Inferring Underlying Manifold of Data by the Use of Persistent Homology Analysis

筑波大学
University of Tsukuba

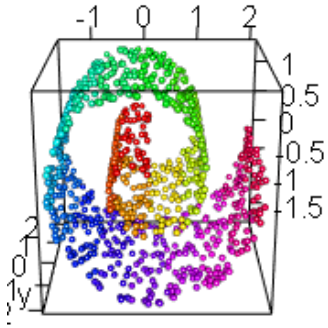
University of Tsukuba, Japan
Rentaro Futagami,
○Noritaka Yamada,
Takeshi Shibuya

CTIC 2019

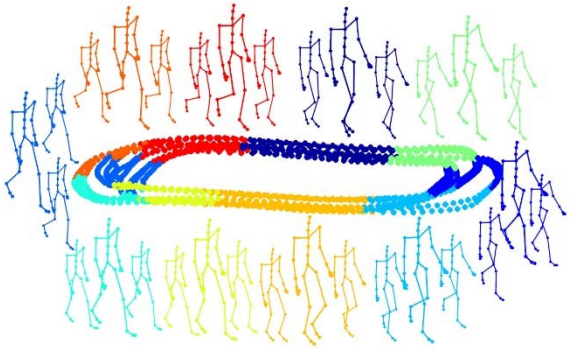
Introduction (1/2)

Many data sets have the property that the data points lie close to a manifold.

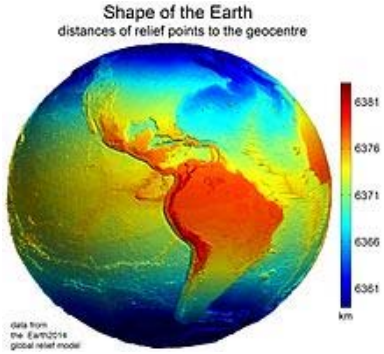
We call this manifold to be “underlying manifold.”



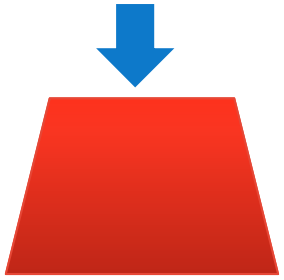
Swiss roll



Annular data

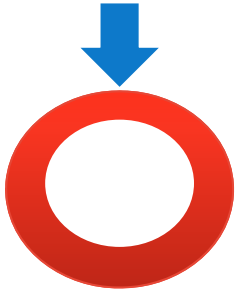


Spherical data



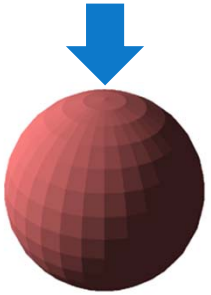
plane

$b_1 = 0$
$b_2 = 0$



1-annulus

$b_1 = 1$
$b_2 = 0$



2-sphere

$b_1 = 0$
$b_2 = 1$

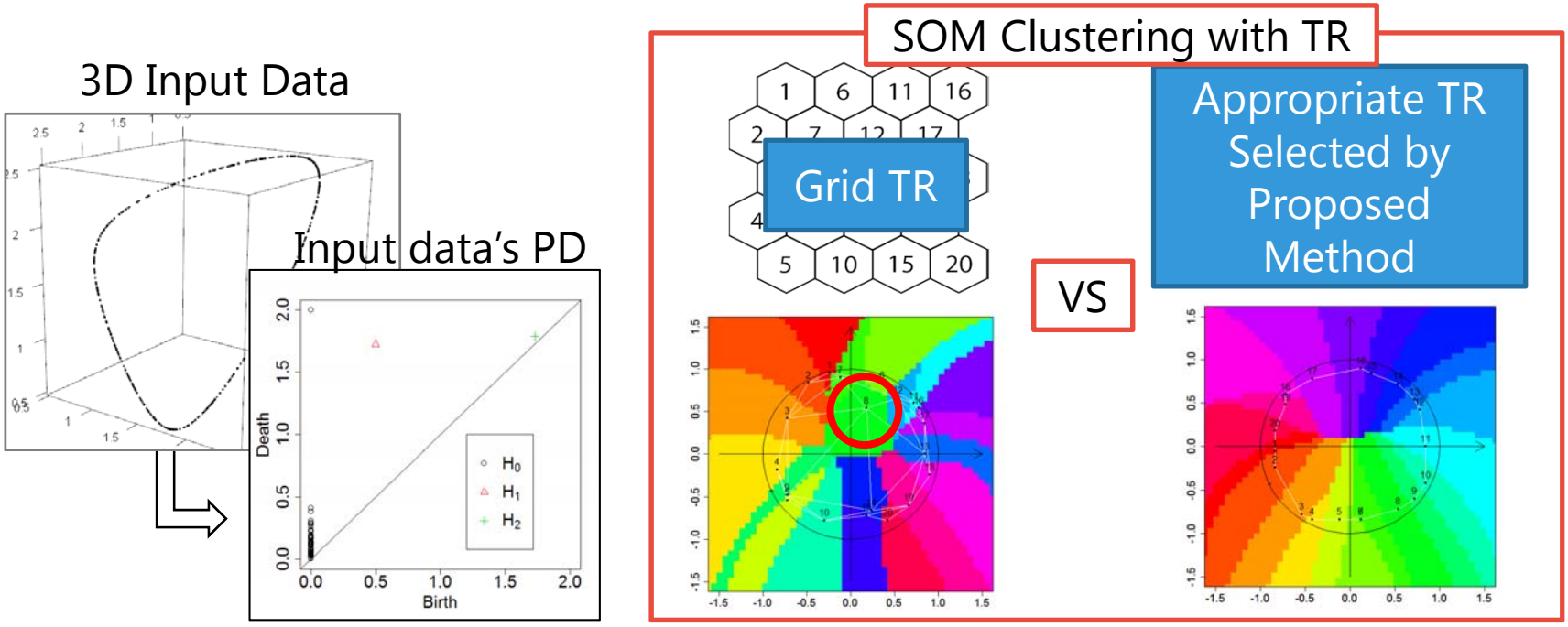
N. Pitelis, et al., “Learning a Manifold as an Atlas,” 2013 IEEE Conference on Computer Vision and Pattern Recognition, p.1643, 2013

https://ja.wikipedia.org/wiki/%E5%9C%B0%E7%90%83#/media/File:Earth2014shape_SouthAmerica_small.jpg

筑波大学 University of Tsukuba

Introduction (2/2)

Inferring underlying manifold improves some dimensionality reduction method[Futagami+ 2016].

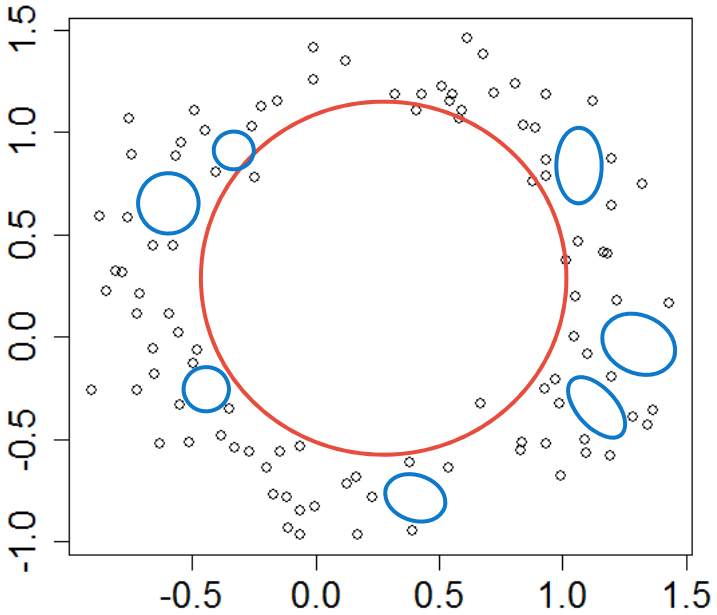


Futagami, R., Shibuya, T.: A method deciding topological relationship for self-organizing maps by persistent homology analysis. In: Proceedings of SICE Annual Conference 2016, pp. 1064–1069 (2016)

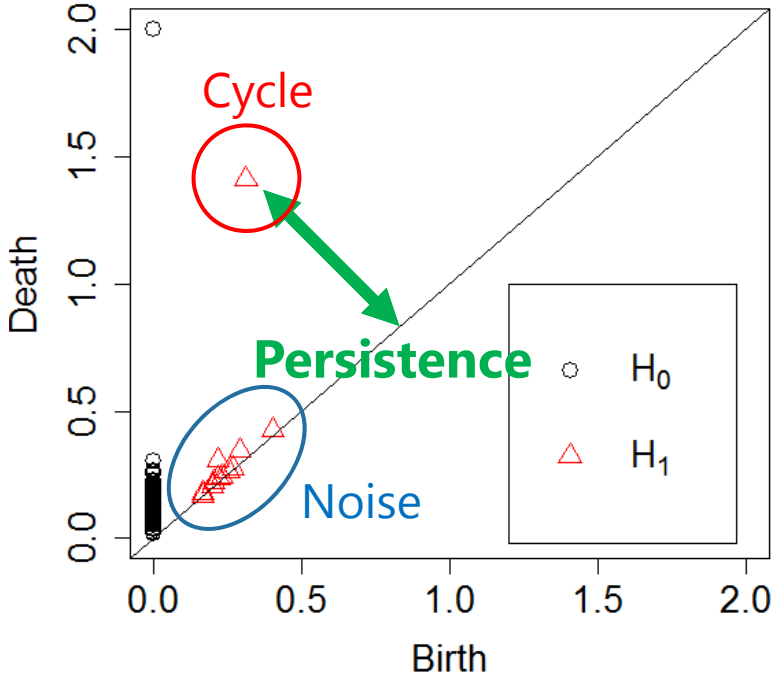
Persistent homology (1/2)

Definitions of cycle and noise on persistent homology

Annular data



Persistence Diagram



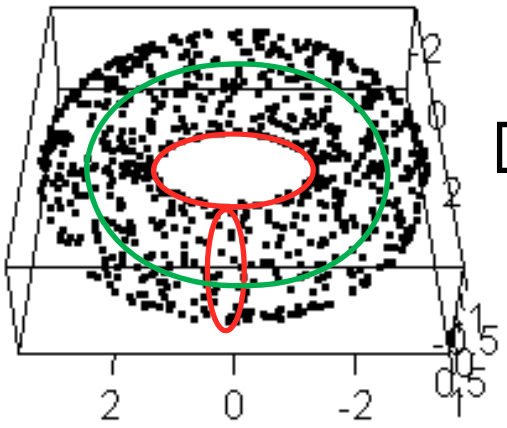
Cycle : Large persistence holes

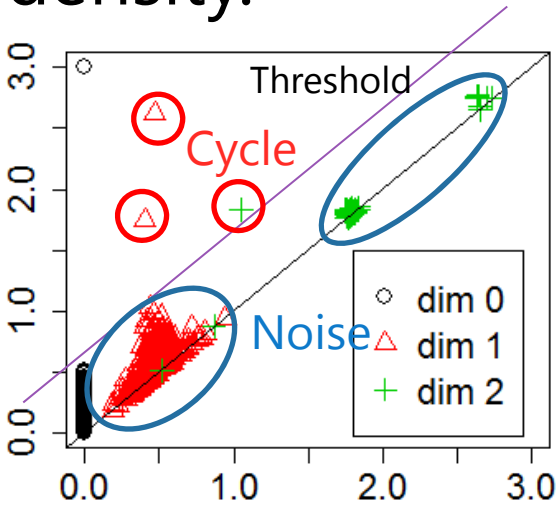
corresponding those in the underlying manifold

Noise : Small persistence holes

Persistent homology (2/2)

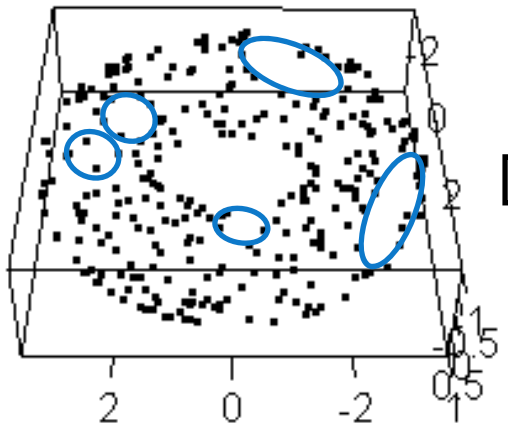
Distinguishing cycles and noises is harder when data is low density.

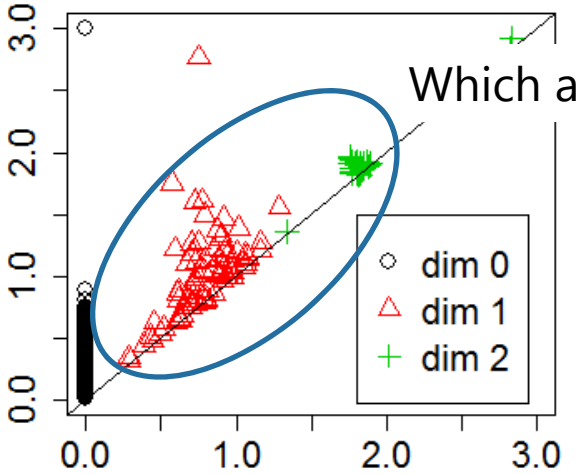




Easy to find
correct Betti numbers
(=the number of cycles).

$b_1 = 2$
$b_2 = 1$





Which are cycles?

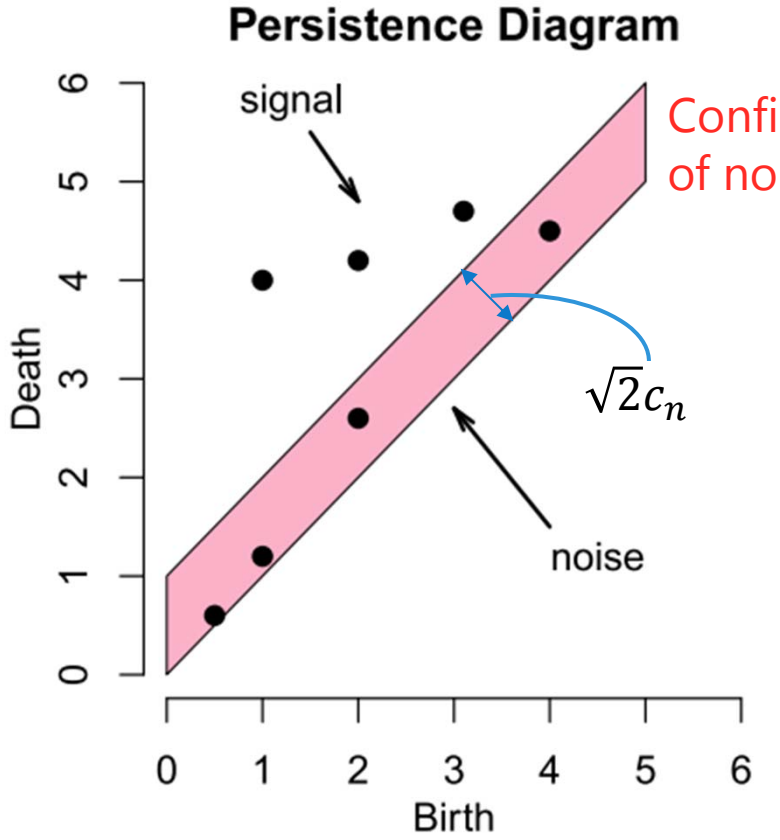
Setting a threshold is difficult.

?

Related works (1/3)

Confidence interval of noise in persistence diagram [Fasy+ 2014]

Holes in confidence interval are noise with α% possibility.



Confidence interval of noise

$$\limsup_{n\to\infty} \mathbb{P}\left(W_\infty\left(\hat{\mathcal{P}}, \mathcal{P}\right) > c_n\right) \leq \alpha$$

Most cycles are considered as noise because this threshold tend to be too large.

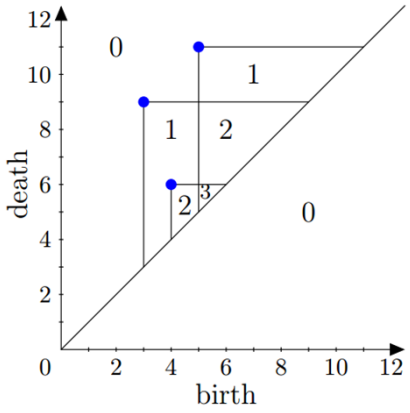
➡ The number of cycle are estimated less than the actual.

Related works (2/3)

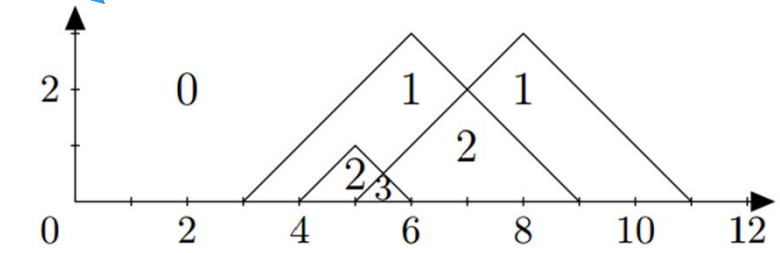
Persistence landscape [Bubenik+ 2015]

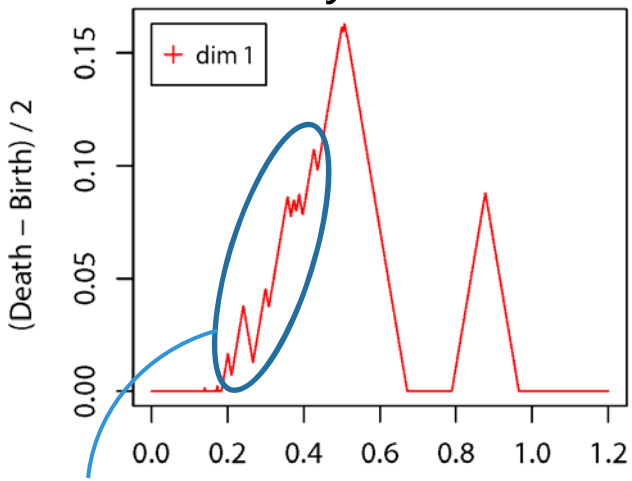
Holes composing the highest landscape are considered as cycles for arbitrary domain.

The number of local maxima
= The number of cycles



45° rotating





Are these zigzags cycles, or noises?

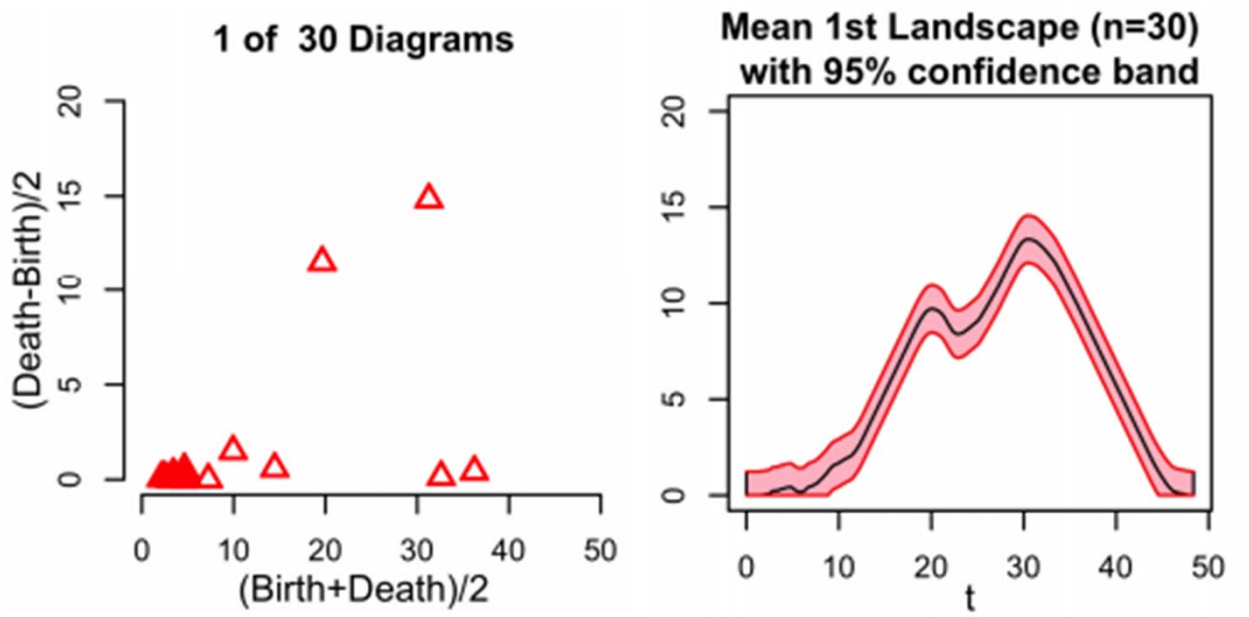
Difficult to find cycles because of zigzags on the landscape.

This method cannot estimate the number of cycles correctly.

Related works (3/3)

Mean landscape [Bubenik+ 2015]

The mean of persistence landscape of some samples

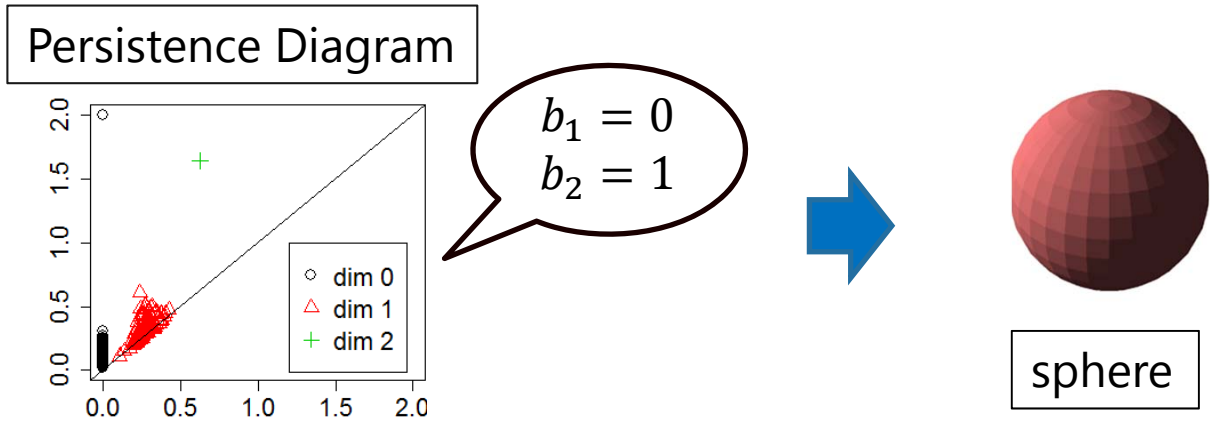


Zigzags decrease, but **still** difficult to find cycles.

This method also cannot estimate the number of cycles correctly.

Purpose of this study

Estimating Betti numbers to infer underlying manifold.



Conventional methods cannot estimate Betti numbers correctly.
So, experts have to analyze calculation results of persistent homology manually.

We propose a method to estimate the Betti numbers of underlying manifold automatically and correctly.

Proposed method (1/5)

Proposed method combined:

- Subsampling
- Persistence landscape
- Interpreting zigzags on persistence landscape

Persistence landscape is effective to reduce noises.
But, interpreting zigzags is difficult.

We interpret zigzags using these two methods.

1. The threshold based on the mean of persistence of noises
2. Analyzing persistence landscape fuzzily with smoothing

Proposed method (2/5)

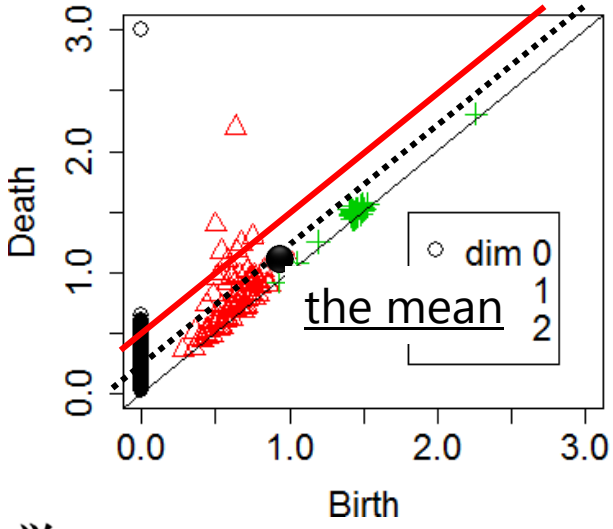
Threshold to distinguish cycles and noises based on persistence

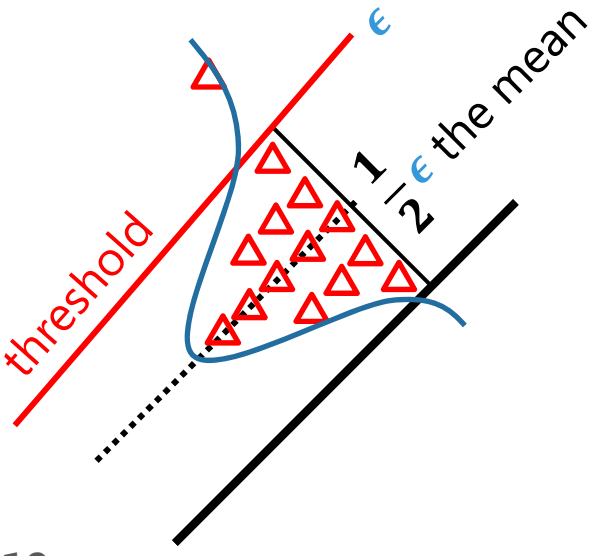
Let us assume that :

The number of noises ≫ The number of cycles

The mean of persistence of all holes ≈ The mean of persistence of noises

We use double the mean of persistence of all holes ϵ as a threshold.



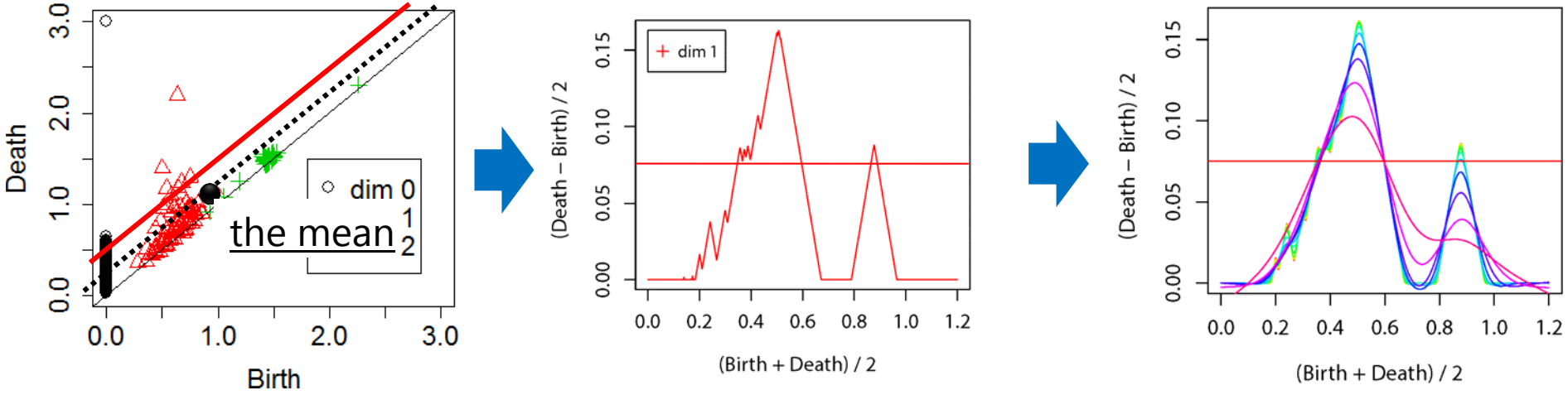


Proposed method (3/5)

Smoothing persistence landscape to interpret zigzags fuzzily

(Analogy of the multi resolution analysis)

Fit a cubic smooth spline based on B-spline
with various smoothing parameter $spar \in \{0, 0.1, 0.2, \cdots, 1\}$.



We find $f(x)$ that minimize σ s.t.

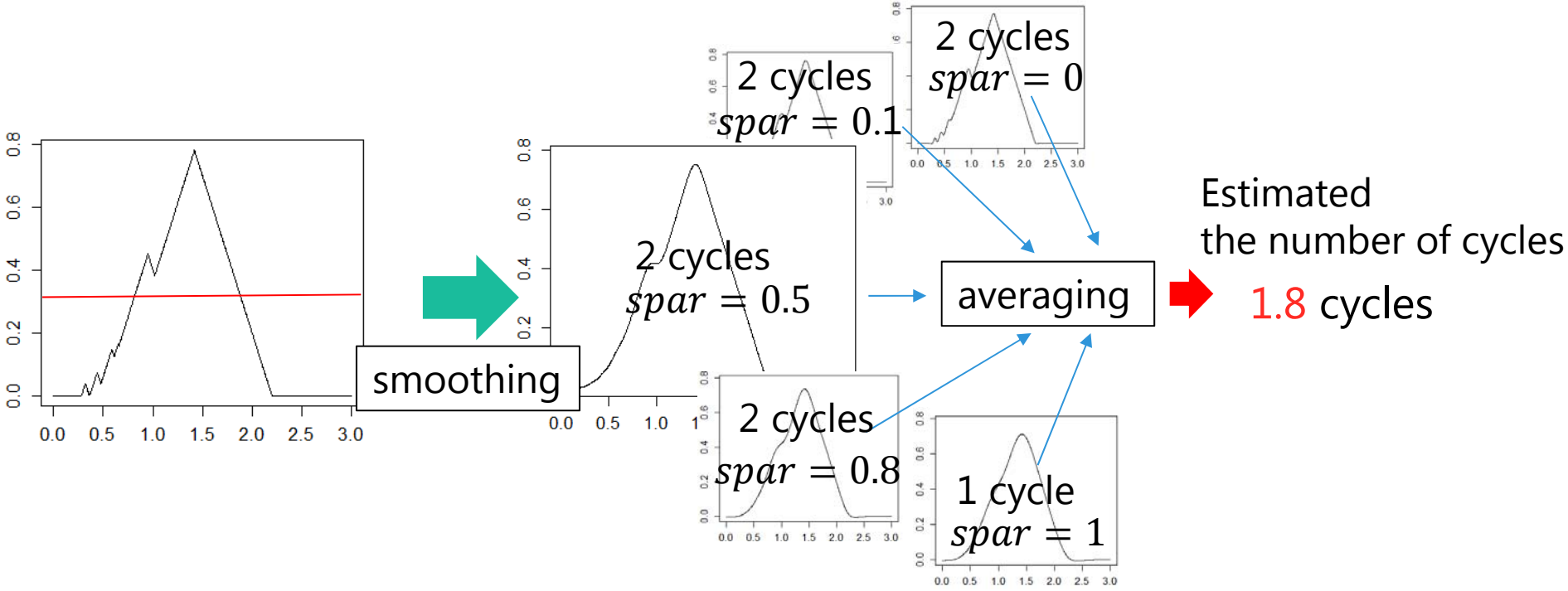
$$\sigma = \sum_{i=1}^{n} (y_i - f(x_i)) + \lambda \int \{f''(x)\}^2 dx$$

$$(\lambda = \rho * 256^{3*spar-1})$$

$$(\rho = \frac{\sum_{i=1}^{n} \{B_i(x_i)\}^2}{\sum_{i=1}^{n} \int \{B''(t)\}^2 dt})$$

Proposed method (4/5)

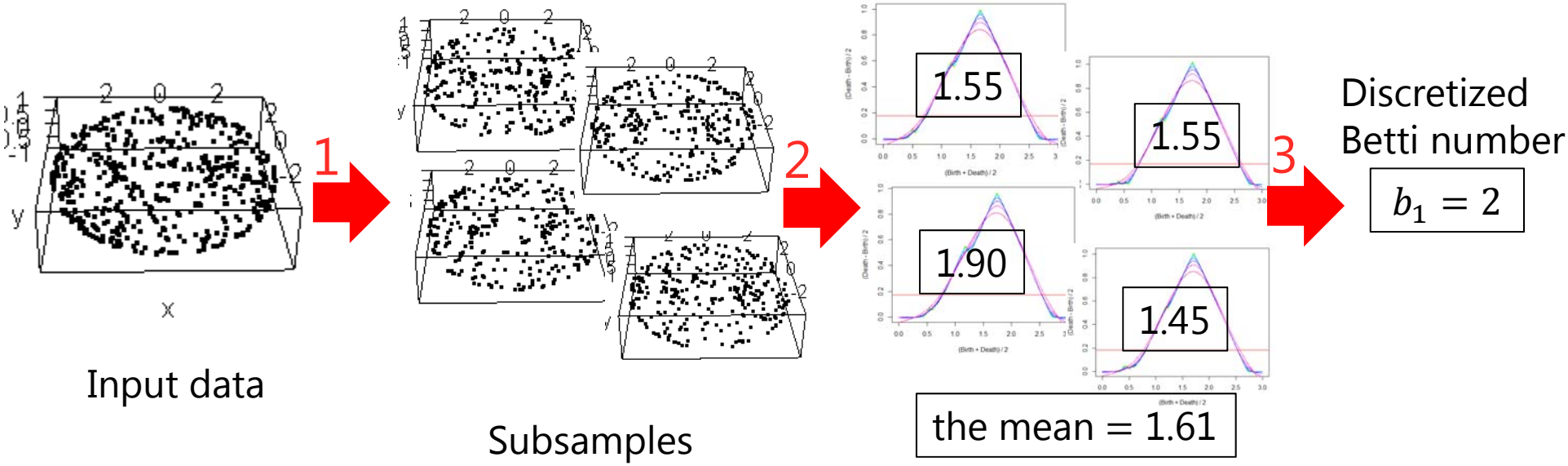
Smoothing persistence landscape to interpret zigzags fuzzily



1. Smooth persistence landscape with various smoothing parameter.
2. Count local maxima above the threshold in each smoothed persistence landscape.
3. Average the number of local maxima.

Proposed method (5/5)

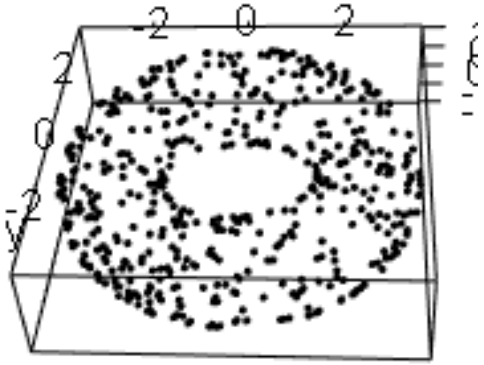
We use subsamples to reduce computational complexity and obtain robust estimate.



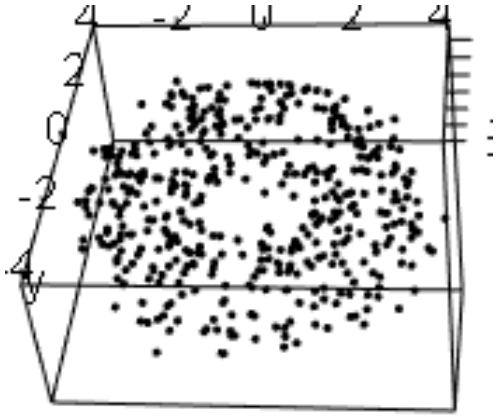
1. Subsampling
2. Estimating the number of cycles in each subsample and averaging them
3. Rounding the mean

Experiments

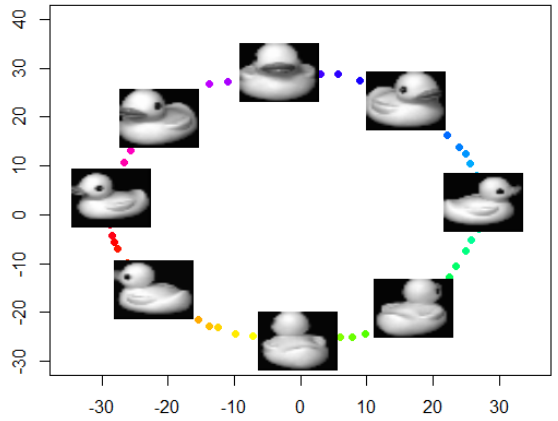
We estimate Betti numbers of three kind of data using proposed and conventional method.



1 Torus



2 Noisy torus



3 High-dimensional image data

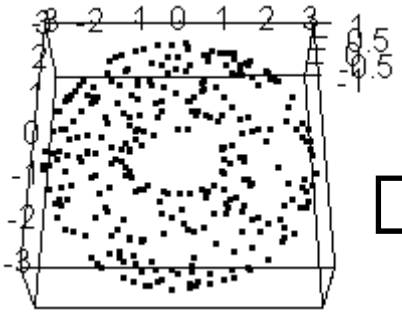
We evaluated these property of the proposed method:

1. Effectiveness for simple shape and low-dimensional data
2. Robustness for noises in data and the number of data points
3. Effectiveness for high-dimensional data

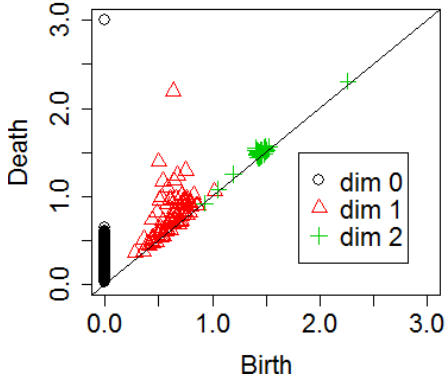
Experiment: torus (1/3)

We estimated Betti numbers of torus shape data.

Experiment setting	
Data sets	100
Data points	500
Subsamples	10
Points of a subsample	300

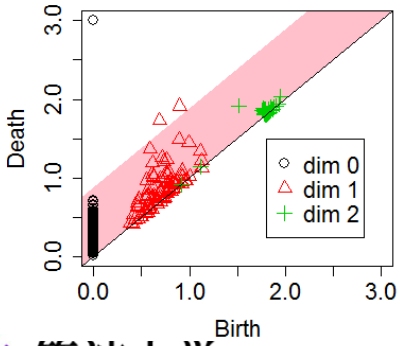


$b_1 = 2$
$b_2 = 1$

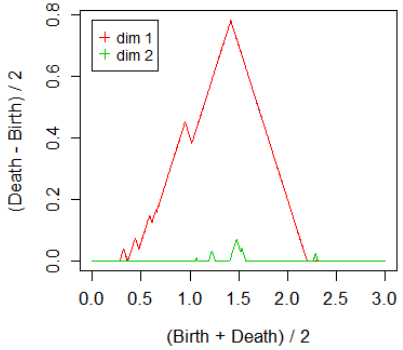


We compared these four methods.

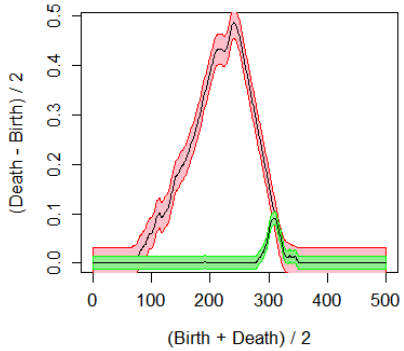
Confidence interval



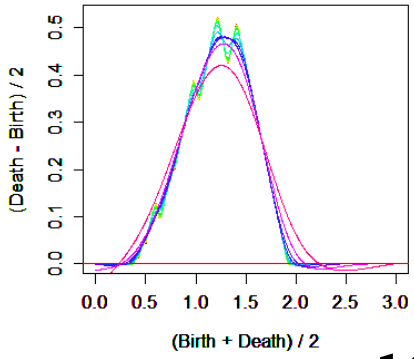
Persistence Landscape



Mean landscape



Proposed method



筑波大学 University of Tsukuba

Experiment: torus (2/3)

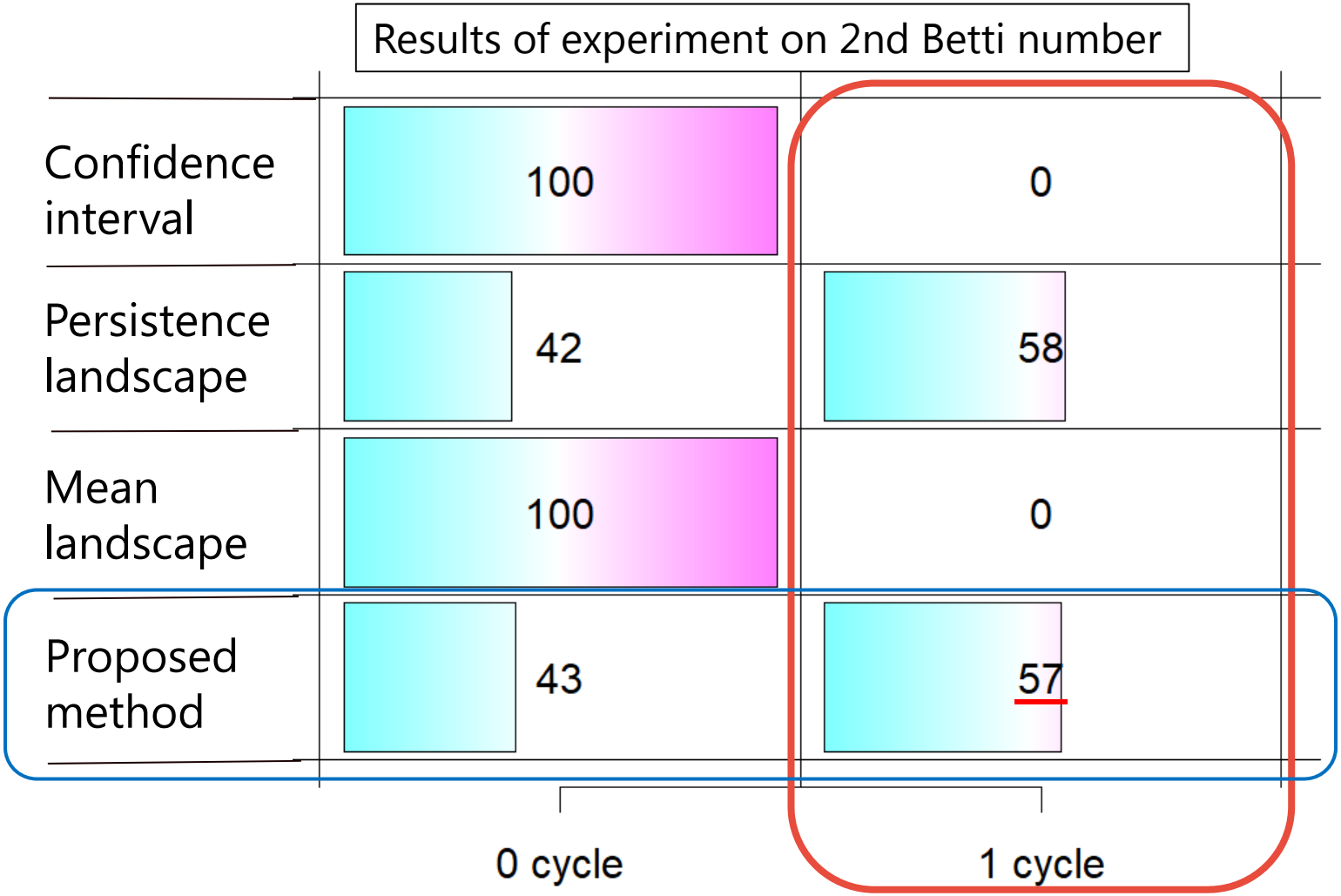
Proposed method estimated correct Betti numbers in the most data set.

Results of experiment on 1st Betti number

	0 cycle	1 cycle	2 cycles	3 cycles	4 cycles
Confidence interval	100	0	0	0	0
Persistence landscape	0	6	81	10	3
Mean landscape	0	11	80	9	0
Proposed method	0	10	90	0	0

Experiment: torus (3/3)

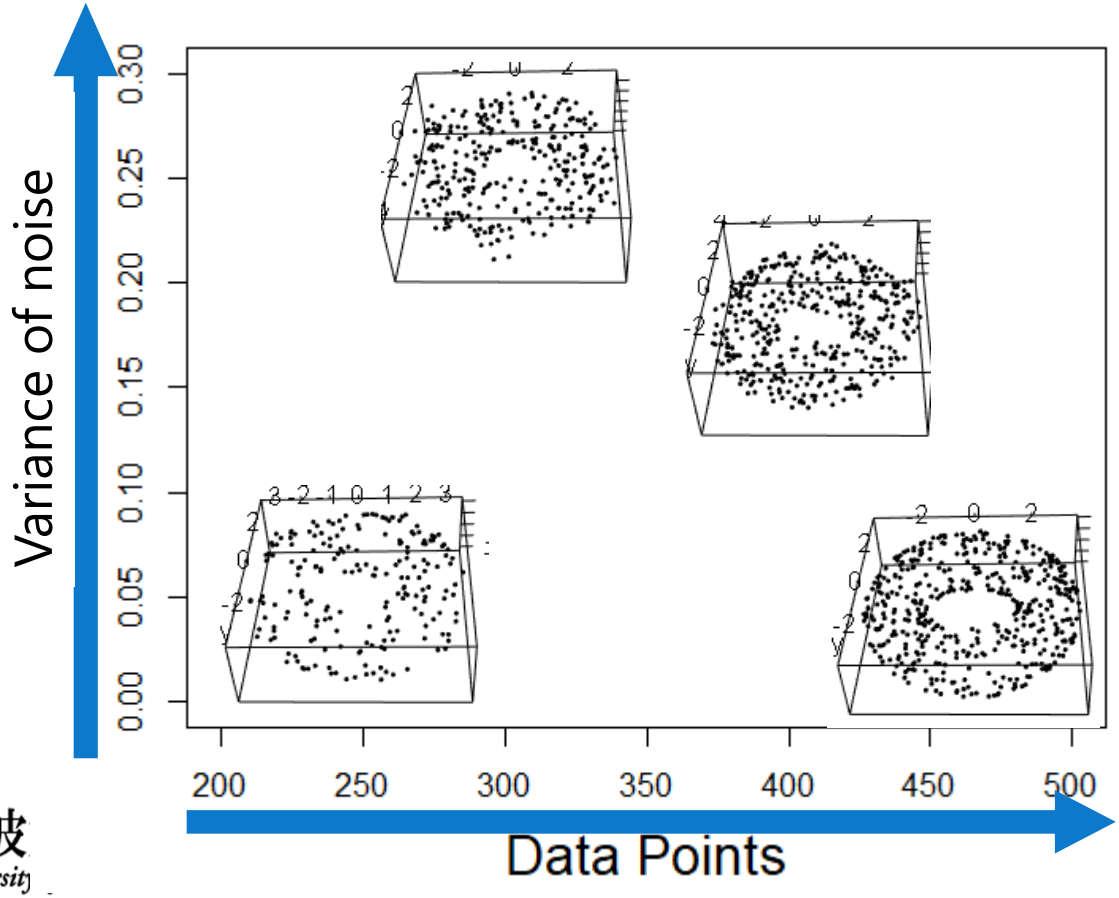
Proposed method estimated correct Betti numbers in many data set.



Experiment: noisy torus (1/3)

Estimate Betti numbers of torus shape data with Gaussian noises.

- The number of data points was determined randomly [200, 600].
- Variance of Gaussian noises was determined randomly [0, 0.3]

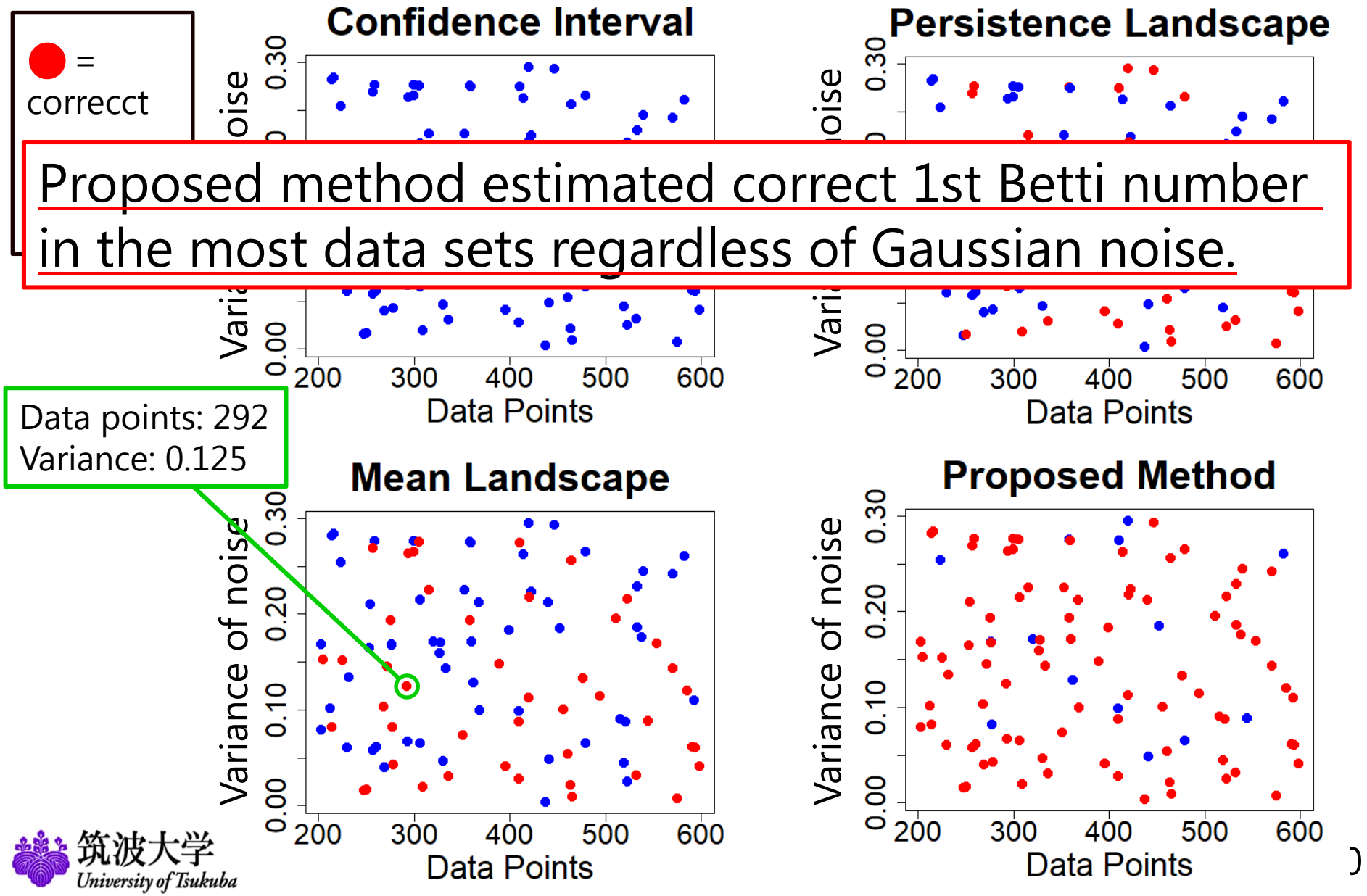


Noises are perturbation adding to data point, not small holes in persistence diagram in this experiment.

Each data point on torus is added Gaussian noise.

筑波 University

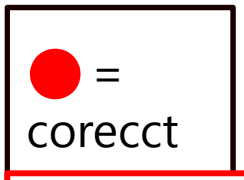
Results on 1st Betti number

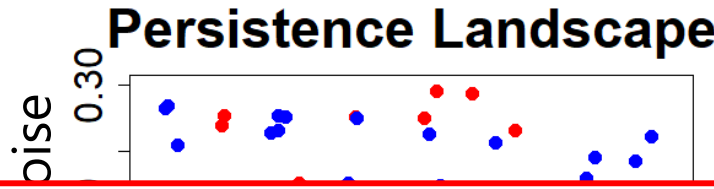


Proposed method estimated correct 1st Betti number in the most data sets regardless of Gaussian noise.

筑波大学 University of Tsukuba

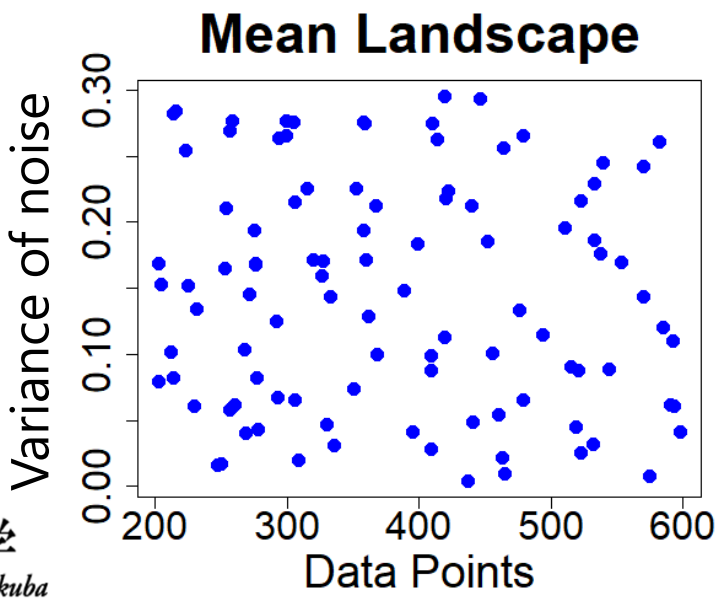
Results on 2nd Betti number

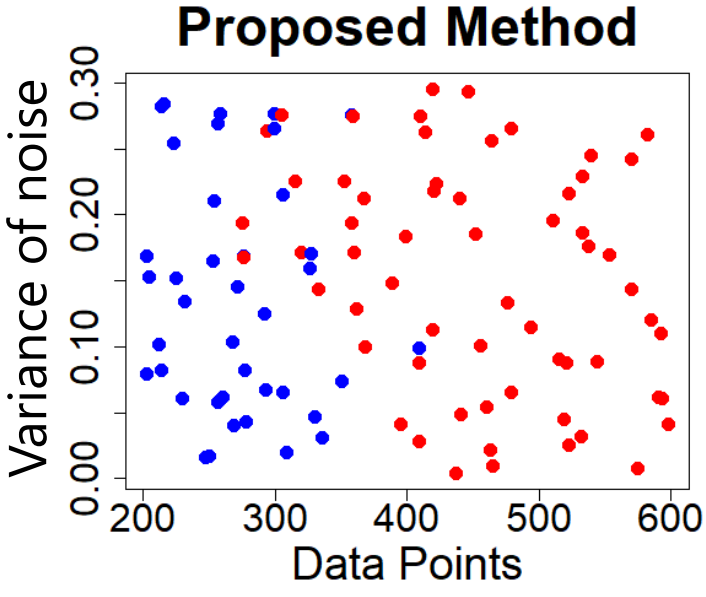




Proposed method estimated correct 2nd Betti number in the most data sets regardless of Gaussian noise.

However,
proposed method estimated incorrectly when data points are few.

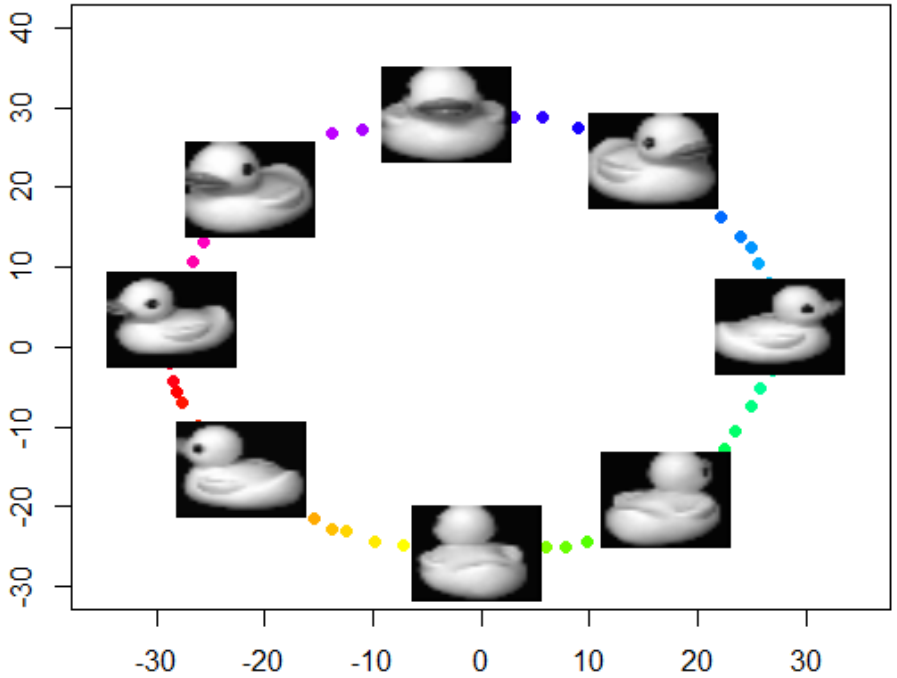




筑波大学 University of Tsukuba

Experiment: high-dimensional data

Estimate Betti numbers of COIL-20 data set.



Nene, S.A., et al.: Columbia object image library (COIL-20). Technical report, No. CUCS-005-96 (1996)

Estimate results of 1st Betti number

Method	Cycle
Confidence interval	0
Persistent landscape	1
Mean landscape	1
Proposed method	1

Proposed method estimated correct Betti number.

Conclusion

Contribution

- We proposed the method to estimate Betti numbers of the underlying manifold of data.
  - Practical threshold to distinguish cycles and noise
  - Estimating the number of cycles using smoothing persistence landscape
  - Using subsamples to reduce computational complexity
  - Robust against noises added to data

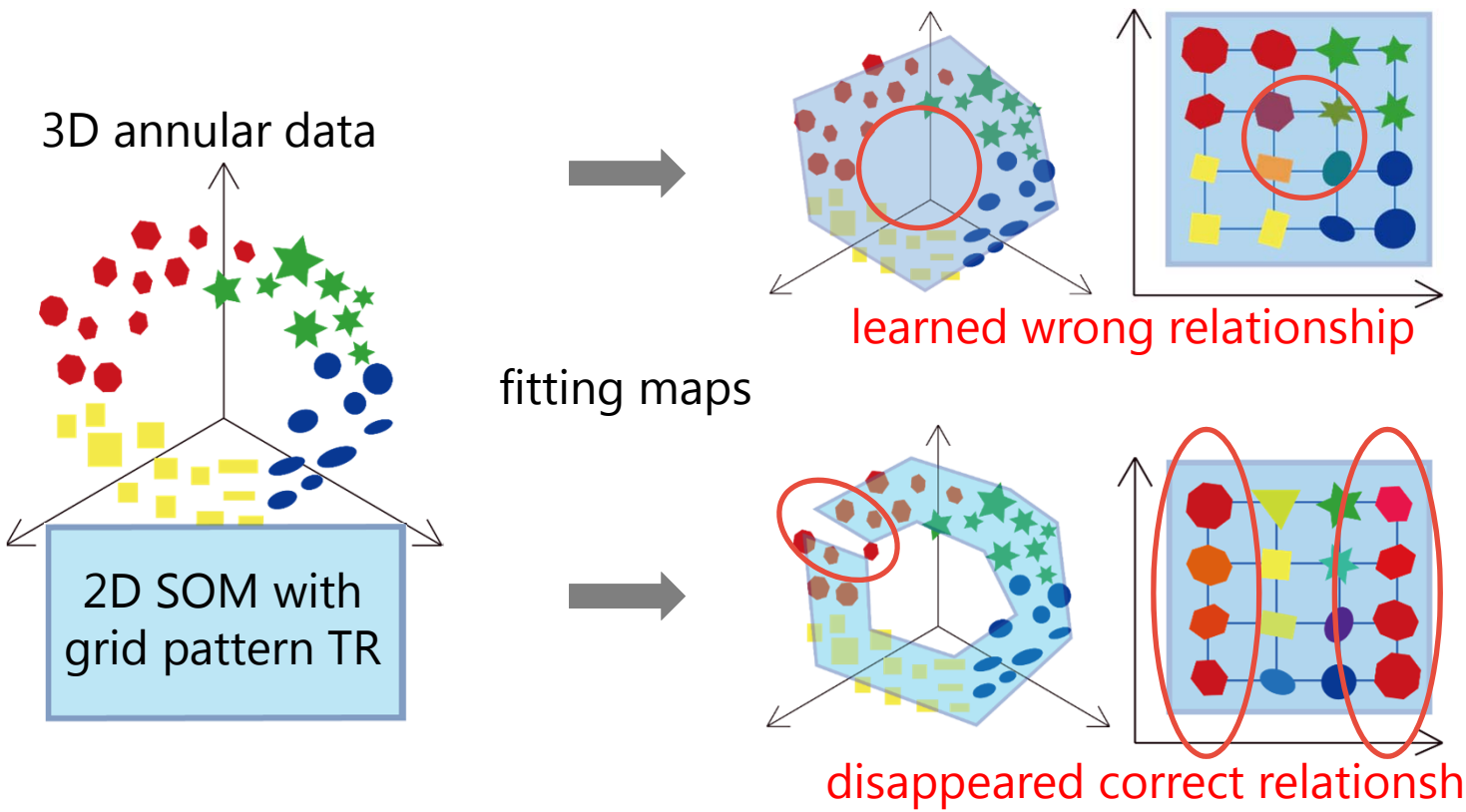
Feature works

- Changing smoothing parameter based on persistence to be more robustly against noises of persistent homology.
- Confirming effectiveness of our proposed method for other shape and various density.

Self-organaizing maps (SOM)

What is caused when TR is inappropriate for data?

Example of mapping 3D **annular data** by SOM which has a 2D **grid pattern topological relationship(TR).**



Smoothing

Find $f(x)$ that minimize σ s.t.

$$\sigma = \sum_{i=1}^{n} (y_i - f(x_i)) + \lambda \int \{f''(x)\}^2 dx$$

B-spline

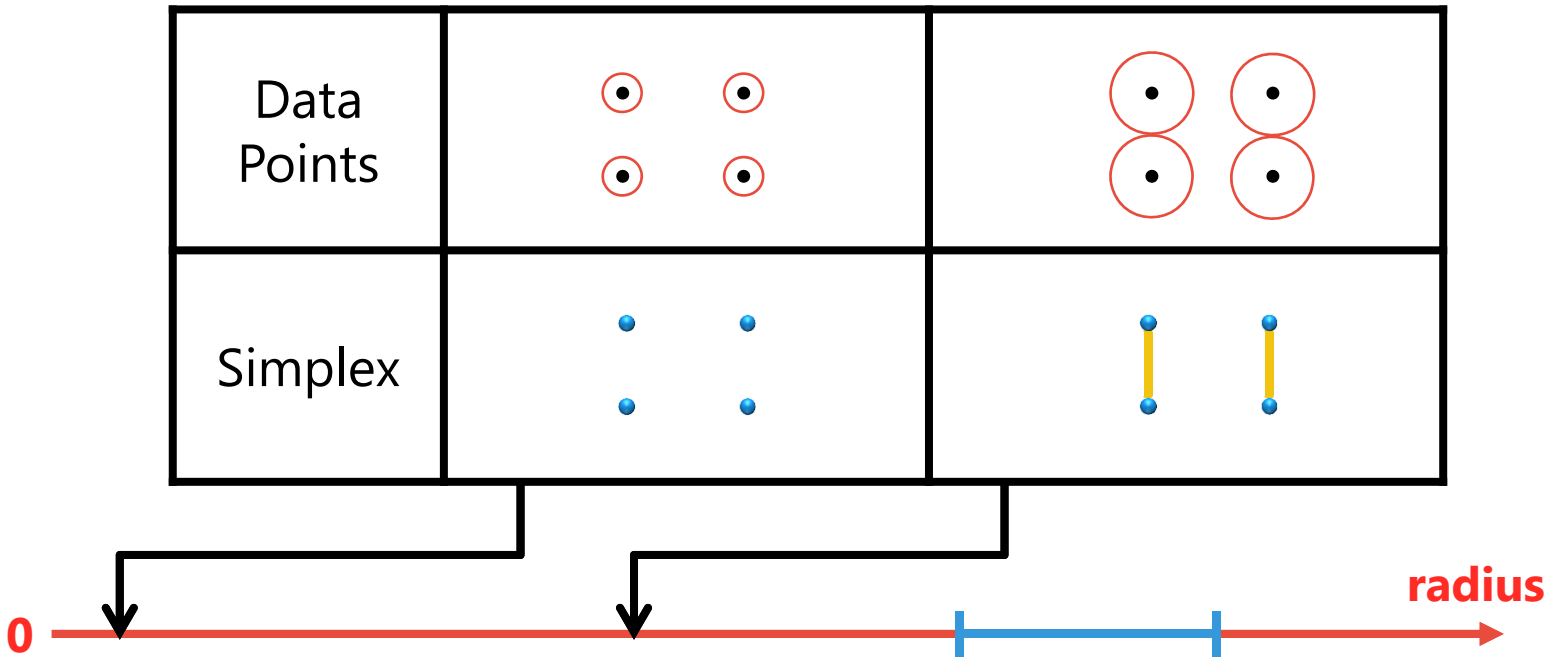
$$f(x) = \sum_{i=1}^{m+3} B_i^{(3)}(x)\beta_i$$

$$B_i^{(d)}(x) = \frac{x - x_i}{x_{i-d-1} - x} B_i^{(d-1)}(x) + \frac{x_{i+d} - x}{x_{i+d} - x_i} B_{i+1}^{d-1}(x) \quad \begin{matrix}(i = 1, \cdots, m) \\ (d = 0, 1, 2)\end{matrix}$$

$$B_i^{(0)}(x) = \begin{cases} 1 \ (x_i \leq x < x_{i-1}) \\ 0 \quad \text{otherwise} \end{cases}$$

Persistent Homology: Algorithm

How to detect **holes** in Persistent Homology?

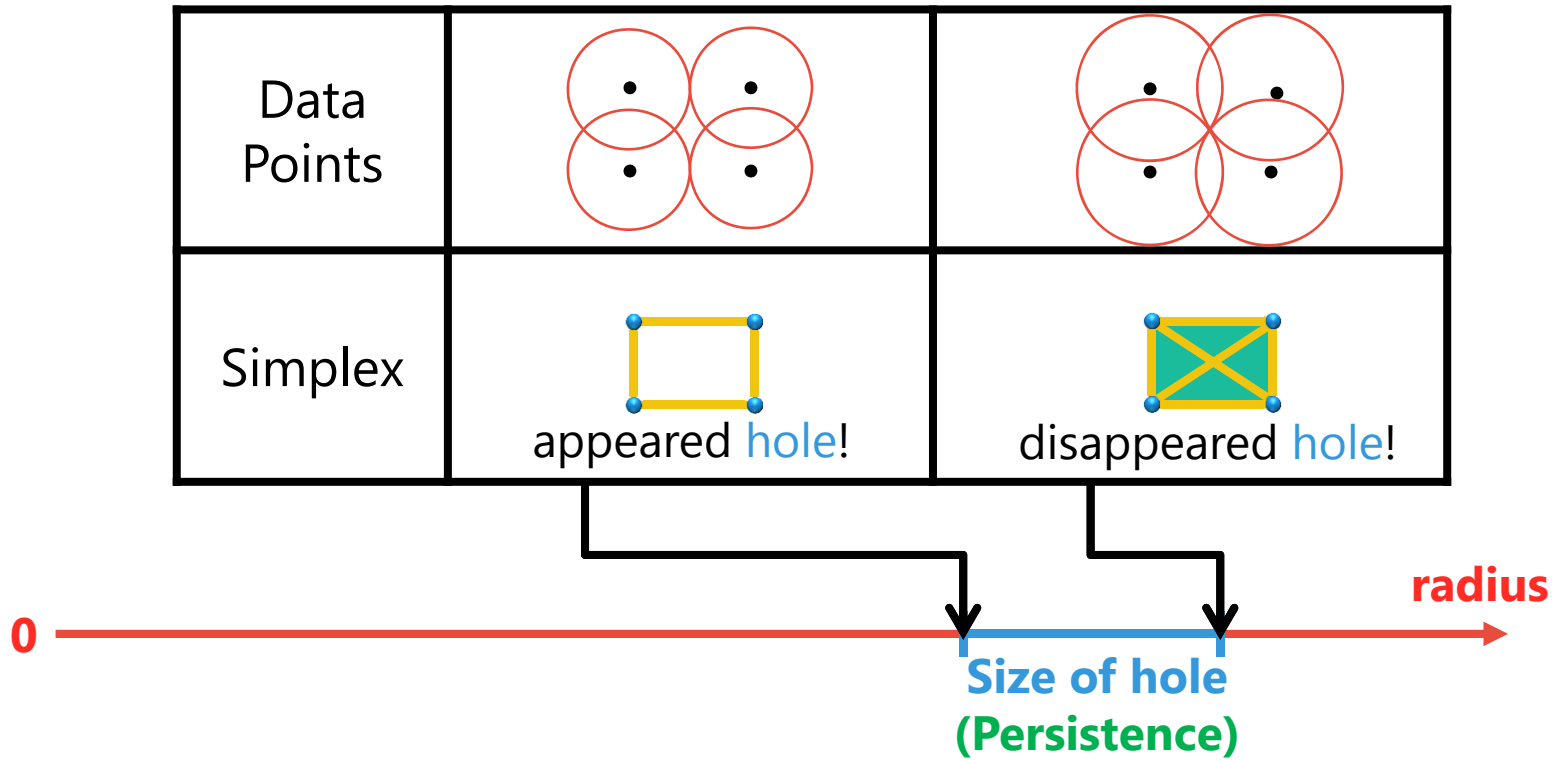


Step1: Expands radius of each data points

Step2: When k points have intersection, gives
$k-1$-simplex (means triangle of $k-1$D)

Persistent Homology: Algorithm

How to detect **holes** in Persistent Homology?



Step3: If k-simplex make loop without $k+1$-simplex,
regards this loop as a $k+1$D hole (k-homology feature)