The Topological Data Analysis Pipeline

Elise Askelsen

University of Iowa
Department of Mathematics

Central College Heartland Talk
November 15, 2023

Table of contents:

My Pipeline

My Pipeline

My Pipeline

My Pipeline

Introduction

Large amounts of data have created a need for new types of analysis, leading to the development of Topological Data Analysis, TDA.

Introduction

Large amounts of data have created a need for new types of analysis, leading to the development of Topological Data Analysis, TDA.

Topological Data Analysis Pipeline:

Data $\rightarrow$ Geometry $\rightarrow$ Algebra $\rightarrow$ Summary $\rightarrow$ Analysis

Data $\rightarrow$ Geometry:

Given a set of data, we build a simplicial complex.

Definition

An **abstract simplicial complex** is a finite collection A of finite non-empty sets such that if $\alpha \in A$, then so is every subset of α.

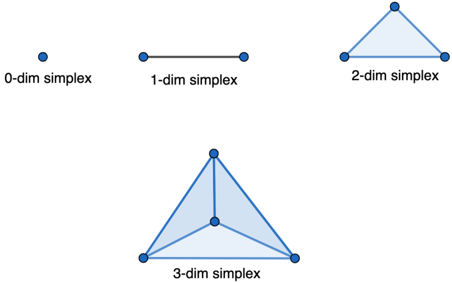
Data $\rightarrow$ Geometry:

Given a set of data, we build a simplicial complex.

Definition

An **abstract simplicial complex** is a finite collection A of finite non-empty sets such that if $\alpha \in A$, then so is every subset of α.

Practically, examples include sets of simplicies include



Data $\rightarrow$ Geometry:

Given a set of data, we can build a simplicial complex in the following way;

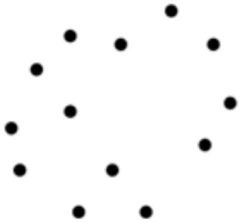


Figure: Sampling of Data

Definition

Fix $r > 0$ and a point set, $P = \{p_1, ..., p_n\} \subset \mathbb{R}^n$. Then the **Vietoris Rips Complex** is $VRips_r = \{\sigma \subset P | max_{p_i, p_j \in \sigma} ||p_i - p_j|| \leq 2r\}$.

Data $\rightarrow$ Geometry:

Given a set of data, we can build a simplicial complex in the following way;

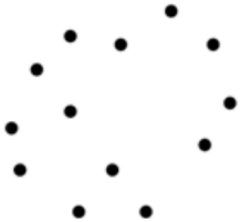


Figure: Sampling of Data

Definition

Fix $r > 0$ and a point set, $P = \{p_1, ..., p_n\} \subset \mathbb{R}^n$. Then the **Vietoris Rips Complex** is $VRips_r = \{\sigma \subset P | max_{p_i, p_j \in \sigma} ||p_i - p_j|| \leq 2r\}$.

There are many different types of complexes that are used in TDA.

Data → Geometry:

In building the Vietoris Rips Complex for our data for increasing radii, we obtain a filtered simplicial complex, namely

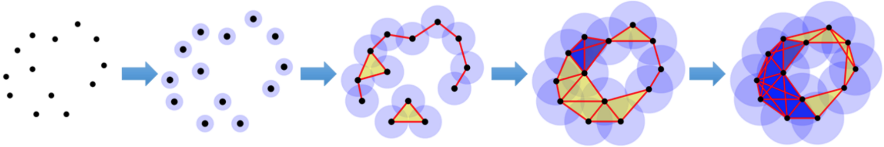


Figure: Building the Vietoris Rips Filtration

Geometry $\to$ Algebra

Now, to translate from geometry and to algebra, we need to learn a little about homology.

In an intuitive sense...
the k^{th} homology group of a simplicial complex X, $H_k(X)$, describes the number of holes in X with a k-dimensional boundary.
A 0-dimensional boundary hole is simply a gap between two components.

Geometry $\rightarrow$ Algebra

Often we use the *Betti Numbers*.

Definition

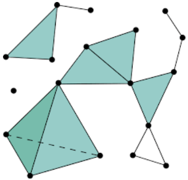
The k^{th} **Betti Number** of a topoplogical space, X, is defined as $\beta_k(X) = rank(H_k(X))$.

Geometry → Algebra

Often we use the *Betti Numbers*.

Definition

The k^{th} **Betti Number** of a topoplogical space, X, is defined as $\beta_k(X) = rank(H_k(X))$.

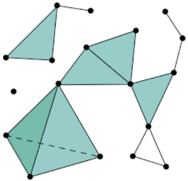


Geometry $\rightarrow$ Algebra

Often we use the *Betti Numbers*.

Definition

The k^{th} **Betti Number** of a topoplogical space, X, is defined as $\beta_k(X) = rank(H_k(X))$.

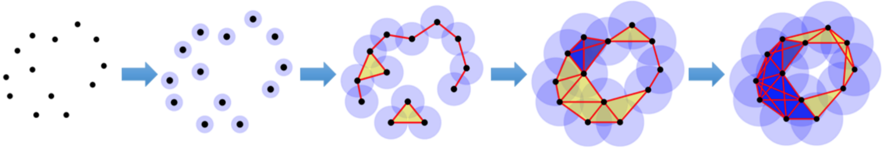


$\beta_0(X)$ =dim(H_0)=3
$\beta_1(X)$ =dim$(H_1) = 1$

Geometry → Algebra:

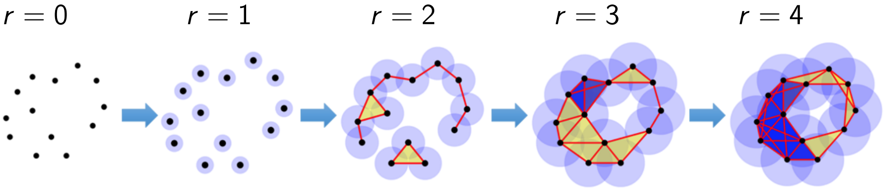
Next we return to our example of the filtered simplicial complex and compute it's Betti Numbers at each index.

$r = 0$ $r = 1$ $r = 2$ $r = 3$ $r = 4$



Geometry → Algebra:

Next we return to our example of the filtered simplicial complex and compute it's Betti Numbers at each index.

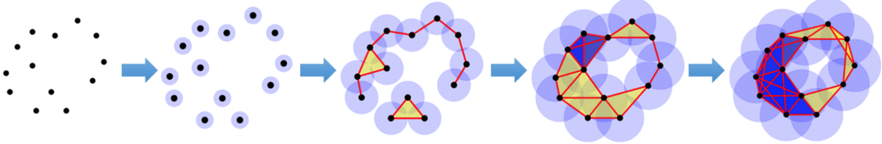


$\beta_0 = 13$
$\beta_1 = 0$

Geometry → Algebra:

Next we return to our example of the filtered simplicial complex and compute it's Betti Numbers at each index.

$r = 0$ $r = 1$ $r = 2$ $r = 3$ $r = 4$



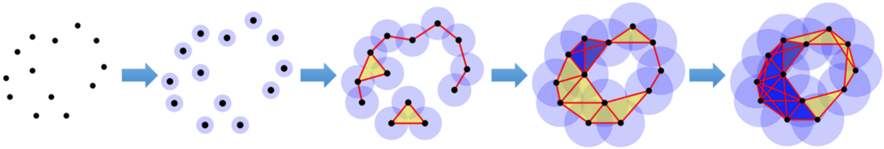
$r = 0$: $\beta_0 = 13$, $\beta_1 = 0$

$r = 1$: $\beta_0 = 13$, $\beta_1 = 0$

Geometry → Algebra:

Next we return to our example of the filtered simplicial complex and compute it's Betti Numbers at each index.

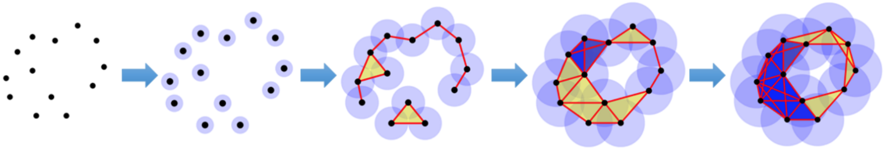
$r = 0$ | $r = 1$ | $r = 2$ | $r = 3$ | $r = 4$



$\beta_0 = 13$, $\beta_1 = 0$ | $\beta_0 = 13$, $\beta_1 = 0$ | $\beta_0 = 2$, $\beta_1 = 0$

Geometry → Algebra:

Next we return to our example of the filtered simplicial complex and compute it's Betti Numbers at each index.

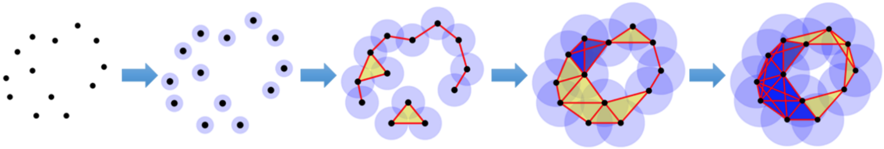


$r = 0$	$r = 1$	$r = 2$	$r = 3$	$r = 4$
$\beta_0 = 13$	$\beta_0 = 13$	$\beta_0 = 2$	$\beta_0 = 1$	
$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 1$	

Geometry → Algebra:

Next we return to our example of the filtered simplicial complex and compute it's Betti Numbers at each index.

$r = 0$	$r = 1$	$r = 2$	$r = 3$	$r = 4$

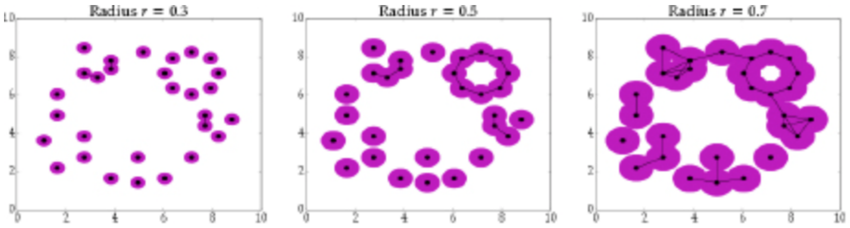


$r = 0$	$r = 1$	$r = 2$	$r = 3$	$r = 4$
$\beta_0 = 13$	$\beta_0 = 13$	$\beta_0 = 2$	$\beta_0 = 1$	$\beta_0 = 1$
$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 0$	$\beta_1 = 1$	$\beta_1 = 1$

Algebra → Summary

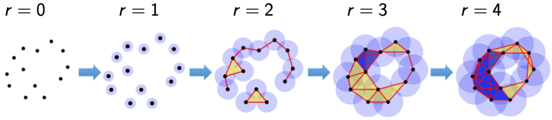
With the persistent homology now computed, we summarize our data in a *barcode* by tracking how long features persist.

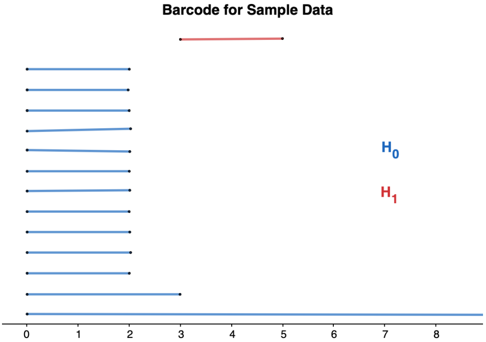
We do this with intervals of the form $[\mathrm{birth}, \mathrm{death})$ for each feature.



Algebra → Summary

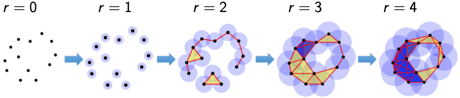
The *barcode* for our example is given by the following.

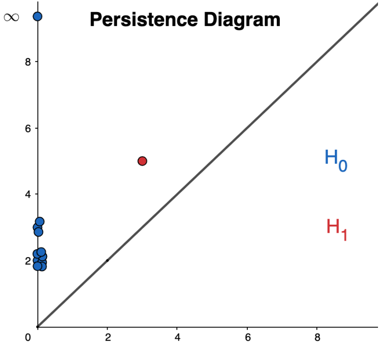




Algebra → Summary

We can also summarize our findings in a *persistent diagram*.





To graph the persistence diagram, we plot information about each feature in the form of points, $(\mathrm{birth}, \mathrm{death})$.

Algebra $\rightarrow$ Summary

Theoretical Diversion: We can view the barcode as a module.

Definition

For an interval, $[a, b)$, we define the **interval module**, $I^{[a,b)}$, to be the following for all i, x, y.

$$I_i^{[a,b)} = \begin{cases} \mathbb{R} & i \in [a, b) \\ 0 & \text{otherwise} \end{cases} \qquad I_{x,y}^{[a,b)} = \begin{cases} id & x \leq y \in [a, b) \\ 0 & \text{otherwise} \end{cases}.$$

The collection of interval modules is a **persistence module**. [Bot]

Algebra → Summary

Theoretical Diversion: We can view the barcode as a module.

Definition

For an interval, $[a, b)$, we define the **interval module**, $I^{[a,b)}$, to be the following for all i, x, y.

$$I_i^{[a,b)} = \begin{cases} \mathbb{R} & i \in [a,b) \\ 0 & \text{otherwise} \end{cases} \qquad I_{x,y}^{[a,b)} = \begin{cases} id & x \leq y \in [a,b) \\ 0 & \text{otherwise} \end{cases}.$$

The collection of interval modules is a **persistence module**. [Bot]

Intuitively, we are assigning $\mathbb{R}$ to each index in the interval. Maps are induced between each copy of $\mathbb{R}$.

Algebra $\rightarrow$ Summary

<u>Theoretical Diversion:</u> We can view the barcode as a module.

Definition

For an interval, $[a, b)$, we define the **interval module**, $I^{[a,b)}$, to be the following for all i, x, y.

$$I_i^{[a,b)} = \begin{cases} \mathbb{R} & i \in [a,b) \\ 0 & \text{otherwise} \end{cases} \qquad I_{x,y}^{[a,b)} = \begin{cases} id & x \leq y \in [a,b) \\ 0 & \text{otherwise} \end{cases}.$$

The collection of interval modules is a **persistence module**. [Bot]

Intuitively, we are assigning $\mathbb{R}$ to each index in the interval. Maps are induced between each copy of $\mathbb{R}$.
For example, in the discrete case,

$$\cdots \longrightarrow 0 \longrightarrow \mathbb{R} \longrightarrow \mathbb{R} \longrightarrow 0 \longrightarrow 0 \longrightarrow \cdots$$

Algebra → Summary

Theoretical Diversion: We can view the barcode as a module.

We use the collection of interval modules to define the direct sum, $\bigoplus_{[a,b)\in B(P)} I^{[a,b)}$ where $B(P)$ is the barcode of P.

Theorem

For V, an $[n]$-module such that $dim V_p < \infty$ for all $p \in [n]$. Then

$$V \cong \bigoplus_{[a,b)\in B(V)} I^{[a,b)}$$

where $B(V)$ is the barcode of V.

Summary → Analysis

This step often depends on the data we are studying and what features within our data we want to consider.

Much study revolves around applications and *stability*, a measure of how similar our results are if we perturb our data slightly.

Summary → Analysis

Stability:

- Requires defining a metric on modules or the barcode modules.
- Sparks the question of what is the best metric

Summary → Analysis

Applications: Audio Detection:

Goal: Use topological descriptors of audio signals for audio identification.

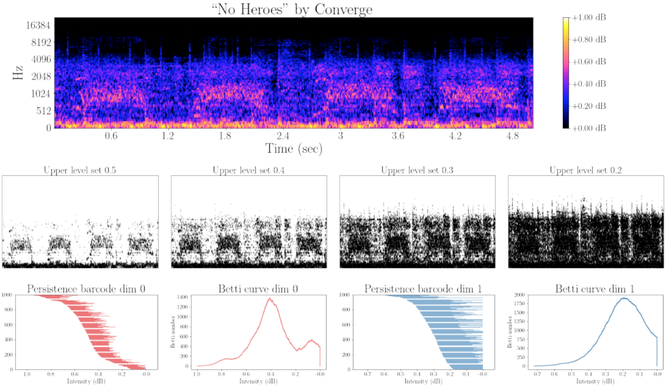


Figure: Song 'No Heroes' from the metal core band *Converge*, with a strong heavy metal rhythm [RFD+23]

Summary → Analysis

Applications: Audio Detection:

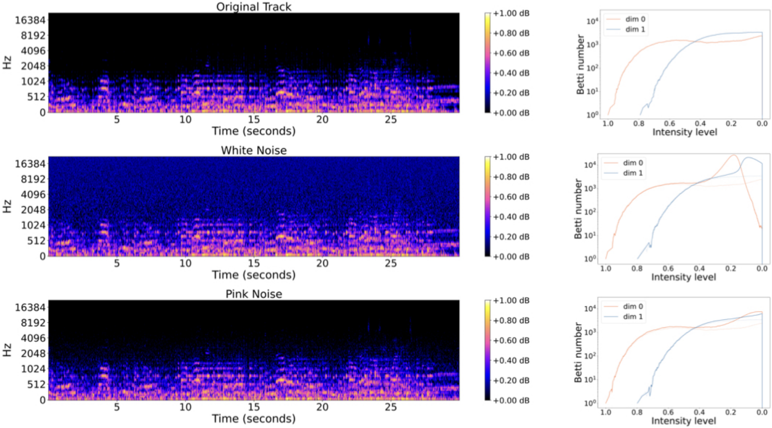


Figure: Data gathered from 'The Morning' by *Le Loup*. [RFD+23]

Summary → Analysis

Applications: Audio Detection:

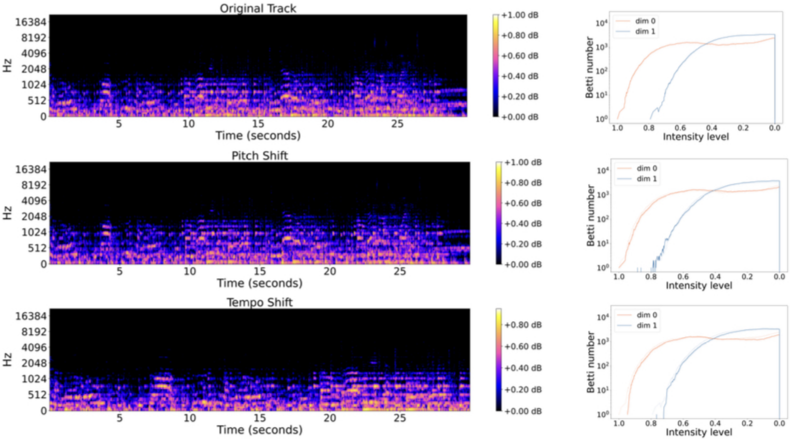


Figure: Data gathered from 'The Morning' by *Le Loup*. [RFD+23]

Future Directions

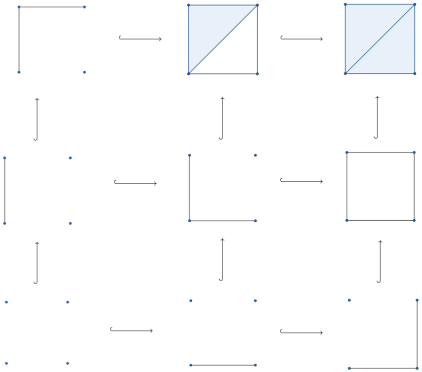
As more complicated data is analyzed, we need to consider multiple parameters.

We call this type of TDA, MultiParameter Persistent Homology.
For n parameters, we can build an n-filtered simplicial complex.

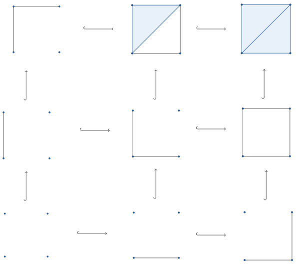
Future Directions

As more complicated data is analyzed, we need to consider multiple parameters.

We call this type of TDA, MultiParameter Persistent Homology.
For n parameters, we can build an n-filtered simplicial complex.



Generalize



Applying homology at each index in the multiparameter case, we get the following;

$$H_0 : \quad \begin{array}{ccccc} k^2 & \longrightarrow & k & \longrightarrow & k \\ \uparrow & & \uparrow & & \uparrow \\ k^3 & \longrightarrow & k^2 & \longrightarrow & k \\ \uparrow & & \uparrow & & \uparrow \\ k^4 & \longrightarrow & k^3 & \longrightarrow & k^2 \end{array}$$

$$H_1 : \quad \begin{array}{ccccc} 0 & \longrightarrow & k & \longrightarrow & 0 \\ \uparrow & & \uparrow & & \uparrow \\ 0 & \longrightarrow & 0 & \longrightarrow & k \\ \uparrow & & \uparrow & & \uparrow \\ 0 & \longrightarrow & 0 & \longrightarrow & 0 \end{array}$$

Future Directions

- Further study of MultiParameter Persistent Homology.
  - No "good" barcode exists in this case with the current generalized definition.
  - Is there another way to summarize the data?
- Work on finding a good measure of stability.
- Continue to develop efficient code for producing results and visualization of data analyzed with TDA.

How do you get started?

For those interested in the theoretical side:

1. Topological Data Analysis Mastermath by Dr. Magnus Bakke Botnan

For those interested in the computational side:

1. TDA package in RStudio
2. giotto-tda

For those interested in both:

1. Dr. Peter Bubenik's webpage

Are there any questions?

Thank you!

Thank you!

Go Dutch!

References:

Magnus Bakke Botnan.
Topological data analysis mastermath.
Course Notes 2022.
https://www.few.vu.nl/~botnan/lecture_notes.pdf.

Wojciech Reise, Ximena Fernández, Maria Dominguez, Heather A. Harrington, and Mariano Beguerisse-Díaz.
Topological fingerprints for audio identification, 2023.

Summary → Analysis

Stability:

In many cases, this requires defining a metric on modules or the barcode modules.

There are many examples of metrics including:

- The **Bottleneck Distance**:
  $d_{\mathcal{B}}(\mathcal{C}, \mathcal{D}) = inf\{c(\chi)|\chi \text{ is a matching between } \mathcal{C} \text{ and } \mathcal{D}\}.$
- The **Interleaving Distance**:
  $d_{\mathcal{I}}(M, N) = inf\{\epsilon|\epsilon\text{-interleaving between } M \text{ and } N\}.$

Generalize

Goal: Find a way to summarize multidimensional data as we did in one dimension with the barcode.

Generalize

Goal: Find a way to summarize multidimensional data as we did in one dimension with the barcode.

Definition

A **good barcode** for an $\mathbb{N}^2$-indexed bipersistence module M is a collection $\mathcal{B}_M$ of subsets of $\mathbb{R}^2$ such that for each $a \leq b \in \mathbb{R}^2$,

$$\mathrm{Rank} M_{a,b} = |\{S \in \mathcal{B}_M | a, b \in S\}|.$$

Generalize

Goal: Find a way to summarize multidimensional data as we did in one dimension with the barcode.

Definition

A **good barcode** for an $\mathbb{N}^2$-indexed bipersistence module M is a collection $\mathcal{B}_M$ of subsets of $\mathbb{R}^2$ such that for each $a \leq b \in \mathbb{R}^2$,

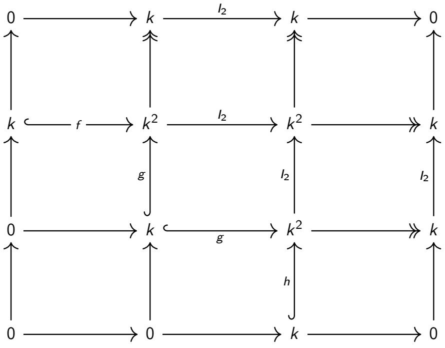
$$\mathrm{Rank} M_{a,b} = |\{S \in \mathcal{B}_M | a, b \in S\}|.$$

The one parameter case satisfies this definition.

Generalize

Claim:

Consider the $\mathbb{N}^2$-indexed persistence module

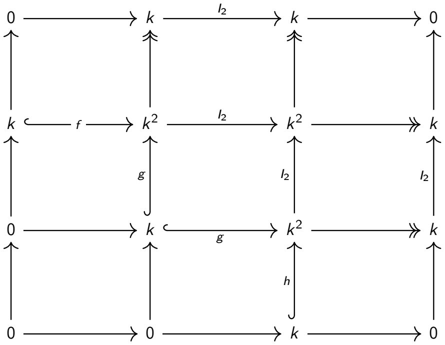


$$f = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad g = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad h = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Generalize

Claim:

Consider the $\mathbb{N}^2$-indexed persistence module



$$f = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad g = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad h = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

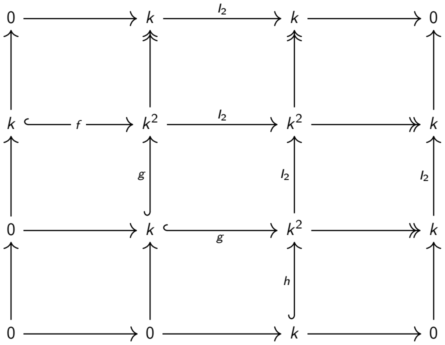
Let $a = (0,0)$ and $b = (2,2)$. Then if $S \subseteq \mathbb{R}^2$ is a region with $a, b \in S$,

$$|\{S \in \mathcal{B}_M | a, b \in S\}| = 3.$$

Generalize

Claim:

Consider the $\mathbb{N}^2$-indexed persistence module



$$f = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad g = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad h = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Let $a = (0, 0)$ and $b = (2, 2)$. Then if $S \subseteq \mathbb{R}^2$ is a region with $a, b \in S$,

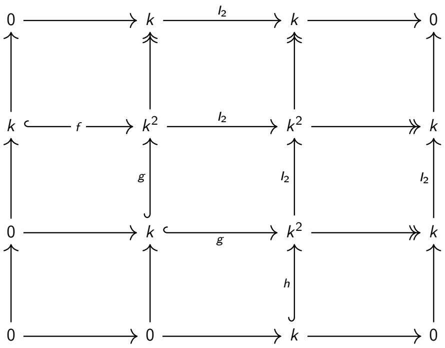
$$|\{S \in \mathcal{B}_M | a, b \in S\}| = 3.$$

However, $\mathrm{Rank} M_a, b = 0$ which shows no such barcode exists for this persistence module.

Generalize

Claim:

Consider the $\mathbb{N}^2$-indexed persistence module



$f = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $g = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $h = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

Let $a = (0,0)$ and $b = (2,2)$. Then if $S \subseteq \mathbb{R}^2$ is a region with $a, b \in S$,

$$|\{S \in \mathcal{B}_M | a, b \in S\}| = 3.$$

However, $\mathrm{Rank} M_a, b = 0$ which shows no such barcode exists for this persistence module.

No good barcode exists for n-parameter persistence modules of any indexing set for $n \geq 2$.