations of Data Science*, 3(4):729–767, 2021.

Larry Wasserman. Topological data analysis, 2016.

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

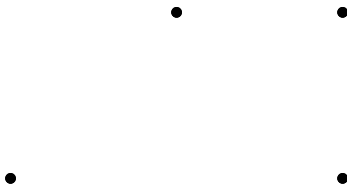
Persistent Homology와 Persistence Landscape

Persistence Homology와 Persistence Landscape의 통계적 추정

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  $E$ 로 이루어져 있으면서  $E \subset \{\{x, y\} \mid x, y \in \mathcal{X}, x \neq y\}$ 를 만족합니다.

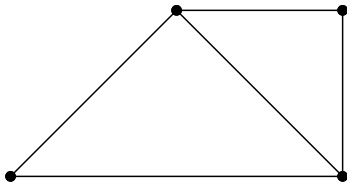
## Graph



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  $E$ 로 이루어져 있으면서  $E \subset \{\{x, y\} \mid x, y \in \mathcal{X}, x \neq y\}$ 를 만족합니다.

## Graph



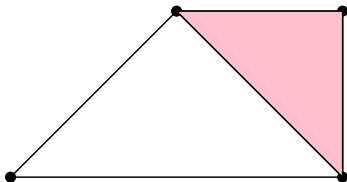
단체 복합체(Simplicial complex)는 고차원으로 일반화한 그래프입니다.

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

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

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

## Simplicial complex

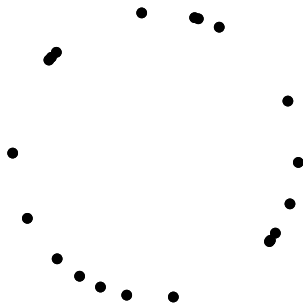


Vietoris-Rips 복합체(Vietoris-Rips complex)는 서로 가까운 꼭지점들을 모아 놓은 복합체입니다.

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

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

### Vietoris-Rips complex

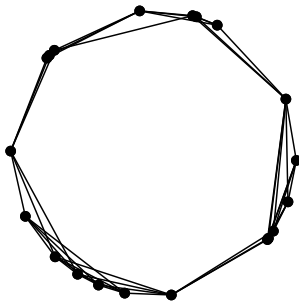


Vietoris-Rips 복합체(Vietoris-Rips complex)는 서로 가까운 꼭지점들을 모아 놓은 복합체입니다.

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

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

### Vietoris-Rips complex

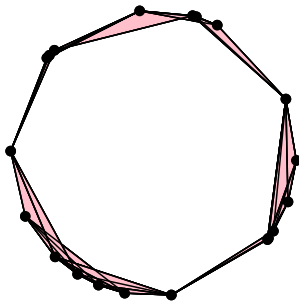


Vietoris-Rips 복합체(Vietoris-Rips complex)는 서로 가까운 꼭지점들을 모아 놓은 복합체입니다.

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

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

### Vietoris-Rips complex



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

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

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

- ▶  $\mathbb{Z}/2\mathbb{Z}$ 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:

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

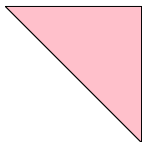
- ▶  $K$ 의  $k$ -연쇄를 모은 집합  $C_k(K)$ 는 선형 공간이 됩니다.

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

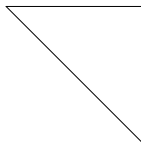
- ▶ 경계 사상(boundary map)  $\partial_k$ 은 각  $k$ -차원 단체  $\sigma$ 를 그의  $k-1$ 차원 면 (face)들의 합으로 보냅니다:

$$\sigma = \{v_0, \dots, v_k\} \mapsto \partial_k \sigma = \sum_{i=0}^k \{v_0, \dots, v_k\} \setminus \{v_i\}.$$

Simplex



Sum of Faces



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

- ▶ 경계 사상(boundary map)  $\partial_k$ 은 각  $k$ -차원 단체  $\sigma$ 를 그의  $k-1$ 차원 면(face)들의 합으로 보냅니다:

$$\sigma = \{v_0, \dots, v_k\} \mapsto \partial_k \sigma = \sum_{i=0}^k \{v_0, \dots, v_k\} \setminus \{v_i\}.$$

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

$$\begin{aligned} C_k(K) &\mapsto C_{k-1}(K) \\ c = \sum a_i \sigma_i &\mapsto \partial_k c = \sum a_i \partial_k \sigma_i \end{aligned}$$



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

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

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

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

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

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

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

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

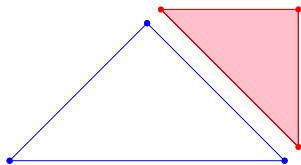
호몰로지(homology)는 cycle을 boundary로 자른 몫공간(quotient space)입니다.

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

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

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

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



▶  $Z_1(K) = \ker \partial_1 = (\mathbb{Z}/2\mathbb{Z})^2 = \langle \text{blue triangle}, \text{red triangle} \rangle$

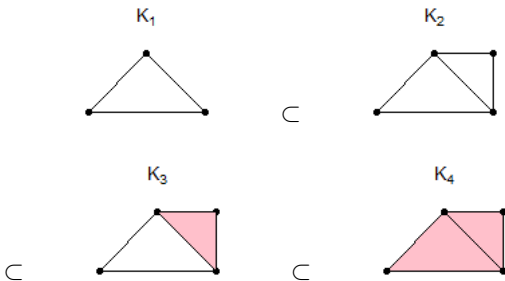
▶  $B_1(K) = \text{im} \partial_2 = \mathbb{Z}/2\mathbb{Z} = \langle \text{red triangle} \rangle$

▶  $H_1(K) = Z_1(K)/B_1(K) = \mathbb{Z}/2\mathbb{Z} = \langle \text{blue triangle} \rangle, \beta_1(K) = 1$

filtration은 증가하는 단체 복합체들의 모임입니다.

- ▶ 단체 복합체  $K$ 가 있을 때, filtration  $\mathcal{F} = \{K_a\}_{a \in \mathbb{R}}$ 는 다음을 만족하는  $K$ 의 부분 복합체(subcomplex)  $K_a$ 들의 모임입니다:

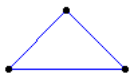
$$a \leq b \implies K_a \subset K_b.$$



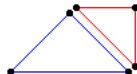
# Persistent Homology는 filtration에서 호몰로지가 어떻게 변화하는지 추적합니다.

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

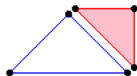
$$H_1(K_1) = \mathbb{Z}/2\mathbb{Z}$$


 $\subset$ 

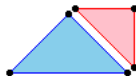
$$H_1(K_2) = (\mathbb{Z}/2\mathbb{Z})^2$$



$$H_1(K_3) = \mathbb{Z}/2\mathbb{Z}$$


 $\subset$ 

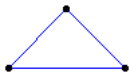
$$H_1(K_4) = 0$$


 $\subset$

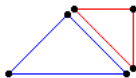
# Persistent Homology는 filtration에서 호몰로지가 어떻게 변화하는지 추적합니다.

- ▶ 단체 복합체  $K$  위의 filtration  $\mathcal{F} = \{K_a \subset K : a \in \mathbb{R}\}$ 가 있을 때,  $k$ -th persistent homology  $\text{PH}_k \mathcal{F}$ 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  $\gamma$ 는  $K_a$ 에서 생기고  $K_b$ 에서  $\gamma = 0$ 이 됩니다. 이 때,  $a$ 를  $\gamma$ 의 birth time이라 하고,  $b$ 를  $\gamma$ 의 death time이라고 합니다.

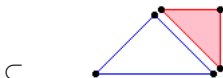
$$H_1(K_1) = \mathbb{Z}/2\mathbb{Z}$$


 $\subset$ 

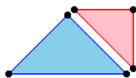
$$H_1(K_2) = (\mathbb{Z}/2\mathbb{Z})^2$$



$$H_1(K_3) = \mathbb{Z}/2\mathbb{Z}$$


 $\subset$ 

$$H_1(K_4) = 0$$


 $\subset$

Persistence Diagram 은 Persistent Homology 를 평면 위의 점들로 나타냅니다.

- ▶ 편의상 filtration  $\mathcal{F} = \{K_a \subset K : a \in \mathbb{R}\}$ 의 부분단체  $K_a$ 들이 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.

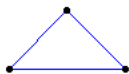
# Persistence Diagram 은 Persistent Homology 를 평면 위의 점들로 나타냅니다.

- ▶ 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:

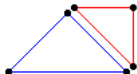
$$\mu_k^{a_i, a_j} = (\beta_k^{a_i, a_j-1} - \beta_k^{a_i, a_j}) - (\beta_k^{a_i-1, a_j-1} - \beta_k^{a_i-1, a_j}).$$

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

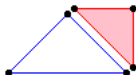
$$H_1(K_1) = \mathbb{Z}/2\mathbb{Z}$$


 $\subset$ 

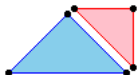
$$H_1(K_2) = (\mathbb{Z}/2\mathbb{Z})^2$$



$$H_1(K_3) = \mathbb{Z}/2\mathbb{Z}$$


 $\subset$ 

$$H_1(K_4) = 0$$



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

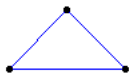
# Persistence Diagram 은 Persistent Homology 를 평면 위의 점들로 나타냅니다.

- ▶ 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:

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

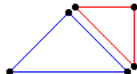
- ▶  $(\mathbb{R} \cup \{\infty\})^2$  위에 점  $(a_i, a_j)$  를 multiplicity  $\mu_k^{a_i, a_j}$  로 찍으면 persistence diagram 이 됩니다.

$$H_1(K_1) = \mathbb{Z}/2\mathbb{Z}$$

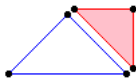


$\subset$

$$H_1(K_2) = (\mathbb{Z}/2\mathbb{Z})^2$$

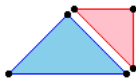


$$H_1(K_3) = \mathbb{Z}/2\mathbb{Z}$$

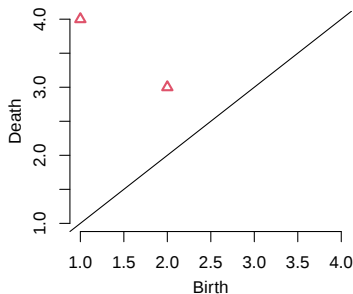


$\subset$

$$H_1(K_4) = 0$$



$\subset$



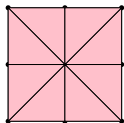
# 함수의 레벨집합으로부터 filtration을 만들 수 있습니다.

- ▶ 단체 복합체  $K$ 와 그 위에서 정의된 함수  $f : K \rightarrow \mathbb{R}$ 가 있을 때,  $f$ 의 sub-level filtration  $\text{sub}(f)$ 를 다음과 같이 정의합니다:

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

Simplicial complex

Function values

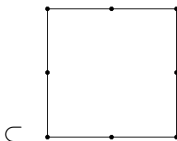


|   |   |   |
|---|---|---|
| 1 | 0 | 1 |
| 0 | 2 | 0 |
| 1 | 0 | 1 |

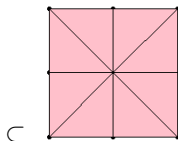
$\{f \leq 0\}$



$\{f \leq 1\}$



$\{f \leq 2\}$



함수의 레벨집합으로부터 filtration을 만들 수 있습니다.

- ▶ 단체 복합체  $K$ 와 그 위에서 정의된 함수  $f : K \rightarrow \mathbb{R}$ 가 있을 때,  $f$ 의 아레벨(sub-level) filtration  $\text{sub}(f)$ 를 다음과 같이 정의합니다:

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

- ▶ 마찬가지로  $f$ 의 윗레벨(super-level) filtration  $\text{super}(f)$ 를 다음과 같이 정의합니다:

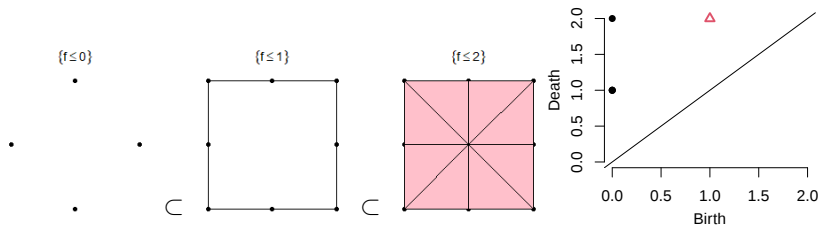
$$\text{super}(f) := \{ \{ \sigma \in K : f(\sigma) \geq L \} \}_{L \in \mathbb{R}}.$$

함수의 레벨집합으로부터 persistent homology를 계산할 수 있습니다.

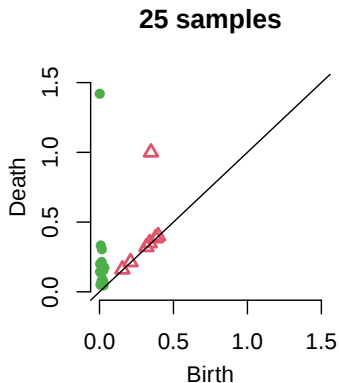
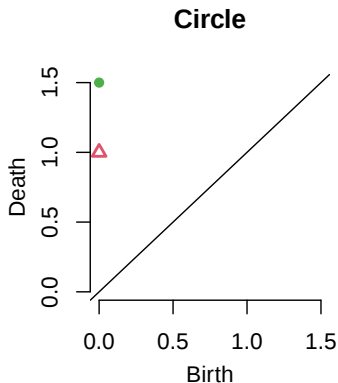
- ▶ 단체 복합체  $K$ 와 그 위에서 정의된 함수  $f : K \rightarrow \mathbb{R}$ 가 있을 때,  $f$ 의 아랫레벨(sub-level) filtration  $\text{sub}(f)$ 를 다음과 같이 정의합니다:

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

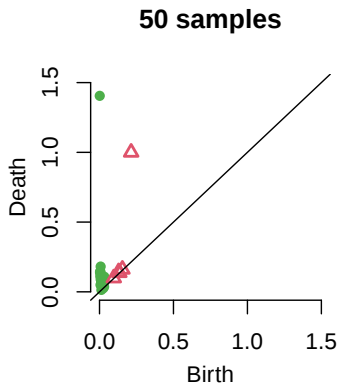
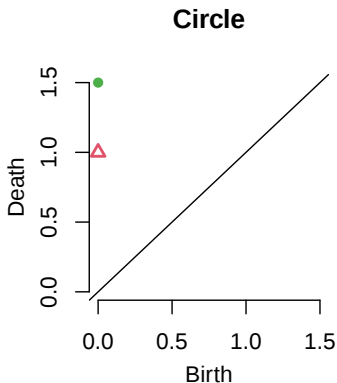
- ▶ 그로부터 계산한 persistent homology 또는 persistence diagram을  $Dgm(f)$ 로 씁니다.



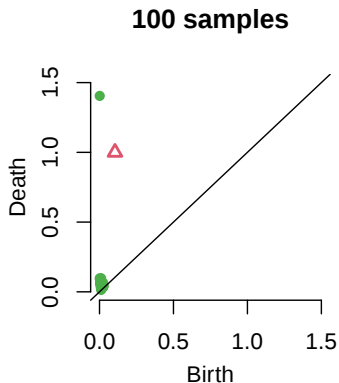
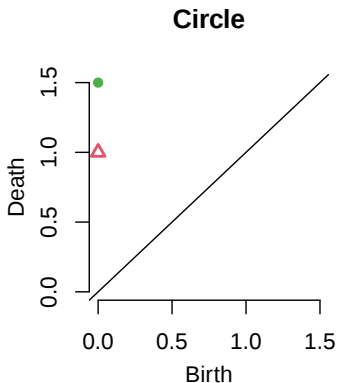
유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.



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



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



## 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

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

### Theorem

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

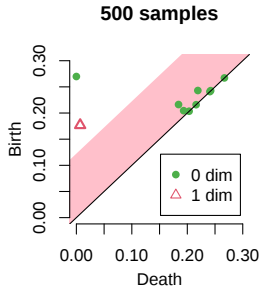
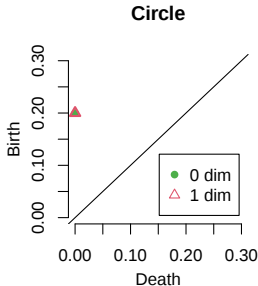
$$\mathbb{P}(W_\infty(Dgm(f_M), Dgm(f_X)) \leq c_n) \geq \mathbb{P}(\|f_M - f_X\|_\infty \leq c_n) \geq 1 - \alpha,$$

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

# Persistent homology의 신뢰되는 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.



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

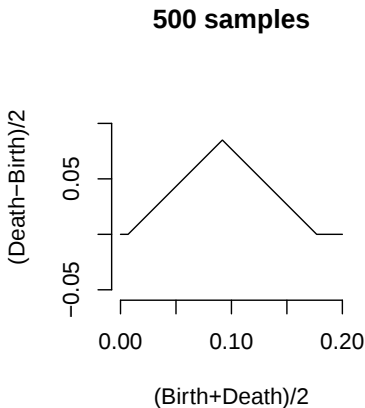
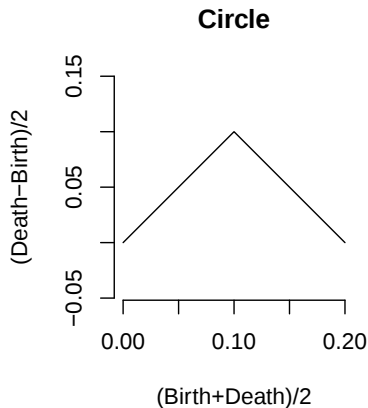
호몰로지(Homology)와 Persistent Homology

Persistent Homology를 통계적으로 추정하기

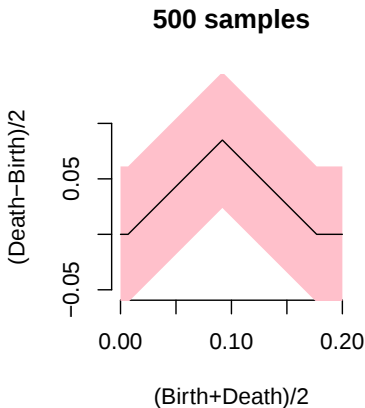
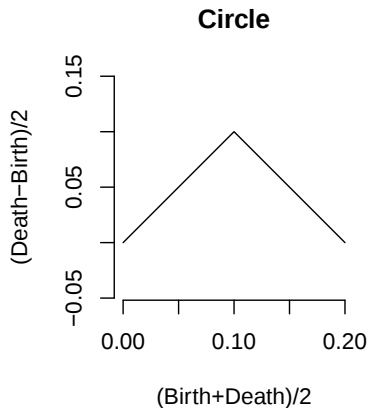
위상 자료 분석(Topological Data Analysis)을 기계학습에 응용  
Persistence Landscape를 이용하여 특성(Feature) 만들기  
Circular Coordinates를 이용하여 특성(Feature) 만들기

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

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



Persistent homology의 신뢰도로 Persistence Landscape의 랜덤성을 정량화할 수 있습니다.

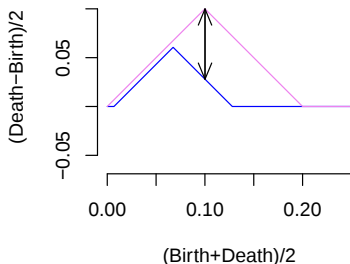


$\infty$ -landscape 거리는 persistence landscape 공간에 거리를 줍니다.

### Definition

[Bubenik, 2015]  $D_1, D_2$ 을 점들의 multiset이라 하고, 그에 해당하는 persistence landscape를  $\lambda_1, \lambda_2$ 라고 놓습니다.  $\infty$ -landscape 거리는 다음과 같이 정의합니다:

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



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

### Theorem

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

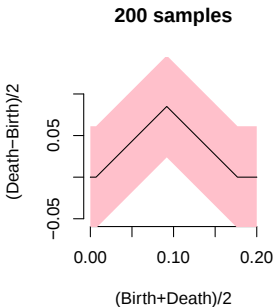
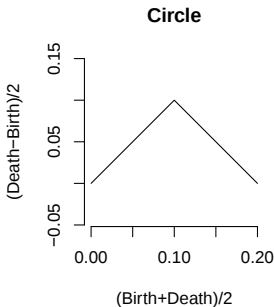
$$\Lambda_{\infty}(\lambda(f), \lambda(g)) \leq \|f - g\|_{\infty}.$$

persistence landscape의 신뢰띠는 بوت스트랩으로 계산할 수 있습니다.

- ▶ 기저  $M$ 과 표본  $X$ 의 persistence landscape를 각각  $\lambda_M$ 과  $\lambda_X$ 로 놓습니다. 안정성 정리(stability theorem)로부터,  
 $\mathbb{P}(\|f_M - f_X\| \leq c_n) \geq 1 - \alpha$ 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  $f_M$ 의 신뢰띠를 persistence landscape  $\lambda_M$ 의 신뢰띠로 사용할 수 있습니다.



persistence landscape의 신뢰도는 부트스트랩으로 계산할 수 있습니다.

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

PLLayer는 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])

PLayer 함수  $S_{\theta, \omega}$ 는 입력  $X$ 에 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  $W_{\infty}$ 는 bottleneck distance입니다.

PLLay는 안정적(stable)입니다.

▶ PLLay는 입력  $X$ 의 변화에 대해 안정적(stable)입니다:

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

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

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

여기서  $W_2$ 는 2-Wasserstein distance입니다.

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

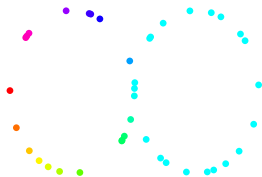
Persistent Homology와 Persistence Landscape

Persistence Homology와 Persistence Landscape의 통계적 추정

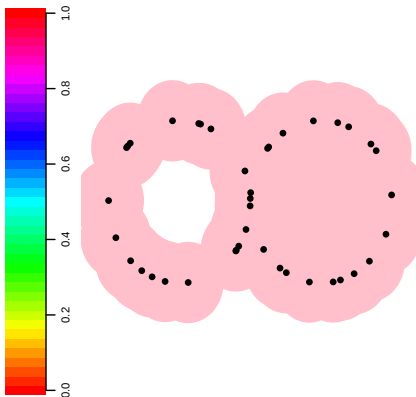
Circular Coordinates 는 자료의 위상 구조를 반영하는 차원 축소 방법입니다.

- ▶ circular coordinate 는 자료  $X$  에서 원  $S^1$  으로 가는 함수입니다.

circular coordinates



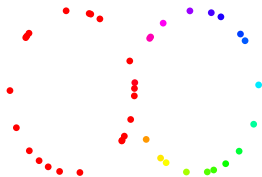
loop



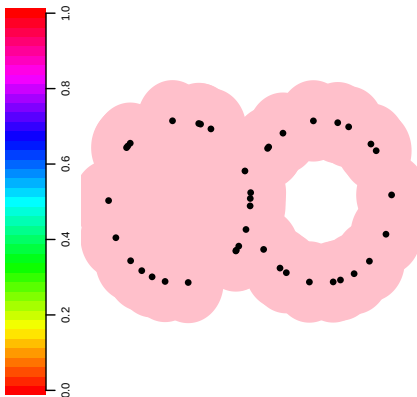
Circular Coordinates 는 자료의 위상 구조를 반영하는 차원 축소 방법입니다.

- ▶ circular coordinate 는 자료  $X$  에서 원  $S^1$  으로 가는 함수입니다.

circular coordinates



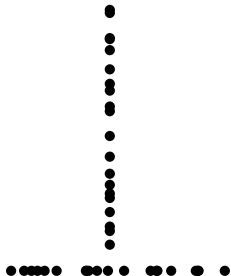
loop



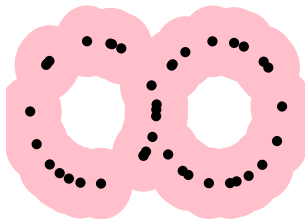
Circular Coordinates 는 자료의 위상 구조를 반영하는 차원 축소 방법입니다.

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

**circular coordinates**

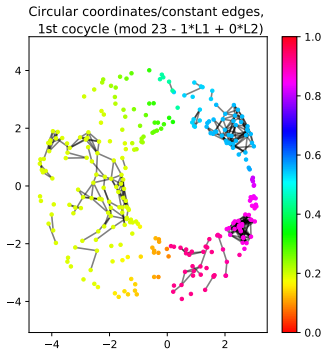
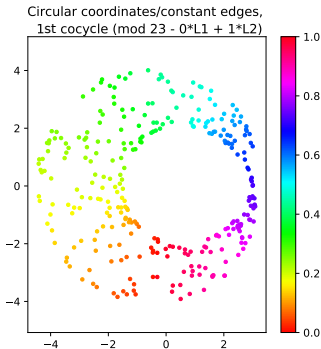


**loop**



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 푼다.
- ▶  $L_2$  손실(loss)을  $L_1$  손실로 바꿈으로써 circular coordinate 값이 더 급격하게 바뀌게 할 수 있습니다: 자료의 위상적인 정보를 더 잘 시각화합니다.



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

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

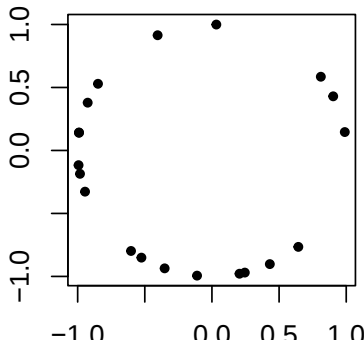
Persistent Homology와 Persistence Landscape

Persistence Homology와 Persistence Landscape의 통계적 추정

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

함수 `circleUnif()`는  $\mathbb{R}^2$  상에 있는 반지름이  $r$ 인 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)  $\hat{p}_h : \mathbb{R}^d \rightarrow [0, \infty)$ 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  $h$ 는 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")
```

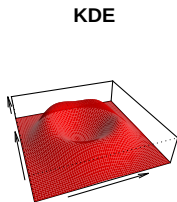
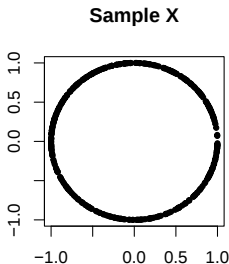
## R 패키지 TDA는 격자 위에서의 핵밀도추정(KDE)을 제공합니다.

가우스 핵밀도추정 (Kernel Density Estimator, KDE)  $\hat{p}_h : \mathbb{R}^d \rightarrow [0, \infty)$ 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  $h$ 는 평활매개변수(smoothing parameter)입니다.

함수 `kde()`는 격자 위의 점에서 핵밀도추정(KDE)  $\hat{p}_h$ 를 계산합니다.



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

## 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.

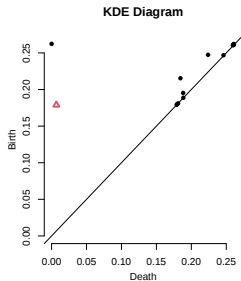
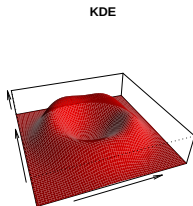
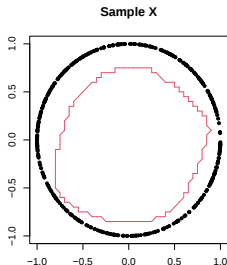
## R 패키지 TDA는 격자 위에서의 Persistent Homology를 계산합니다.

```
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.

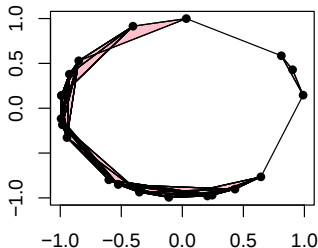
- ▶ 함수 `gridDiag()`는 입력함수의 아랫레벨(sublevel) 및 윗레벨(superlevel) 집합들의 persistence diagram을 계산합니다.
  - ▶ `gridDiag()`는 격자 위에서 실수값 입력함수를 계산합니다.
  - ▶ `gridDiag()`는 입력함수의 값으로 단체(simplex)들의 filtration을 만듭니다.
  - ▶ `gridDiag()`는 filtration의 persistent homology를 계산합니다.
- ▶ 사용자는 persistent homology를 계산하는 데에 C++ 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,

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



- ▶ Vietoris-Rips filtration은 Vietoris-Rips 복합체에서  $r$ 을 서서히 증가시키면서 만들어집니다.

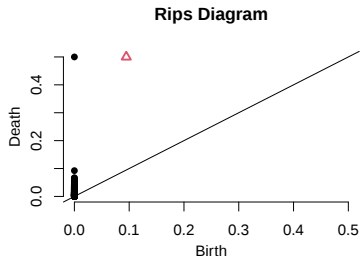
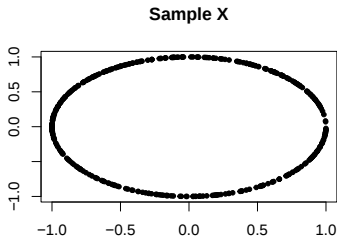
## R 패키지 TDA는 Vietoris-Rips Persistent Homology를 계산합니다.

- ▶ 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  $D$ 의 birth-death 쌍  $(b, d)$ 로부터 점  $p = (x, y) = (\frac{b+d}{2}, \frac{d-b}{2})$ 를 생각하고, 이  $p$ 를 꼭지점으로 한 텐트 모양의 함수  $\Lambda_p$ 를 생각합니다.
- ▶  $D$ 의 persistence landscape는 다음과 같은 함수들의 모임입니다:

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

여기서  $k \max$ 는 집합에서  $k$ 번째로 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)
```

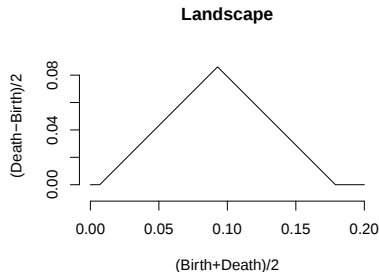
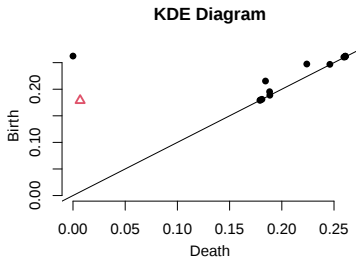
## R 패키지 TDA는 Persistence Landscape를 계산합니다.

- ▶ persistence diagram  $D$ 의 birth-death 쌍  $(b, d)$ 로부터 점  $p = (x, y) = (\frac{b+d}{2}, \frac{d-b}{2})$ 를 생각하고, 이  $p$ 를 꼭지점으로 한 텐트 모양의 함수  $\Lambda_p$ 를 생각합니다.
- ▶  $D$ 의 persistence landscape는 다음과 같은 함수들의 모임입니다:

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

여기서  $k\max$ 는 집합에서  $k$ 번째로 큰 값을 줍니다.

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



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

## 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.

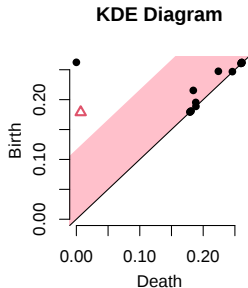
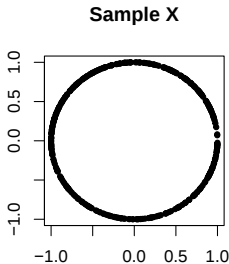
```
bandKDE <- bootstrapBand(X = X, FUN = kde, Grid = Grid, B = 20,  
  parallel = FALSE, alpha = 0.1, h = h)  
print(bandKDE[["width"]])
```

```
##          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.

