Topological data analysis
Lecture 4

Anton Ayzenberg

ATA Lab, FCS NRU HSE
Noeon Research

Spring 2024
Faculty of Computer Science / Yandex Data School

A principal incompatibility between “topology” and “applied”.

- Data analysis and machine learning deal with real numbers and real optimization.
- Topological invariants are discrete. There is no space with 2.3457 many connected components or $\frac{5}{6}$ many holes.
- How can one make Topology “applied”?

A principal incompatibility between "topology" and "applied".

- Data analysis and machine learning deal with real numbers and real optimization.
- Topological invariants are discrete. There is no space with 2.3457 many connected components or $\frac{5}{6}$ many holes.
- How can one make Topology "applied"?

Introduce "topological processes"!

Let X_t be a space depending on time $t \in \mathbb{R}$. If $t_1 \leqslant t_2$, we assume there is a map

$$f_{t_1 \leqslant t_2} \colon X_{t_1} \to X_{t_2},$$

such that $f_{t \leqslant t} = \mathrm{id}_{X_t}$ and $f_{t_2 \leqslant t_3} \circ f_{t_1 \leqslant t_2} = f_{t_1 \leqslant t_3}$.

Compare this with stochastic processes...

Idea of applied topology

Topological process

Let X_t be a space depending on time $t \in \mathbb{R}$ and there are maps $f_{t_1 \leqslant t_2}\colon X_{t_1} \to X_{t_2}$.

Usually, all connecting maps $f_{t_1 \leqslant t_2}$ are inclusions. In this case the process is called a **filtration**.

Idea

- We may average topological invariants along all values of time t.
- This gives real-valued invariants which can be optimized using methods of machine learning.

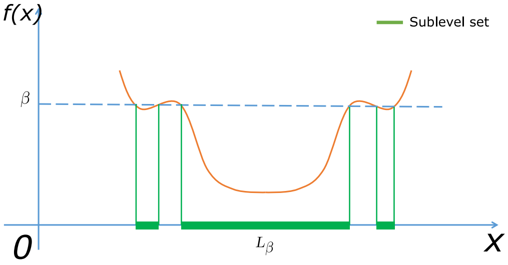
Important construction

Sublevel set filtration

Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a function. Consider sublevel sets of f

$$X_t^f = \{x \in \mathbb{R}^d \mid f(x) \leqslant t\}$$

This is a filtration.



Another important construction

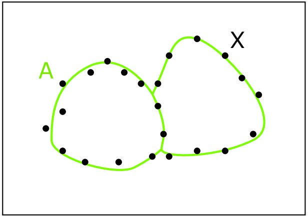
Čech filtration

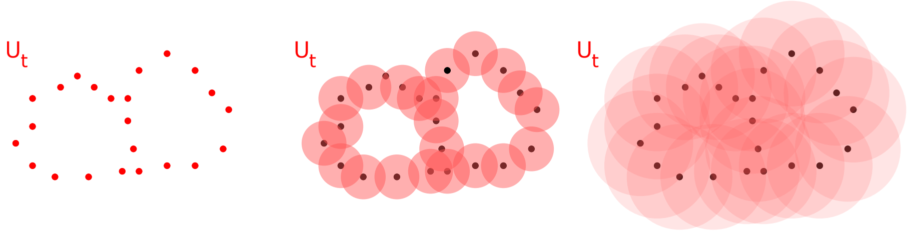
Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ be a finite set (point cloud). Itself, the space X is not interesting topologically. But we may surround each point with a ball of variable radius $t/2$, and see how topology evolve:

$$X_t = \bigcup_{i=1}^{m} B_{t/2}(x_i)$$

This is a filtration defined for $t \geqslant 0$.

Čech filtration





$t \approx 0$ t is nice $t >> 0$

Toy example: average number of components

Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ be a point cloud and X_t its Čech filtration. Let $\mathrm{nc}(X_t)$ be the number of connected components of X_t.

A new invariant

Define the number

$$\overline{\mathrm{nc}}(X) = \int_0^{+\infty} (\mathrm{nc}(X_t) - 1)dt.$$

Toy example: average number of components

Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ be a point cloud and X_t its Čech filtration. Let $\mathrm{nc}(X_t)$ be the number of connected components of X_t.

A new invariant

Define the number

$$\overline{\mathrm{nc}}(X) = \int_0^{+\infty} (\mathrm{nc}(X_t) - 1)dt.$$

Question: any guess what $\overline{\mathrm{nc}}(X)$ is?

Demonstration: press to play in browser

Toy example: evolution of components

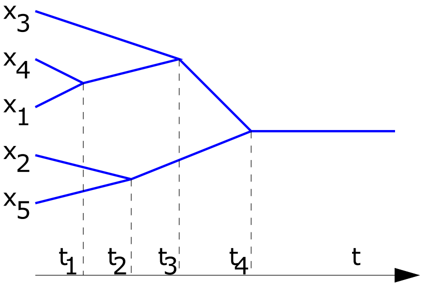
Answer:

$\overline{\mathrm{nc}}(X)$ equals the length of the minimal spanning tree of X. Guess why.

Toy example: evolution of components

Answer:

$\overline{\mathrm{nc}}(X)$ equals the length of the minimal spanning tree of X. Guess why.



Dendrogram

Open question: how can we encode such dendrograms?

Persistent homology

Homology

Homology = higher dimensional analogue of counting connected components.
$\beta_i(X)$ = number of i-dimensional holes in X.

Persistent homology

How the number of holes in a filtration changes in time.

Filtrations with discrete time

Filtration

A chain of simplicial complexes

$$K_0 \subset K_1 \subset K_2 \subset \cdots \subset K_m = K$$

is called **a filtration**.

For each j, it induces the chain of linear maps

$$H_j(K_0) \to H_j(K_1) \to H_j(K_2) \to \cdots \to H_j(K_m)$$

of $\Bbbk$-vector spaces.

Persistence modules

Definition

A persistence module is a chain of finite dimensional $\Bbbk$-vector spaces and linear maps

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

If, for some m, $V_m = V_{m+1} = \cdots$, we say that persistence module **stabilizes**.

Main example: j-th homology of a filtration is a stabilizing persistence module. It is called **the persistence homology module** of a filtration:

$$H_j(K_0) \to H_j(K_1) \to H_j(K_2) \to \cdots \to H_j(K_m) \xrightarrow{=} \cdots$$

Persistence modules

Definition

A persistence module is a chain of finite dimensional $\Bbbk$-vector spaces and linear maps

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

If, for some m, $V_m = V_{m+1} = \cdots$, we say that persistence module **stabilizes**.

Main example: j-th homology of a filtration is a stabilizing persistence module. It is called **the persistence homology module** of a filtration:

$$H_j(K_0) \to H_j(K_1) \to H_j(K_2) \to \cdots \to H_j(K_m) \xrightarrow{=} \cdots$$

Exercise: prove that persistence module is the synonym for "graded module over the polynomial ring $\Bbbk[x]$" if you understand this phrase.

Interval modules and the structure theorem

Example: **An interval module** $I_{[b;d)}$ is the following module

$$0 \to \cdots \to 0 \to \underset{b}{\Bbbk} \xrightarrow{=} \cdots \xrightarrow{=} \Bbbk \to \underset{d}{0} \to 0 \to \cdots$$

where $b \in \mathbb{Z}_+$ is called the birth-time of a module and $d \in \mathbb{Z}_+ \sqcup \{+\infty\}$ is the death-time.

Interval modules and the structure theorem

Example: **An interval module** $I_{[b;d)}$ is the following module

$$0 \to \cdots \to 0 \to \underset{b}{\Bbbk} \xrightarrow{=} \cdots \xrightarrow{=} \Bbbk \to \underset{d}{0} \to 0 \to \cdots$$

where $b \in \mathbb{Z}_+$ is called the birth-time of a module and $d \in \mathbb{Z}_+ \sqcup \{+\infty\}$ is the death-time.

Main Structural Theorem (about persistence modules)

Every stabilizing persistence module is isomorphic to a direct sum of interval modules. The summands are determined uniquely up to permutation.

Interval modules and the structure theorem

Example: **An interval module** $I_{[b;d)}$ is the following module

$$0 \to \cdots \to 0 \to \underset{b}{\Bbbk} \xrightarrow{=} \cdots \xrightarrow{=} \Bbbk \to \underset{d}{0} \to 0 \to \cdots$$

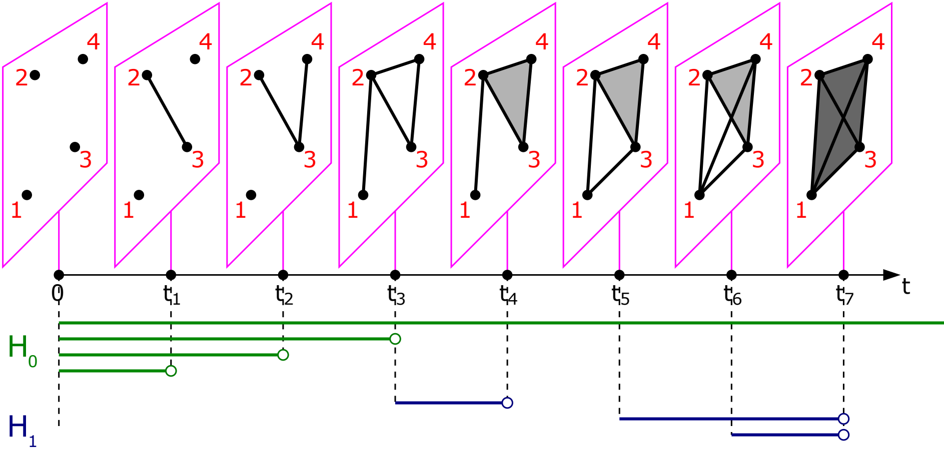
where $b \in \mathbb{Z}_+$ is called the birth-time of a module and $d \in \mathbb{Z}_+ \sqcup \{+\infty\}$ is the death-time.

Main Structural Theorem (about persistence modules)

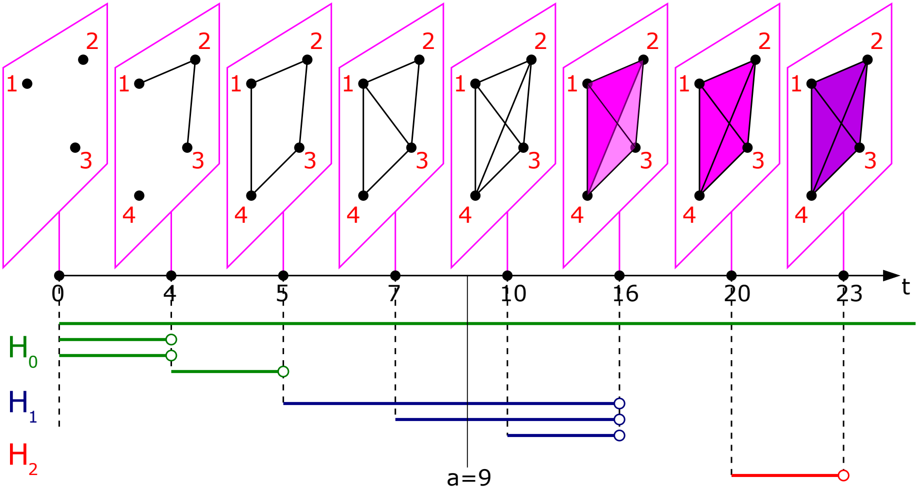
Every stabilizing persistence module is isomorphic to a direct sum of interval modules. The summands are determined uniquely up to permutation.

Remark: This is actually an instance of the classification theorem for finitely generated modules over PID (the ring $\Bbbk[x]$ is a principal ideal domain).

Persistence homology decomposed into interval summands



Persistence homology decomposed into interval summands



Persistent homology of a filtration

Filtration:

$$K_0 \subset K_1 \subset K_2 \subset \cdots \subset K_m = K$$

j-th persistent homology module:

$$H_j(K_0) \to H_j(K_1) \to H_j(K_2) \to \cdots \to H_j(K_m).$$

Persistent homology of a filtration

Filtration:

$$K_0 \subset K_1 \subset K_2 \subset \cdots \subset K_m = K$$

j-th persistent homology module:

$$H_j(K_0) \to H_j(K_1) \to H_j(K_2) \to \cdots \to H_j(K_m).$$

Main question

Should we compute all homology $H_j(K_i)$ separately, and then merge them to get interval decomposition?

Persistent homology of a filtration

Filtration:

$$K_0 \subset K_1 \subset K_2 \subset \cdots \subset K_m = K$$

j-th persistent homology module:

$$H_j(K_0) \to H_j(K_1) \to H_j(K_2) \to \cdots \to H_j(K_m).$$

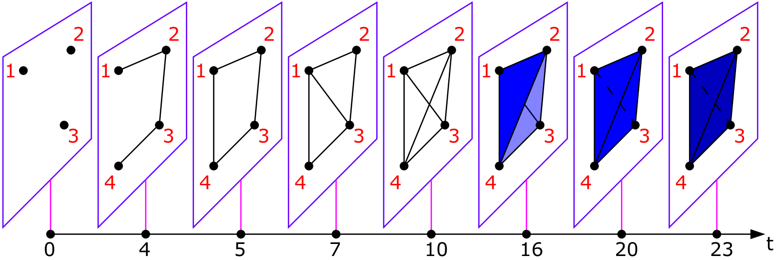
Main question

Should we compute all homology $H_j(K_i)$ separately, and then merge them to get interval decomposition?

Luckily, no! We need to store our filtration in an optimal form.

How to treat filtrations

Instead of $K_0 \subset K_1 \subset K_2 \subset \cdots \subset K_m = K$ let us store the list of all simplices of K together with their birth times.



{1}:0, {2}:0, {3}:0, {4}:4

{1,2}:4, {1,3}:7, {1,4}:5, {2,3}:4, {2,4}:10, {3,4}:4

{1,2,3}:16, {1,2,4}:16, {1,3,4}:16, {2,3,4}:20

{1,2,3,4}:23

How to treat filtrations

We have two lists: **BirthTimes** and **Simplices**. We assume they satisfy the following:

- Their indices agree: BirthTimes[i] is the time of appearance of Simplices[i] in the filtration.

How to treat filtrations

We have two lists: **BirthTimes** and **Simplices**. We assume they satisfy the following:

- Their indices agree: BirthTimes[i] is the time of appearance of Simplices[i] in the filtration.
- BirthTimes is sorted.

How to treat filtrations

We have two lists: **BirthTimes** and **Simplices**. We assume they satisfy the following:

- Their indices agree: BirthTimes[i] is the time of appearance of Simplices[i] in the filtration.
- BirthTimes is sorted.
- For each i all subsets of Simplices[i] have indices $< i$.

How to treat filtrations

We have two lists: **BirthTimes** and **Simplices**. We assume they satisfy the following:

- Their indices agree: BirthTimes[i] is the time of appearance of Simplices[i] in the filtration.
- BirthTimes is sorted.
- For each i all subsets of Simplices[i] have indices $< i$.

Exercise: prove that BirthTimes and Simplices can be simultaneously sorted this way.

How to treat filtrations

We have two lists: **BirthTimes** and **Simplices**. We assume they satisfy the following:

- Their indices agree: BirthTimes[i] is the time of appearance of Simplices[i] in the filtration.
- BirthTimes is sorted.
- For each i all subsets of Simplices[i] have indices $< i$.

Exercise: prove that BirthTimes and Simplices can be simultaneously sorted this way.

Last condition assures that $K^i = \{\text{Simplices[j]} \mid j \leqslant i\}$ is always a simplicial complex.

How to treat filtrations

Last condition assures that $K^i = \{\text{Simplices[j]} \mid j \leqslant i\}$ is always a simplicial complex. We get new filtration

$$K^0 \subset K^1 \subset K^2 \subset \cdots \subset K^N$$

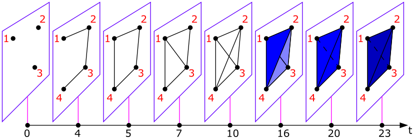
where N is the total number of simplices.

What is good

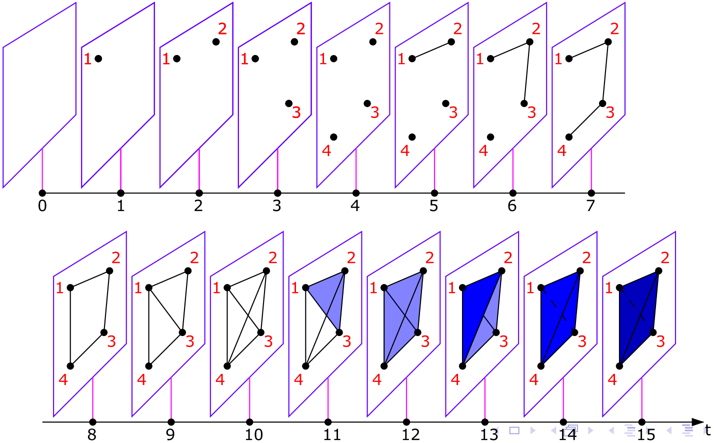
At each step of this new filtration, exactly one simplex is added. Namely Simplices[i] is added at i-th step.

How to treat filtrations

Old filtration



New filtration



What happens with homology at each step

Proposition

Assume that $L \subset K$ and $K \backslash L$ is a single j-dim simplex. Then we have an alternative:

- $(j-1)$-th Betti number reduces by 1.
- j-th Betti number increases by 1.

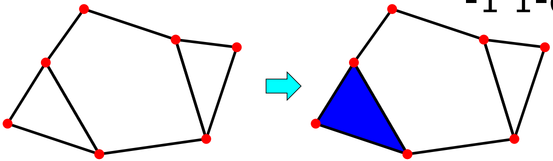
Other Betti numbers do not change.

Adding a j-simplex, we either seal up a $(j-1)$-hole, or create a j-hole.

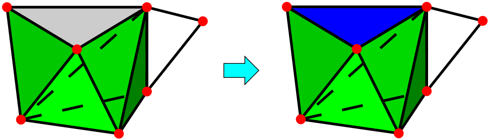
Exercise: prove it.

Adding a 2-simplex

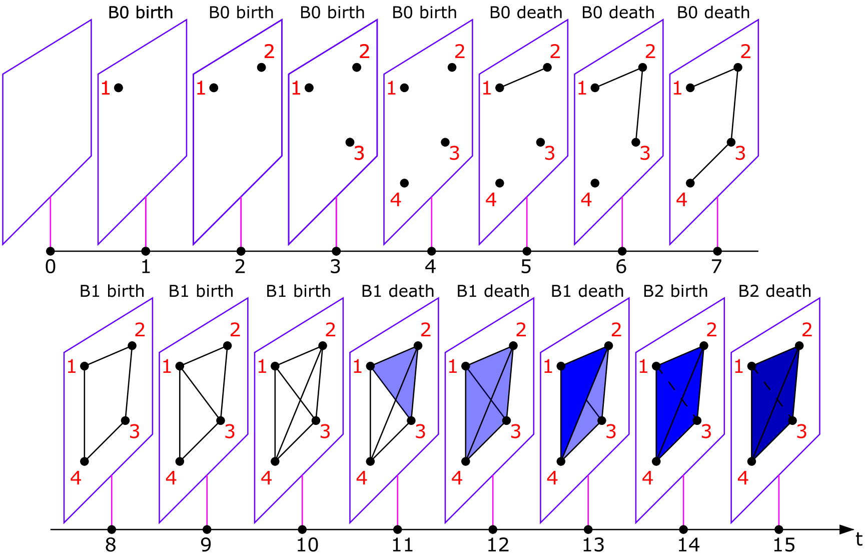
-1 1-dim hole



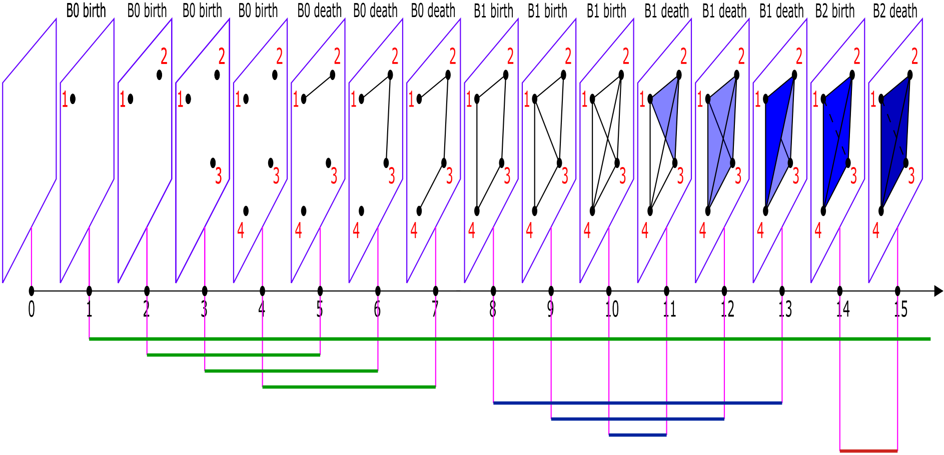
+1 2-dim hole



Our detailed filtration



Our detailed filtration



Sources

- S. Barannikov, *Framed Morse complex and its invariants*, Advances in Soviet Mathematics. Vol.21 (1994), pp. 93–115.
- H. Y. Cheung, T. C. Kwok, L. C. Lau, *Fast matrix rank algorithms and applications*, J. ACM 60:5 (2013), Article 31.
- H. Edelsbrunner, J. L. Harer, Computational Topology: An Introduction, 2010.
- Morozov, Dionysus2 library https://mrzv.org/software/dionysus2/
- A. J. Zomorodian, Topology for computing, 2005.