321.621A: 통계이론세미나 - 위상구조의 통계적 추정

321.621A: Seminar in Recent Development of Statistical Theories - Statistics on Topological Structure

2023 가을학기

김지수

2023-09-05

위상 구조(Topological Structure)의 통계적 추정(Statistical Inference)

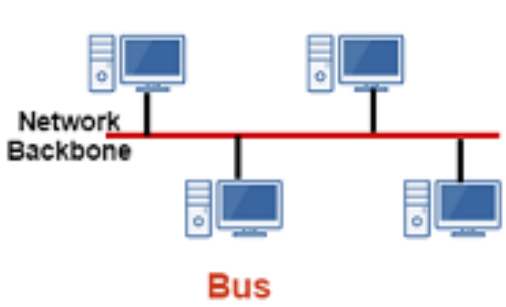
2581 = ?

This problem can be solved by pre-school children in five to ten minutes, by programmers in an hour and by people with higher education... well, check it yourself!

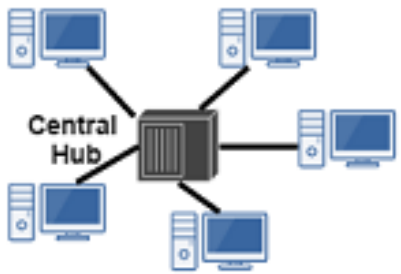
8809 = 6	5555 = 0
7111 = 0	8193 = 3
2172 = 0	8096 = 5
6666 = 4	1012 = 1
1111 = 0	7777 = 0
3213 = 0	9999 = 4
7662 = 2	7756 = 1
9313 = 1	6855 = 3
0000 = 4	9881 = 5
2222 = 0	5531 = 0
3333 = 0	2581 = ???

위상: 국소적인 부분들이 대역적으로 어떻게 연결되어 있는지

- 보통은 기하의 일부분으로 취급합니다.



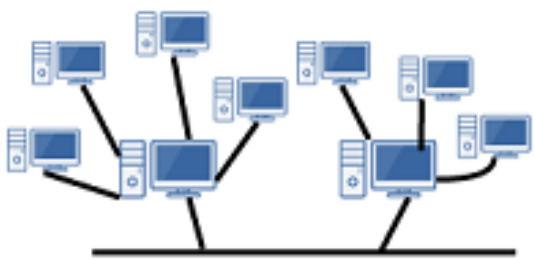
Bus



Star

Ring

Mesh



Tree

Point-to-Point

1

[1] https://systemzone.net/computer-network-topology-outline/

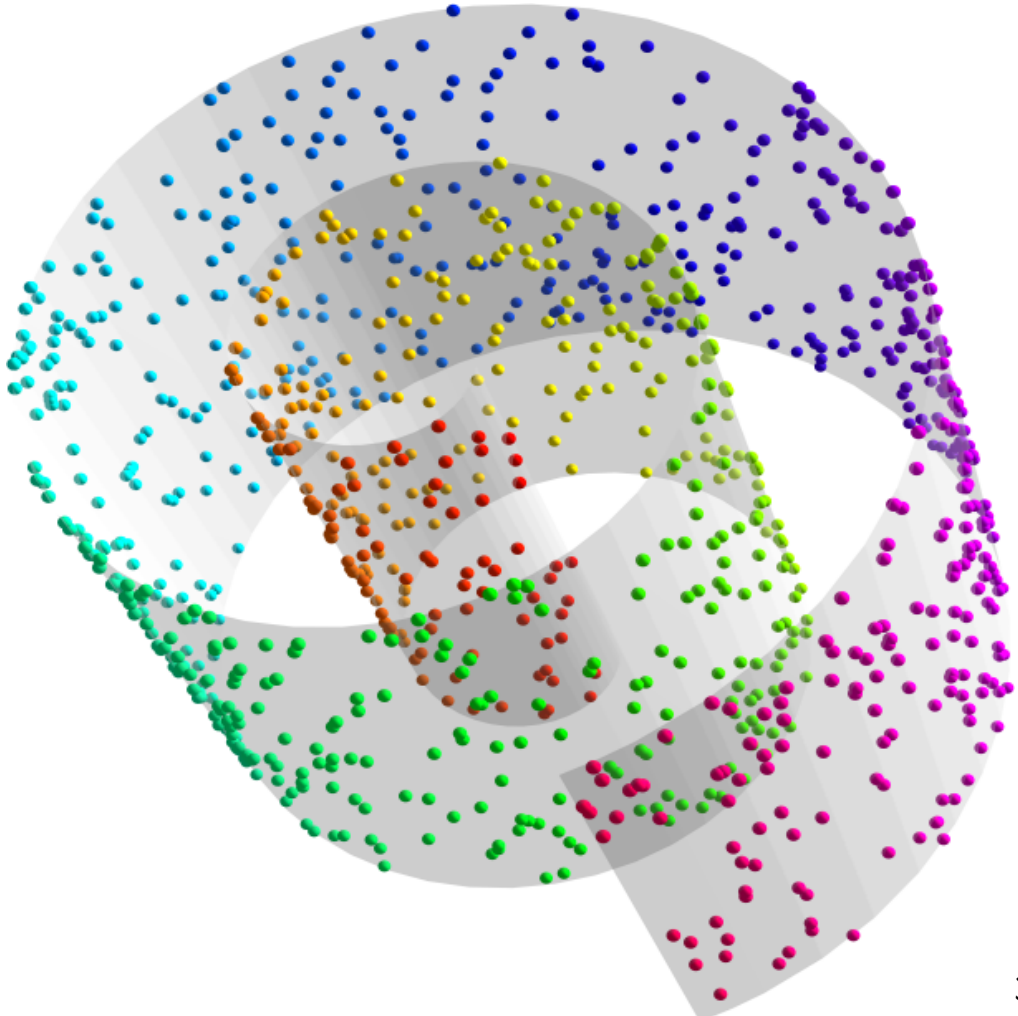
위상의 시초: 한붓그리기

- 쾨니히스베르크(현재 칼리닌그라드)의 7개의 다리를 한 번씩만 건널 수 있을까요?
- 레온하르트 오일러가 그래프로 도식화하여 접근: 차수가 홀수인 꼭지점의 개수가 0개이거나 2개여야만 한붓그리기가 가능합니다.

[2]

[2] https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg

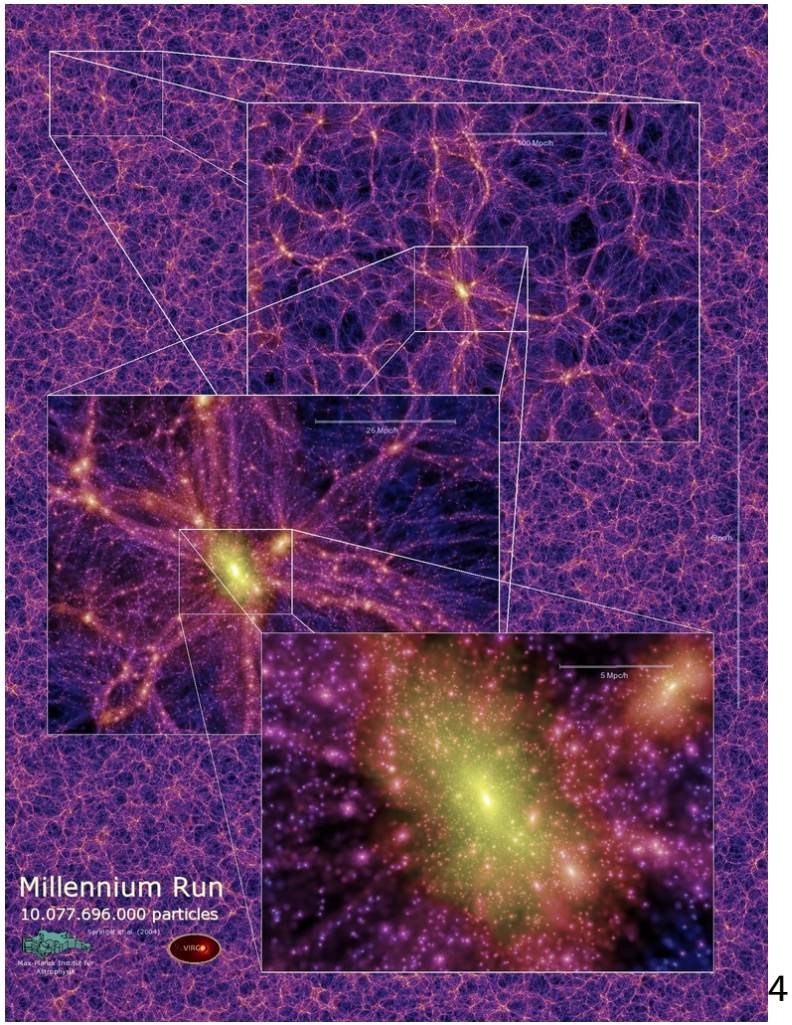
자료에 저차원 기하 구조를 줌으로써 자료의 차원이 높은 데서 오는 문제를 피할 수 있습니다.



3

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

자료의 위상학적 구조로부터 정보를 얻을 수 있습니다.

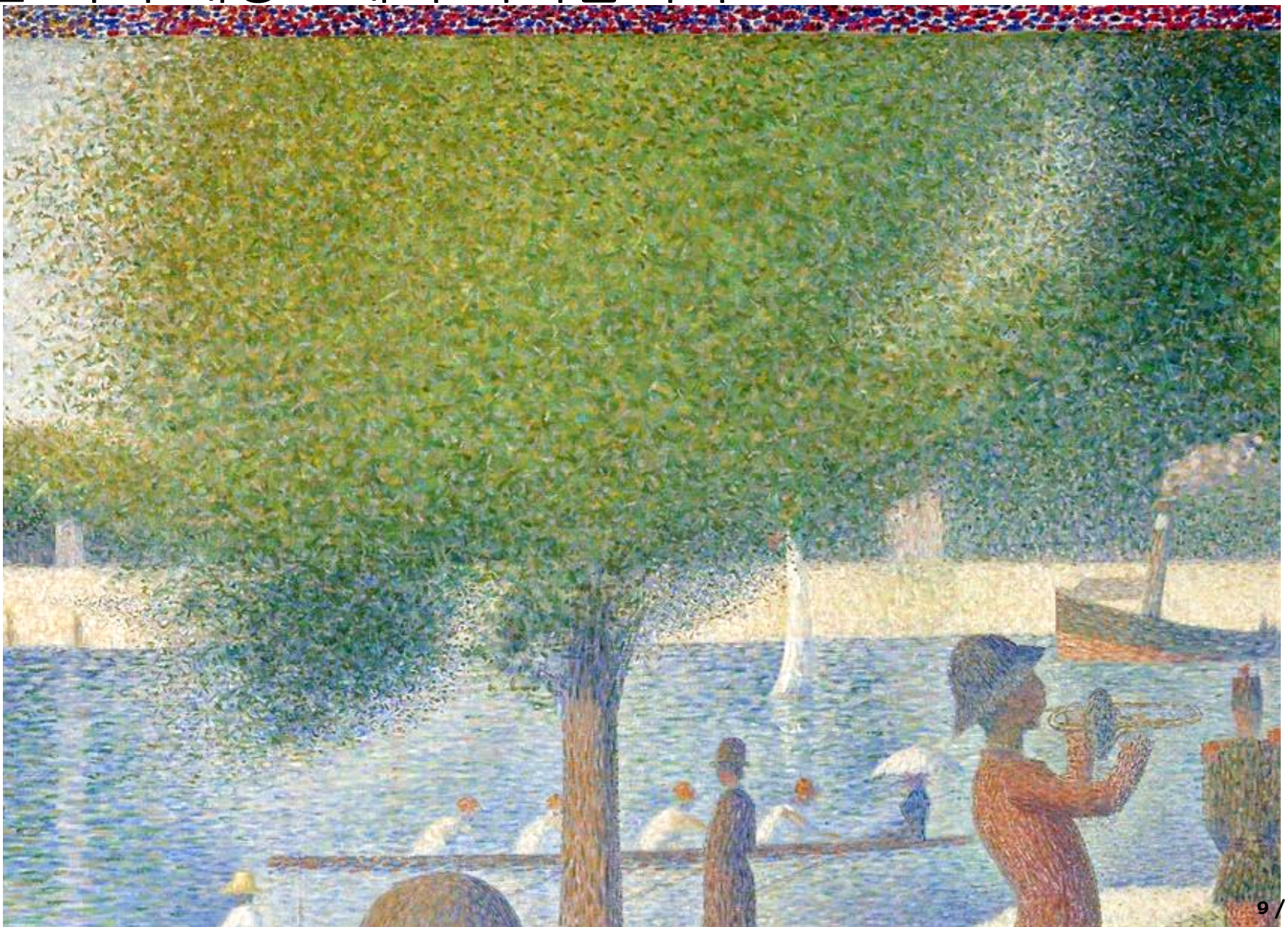


[4]

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

자료로부터 위상 구조 계산하기: 위상 구조를 하나의, 또는 여러 해상도에서 바라봅니다.

자료로부터 위상 구조 계산하기: 위상 구조를 하나의, 또는 여러 해상도에서 바라봅니다.



자료로부터 위상 구조 계산하기: 위상 구조를 하나의, 또는 여러 해상도에서 바라봅니다.

자료로부터 위상 구조 계산하기: 위상 구조를 하나의, 또는 여러 해상도에서 바라봅니다.

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

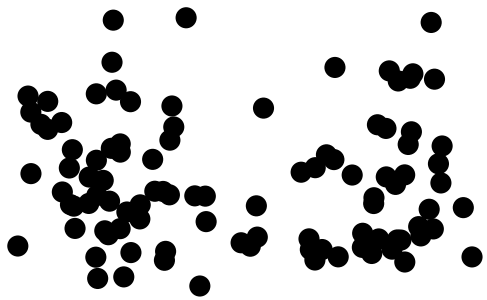
자료의 위상 구조(Topological Structure)를 통계적으로 어떻게 활용하고 추정하는지 소개합니다.

- 위상학적 자료 분석(Topological Data Analysis) 의 다음 두 서베이 논문을 주로 참조합니다.
  - Topological Data Analysis (Wasserman, 2016)
  - An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists (Chazal, Michel, 2021)
- 수업 전반적으로 다음 문헌들을 참조합니다.
  - Computational Topology: An Introduction (Edelsbrunner, Harer, 2010)
  - Algebraic topology (Hatcher, 2002)
- 그 외 각 주제에 맞춰서 많은 문헌들을 참조합니다.

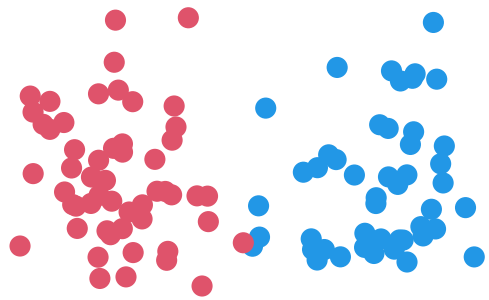
군집분석은 자료들을 같은 군집에 속하는 자료들끼리 좀 더 유사하도록 몇 개의 군집으로 묶는 과정입니다.

- 자료에 목표가 되는 군집이 표기되어 있지 않기 때문에, 군집분석은 비지도학습으로 분류됩니다.
- 군집분석을 모집단의 특성을 배우는 것으로 해석할 수 있습니다.

Sample

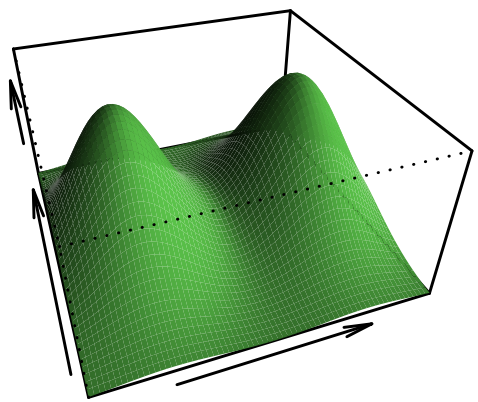


Clustered sample

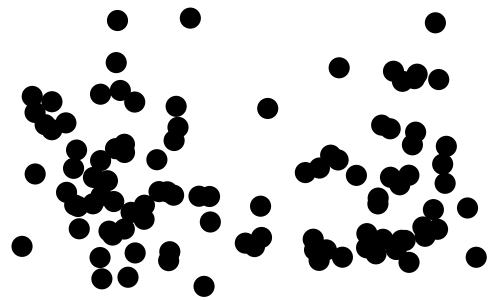


밀도 군집(Density Clustering)은 밀도함수를 사용하여 군집으로 묶습니다.

True density



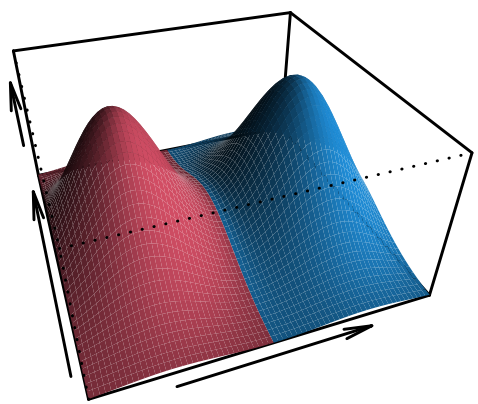
Sample



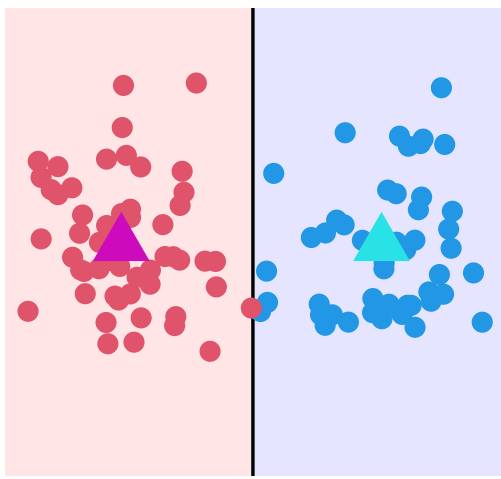
Mean Shift 알고리즘은 최빈값의 basins of attraction을 목표 군집으로 삼습니다.

- x로부터의 적분 경로 (또는 기울기 상승 경로) 는 경로 $\pi_x : \mathbb{R} \to \mathbb{R}^d$ 로써 $\pi_x(0) = x$ 와 $\frac{d}{dt}\pi_x(t) = \nabla p(\pi_x(t))$ 를 만족합니다.
- 각 국소적 최빈값 m_j의 basin of attraction을 집합 $\mathcal{A}_j = \left\{x : \lim_{t\to\infty} \pi_x(t) = m_j\right\}$ 로 정의합니다.
- Mean Shift 알고리즘은 최빈값의 basins of attraction을 자료로부터 알고리즘으로 근사하여 계산합니다.

True density



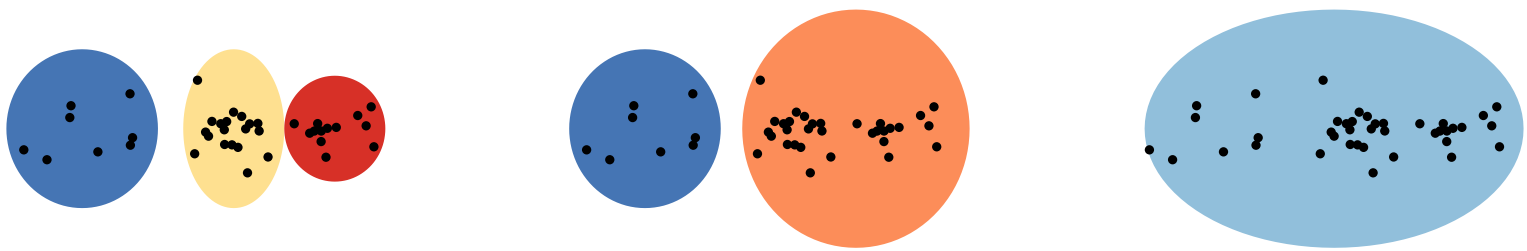
Basin of Atraction



자료를 해상도에 따라 다른 군집으로 묶고자 합니다.

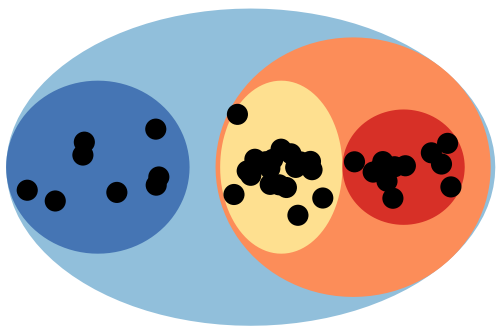
원하는 해상도에 따라 다른 군집이 생길 수 있습니다.

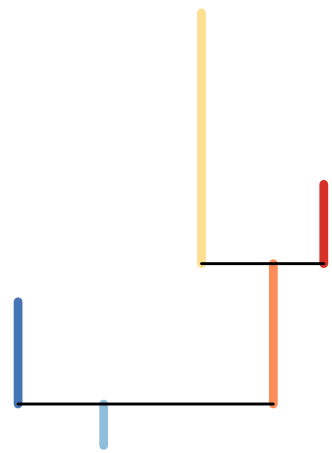
- 국소적(local)이고 상세한 정보를 묘사하고 싶으면 (높은 해상도), 작은 규모의 많은 군집이 생깁니다.
- 대역적(global)이고 개략적인 정보를 묘사하고 싶으면 (낮은 해상도), 큰 규모의 적은 군집이 생깁니다.



군집들의 네트워크가 나무를 형성합니다: 군집 나무 (cluster tree)

- 다른 수준의 해상도로부터 얻어지는 군집들은 포함 관계에 의해 자연스러운 네트워크가 생깁니다.
- 포함 관계 네트워크는 나무로 표현될 수 있습니다: 군집 나무(cluster tree)

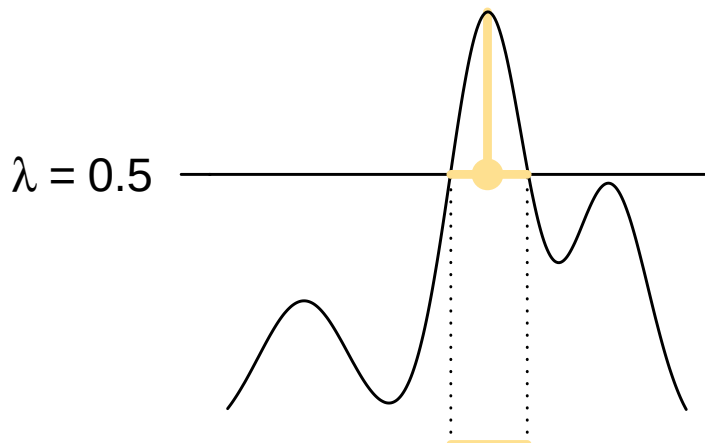




군집 나무(Cluster Tree)는 고밀도 군집들의 계층 구조입니다.

Definition

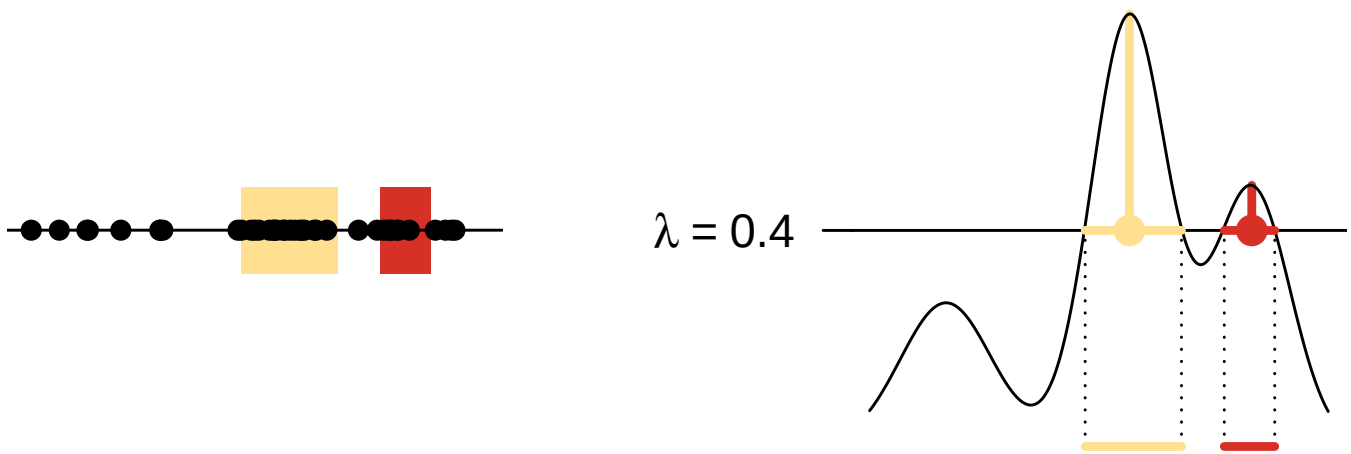
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



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

Definition

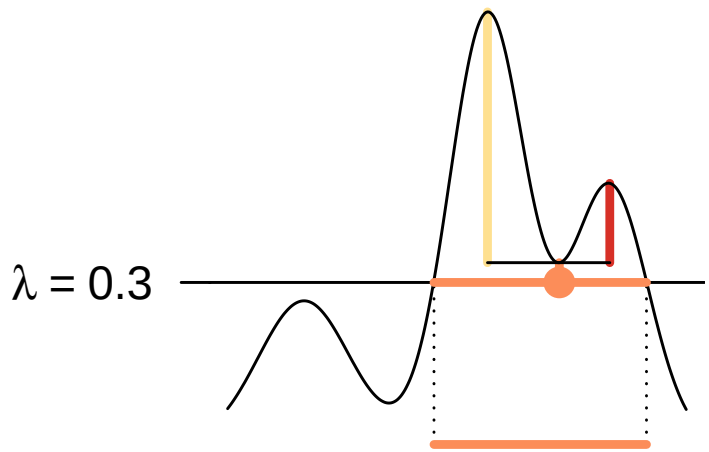
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



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

Definition

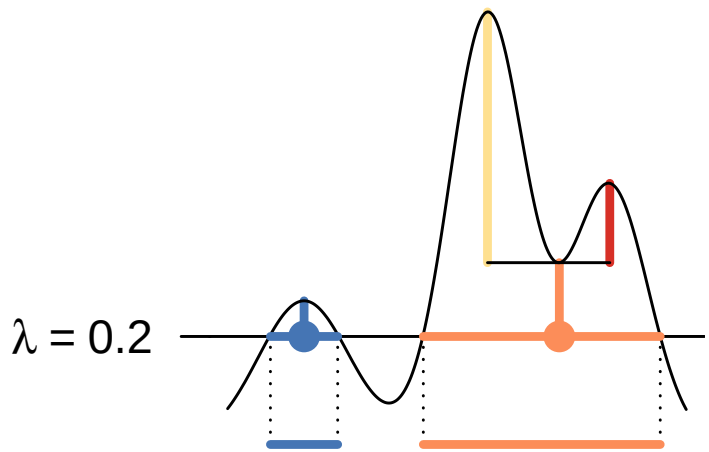
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



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

Definition

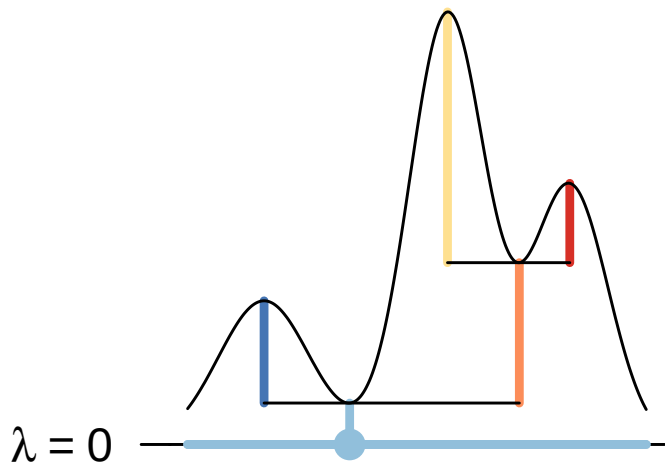
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



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

Definition

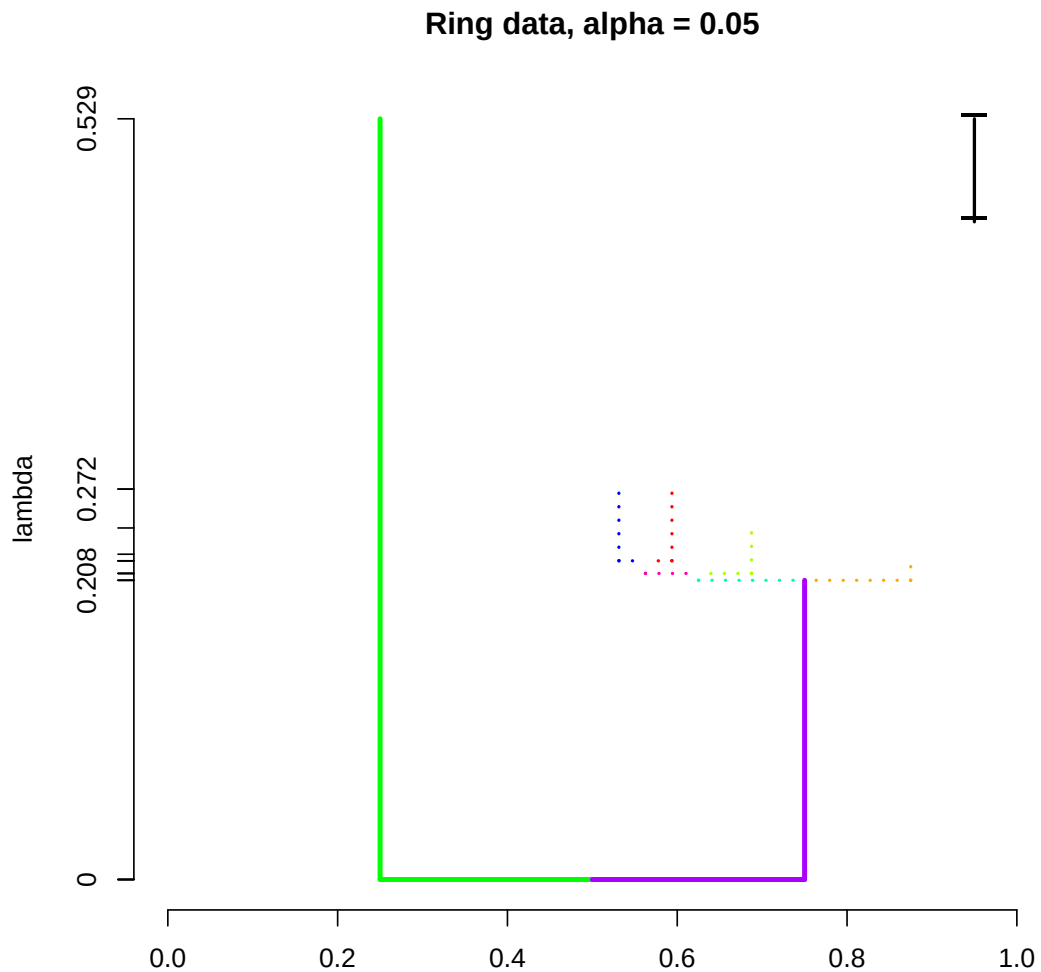
주어진 밀도함수 p의 군집 나무(Cluster Tree) $T_p : \mathbb{R} \to \mathcal{P}(\mathcal{X})$는 각 실수값 λ마다 윗레벨 집합 $\{x : p(x) \geq \lambda\}$의 연결부분들의 집합이 $T_p(\lambda)$로써 대응되는 함수입니다.



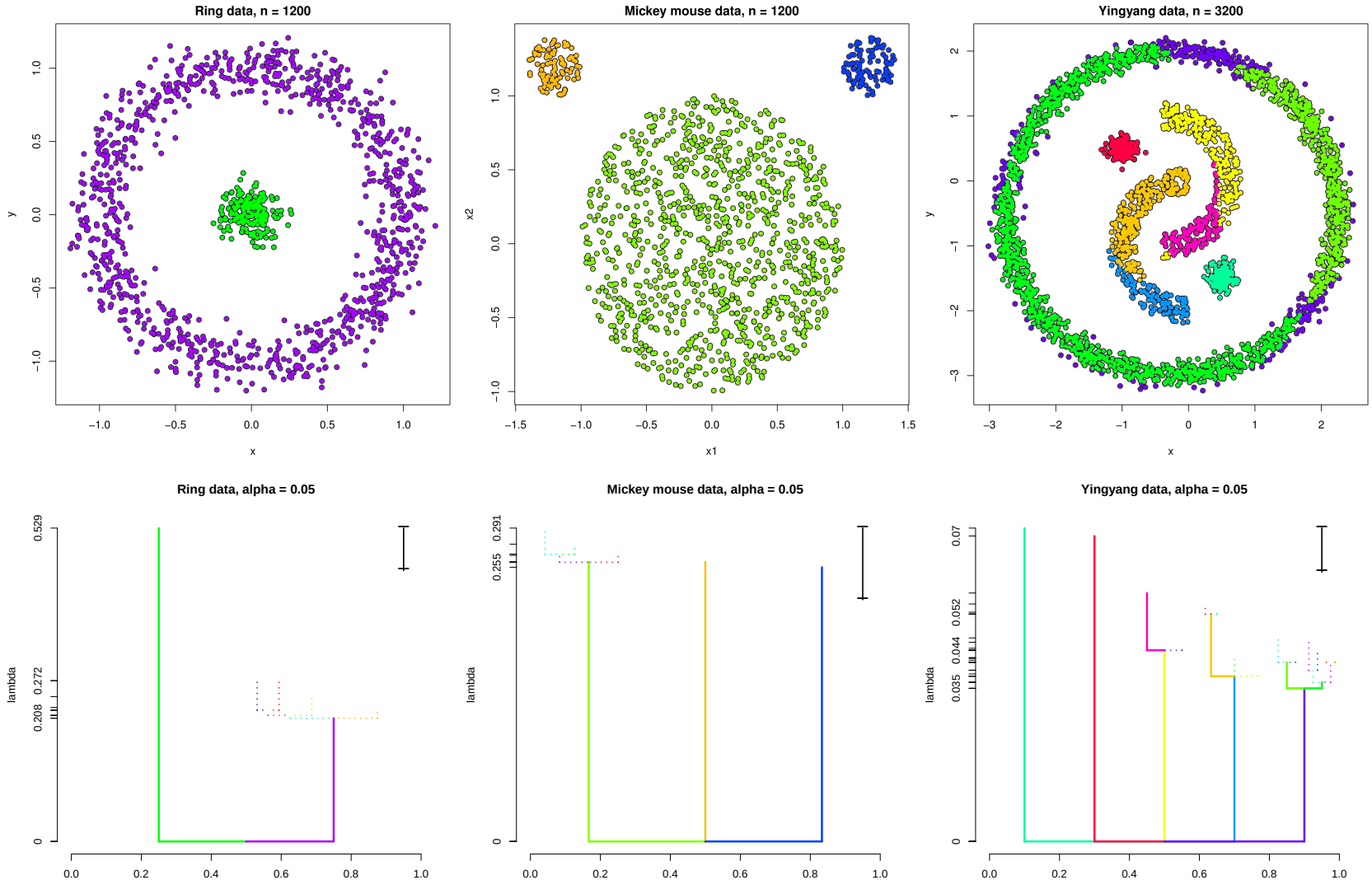
신뢰집합은 경험적 군집 나무에서 잡음을 줄이는 데에 도움을 줍니다.

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

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

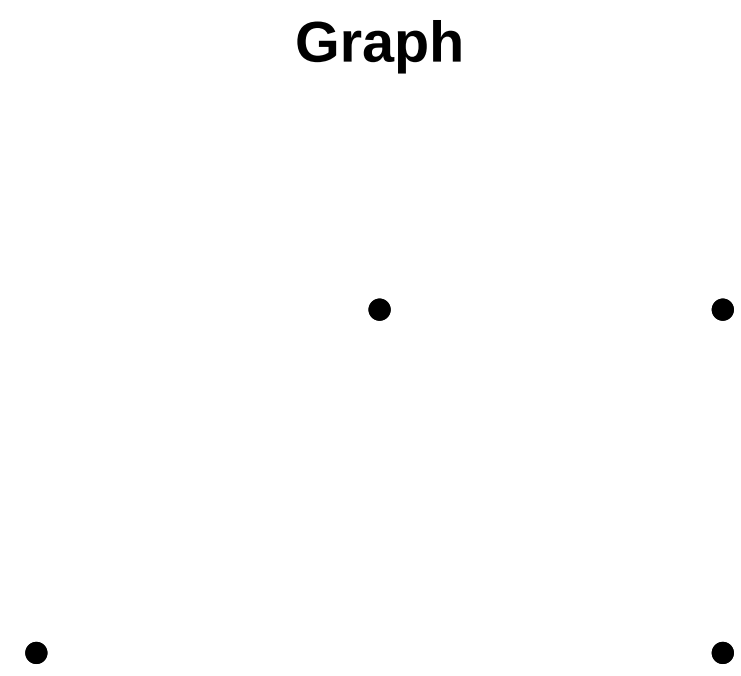


신뢰집합을 이용하여 가지치기한 군집나무로 실제 군집나무를 찾을 수 있습니다.



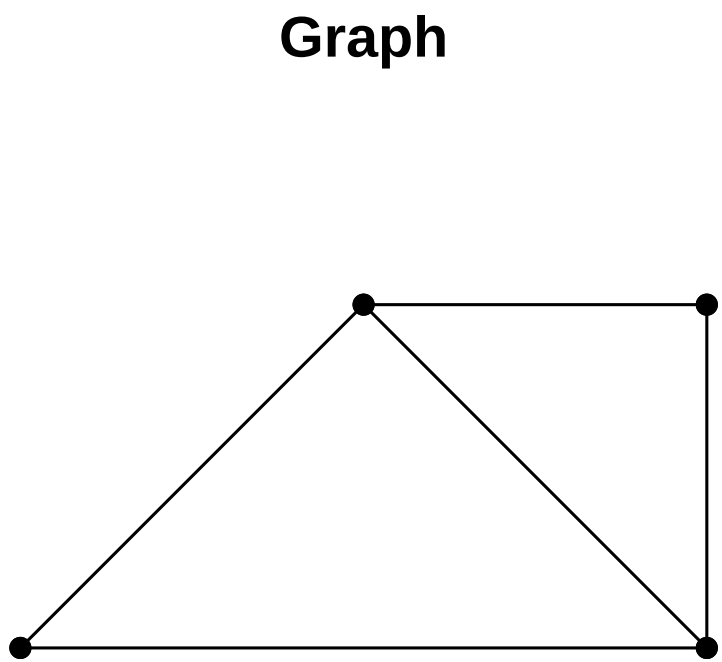
그래프(graph)는 꼭지점(vertex)과 변(edge)로 이루어진 이산 구조입니다.

- 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.



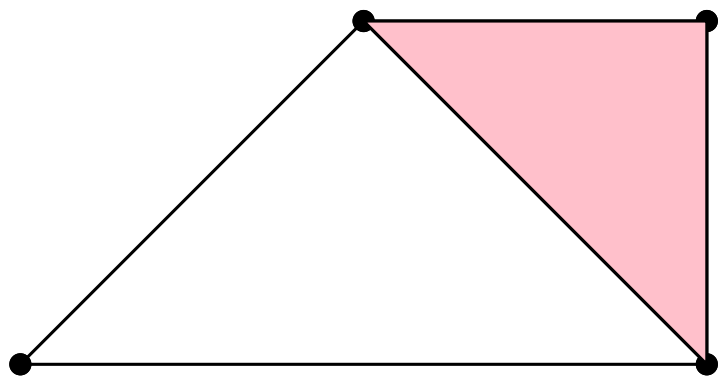
단체 복합체(Simplicial complex)는 고차원으로 일반화한 그래프입니다.

- 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.

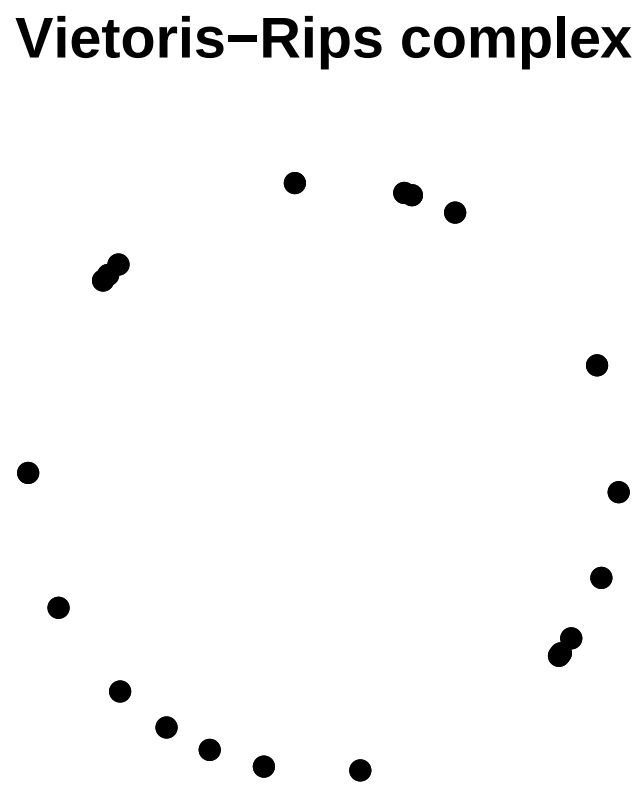
Simplicial complex



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:

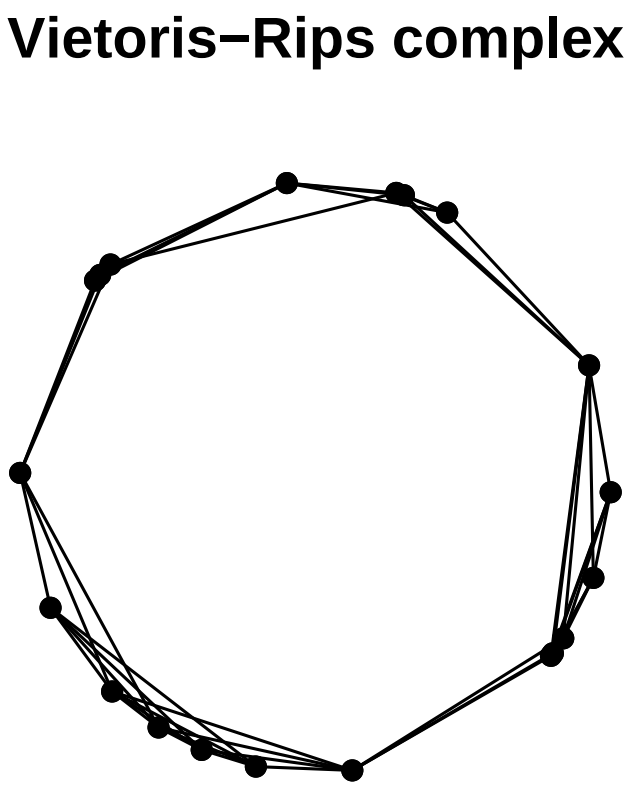
$$\mathrm{Rips}(\mathcal{X}, r) = \{\{x_1, \ldots, x_k\} \subset \mathcal{X} : \, d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k\}.$$



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:

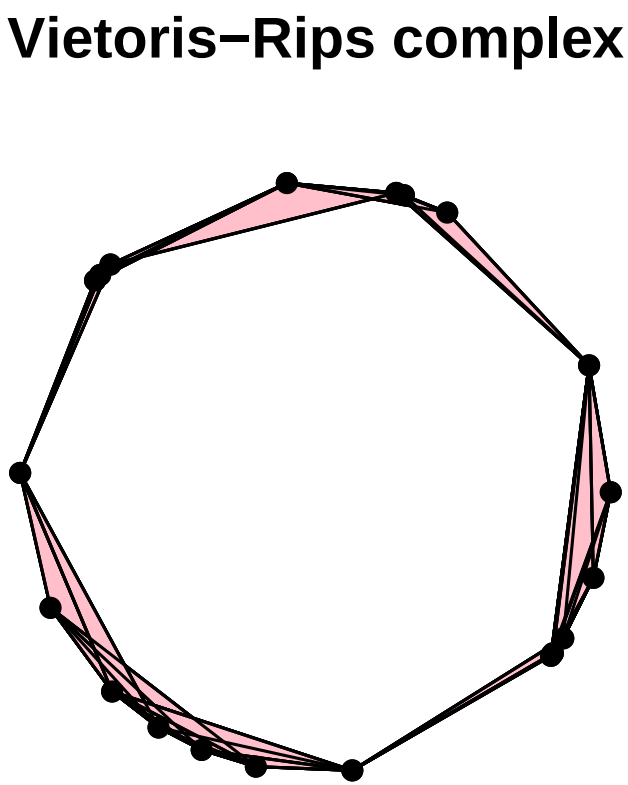
$$\mathrm{Rips}(\mathcal{X}, r) = \{\{x_1, \ldots, x_k\} \subset \mathcal{X} : \, d(x_i, x_j) < 2r, \text{ for all } 1 \leq i, j \leq k\}.$$



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:

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



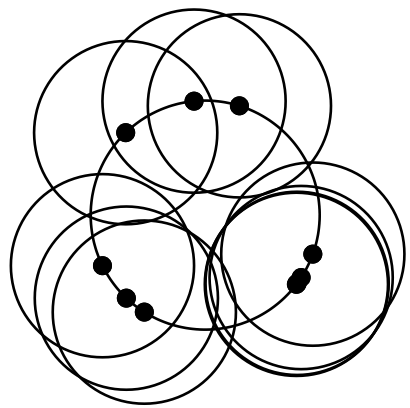
Čech 복합체(Čech complex)는 공들의 비지 않은 교집합으로 만들어진 복합체입니다.

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

$$\check{C}ech_{\mathbb{X}}(\mathcal{X}, r) = \left\{ \{x_1, \ldots, x_k\} \subset \mathcal{X} : \bigcap_{j=1}^{k} \mathbb{B}_{\mathbb{X}}(x_j, r) \neq \emptyset \right\},$$

여기서 $\mathbb{B}_{\mathbb{X}}(x, r) = \{y \in \mathbb{X} : \|y - x\| < r\}$ 는 반지름이 r이고 중심이 x인 공입니다.

Cech complex



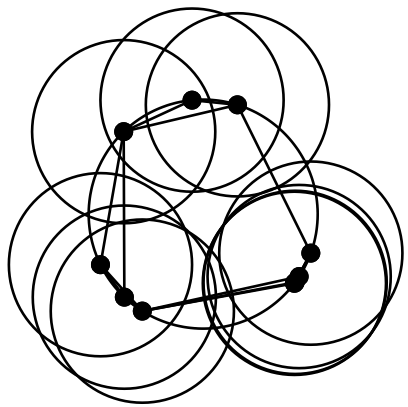
Čech 복합체(Čech complex)는 공들의 비지 않은 교집합으로 만들어진 복합체입니다.

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

$$\check{C}ech_{\mathbb{X}}(\mathcal{X}, r) = \left\{ \{x_1, \ldots, x_k\} \subset \mathcal{X} : \bigcap_{j=1}^{k} \mathbb{B}_{\mathbb{X}}(x_j, r) \neq \emptyset \right\},$$

여기서 $\mathbb{B}_{\mathbb{X}}(x, r) = \{y \in \mathbb{X} : \|y - x\| < r\}$ 는 반지름이 r이고 중심이 x인 공입니다.

Cech complex



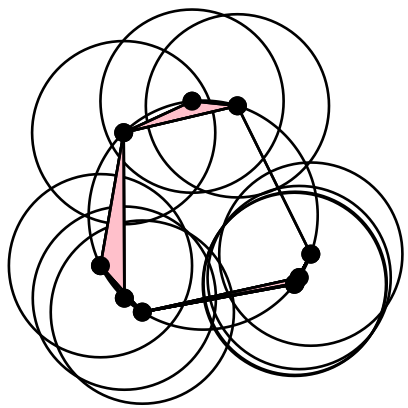
Čech 복합체(Čech complex)는 공들의 비지 않은 교집합으로 만들어진 복합체입니다.

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

$$\check{C}ech_{\mathbb{X}}(\mathcal{X}, r) = \left\{ \{x_1, \ldots, x_k\} \subset \mathcal{X} : \bigcap_{j=1}^{k} \mathbb{B}_{\mathbb{X}}(x_j, r) \neq \emptyset \right\},$$

여기서 $\mathbb{B}_{\mathbb{X}}(x, r) = \{y \in \mathbb{X} : \|y - x\| < r\}$ 는 반지름이 r이고 중심이 x인 공입니다.

Cech complex

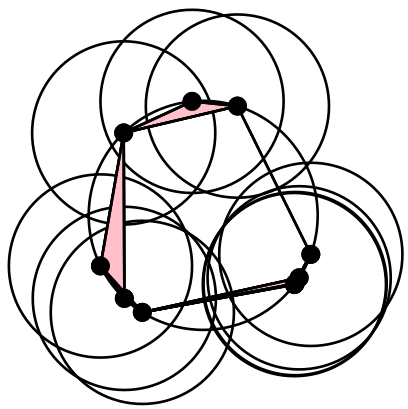


Čech 복합체(Čech complex)는 덮개 복합체(Nerve compelx)의 일종입니다.

- 열린 집합들의 모임 $\mathcal{U} = \{U_i\}_{i \in I}$가 주어졌을 때, $\mathbb{B}_{\mathbb{X}}(x, r)$을 더 일반적으로 열린집합 U_i들로 바꾸면 덮개 복합체(Nerve complex)가 됩니다.

$$Nerve(\mathcal{U}) = \left\{ \{U_1, \ldots, U_k\} : \bigcap_{j=1}^{k} U_j \neq \emptyset \right\}.$$

Cech complex



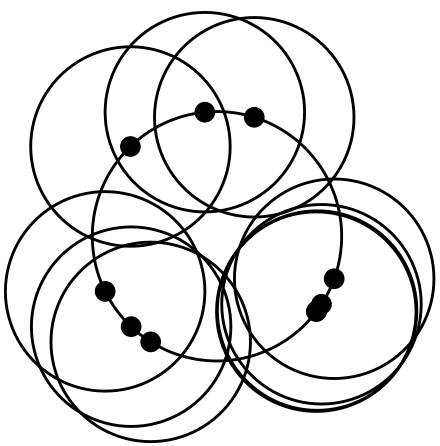
공들의 비지 않은 교집합이 항상 축약가능할(contractible) 때, 공들의 합집합과 Čech 복합체는 호모토피 동치(homotopy equivalent)입니다.

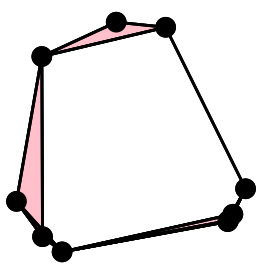
Theorem (Nerve Theorem)

만약 공들의 집합 $\{\mathbb{B}_{\mathbb{X}}(x, r_x) : x \in \mathcal{X}\}$의 임의의 비지 않은 교집합이 축약가능할(contractible) 때, 공들의 합집합 $\bigcup_{x\in\mathcal{X}} \mathbb{B}_{\mathbb{X}}(x, r)$과 Čech 복합체 $\check{C}ech_{\mathbb{X}}(\mathcal{X}, r)$는 호모토피 동치(homotopy equivalent)입니다.

- $\mathbb{X} = \mathbb{R}^d$ 이면 축약가능(contractible) 조건을 항상 만족합니다.

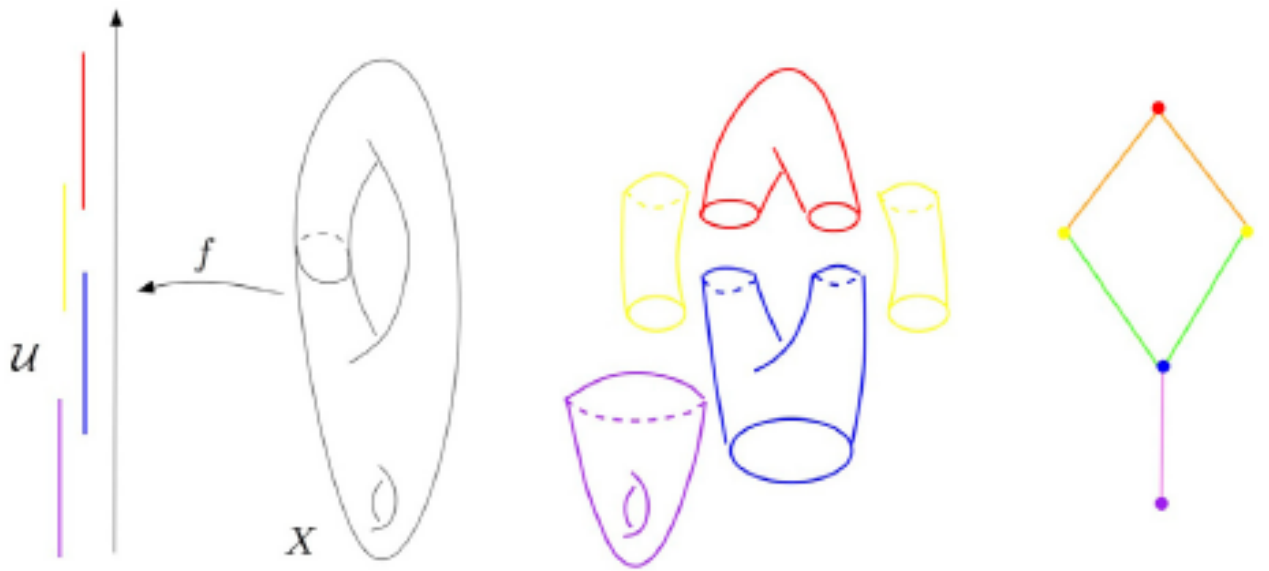
Union of balls $\simeq$ **Cech complex**





Mapper는 덮개들의 연결 성분으로 자료를 요약하고 시각화합니다.

- 연속함수 $f : X \to \mathbb{R}^d$, $d \geq 1$와 $\mathbb{R}^d$의 덮개(cover) $\mathcal{U} = \{U_i\}_{i \in I}$가 주어졌을 때, 당김 덮개(pull-back cover)는 열린 집합들의 모임 $\{f^{-1}(U_i)\}_{i \in I}$입니다. 세분화된 당김 덮개(refined pull-back cover)는 각 열린집합 $f^{-1}(U_i)$의 연결성분(connected component)들의 모임입니다.
- Mapper 알고리즘은 세분화된 당김 덮개의 덮개 복합체(nerve compelx)를 자료로부터 계산합니다.



5

[5] Chazal, Michel, 2021

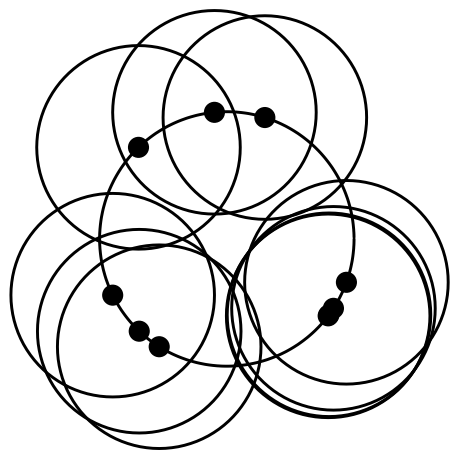
목표 공간을 각 자료를 중심으로 한 공들의 합집합으로 근사합니다.

- $r > 0$이 주어졌을 때, 목표공간 $\mathbb{X}$를 각 자료를 중심으로 한 공들의 합집합으로 근사합니다:

$$\bigcup_{x \in \mathcal{X}} \mathbb{B}_{\mathbb{X}}(x, r),$$

여기서 $\mathbb{B}_{\mathbb{X}}(x, r) = \{y \in \mathbb{X} : \|y - x\| < r\}$ 는 반지름이 r이고 중심이 x인 공입니다.

Union of balls

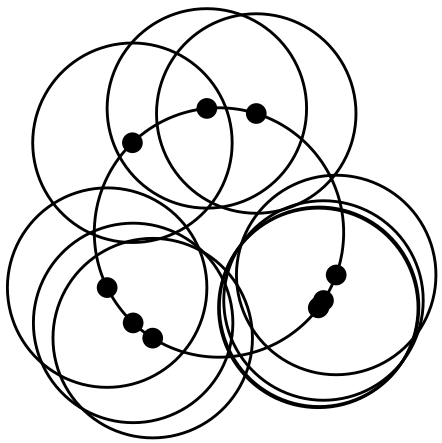


Reach가 양수인 집합의 호모토피(homotopy)는 Čech 복합체로부터 재구성할 수 있습니다.

Corollary (Kim et al. [2020, Corollary 10])

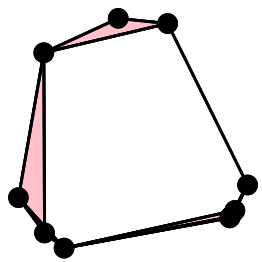
$\mathbb{X} \subset \mathbb{R}^d$ 가 reach $\tau > 0$인 집합이고, $\mathcal{X} \subset \mathbb{X}$가 유한개의 점이 주어졌다. 이 때, $r \leq \tau$를 만족하면, $\bigcup_{x \in \mathcal{X}} \mathbb{B}_{\mathbb{X}}(x, r)$는 Čech 복합체 $\check{C}ech_{\mathbb{X}}(\mathcal{X}, r)$와 호모토피 동치(homotopy equivalent)이다.

Underlying circle



$\simeq$

Cech complex



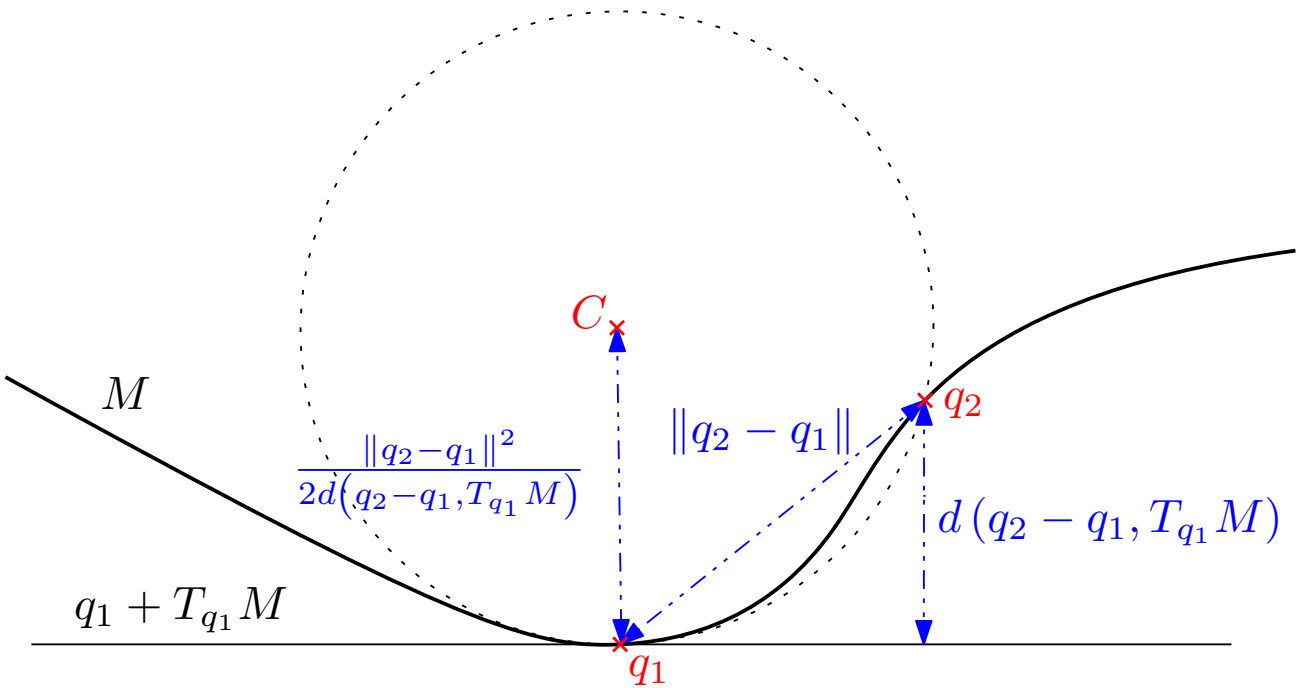
reach는 다양체(manifold) 위에서 구를 수 있는 공의 최대 반지름입니다.

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)

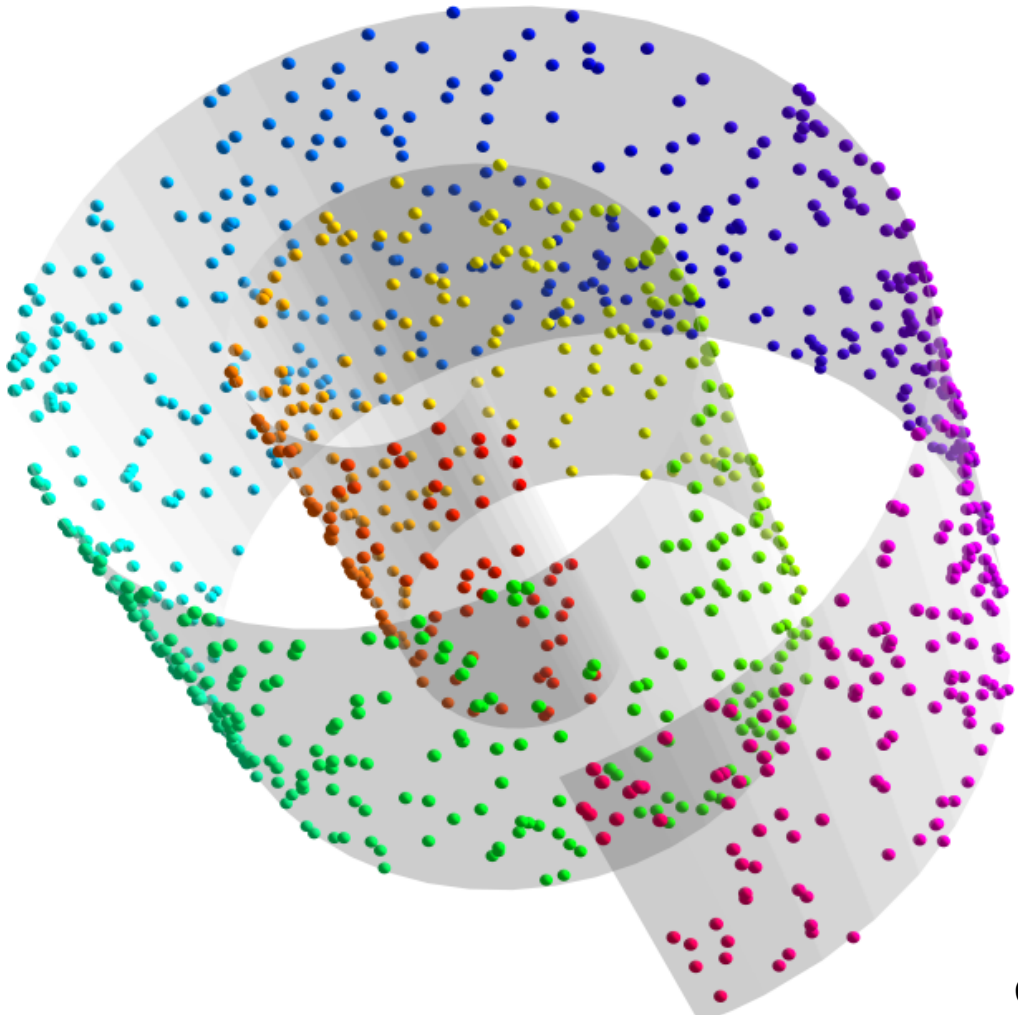
- $$R_n = \inf_{\hat{\tau}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{P^{(n)}} \left[\left| \frac{1}{\tau(P)} - \frac{1}{\hat{\tau}_n(X)} \right|^q \right]$$
  - $X = (X_1, \cdots, X_n)$는 고정된 분포 P에서 추출하고, P는 확률분포의 집합 $\mathcal{P}$에 속합니다.
  - 추정량 $\hat{\tau}_n$은 자료 X의 임의의 함수입니다.
  - 역-l_q 손실함수를 사용합니다, 따라서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

reach 추정의 미니맥스 위험(minimax risk)

Theorem

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

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

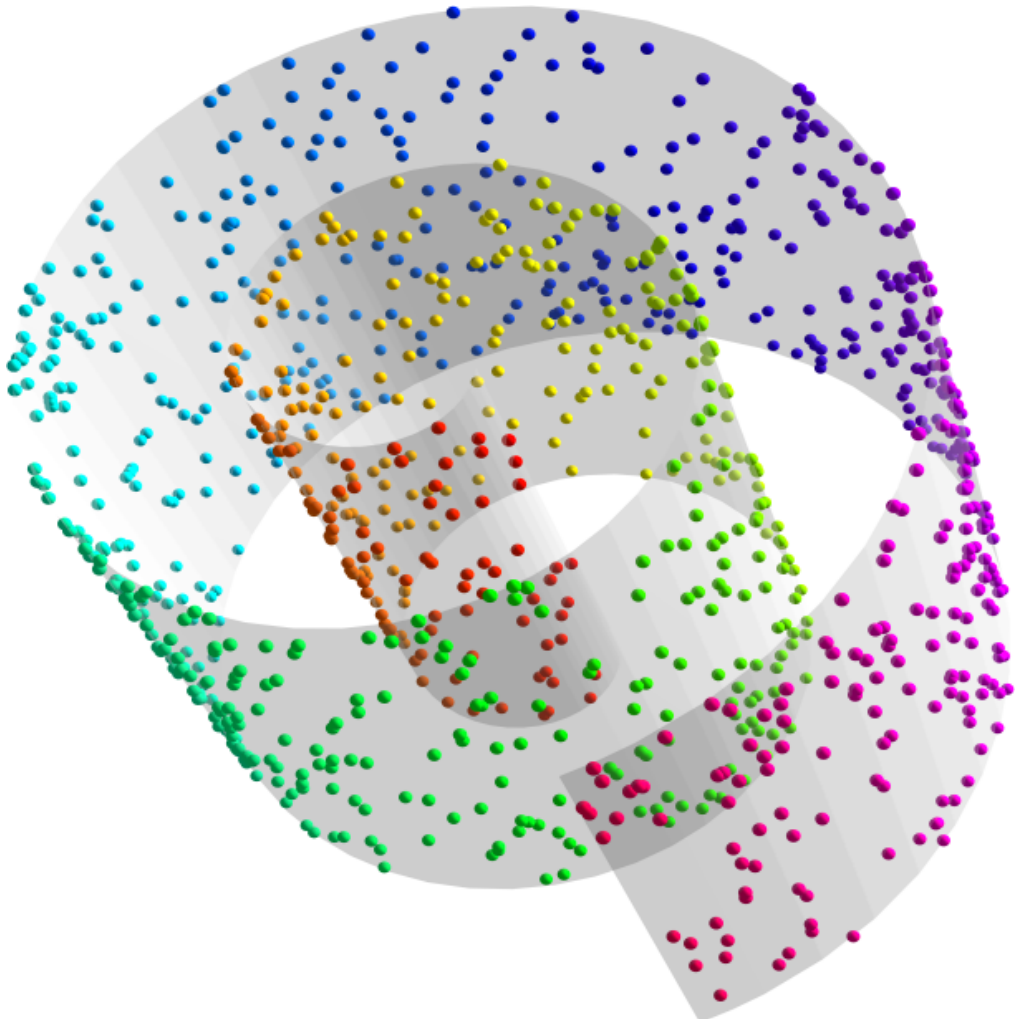


6

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

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)

Theorem

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

구멍의 개수로 기하학적 대상들을 분류할 수 있습니다.

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

예제: 대상들을 호몰로지에 따라 분류합니다.

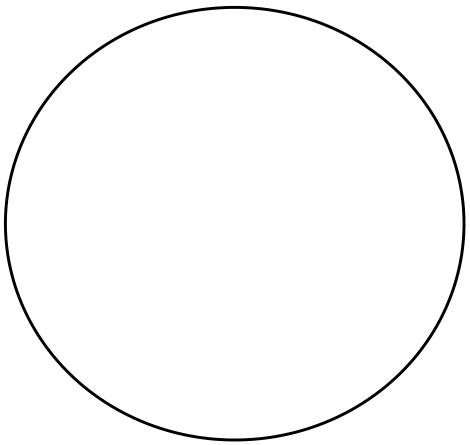
1. β_0 =연결된 성분의 개수 ●
2. β_1 =고리의 개수 ○

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

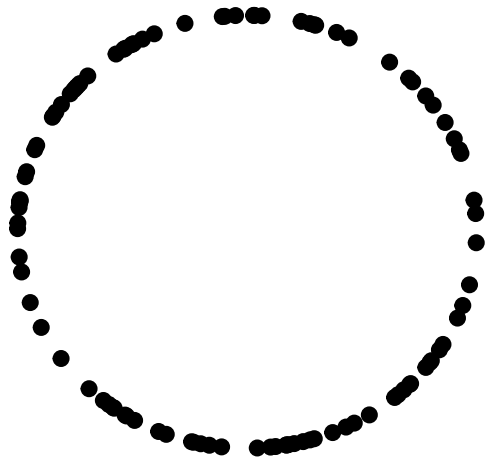
유한한 자료의 호몰로지는 기저 구조의 호몰로지와 다르기 때문에, 유한한 자료로 직접 기저 구조의 호몰로지를 추정할 수는 없습니다.

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

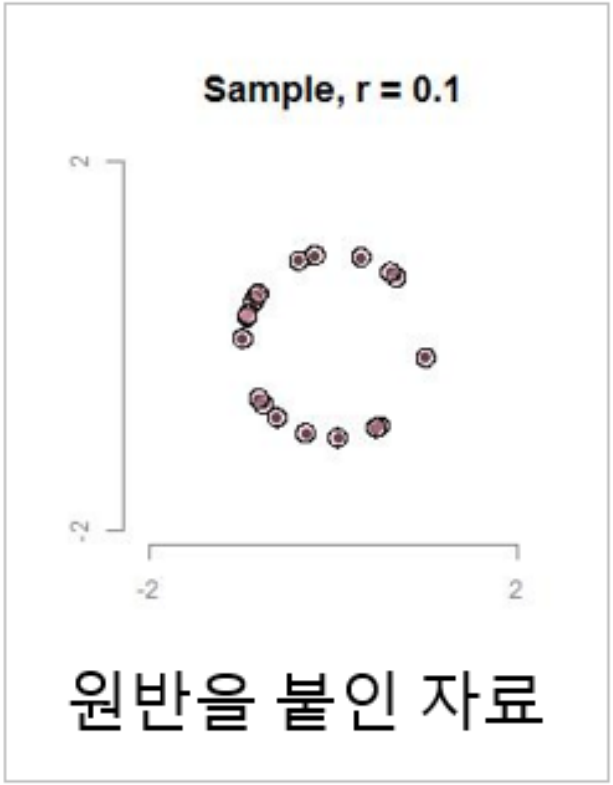
Underlying circle: $\beta_0 = 1, \beta_1 = 1$



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

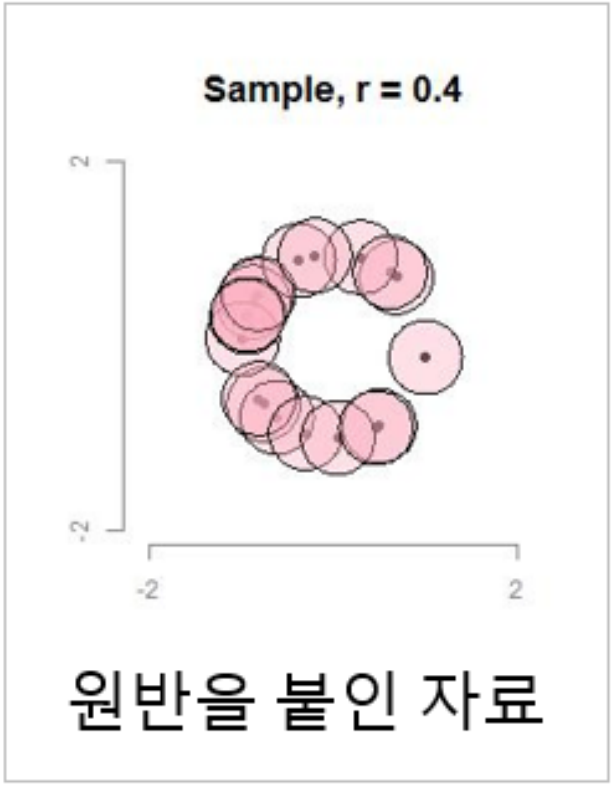


Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



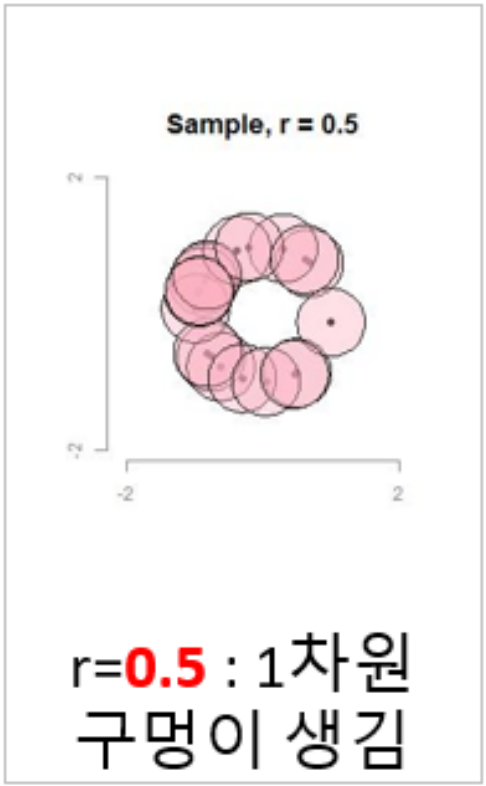
원반을 붙인 자료

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

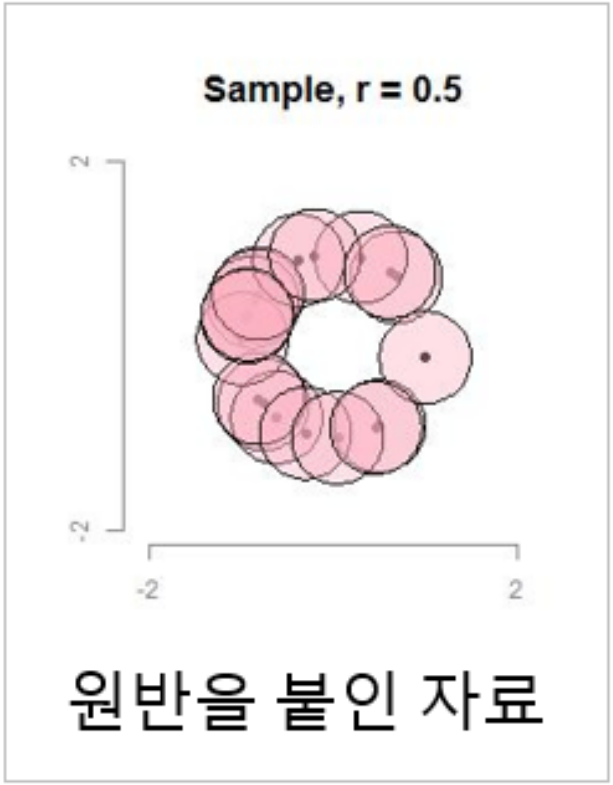


원반을 붙인 자료

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

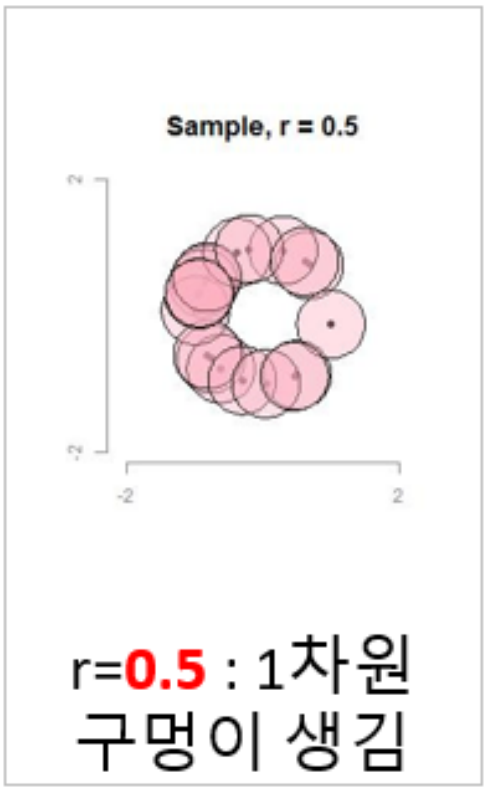


r=**0.5** : 1차원 구멍이 생김

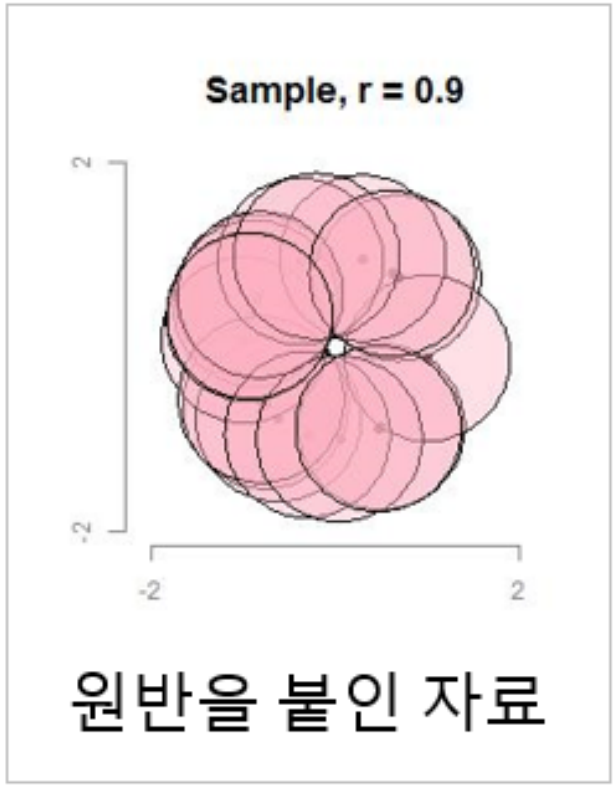


원반을 붙인 자료

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

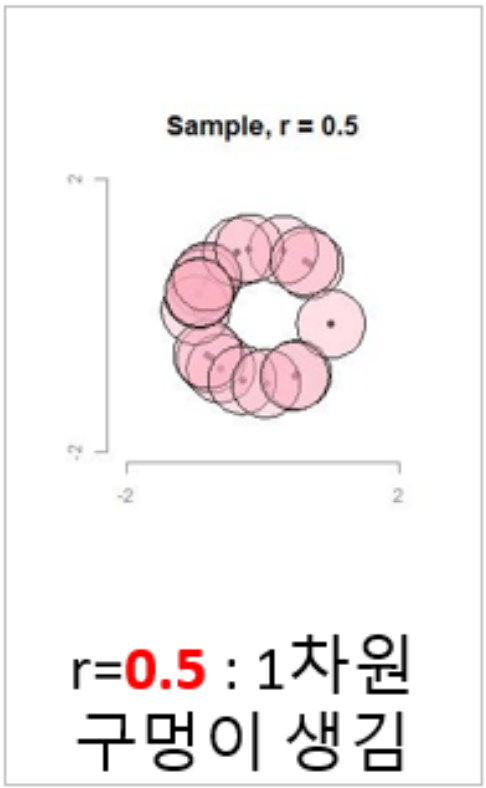


r=**0.5** : 1차원 구멍이 생김

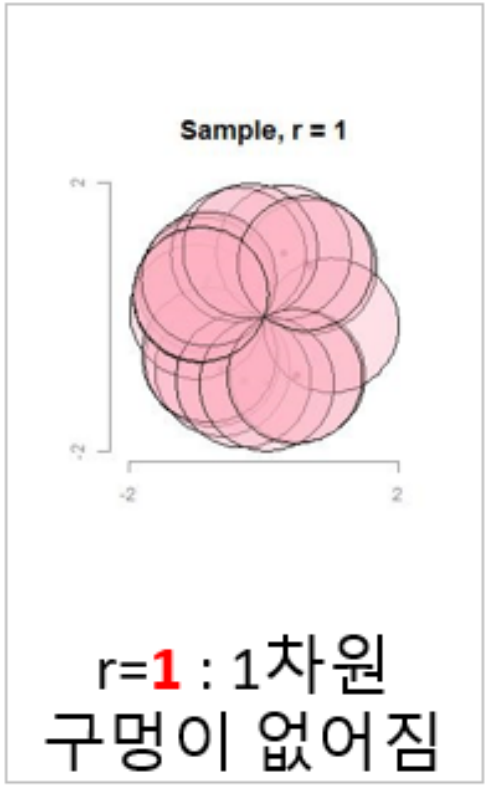


원반을 붙인 자료

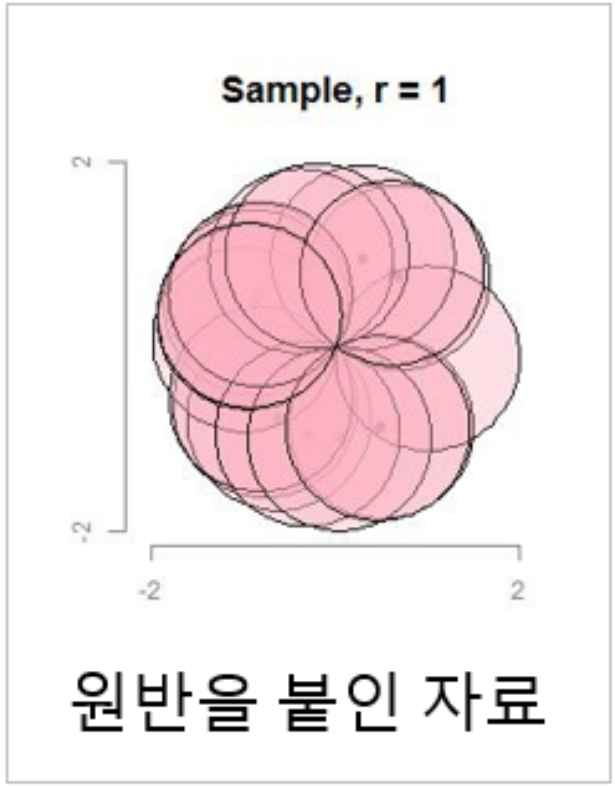
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



r=**0.5** : 1차원 구멍이 생김

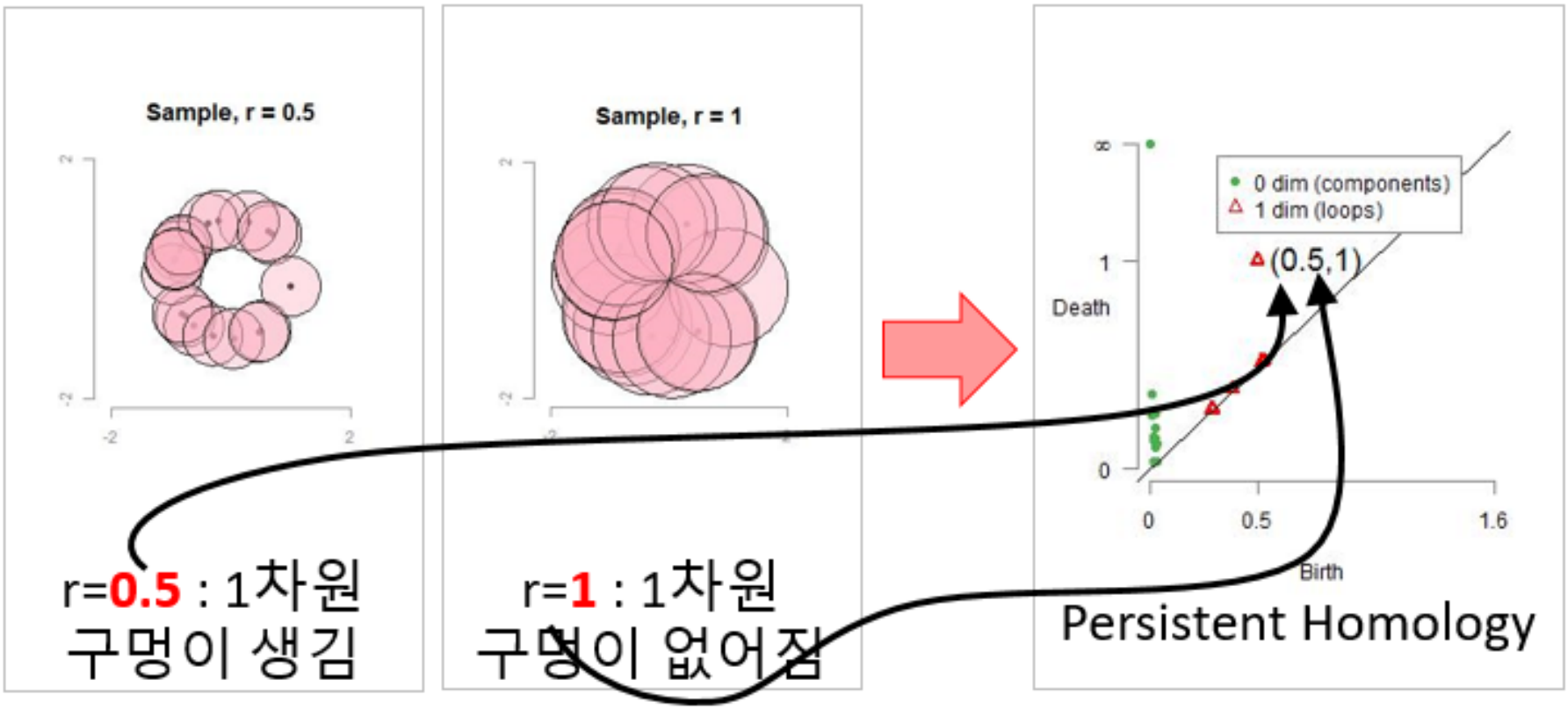


r=**1** : 1차원 구멍이 없어짐



원반을 붙인 자료

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

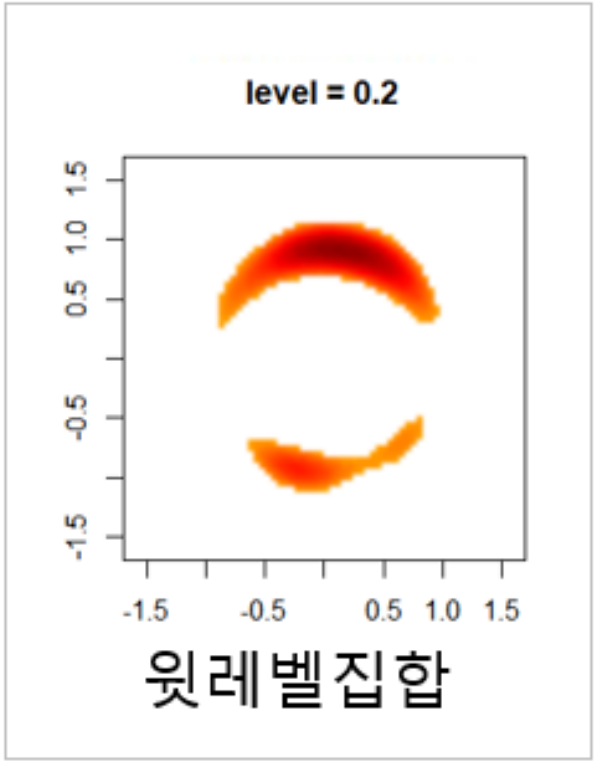


기저 구조의 위상학적 정보를 추출하는 데에 핵밀도추정(kernel density estimator)을 사용합니다.

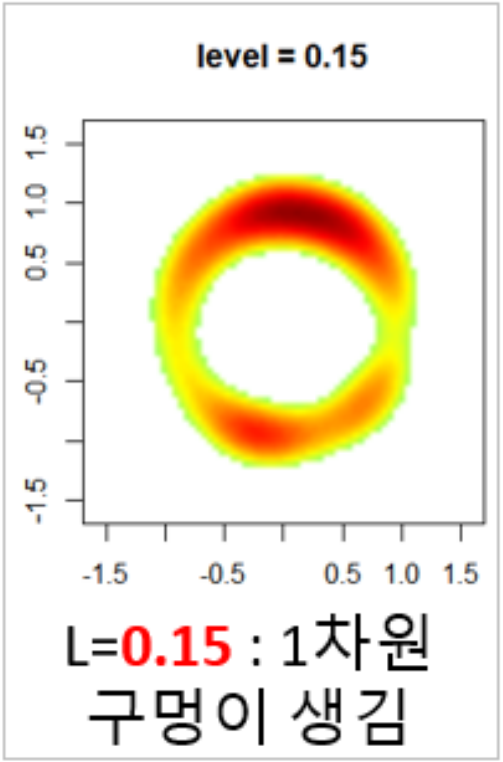
- 핵밀도추정(kernel density estimator)은 다음과 같습니다:

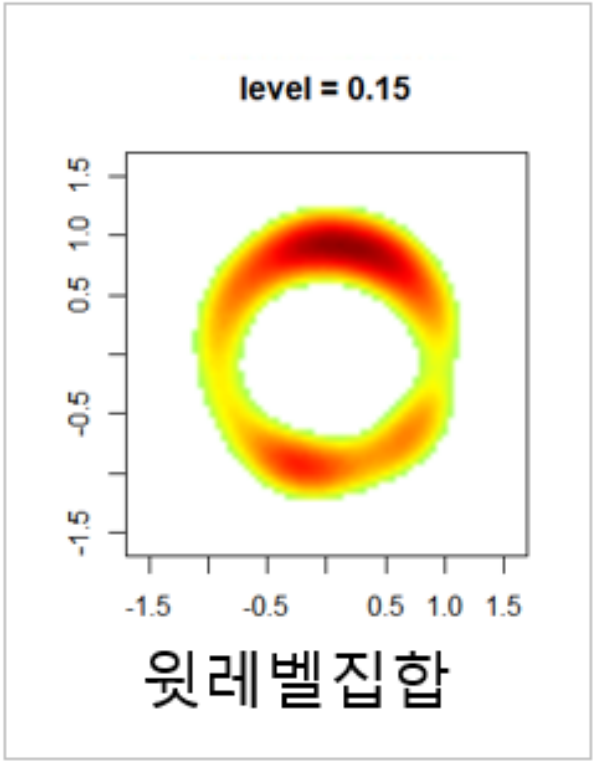
$$\hat{p}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

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

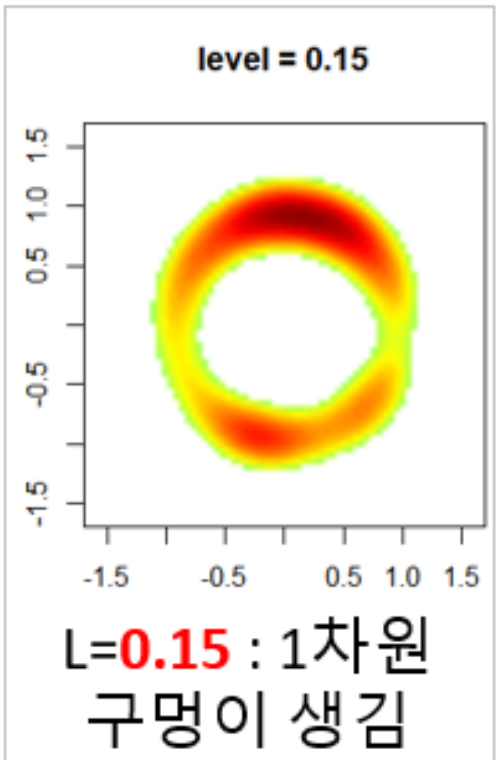


Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.

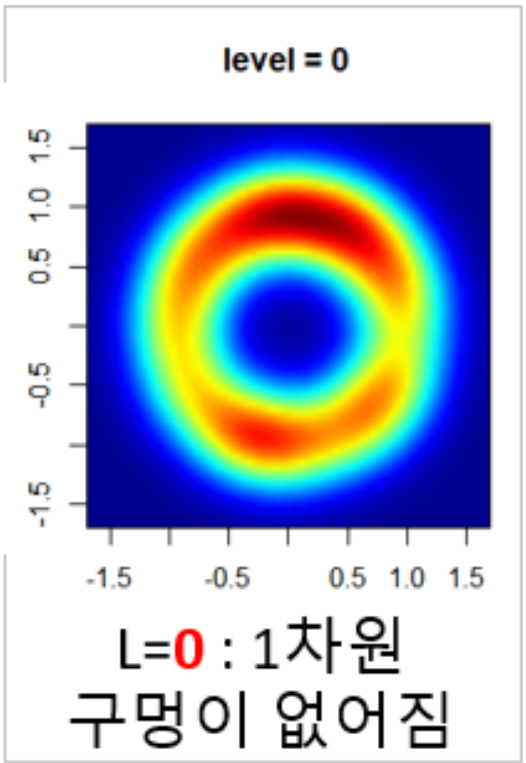




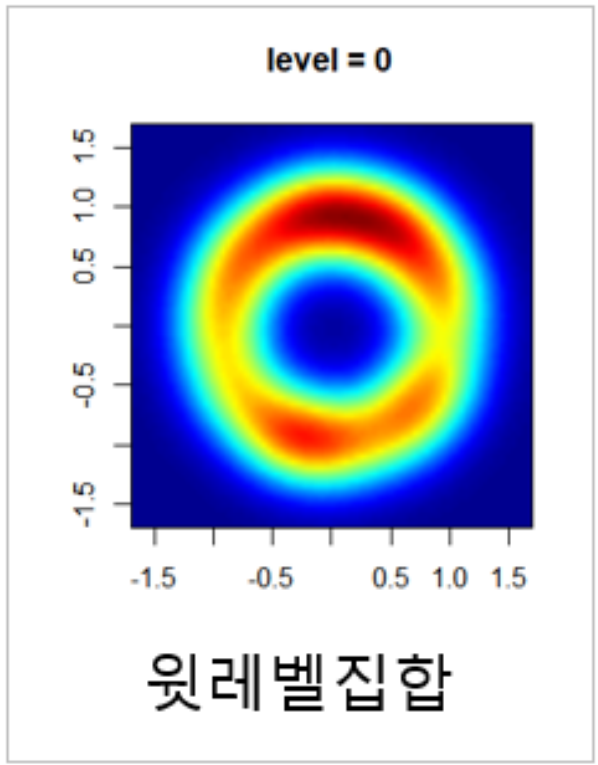
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



L=**0.15** : 1차원 구멍이 생김

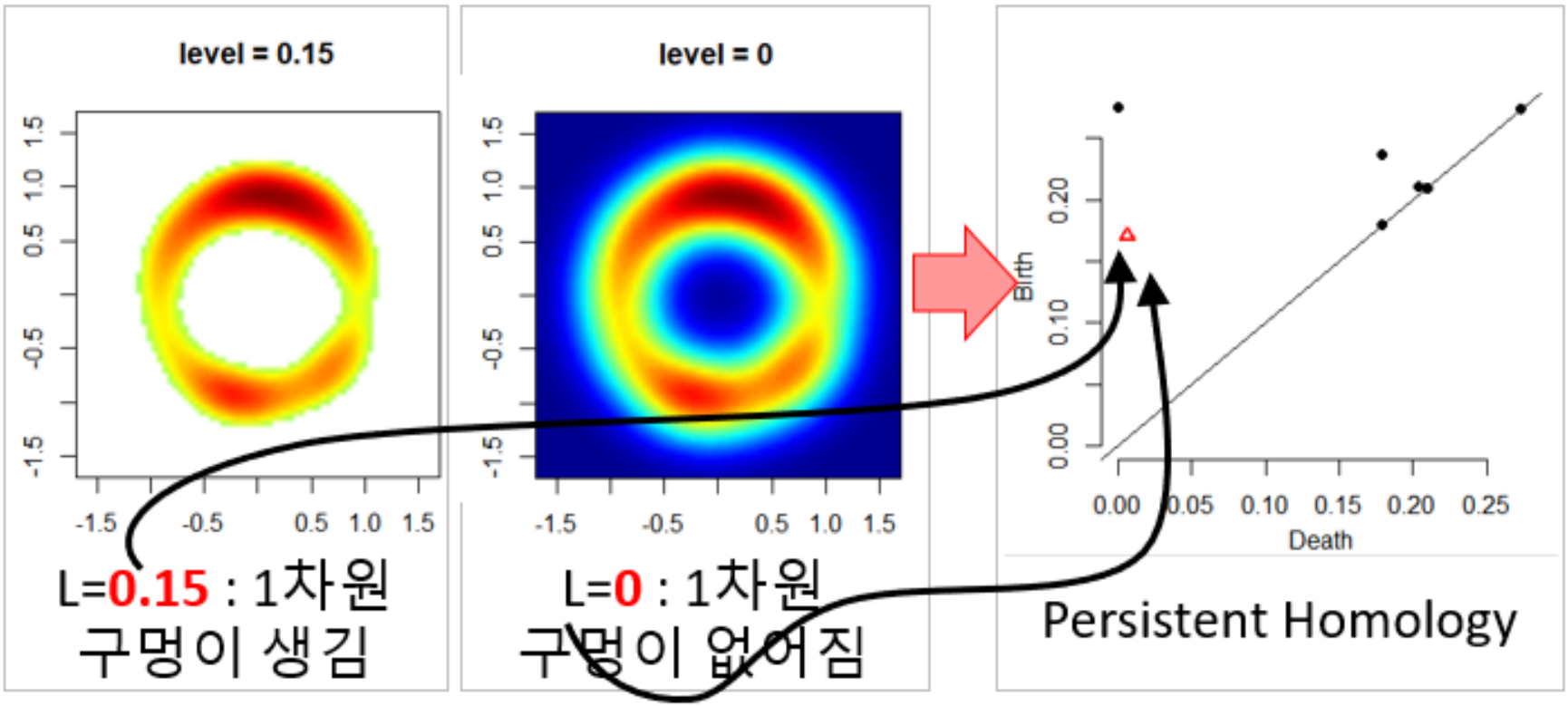


L=**0** : 1차원 구멍이 없어짐

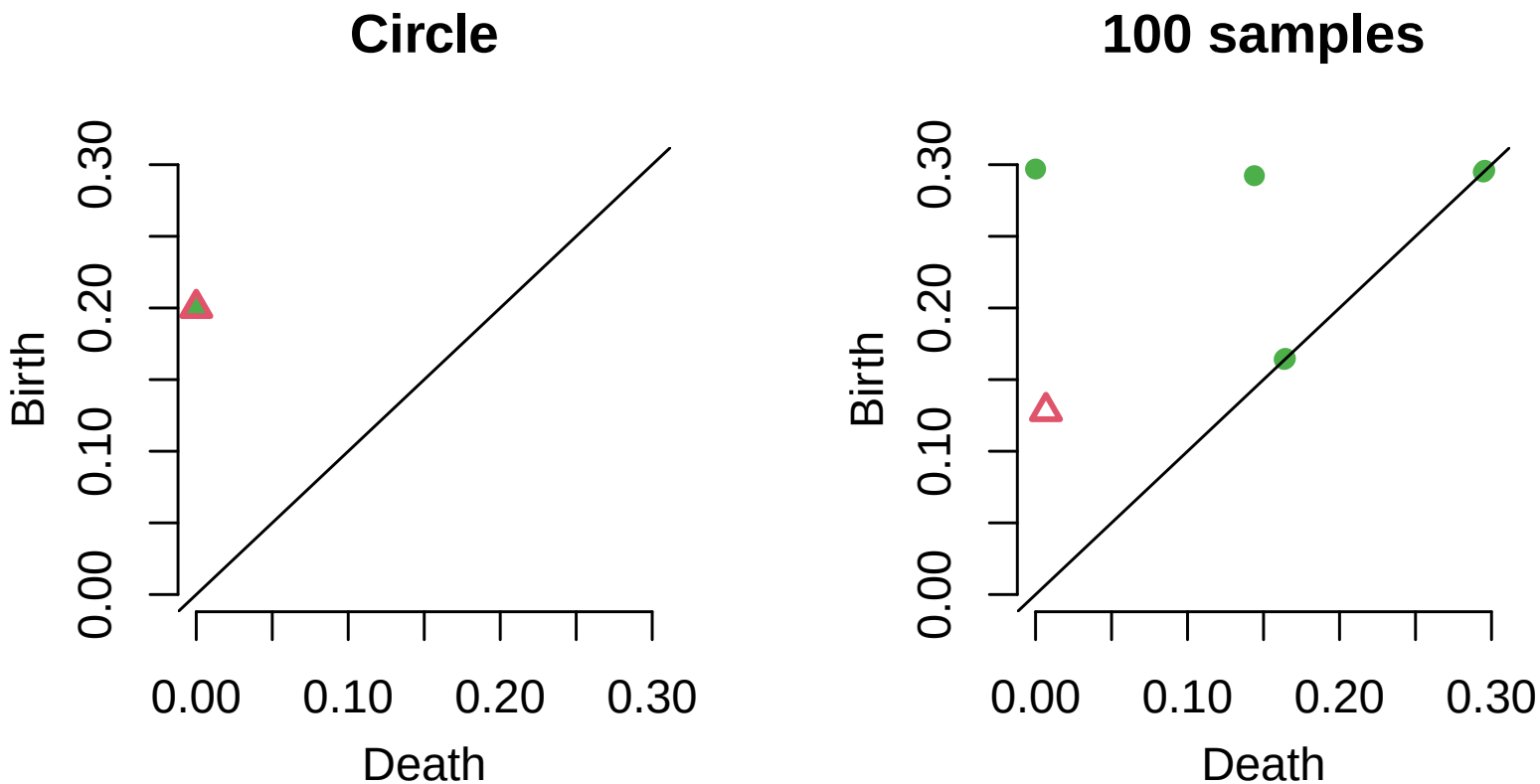


윗레벨집합

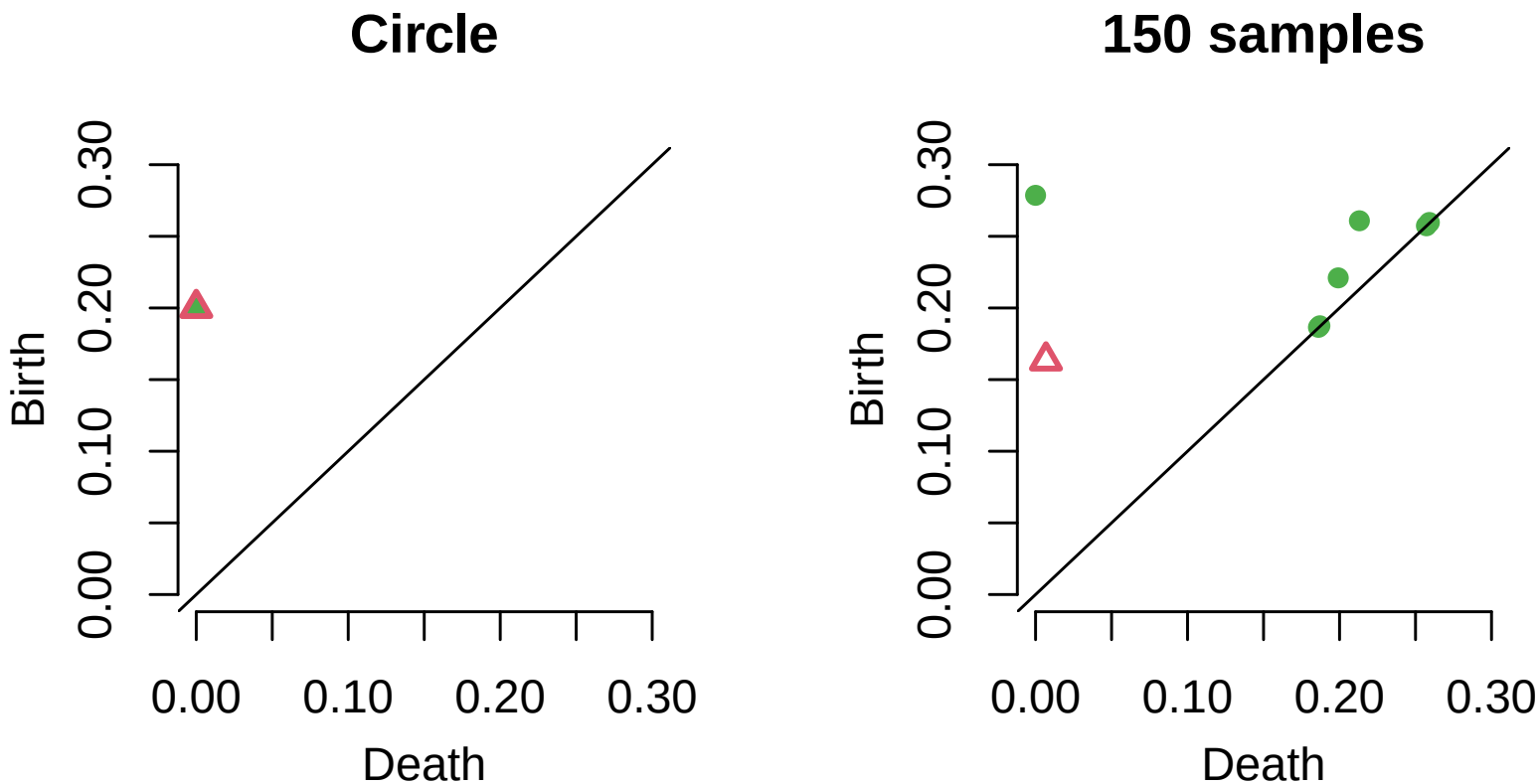
Persistent homology는 집합들의 모임에서 호몰로지를 계산하고, 호몰로지가 언제 나타나고 사라지는지 기록합니다.



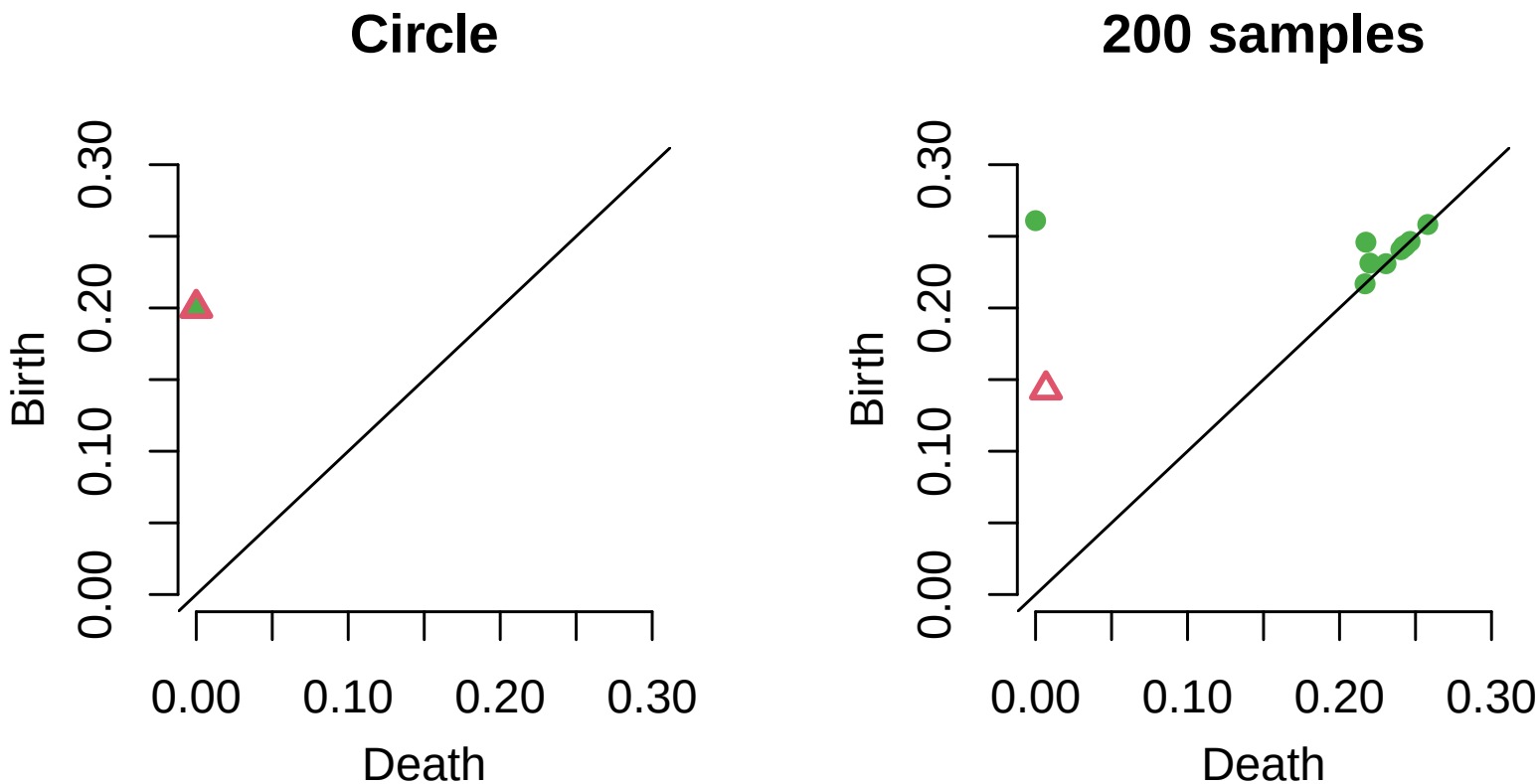
유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.



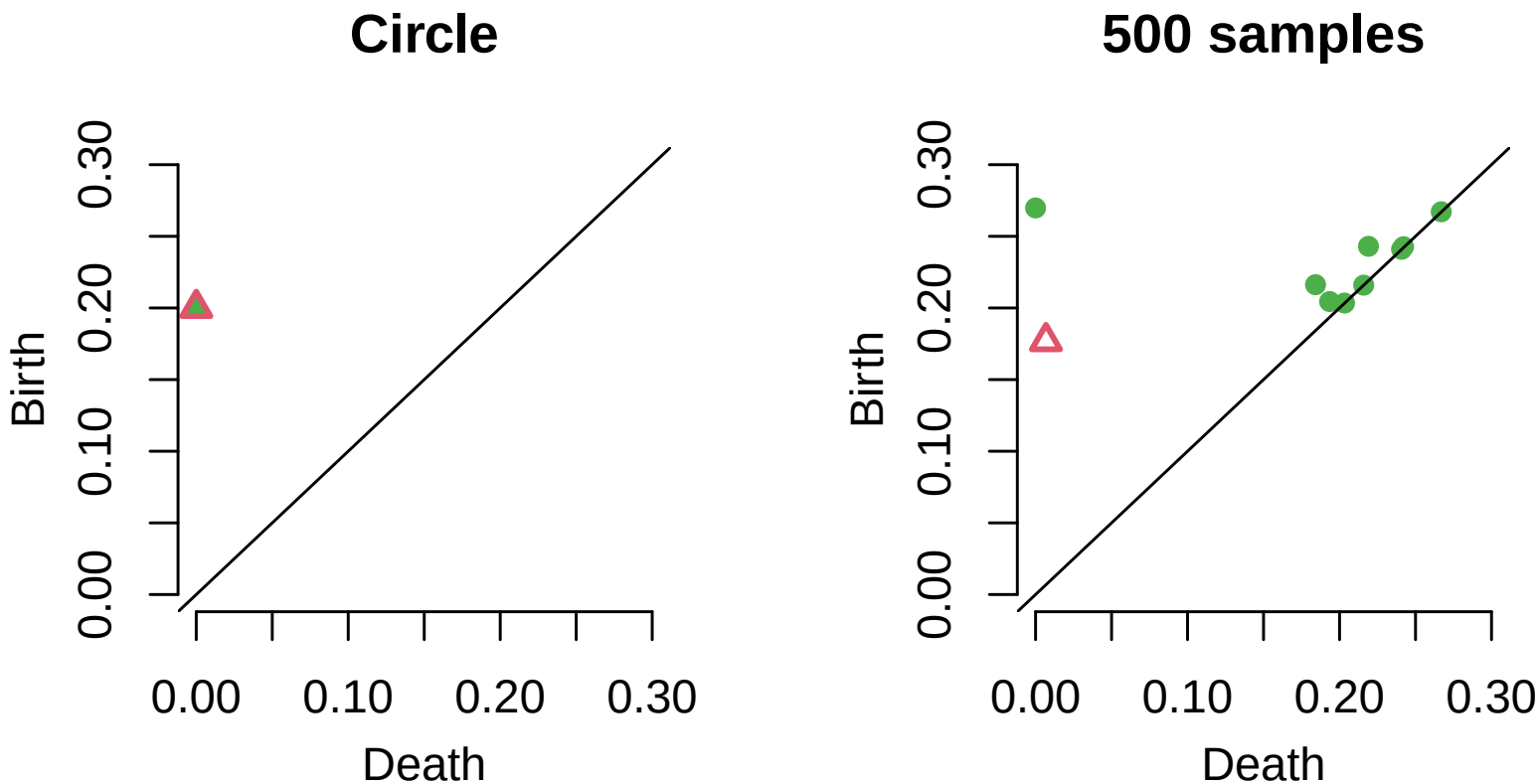
유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.



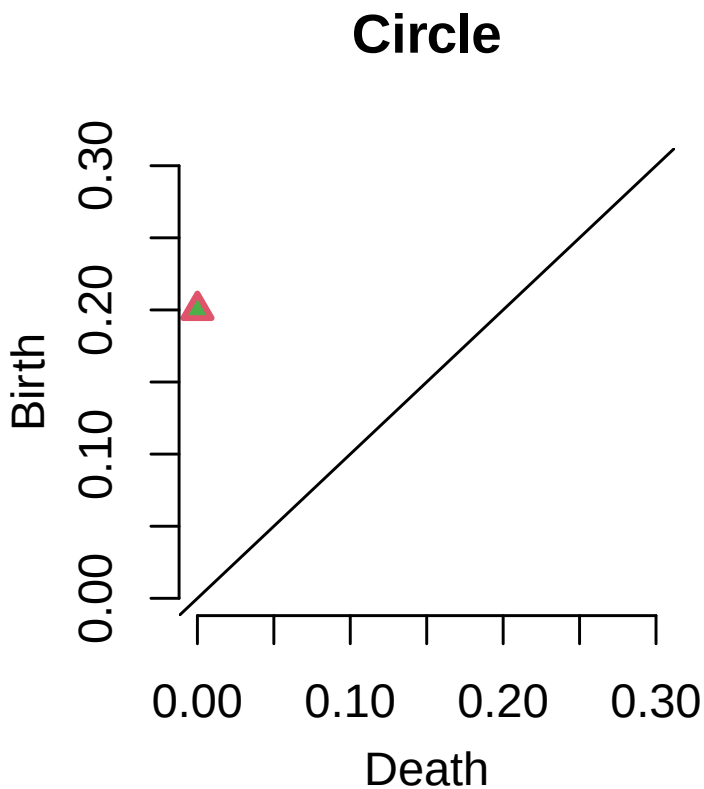
유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.

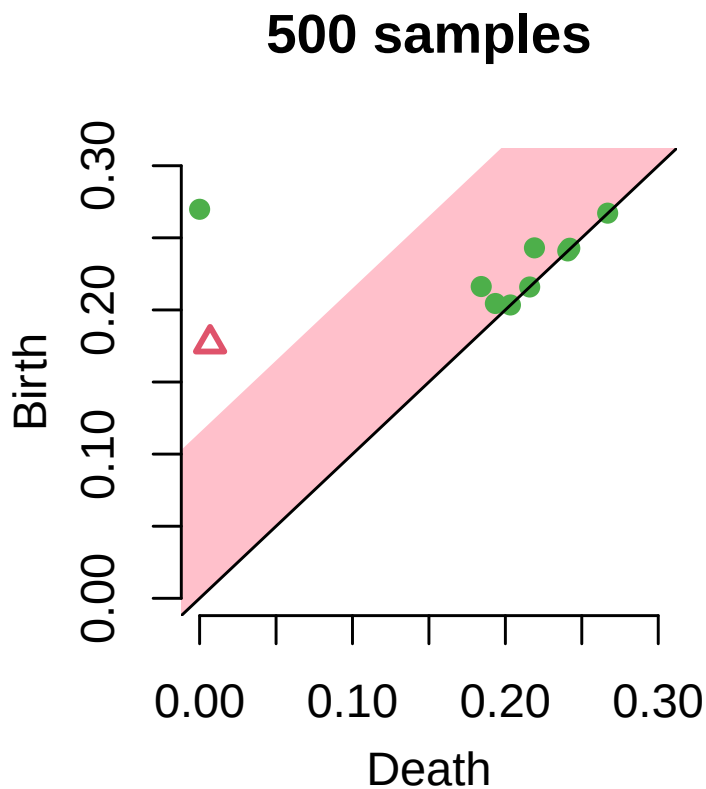


유한한 자료의 Persistent homology로부터 기저 구조의 Persistent homology를 추정할 수 있습니다.



통계적으로 유의한 호몰로지 특성과 그렇지 않은 호몰로지 특성을 어떻게 구분할까요?





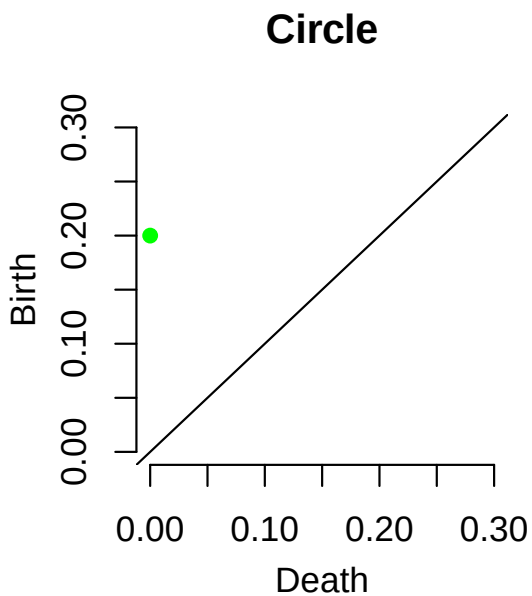
Bottleneck distance는 Persistent homology 공간에 거리 함수를 줍니다.

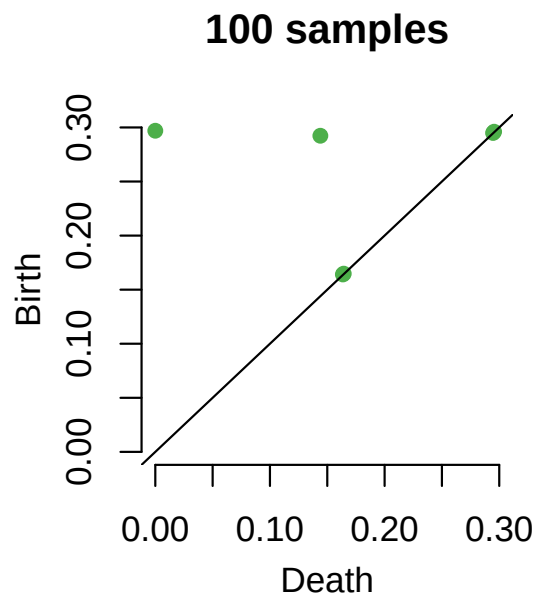
Definition

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

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

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





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

Theorem

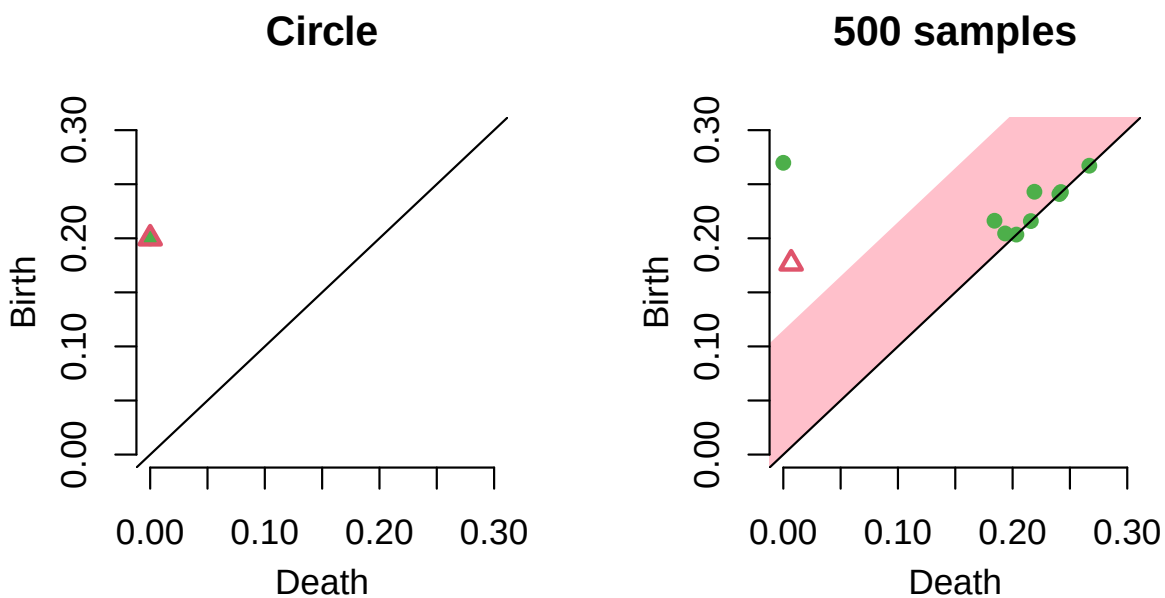
[Edelsbrunner and Harer, 2010][Chazal, de Silva, Glisse, and Oudot, 2012] K를 단체 복합체(simplicial complex)라 하고 $f, g : K \to \mathbb{R}$를 두 함수라 합니다. $Dgm(f)$와 $Dgm(g)$를 그에 상응하는 persistent homology라고 할 때, 다음이 성립합니다:

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

Persistent homology의 신뢰띠는 Persistent homology를 높을 확률로 포함하는 확률변수입니다.

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

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



Persistent homology의 신뢰띠는 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.

기계학습(machine learning) (아주) 대충 보기

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

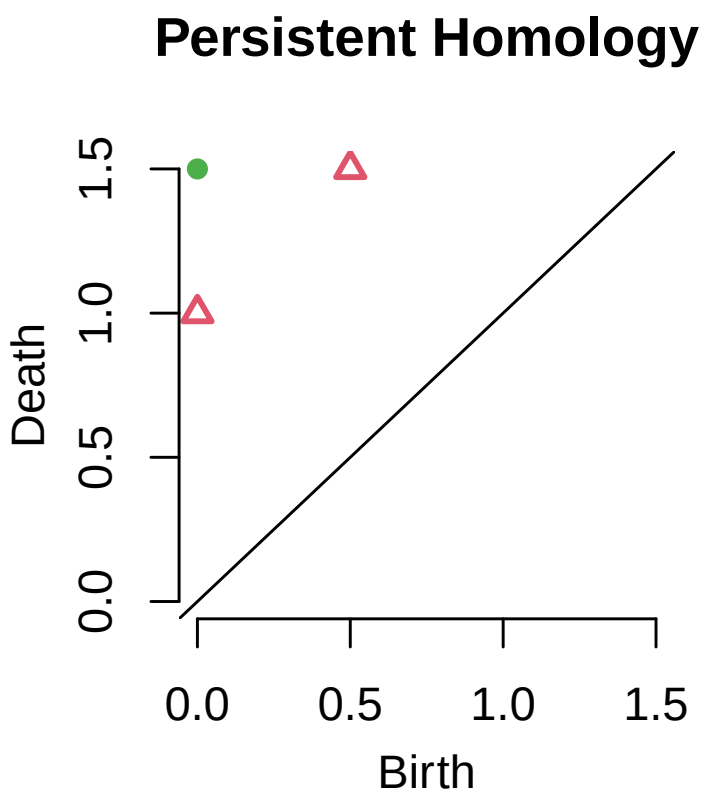
$$\theta_{n+1} = \theta_n - \lambda \nabla_\theta \mathcal{L}(f_\theta, \mathcal{X}).$$

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

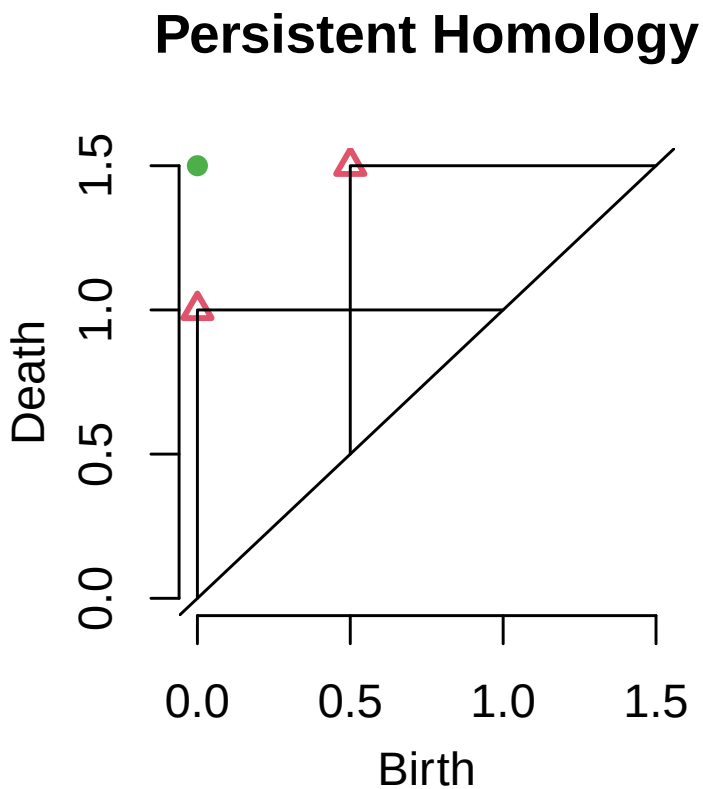
- A Survey of Topological Machine Learning Methods (Hensel, Moor, Rieck, 2021)
- 위상학적 자료 분석을 기계학습에 응용하는 데에는 크게 두 가지 방향이 있습니다:
  - 위상학적 자료 분석을 이용하여 특성(feature)을 만들어, 자료 X에 위상학적 특성을 추가하기: 더 흔한 방식
  - 손실함수(loss function) $\mathcal{L}$에 위상학적 손실 고려하기: 최근 주목

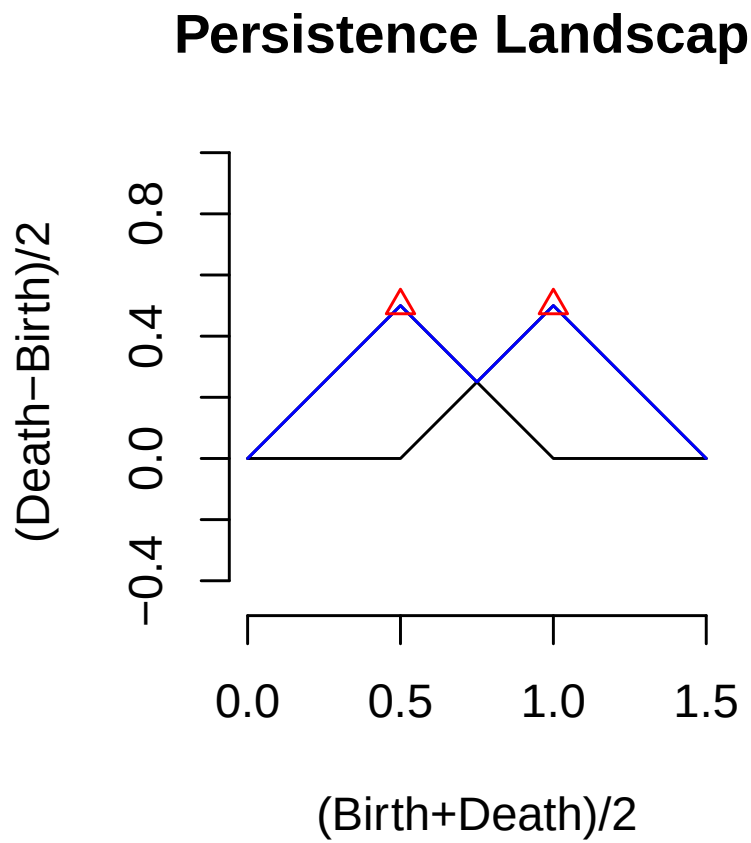
Persistent homology를 한 번 더 요약해서 유클리드 공간 또는 함수 공간에 넣습니다.

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



Persistence Landscape은 Persistent homology의 함수 요약입니다.



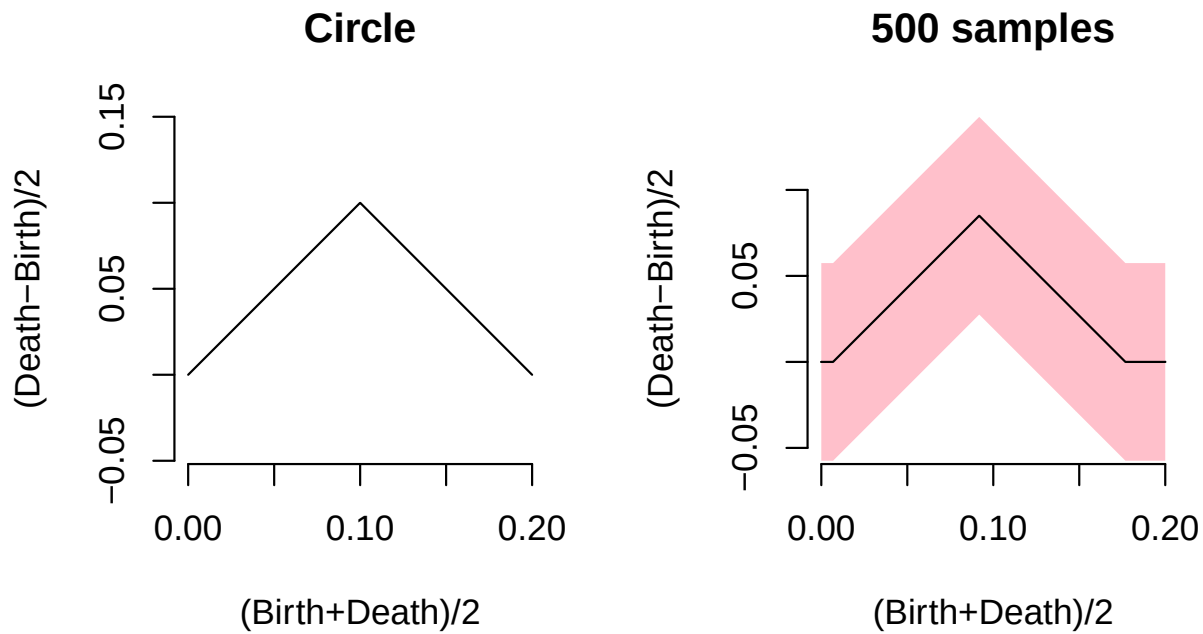


persistence landscape의 신뢰띠는 붓스트랩으로 계산할 수 있습니다.

- 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:

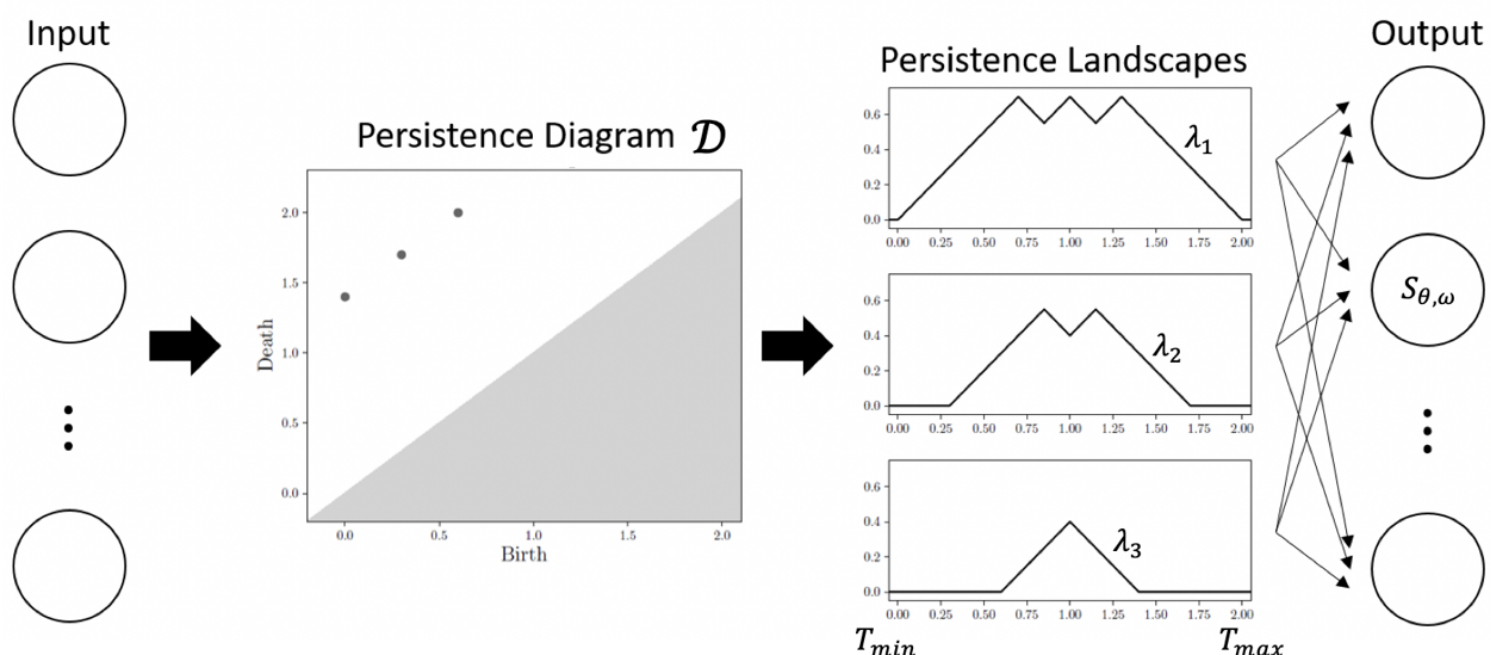
$$\mathbb{P}(\lambda_X(t) - c_n \leq \lambda_M(t) \leq \lambda_X(t) + c_n \, \forall t) \geq \mathbb{P}(||f_M - f_X|| \leq c_n) \geq 1 - \alpha,$$

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

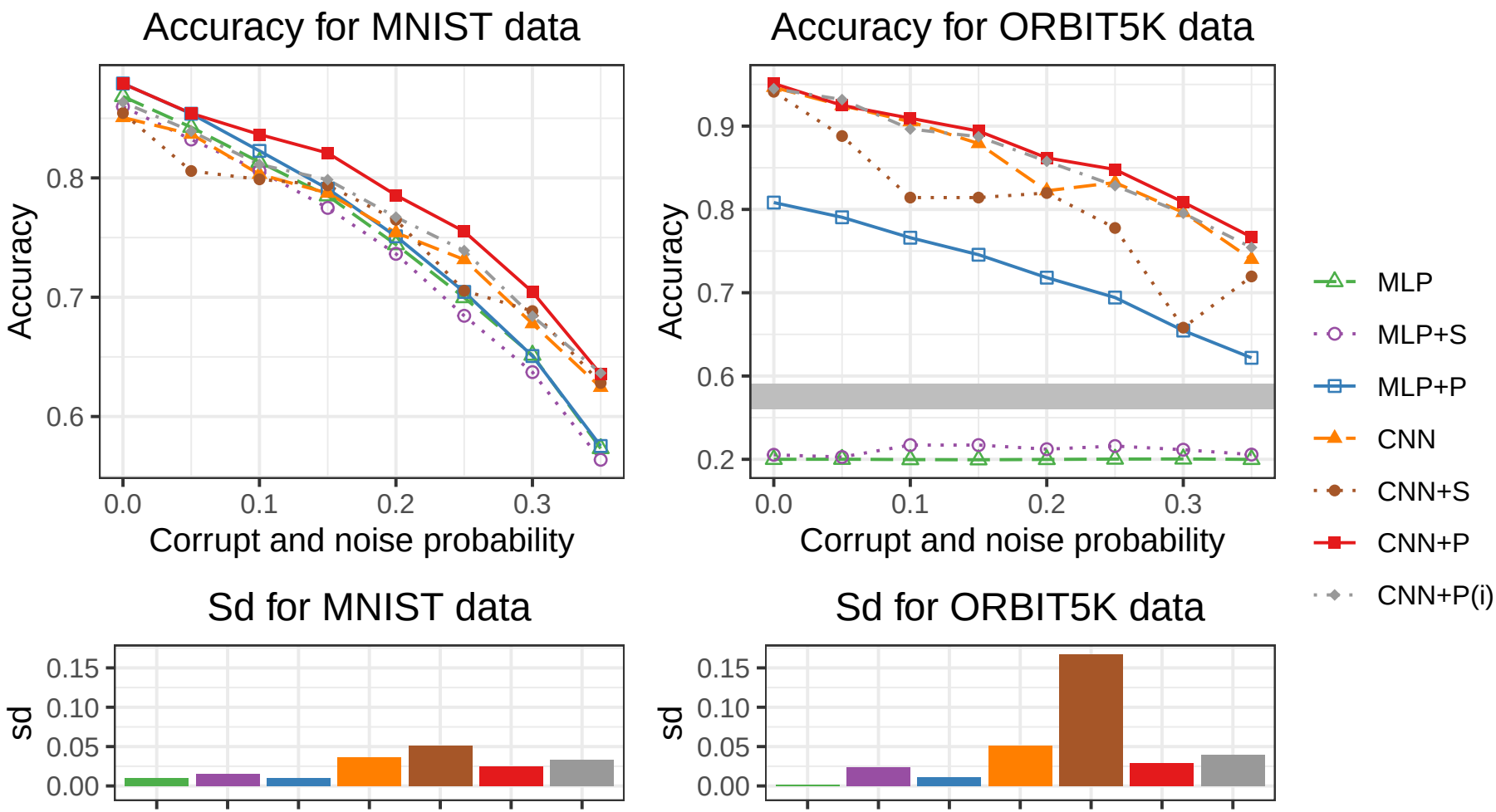


Persistence Landscape으로 위상학적 층(topological layer) 만들기

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



Persistence Landscape으로 위상학적 층(topological layer) 만들기



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

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

R 패키지 TDA는 위상학적 자료 분석을 해주는 C++ 라이브러리의 R 인터페이스를 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 위상학적 자료 분석을 해주는 C++ 라이브러리인 GUDHI/Dionysus/PHAT의 R 인터페이스를 제공합니다.

위상 구조(Topological Structure)의 통계적 추정(Statistical Inference)

밀도 군집 (Density Clustering)

거리 공간(Metric Spaces), 덮개(Covers), 단체 복합체(Simplicial Complex)

Mapper

Reach와 기하학적 재구성(Geometric Reconstruction)

내재적 차원(Intrinsic Dimension) 추정

호몰로지(homology)와 그의 추정

Persistent Homology와 통계적 추정

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

R 패키지 TDA: 위상학적 자료 분석을 위한 통계 계산 도구

참조문헌

참조문헌 I

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers Artif. Intell.*, 4:667963, 2021. doi: 10.3389/frai.2021.667963. URL https://doi.org/10.3389/frai.2021.667963.

Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.

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.

Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-79160-X; 0-521-79540-0.

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 II

Jisu Kim, Jaehyeok Shin, Frédéric Chazal, Alessandro Rinaldo, and Larry Wasserman. Homotopy Reconstruction via the Cech Complex and the Vietoris-Rips Complex. In Sergio Cabello and Danny Z. Chen, editors, *36th International Symposium on Computational Geometry (SoCG 2020)*, volume 164 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 54:1–54:19, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. ISBN 978-3-95977-143-6. doi: 10.4230/LIPIcs.SoCG.2020.54. URL `https://drops.dagstuhl.de/opus/volltexte/2020/12212`.

Larry Wasserman. Topological data analysis, 2016.

감사합니다!