TDA III

Université Lyon 2

2024/2025

Rémi Vaucher

HALIAS & ERIC Lab.

Table of Contents

On a vu dans les slides précédents:

- Les complexes simpliciaux: c'est un objet géométrique permettant d'estimer des caractéristiques géométriques sur un "nuage de point" (au sens large).
- Les nombres de Betti: ce sont des quantités qui permettent de caractériser les espaces topologiques à homotopie près.
- Les courbes de Betti: cela permets de caractériser l'évolution des complexes de Čech ou Rips à un rayon donné.

Il est maintenant temps de voir des outils pouir caractériser les constructions de types Čech ou Rips

Filtration

Définition

Soit $\mathcal{C}$ un complexe simplicial. Une **filtration** de $\mathcal{C}$ est une collection de complexe $\mathcal{F} = (\mathcal{C}_r)_{r\in[0;T]}$ telle que:

- Pour tout $r \leq r'$, on a $\mathcal{C}_r \subseteq \mathcal{C}_{r'}$
- $\bigcup_{r\in[0;T]} \mathcal{C}_r = \mathcal{C}$

Bien évidemment, pour un complexe simplicial donné, il existe de nombreuse filtration possible.

Remarque(s)

Les complexes de Čech et Rips sont des filtrations pour leur complexe de plus haut rayon.

Sur cette filtration, on a les courbes de Betti à disposition. Elles donnent donc les nombres de Betti à un instant donné. Toutefois si ces nombres ne bougent pas au cours de la filtration, cela ne veux pas dire que le complexe n'a pas foncièrement changé:

Figure 1: A gauche: un complexe d'une filtration au temps t_1. A droite, un complexe de la même filtration au temps $t_2 > t_1$.

Les nombres de Betti sont les mêmes, pourtant, la structure a bien changée.

Revenons un peu aux mathématiques de boeufs: les groupes d'homologies ont évolués entre t_1 et t_2, mais pas leur dimension. En effet, les classes d'homologies ne sont plus les mêmes, mais il y en a toujours le même nombre.
Il nous faut un moyen de quantifier ces changements autrement qu'avec les courbes de Betti.
Ce moyen, c'est l'**homologie de persistance**.

Définition

Soit $\mathcal{F} = (\mathcal{C}_r)$ une filtration sur $\mathcal{C}$. On définit $H_k^{s,t}(\mathcal{C})$ comme le **groupe d'homologie de persistance** de $\mathcal{C}$ entre les temps s et t. Ces groupes contiennent les classes d'homologie communes à tous les complexes de la filtration dont les temps sont compris entre s et t.

On rappelle brièvement que les classes d'homologies représentent les **cycles** à une ou plusieurs frontière près. On a l'inclusion triviale: $H_k^{s,t} \subseteq H_k(\mathcal{C}_s)$

Dans la filtration suivante, qualifions les classes d'homologies à chaque instant puis regardons lesquelles **persistent** dans le temps.

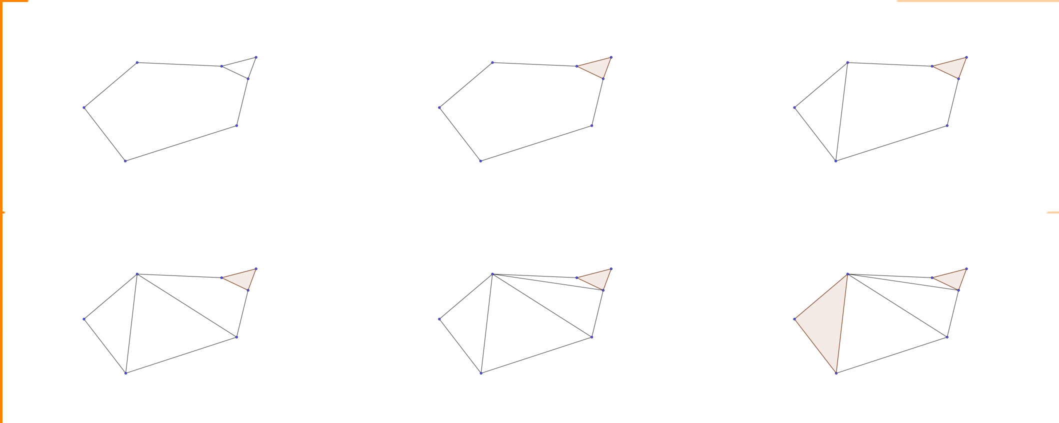


Figure 2: Le temps évolue en ligne

La question désormais, c'est comment quantifier cette persistance homologique pour l'utiliser statistiquement? Nous possédons deux outils à notre disposition:

- Le **code-barre**.
- Le **diagramme de persistance**.

Table of Contents

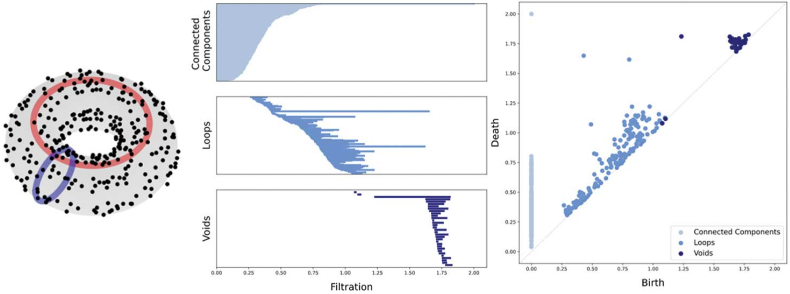
Deux aspects des classes d'homologies de persistance sont importants:

- Leur temps de naissance: Le plus petit s pour lequel la classe apparaît dans un $H_k^{s,t}$.
- Leur temps de mort: Le plus grand t pour lequel la classe apparaît dans un $H_k^{s,t}$.

Si l'on arrive à déterminer le temps de vie et de mort de chaque classe d'homologie, on peut leur associer un **intervalle de vie**.

Définition

Le **code-barre** d'une filtration est une représentation graphiques de tous les intervalles de vie.



(a) 3D Scatterplot (b) Barcode Plot (c) Persistence Diagram

Figure 3: https://chance.amstat.org/2021/04/topological-data-analysis/

- Une petite barre correspond à une classe qui disparait aussi vite qu'elle est née: c'est du **bruit topologique**.
- Une très longue barre de vie correspond à une classe persistante: ce sont elles qui nous intéressent, et qui caractériseront notre jeu de données.

Définition

Le **diagramme de persistance** est un nuage de point représentant chaque classe d'homologie par son temps de naissance et de mort (voir Fig3).

- Le *temps de mort* étant logiquement supérieur au *temps de naissance*, ce nuage de point est toujours au dessus de la droite $x = y$.
- Un point situé près de la droite $x = y$ est un petit intervalle: c'est du bruit.
- Un point éloigné de la droite $x = y$ est un grand intervalle, et donc une classe d'homologie persistante importante.

Animation complexe de Rips, code-barre et diagramme de persistance.

Table of Contents

Dans l'état actuel de nos connaissance, chaque nuage de points dans $\mathbb{R}^n$ (éventuellement dans des espaces de dimension infinie) peut être envoyé dans l'espace des complexes simpliciaux, ou dans celui des filtrations, ou bien dans celui des diagrammes de persistance.
Nous allons nous concentrer sur ce dernier, car nous pouvons le munir d'une **métrique**.

Bottleneck Distance

Sur la figure suivante:

- Déterminer une bijection entre le diagramme rouge et le diagramme bleu. Si ce n'est pas possible, compléter l'un des deux diagrammes par des points sur la diagonale.
- Calculer les distances en norme ∞ sur chaque paire formée.
- Déterminer le maximum de ces distances.
- Recommencer pour un autre matching.

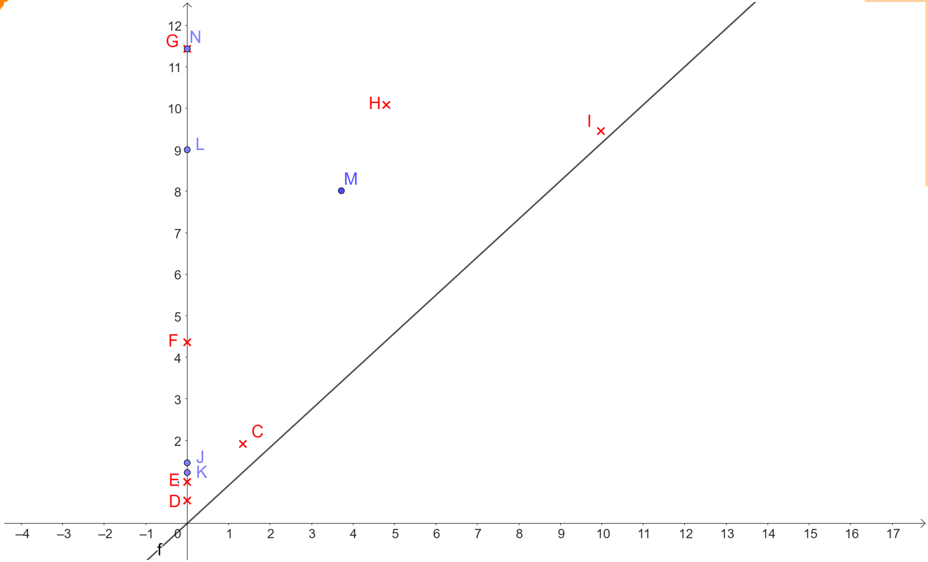


Figure 4: En rouge: diagramme rouge. En bleu: diagramme bleu.

Définition

La distance **bottleneck** entre deux diagrammes de persistance D_1 et D_2 es tdéfinie par:

$$d_B(D_1, D_2) = \inf_{\text{matching } m} \max_{(P,Q)\in m} \|P - Q\|_\infty$$

Sous entendu, on regarde toutes les bijections entre les diagrammes (quitte à compléter par la projection sur l'axe diagonal), on calcule l'écart maximal entre ces diagrammes **en respectant la bijection**, et on garde la valeur minimal sur toutes les bijections.

Remarque(s)

- L'écart entre les points est calculé grâce à la norme infinie. Quid des autres normes?

En fait, nous allons continuer à utiliser $\|.\|_\infty$ pour décliner la Bottleneck

Définition

Soit D_1 et D_2 deux diagrammes de persistance. On définit la distance p-Wasserstein par

$$d_B(D_1, D_2) = \inf_{\text{matching } m} \sum_{(P,Q)\in m} \|P - Q\|_\infty^p$$

Remarque(s)

La distance Bottleneck est le cas particulier pour $p = 1$.