Topology for Data Science 1: An Introduction to Topological Data Analysis

Peter Bubenik

University of Florida
Department of Mathematics,
peter.bubenik@ufl.edu
http://people.clas.ufl.edu/peterbubenik/

January 23, 2017

Tercera Escuela de Análisis Topológico de Datos
y Topología Estocástica
ABACUS, Estado de México

Topological Data Analysis

What is topology and why use it to analyze data?

Topology is a branch of mathematics which is good at extracting global qualitative features from complicated geometric structures.

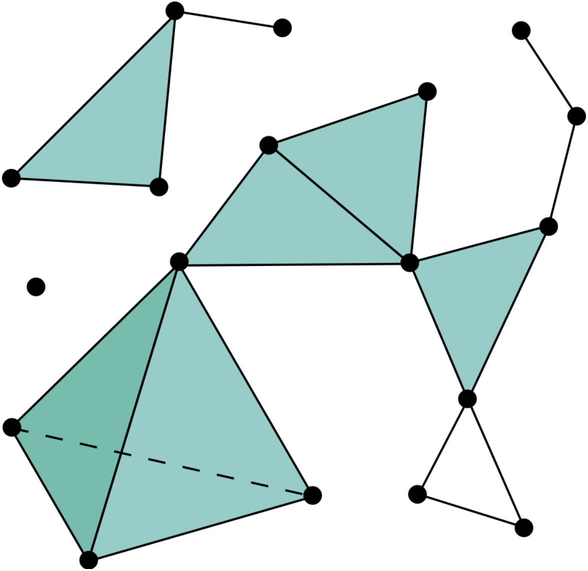
Example of a topological question

Is a given graph connected?

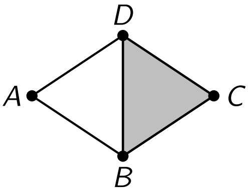
Topological Data Analysis

uses topology to summarize and learn from the "shape" of data.

Simplicial complexes

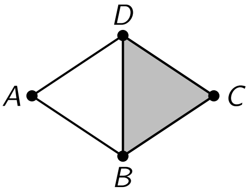


Exercise 1: Simplicial complexes for computers



What is the corresponding abstract simplicial complex?

Exercise 1: Simplicial complexes for computers

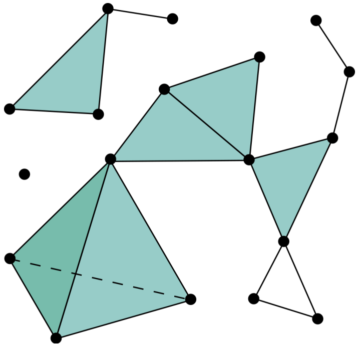


What is the corresponding abstract simplicial complex?

$$\{\{A\},\{B\},\{C\},\{D\},\{A,B\},\{A,D\},\{B,C\},\{B,D\},$$
$$\{C,D\},\{B,C,D\}\}$$

Exercise 2: Betti numbers of simplicial complexes

$$\beta_0 = \# \text{ of connected components}$$
$$\beta_1 = \# \text{ of holes}$$
$$\beta_2 = \# \text{ of voids}$$



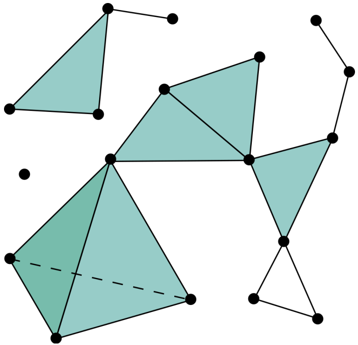
$\beta_0 =$

$\beta_1 =$

$\beta_2 =$

Exercise 2: Betti numbers of simplicial complexes

$$\begin{aligned}\beta_0 &= \#\text{ of connected components}\\ \beta_1 &= \#\text{ of holes}\\ \beta_2 &= \#\text{ of voids}\end{aligned}$$

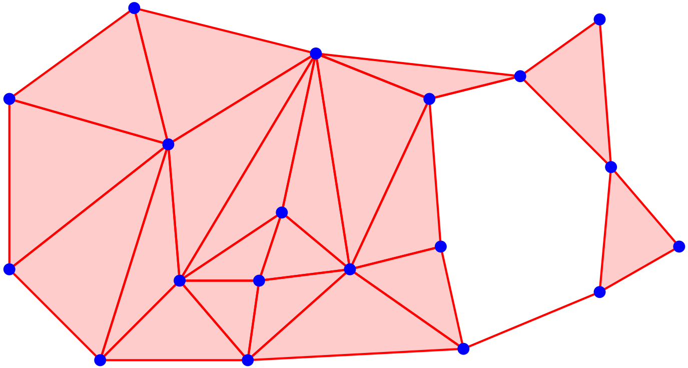


$$\begin{aligned}\beta_0 &= 3\\ \beta_1 &= 1\\ \beta_2 &= 1\end{aligned}$$

Homology of simplicial complexes

Definition

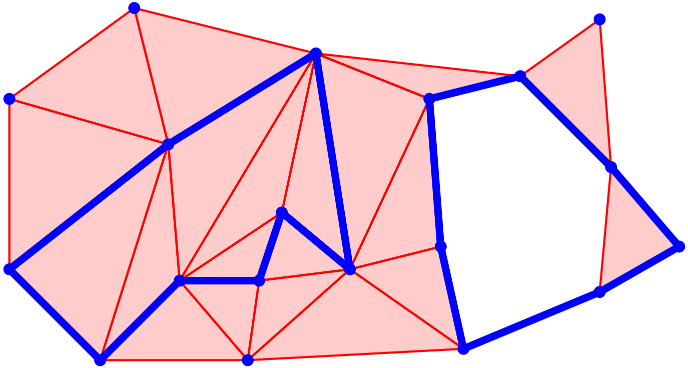
Homology in degree k is given by k-cycles modulo the k-boundaries.



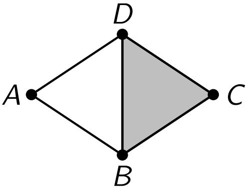
Homology of simplicial complexes

Definition

Homology in degree k is given by k-cycles modulo the k-boundaries.



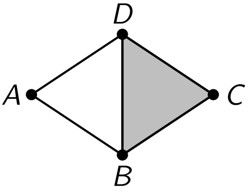
Exercise 3: Homology via linear algebra



Dimensions of vectors spaces of k-chains:

$$\dim(C_0) =$$
$$\dim(C_1) =$$
$$\dim(C_2) =$$

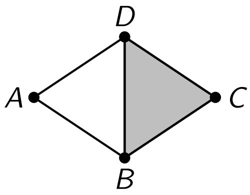
Exercise 3: Homology via linear algebra



Dimensions of vectors spaces of k-chains:

$$\begin{aligned} \dim(C_0) &= 4 \\ \dim(C_1) &= 5 \\ \dim(C_2) &= 1 \end{aligned}$$

Exercise 3: Homology via linear algebra



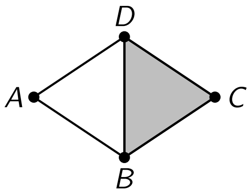
Dimensions of vectors spaces of k-chains:

$$\dim(C_0) = 4$$
$$\dim(C_1) = 5$$
$$\dim(C_2) = 1$$

Boundary matrices:

$$\partial_1 = \left[\quad \right] \qquad \partial_2 = \left[\quad \right]$$

Exercise 3: Homology via linear algebra



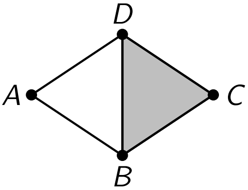
Dimensions of vectors spaces of k-chains:

$$\begin{aligned}\dim(C_0) &= 4\\ \dim(C_1) &= 5\\ \dim(C_2) &= 1\end{aligned}$$

Boundary matrices:

$$\partial_1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix} \quad \partial_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Exercise 3: Homology via linear algebra



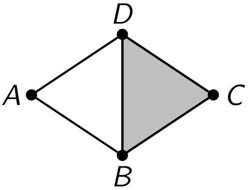
Dimensions of vectors spaces of k-chains:

$$\dim(C_0) = 4$$
$$\dim(C_1) = 5$$
$$\dim(C_2) = 1$$

Boundary matrices:

$$\partial_0 = 0 \quad \partial_1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix} \quad \partial_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \partial_3 = 0$$

Exercise 3: Homology via linear algebra



Dimensions of vectors spaces of k-chains:

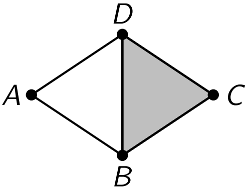
$$\dim(C_0) = 4$$
$$\dim(C_1) = 5$$
$$\dim(C_2) = 1$$

Boundary matrices:

$$\partial_0 = 0 \quad \partial_1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix} \quad \partial_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \partial_3 = 0$$

$$\beta_0 = \text{nullity}(\partial_0) - \text{rank}(\partial_1) =$$
$$\beta_1 = \text{nullity}(\partial_1) - \text{rank}(\partial_2) =$$
$$\beta_2 = \text{nullity}(\partial_2) - \text{rank}(\partial_3) =$$

Exercise 3: Homology via linear algebra



Dimensions of vectors spaces of k-chains:

$$\begin{aligned} \dim(C_0) &= 4 \\ \dim(C_1) &= 5 \\ \dim(C_2) &= 1 \end{aligned}$$

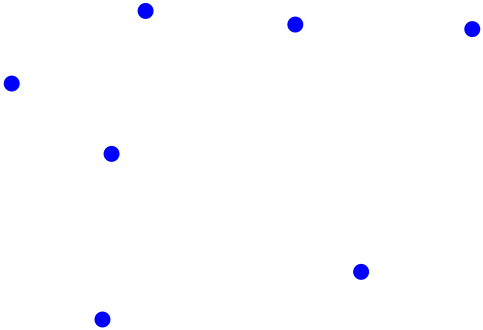
Boundary matrices:

$$\partial_0 = 0 \quad \partial_1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix} \quad \partial_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \partial_3 = 0$$

$$\begin{aligned} \beta_0 &= \text{nullity}(\partial_0) - \text{rank}(\partial_1) = 4 - 3 = 1 \\ \beta_1 &= \text{nullity}(\partial_1) - \text{rank}(\partial_2) = 2 - 1 = 1 \\ \beta_2 &= \text{nullity}(\partial_2) - \text{rank}(\partial_3) = 0 - 0 = 0 \end{aligned}$$

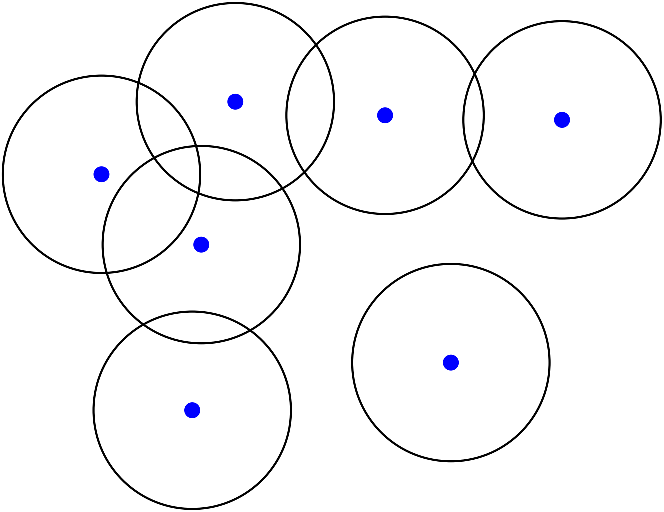
Simplicial complexes from point data

The Čech construction



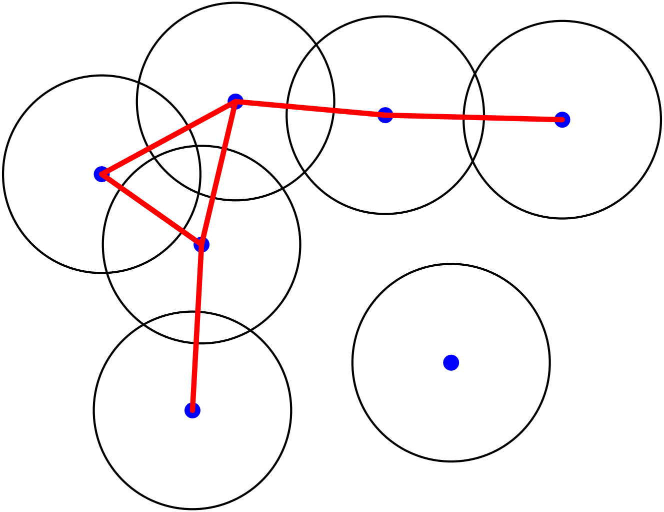
Simplicial complexes from point data

The Čech construction



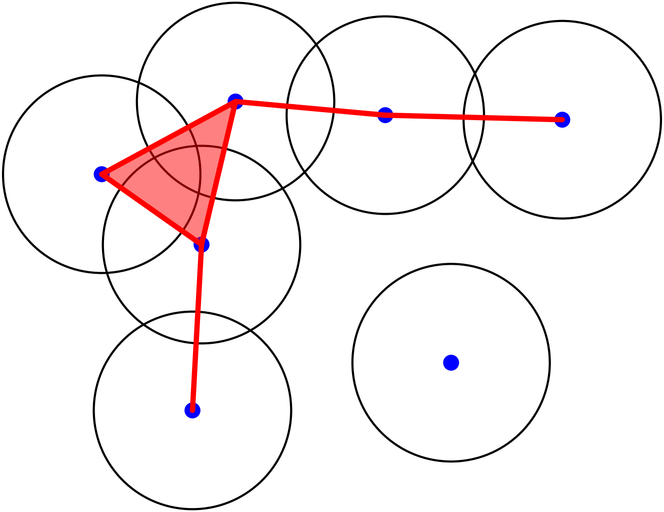
Simplicial complexes from point data

The Čech construction



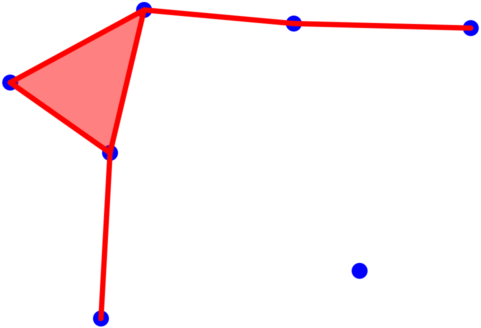
Simplicial complexes from point data

The Čech construction



Simplicial complexes from point data

The Čech construction

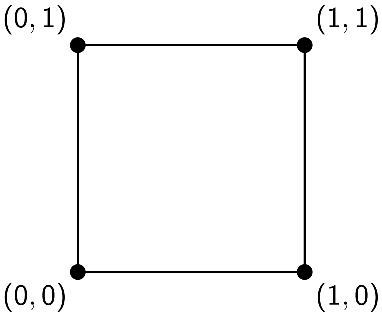


Exercise 4: Constructing a Čech complex

Draw a picture of $\check{C}_{\frac{1}{2}}(\{(0,0),(0,1),(1,0),(1,1)\})$.

Exercise 4: Constructing a Čech complex

Draw a picture of $\check{C}_{\frac{1}{2}}(\{(0,0),(0,1),(1,0),(1,1)\})$.



The parameter

Question

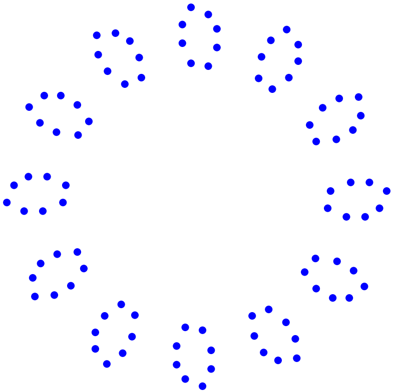
What is the right value for the parameter in the Čech construction?

The parameter

Question

What is the right value for the parameter in the Čech construction?

Often, there is no one "right" choice.

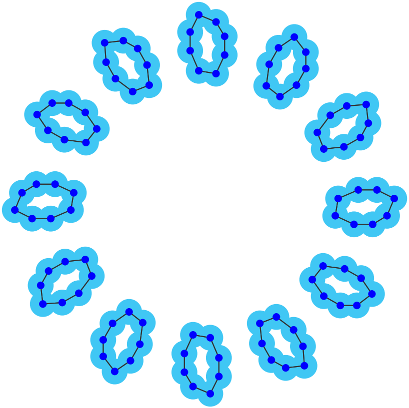


The parameter

Question

What is the right value for the parameter in the Čech construction?

Often, there is no one "right" choice.

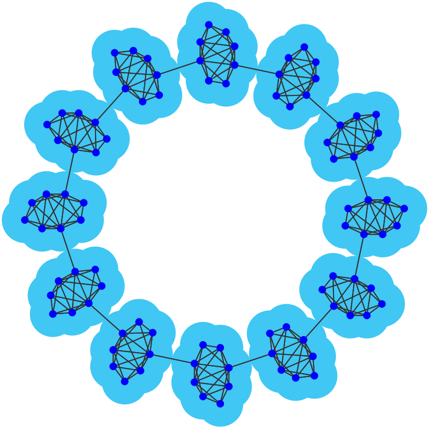


The parameter

Question

What is the right value for the parameter in the Čech construction?

Often, there is no one "right" choice.



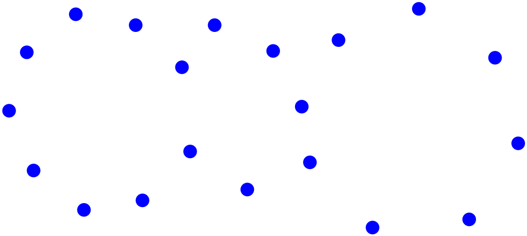
Persistence

Main idea: persistence

Vary the parameter and keep track of when features appear and disappear.

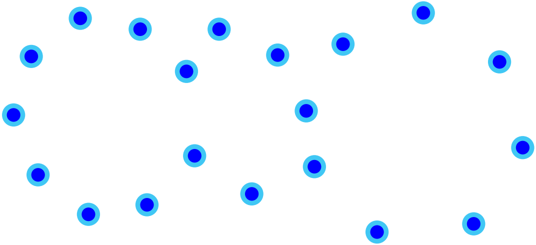
Varying the radii of the spheres in the Čech construction we get an increasing family of simplicial complexes.

Filtered simplicial complex from points in $\mathbb{R}^2$



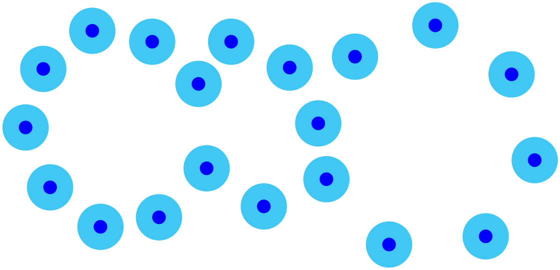
radius $= 0$

Filtered simplicial complex from points in $\mathbb{R}^2$



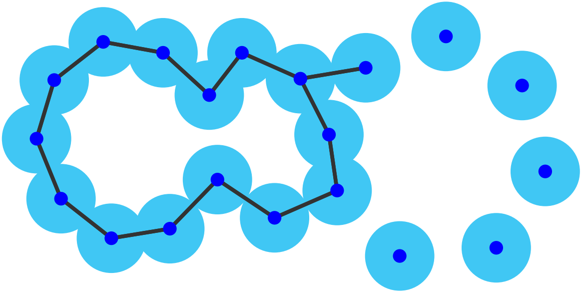
radius = 1

Filtered simplicial complex from points in $\mathbb{R}^2$



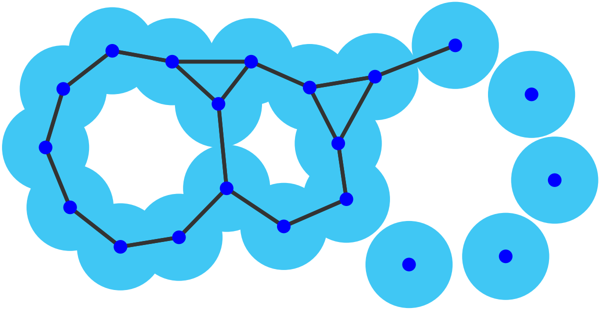
radius $= 2$

Filtered simplicial complex from points in $\mathbb{R}^2$



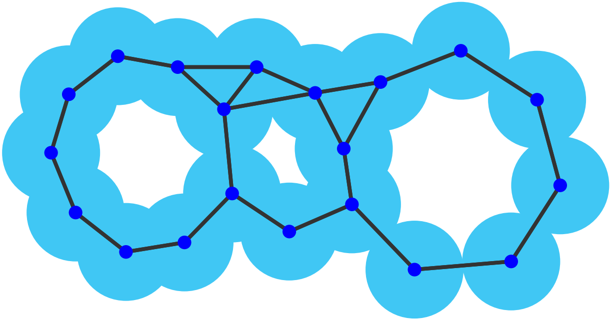
radius = 3

Filtered simplicial complex from points in $\mathbb{R}^2$



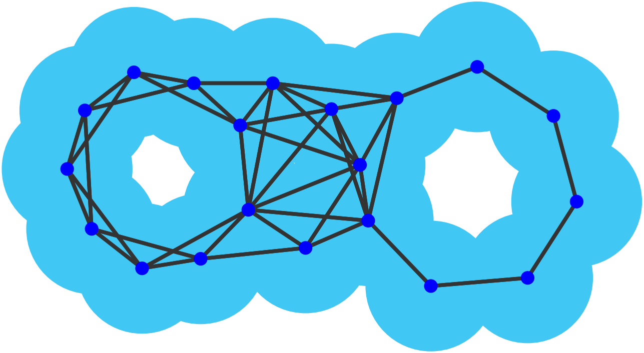
radius = 4

Filtered simplicial complex from points in $\mathbb{R}^2$



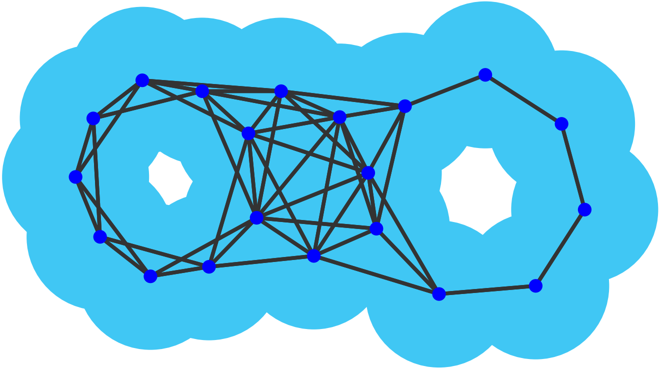
radius = 5

Filtered simplicial complex from points in $\mathbb{R}^2$



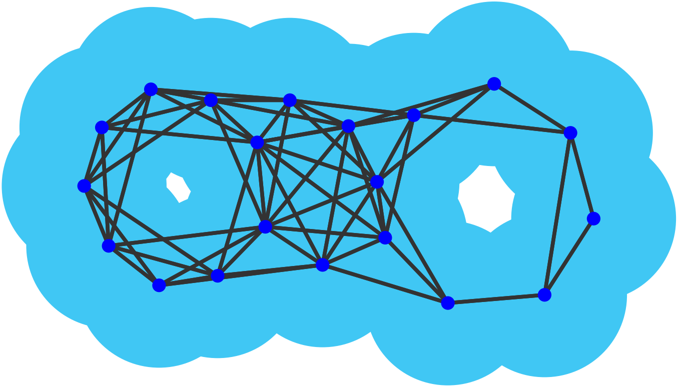
radius = 6

Filtered simplicial complex from points in $\mathbb{R}^2$



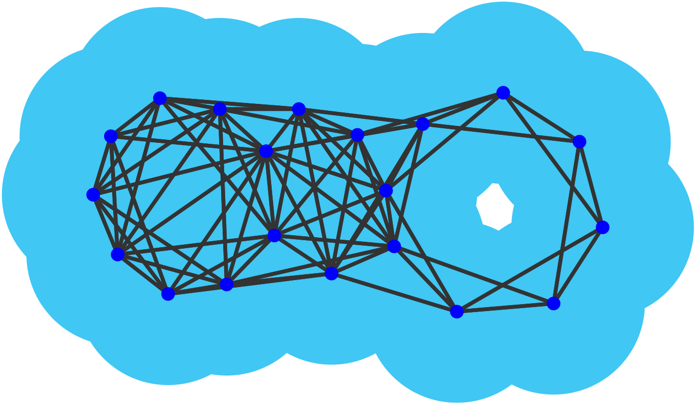
radius = 7

Filtered simplicial complex from points in $\mathbb{R}^2$



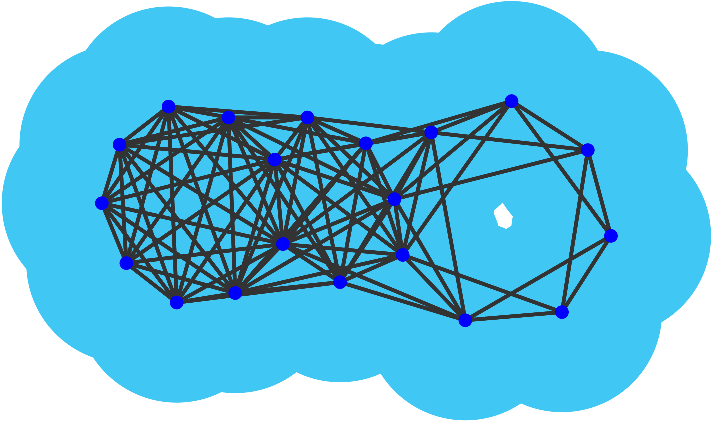
radius = 8

Filtered simplicial complex from points in $\mathbb{R}^2$



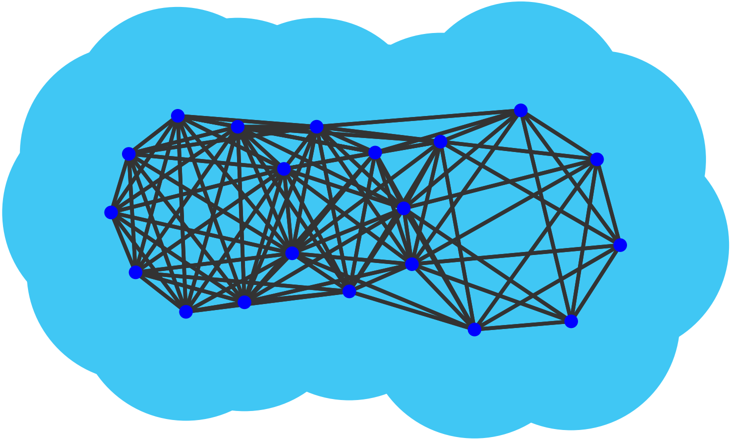
radius = 9

Filtered simplicial complex from points in $\mathbb{R}^2$



radius = 10

Filtered simplicial complex from points in $\mathbb{R}^2$



radius = 11

Mathematical encoding

We have an increasing sequence of simplicial complexes

$$X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots \subseteq X_m$$

called a filtered simplicial complex.

Apply homology.

We get a sequence of vector spaces and linear maps

$$V_0 \to V_1 \to V_2 \to \cdots \to V_m$$

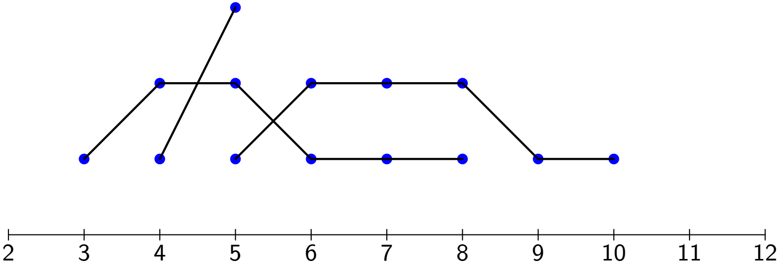
called a persistence module.

Graph of a persistence modules

$$V_0 \to V_1 \to V_2 \to V_3 \to V_4 \to V_5 \to V_6 \to V_7 \to \cdots \to V_m$$

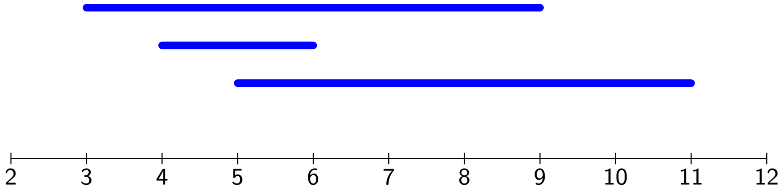
Fundamental Theorem of Persistent Homology

There exists a choice of bases for the vector spaces V_i such that each map is determined by a bipartite matching of basis vectors.

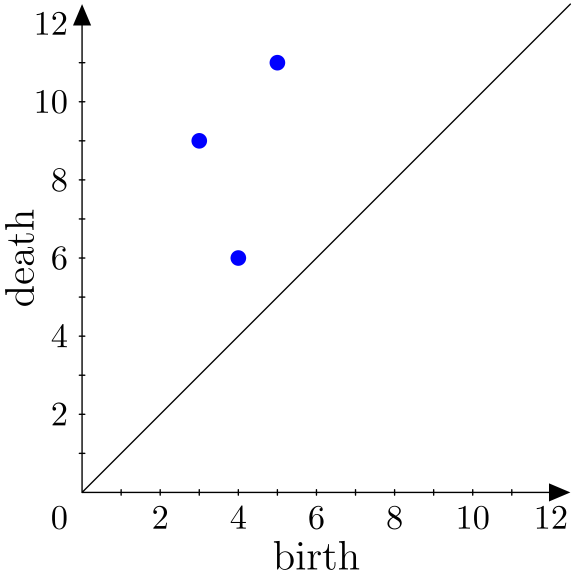


Barcode from our points in $\mathbb{R}^2$

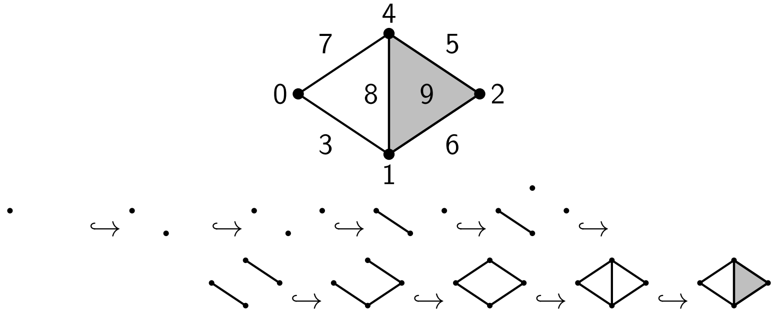
Straightening out the previous graph, we get a barcode.



Persistence diagram from our points in $\mathbb{R}^2$



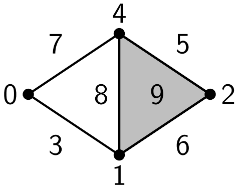
Exercise 5: Barcodes and persistence diagrams

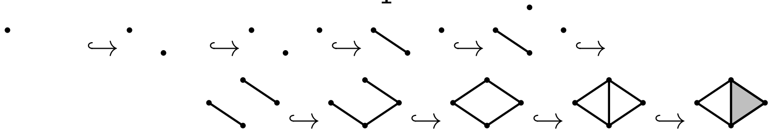


Time	0	1	2	3	4	5	6	7	8	9
Betti number	β_0									
effect	$+$									

Birth–Death pairs for H_0:
Birth–Death pairs for H_1:

Exercise 5: Barcodes and persistence diagrams



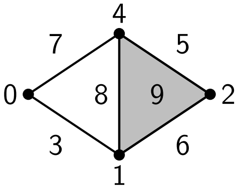


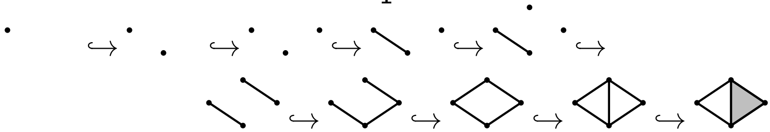
Time	0	1	2	3	4	5	6	7	8	9
Betti number	β_0	β_0	β_0	β_0	β_0	β_0	β_0	β_1	β_1	β_1
effect	$+$	$+$	$+$	$-$	$+$	$-$	$-$	$+$	$+$	$-$

Birth–Death pairs for H_0:

Birth–Death pairs for H_1:

Exercise 5: Barcodes and persistence diagrams



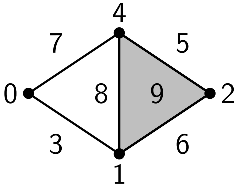


Time	0	1	2	3	4	5	6	7	8	9
Betti number	β_0	β_0	β_0	β_0	β_0	β_0	β_0	β_1	β_1	β_1
effect	$+$	$+$	$+$	$-$	$+$	$-$	$-$	$+$	$+$	$-$

Birth–Death pairs for H_0: $(0,\infty)$, $(1,3)$, $(2,6)$, $(4,5)$

Birth–Death pairs for H_1: $(7,\infty)$, $(8,9)$

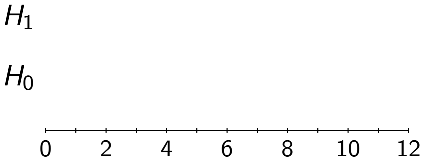
Exercise 5: Barcodes and persistence diagrams



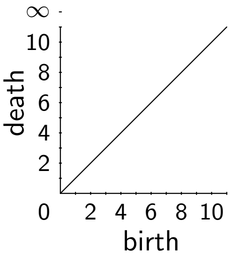
Birth–Death pairs for H_0: $(0, \infty)$, $(1, 3)$, $(2, 6)$, $(4, 5)$

Birth–Death pairs for H_1: $(7, \infty)$, $(8, 9)$

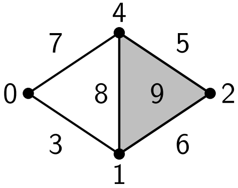
Barcode



Persistence diagram

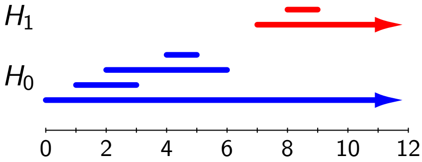


Exercise 5: Barcodes and persistence diagrams

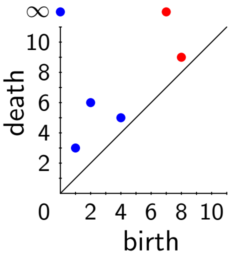


Birth–Death pairs for H_0: $(0, \infty)$, $(1, 3)$, $(2, 6)$, $(4, 5)$
Birth–Death pairs for H_1: $(7, \infty)$, $(8, 9)$

Barcode

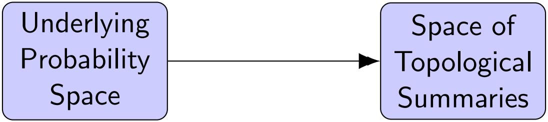


Persistence diagram



Statistical viewpoint

The barcode/persistence diagram is a random variable;
it is a summary statistic.

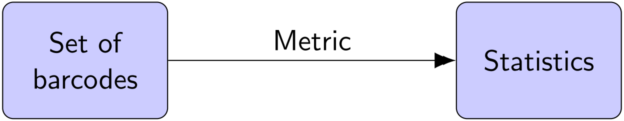


Challenges

For example:

- calculate averages
- understand variances
- test hypotheses
- cluster and classify

Statistics with barcodes/persistence diagrams



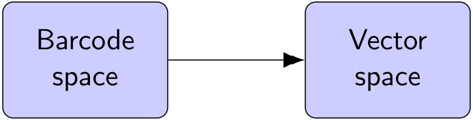
Easy:

- clustering
- certain hypothesis tests

Hard:

- calculating averages
- understanding variances
- classification

Making life easier



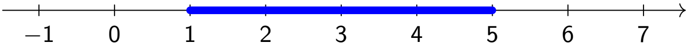
One way to turn a barcode or persistence diagram into a vector is the persistence landscape.

Advantages:

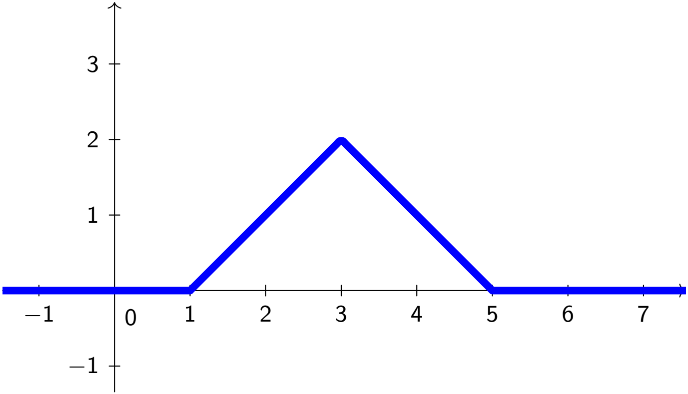
- it does not lose information
- it is stable
- it has a discrete and a continuous version

Persistence landscape from a barcode

Replace

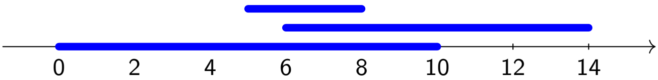


with

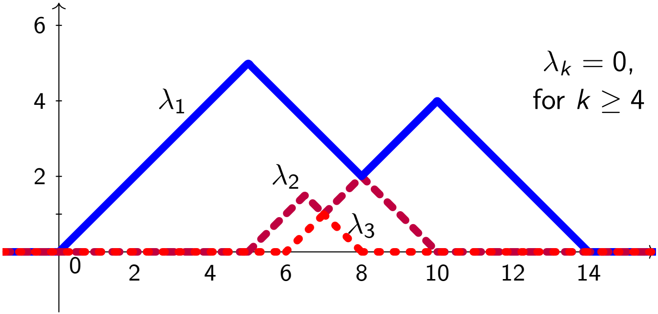


Persistence landscape from a barcode

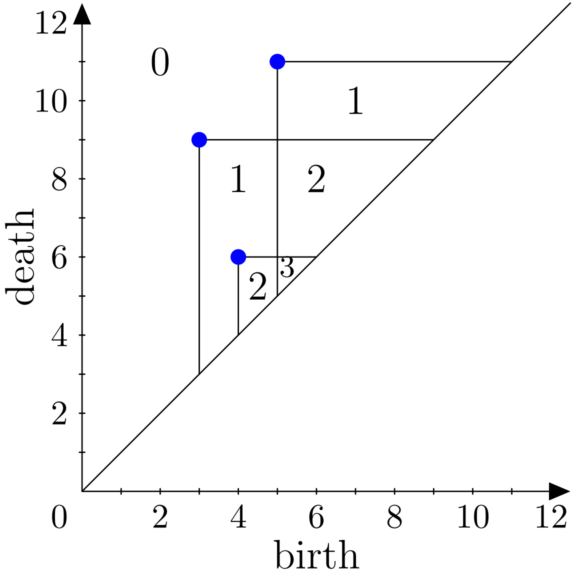
Barcode:



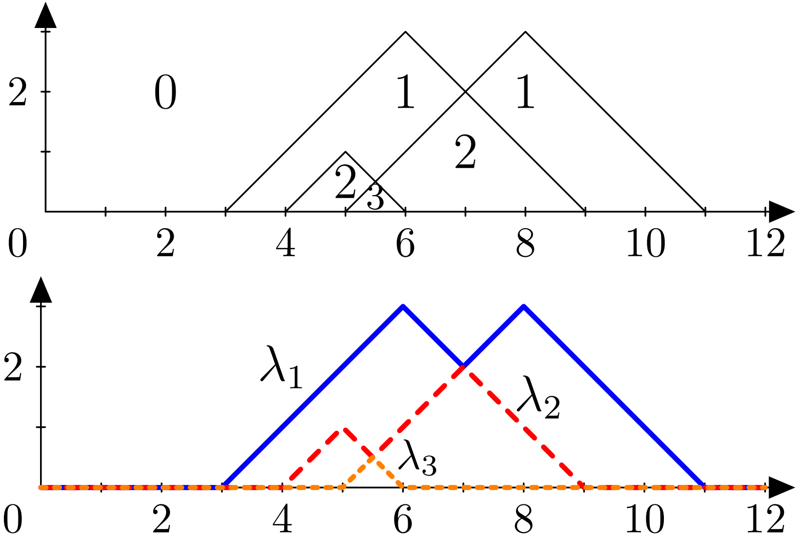
Persistence Landscape:



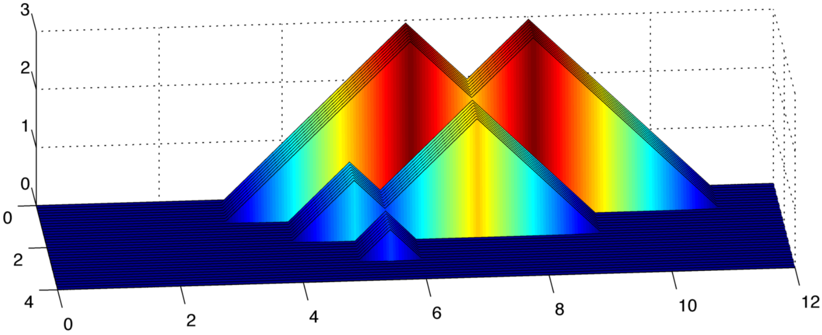
Persistence landscape from a persistence diagram



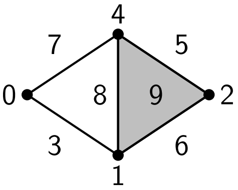
Persistence landscape from a persistence diagram



Persistence landscape from a persistence diagram

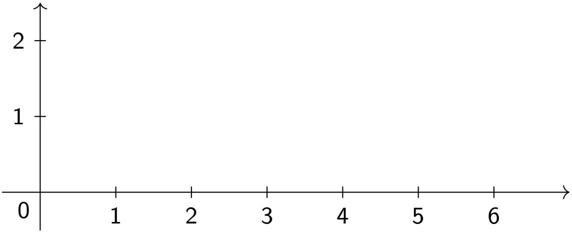


Exercise 6: Graphing the persistence landscape

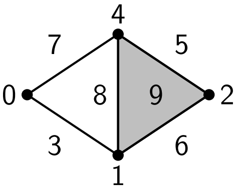


Birth–Death pairs for $\tilde{H}_0$:
$(1, 3),\ (2, 6),\ (4, 5)$

Graph the corresponding persistence landscape.

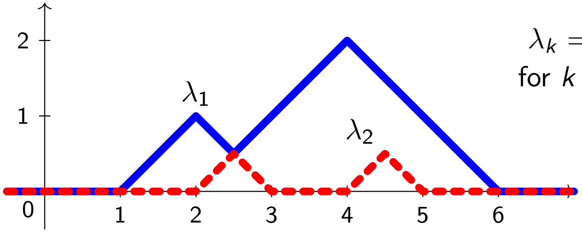


Exercise 6: Graphing the persistence landscape



Birth–Death pairs for $\tilde{H}_0$:
$(1,3),\ (2,6),\ (4,5)$

Graph the corresponding persistence landscape.



$\lambda_k = 0,$
for $k \geq 3$

Making life easier

Choices for the vector space

- continuous version: $L^2(\mathbb{R}^2)$
- discrete version: $\mathbb{R}^n$

Making life easier

Choices for the vector space

- continuous version: $L^2(\mathbb{R}^2)$
- discrete version: $\mathbb{R}^n$

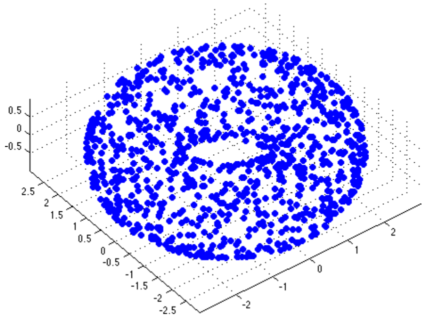
What is great about $\mathbb{R}^n$ and $L^2(\mathbb{R}^2)$?

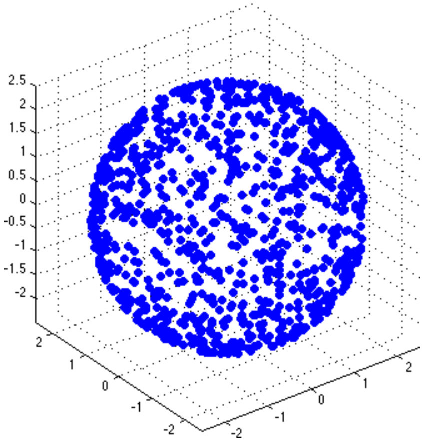
- are vector spaces (easy to measure distances, averages)
- have inner products (easy to measure angles)
- are complete (good for studying convergence)

Thus we can

- apply tools from probability, statistics and machine learning

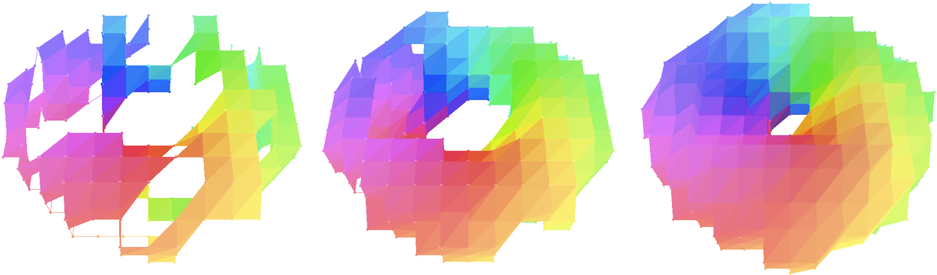
Topological hypothesis testing



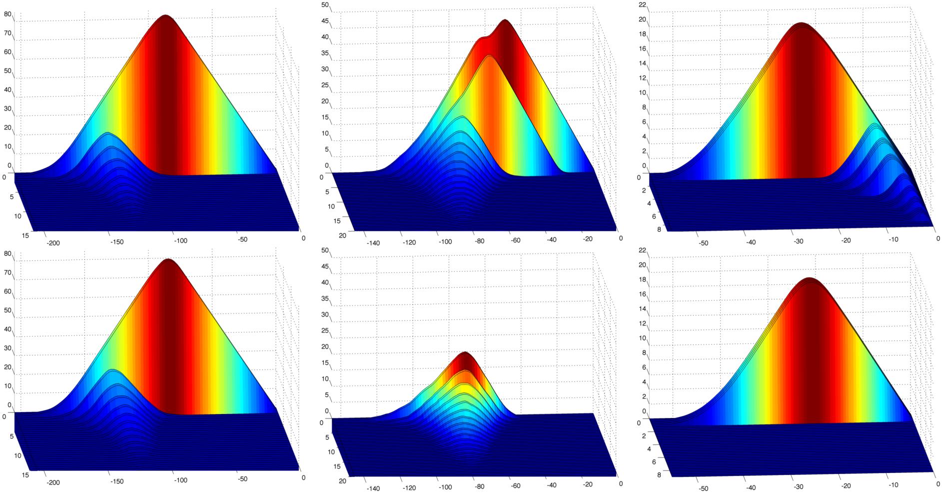


Topological hypothesis testing

Points → kernel density estimator → filtered simplicial complex



Topological hypothesis testing



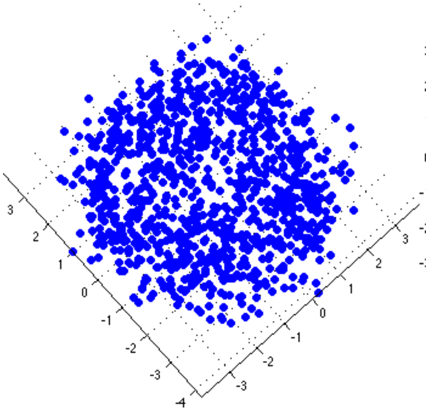
Topological hypothesis testing

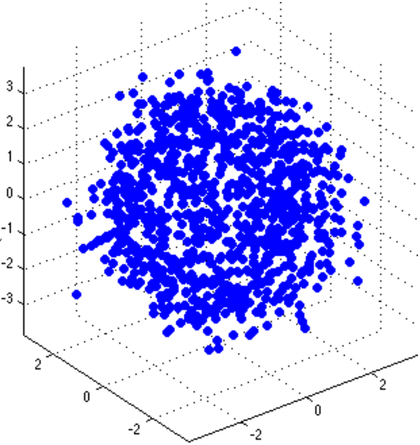
Null hypothesis: $\|\overline{\lambda_S}\|_1 = \|\overline{\lambda_T}\|_1$.

two-sample z-test:

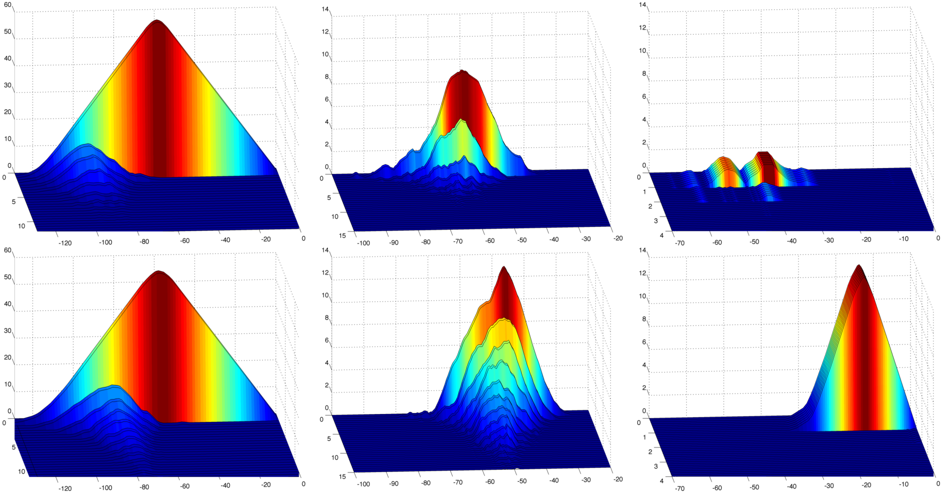
degree	decision	p value
0	cannot reject	
1	reject	3×10^{-6}
2	cannot reject	

Topological hypothesis testing, noisy





Topological hypothesis testing, noisy



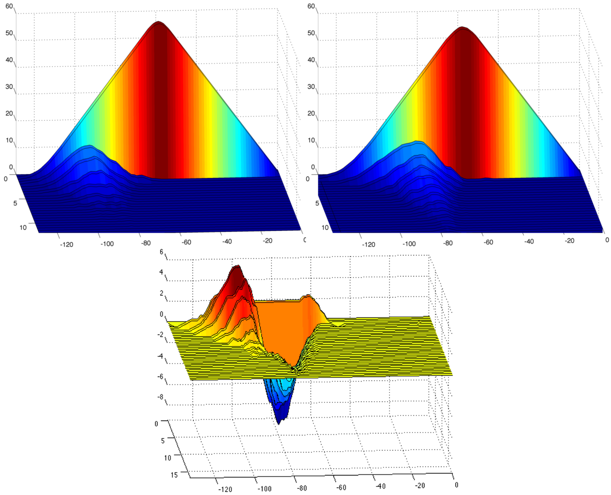
Topological hypothesis testing, noisy

Null hypothesis: $\|\overline{\lambda_S} - \overline{\lambda_T}\|_2 = 0$.

Permutation test:

dim	decision	p value
0	reject	0.0111
1	reject	0.0000
2	reject	0.0000

Topological hypothesis testing, noisy



Software

Persistent Homology:

- CHOMP, Dionysus, DIPHA, Eirene, GUDHI, JavaPlex, Perseus, PHAT, Ripser, SimBa, SimPers

Persistence Landscape:

- The Persistence Landscape Toolbox

Topological Data Analysis:

- the R package TDA
- my R code

Stability

Given $f : X \to \mathbb{R}$,
let $\lambda(f)$ the persistence landscape of sublevel sets of f.

Landscape Stability Theorem (B)

Let $f, g : X \to \mathbb{R}$.

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

If X is nice and f and g are tame and Lipschitz then

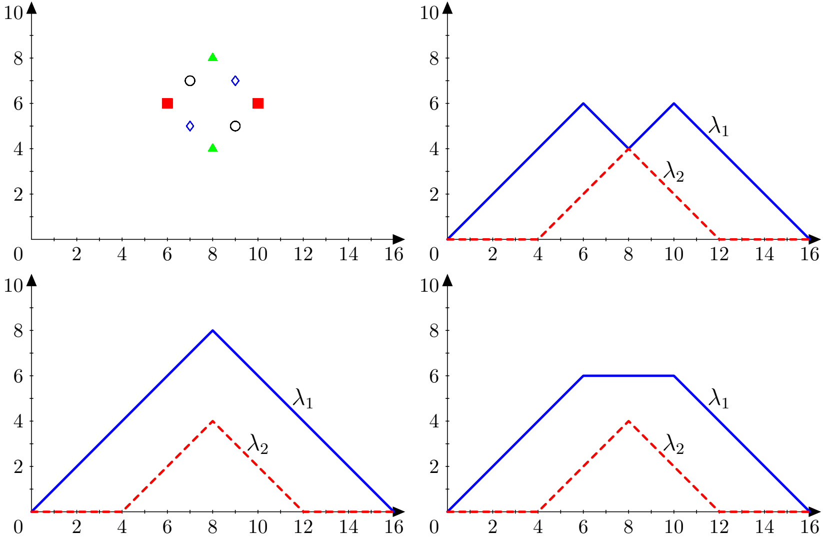
$$\|\lambda(f) - \lambda(g)\|_2^2 \leq C\|f - g\|_\infty^{2-k}.$$

Average landscapes

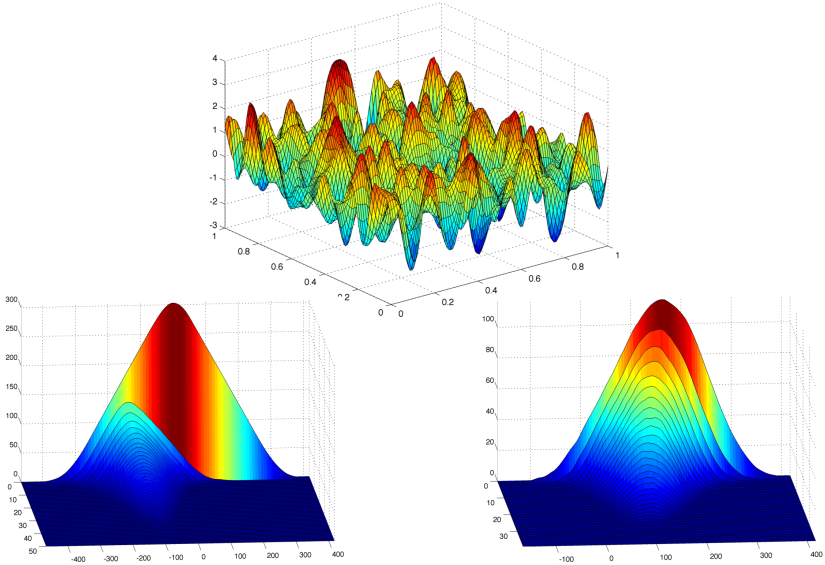
Persistence landscapes, $\lambda^{(1)}, \dots, \lambda^{(n)}$, have a pointwise average,

$$\overline{\lambda}(k,t) = \frac{1}{n}\sum_{i=1}^{n} \lambda^{(i)}(k,t)$$

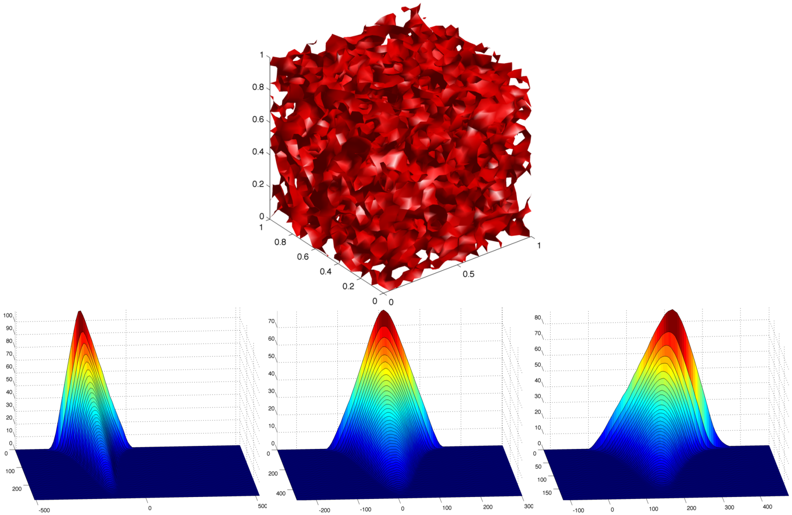
Average diagram vs average landscape



Average landscapes for Gaussian random fields



Average landscapes for Gaussian random fields



Asymptotics for persistence landscapes

λ is a random variable in $L^2(\mathbb{R}^2)$, $\|\lambda\|$ is a real random variable.

If $E\|\lambda\| < \infty$ then there exists $E(\lambda) \in L^2(\mathbb{R}^2)$ such that $E(f(\lambda)) = f(E(\lambda))$ for all continuous linear functionals f.

Strong Law of Large Numbers (B, 2015)

$\overline{\lambda}^{(n)} \to E(\lambda)$ almost surely

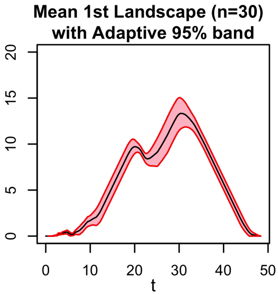
Central Limit Theorem (B, 2015)

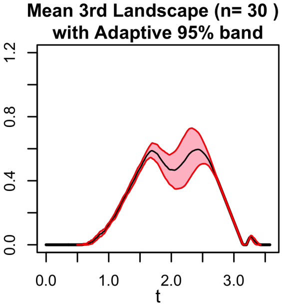
$\sqrt{n}[\overline{\lambda}^{(n)} - E(\lambda)]$ converges weakly to a Gaussian random variable

Understanding variance

Two approaches:

- Bootstrap and confidence intervals for persistence landscapes [Chazal, Fasy, Lecci, Rinaldo, Singh, Wasserman]





- Principal component analysis (coming in Talk 2)