Topological Data Analysis

D'après Julien Tierny et Frédéric Chazal

30 septembre 2025

Table des matières

mailto:julien.tierny@sorbonne-universite.fr mailto:frederic.chazal@inria.fr https://julien-tierny.github.io/topologicalDataAnalysisClass.html

Introduction

Méthodes algorithmiques d'analyse topologique de données, particulièrement en science et en ingénierie.

Le but est de partir de données, sous forme de maillages et maillables, et de retrouver des structures au sein de jeux de données. Partant d'une carte (considérée comme jeu de données brutes), avec des features intéressantes, pour pouvoir raisonner sur l'espace, on passe à une représentation abstraite, par exemple comme un graphe, et c'est sur cette structure de données sous-jacente qu'on va raisonner. Ici, on peut ajouter des filtres pour redéfinir le maillage et donc redéfinir le résultat du raisonnement. Plus généralement, on veut construire une carte à partir d'un jeu de données. En astrophysique, par exemple, on modélise la croissance de l'univers à une grande échelle, on la simule par une grille de voxel, on estime la densité de matière noire sur chaque voxel, et on découvre une sorte de géométrie ressemblant à des neurones lorsqu'on trouve aussi des groupes de galaxies, formant une "toile cosmique". On peut calculer les connexions avec des complexes simpliciaux dits de Morse-Smale, dont on peut extraire une structure de graphes.

Ainsi, on extrait de la structure d'un ensemble de données, de manière robuste et indépendante de l'échelle, par comparaison et extraction de propriétés. Sous le capot, on fait :

- Homologie Simpliciale
- Théorie de Morse
- Homologie Persistente

Pour des données numériques, étant données un échantillon de points dans un espace euclidien, par exemple, on peut les représenter et objectiver des représentations géométriques apparaissant. On a des manières de mailler l'ensemble (triangulation de Delaunay, par exemple) qui amènent à des indicateurs qui nous expliquent où sont répartis les données, par exemple avec des noyaux pour estimer la densité. Avec une fonction scalaire sur un maillage, on définit une filtration, et on regarde les propriétés de la fonction, comme les optima locaux et on en extrait une structure algébrique (complexe de Morse-Smale) qui nous donne une structure algébrique. On obtient des générateurs, et des composantes "connexes".

On a ce genre de densité de pixels, par exemple la hauteur de surface de la mer qui permet de remarquer les vortexs, en chimie quantique ou des spectrogrammes d'enregistrement vocaux. On part d'un domaine géométrique et d'un signal sur ce domaine, signal qui exhibe des patternes géométriques qu'on souhaite quantifier. Ceci permet l'extraction de propriétés, la segmentation, la réduction de dimension et autres. Dans le cas de points en grande dimension, on a une unique théorie qui s'applique très généralement.

En terme de logiciels, on a le TTK (ParaView >= 5.10) et Gudhi (bibliothèque python).

1 Homologie Simplicale

Les données reçues, parfois, vont contenir explicitement la géométrie avec une construction combinatoire. On supposera qu'on aura une donnée d'entrée linéaire par morceau sur un complexe simplical.

1.1 Complexe Simpliciaux

Definition 1.1 Un d-simplexe est l'enveloppe convexe σ de $d+1$ points affinement indépendants dans l'espace euclidien $\mathbb{R}^n$ avec $0 \leq d \leq n$. On dit que d est la dimension du simplexe.

Une ligne est un 1-simplexe, un triangle un 2-simplexe et un tétrahèdre un 3-simplexe.

Definition 1.2 Une face τ d'un simplexe σ est un simplexe construit par un ensemble non vide des $d+1$ poits définissant σ. On note $\tau \leq \sigma$ et τ_i une face de dimension i. On dit aussi que σ est une coface de τ.

Selon la définition, on a $\sigma \leq \sigma$.

Definition 1.3 Un complexe simplicial $\mathcal{K}$ est une collection finie non-vide de simplexes $\{\sigma_i\}$, telle que :

1. $\tau \leq \sigma \Rightarrow \tau \in \mathcal{K}$;
2. $\sigma_i \cap \sigma_j$ est soit une face, soit vide.

Definition 1.4 L'étoile d'un simplexe $\sigma \in \mathcal{K}$ est l'ensemble des simplexes de $\mathcal{K}$ qui contiennent σ :

$$\mathrm{St}(\sigma) = \{\tau \in \mathcal{K} | \sigma \leq \tau\}$$

On note $\mathrm{St}_d(\sigma)$ les d-simplexes de $\mathrm{St}(\sigma)$.

C'est l'ensemble des cofaces de σ dans $\mathcal{K}$. C'est le plus petit voisinage combinatoire autour d'un simplexe.

Definition 1.5 Le lien d'un simplexe σ est l'ensemble des faces de $\mathrm{St}(\sigma)$ disjointes de σ :

$$\mathrm{Lk}(\sigma) = \{\tau \leq \sigma' \,|\, \sigma' \in \mathrm{St}(\sigma) \wedge \tau \cap \sigma = \varnothing\}$$

On définit de même le d-lien $\mathrm{Lk}_d(\sigma)$ en remplaçant St par St_d dans la définition

C'est en quelque sorte la bordure du voisinage combinatoire de lui-même.

En réalité on va considérer que les sommets (ou 0-simplexes) sont des points, et que les d-simplexes sont des ensembles de points. Ceci définit une notion de complexe simplicial abstrait, utile lorsqu'on n'a pas d'immersion dans un espace euclidien, ou alors dans un espace euclidien en trop grande dimension. On relâche ici la condition d'intersection, puisqu'on n'a plus de structure géométrique de l'espace.

Un exemple de complexe simplicial abstrait est le complexe de Rips ou complexe de Vietori-Rips. Étant donné un nuage de points avec une métrique :

- Le diamètre d'un ensemble P est défini par : $\varnothing(P) = \sup\{d(x,y) \mid x,y \in P\}$
- On construit un complexe simplicial $p \leq p_{max}$ de sorte que tous $p+1$ points dont le diamètre est plus petit qu'une valeur seuil d_{max}.

Le complexe de Rips est une généralisation de la notion de graphe de voisinage.

Definition 1.6 L'espace sous-jacent à un complexe simplicial est l'union des simplexes du complexe.

Definition 1.7 La triangulation $\mathcal{T}$ d'un espace topologique X est un complexe simplicial $\mathcal{K}$ dont l'espace sous-jacent $|\mathcal{K}|$ est homéomorphe à X.

Une triangulation d'un espace est donc un complexe simplicial abstrait.

Definition 1.8 Une d-variété M est un espace topologique dans lequel tout élément m a un voisinage ouvert homéomorphe à une d-boule euclidienne.

Definition 1.9 La triangulation d'une d-variété est appelée d-variété linéaire par morceaux.

Représenter en mémoire un complexe simplicial est très couteux : il faut, pour chaque dimension, une liste des hyperarêtes (ou d-simplexes) en les représentant par un indice de sommet.

1.2 Homologie simpliciale

1.2.1 Rappels d'homotopie

Definition 1.10 Un chemin p dans C est un homéomorphisme d'un intervalle réel vers l'objet C. On dit que C est connexe (par arcs) si pour tous deux points il existe un chemin dans C les reliant.

Definition 1.11 Une composante connexe d'un objet est un sous-ensemble connexe (par arcs) maximal de l'objet.

Definition 1.12 Une homotopie entre deux fonctions continues f et g de X vers Y est une fonction continue $H : X \to [0,1] \to Y$ du produit d'un espace topologique X par l'intervalle unité vers un espace topologique Y de sorte que $H(x,0) = f(x)$ et $H(x,1) = g(x)$ pour tout $x \in X$. S'il existe une homotopie entre f et g on dit que f et g sont homotopes.

Definition 1.13 Si dans un espace X, tous les chemins entre tous deux points sont homotopes, on dit que X est simplement connexe.

Le disque est simplement connexe, mais pas le disque privé de 0.

Definition 1.14 La caractéristique d'Euler d'une triangulation T d'un espace topologique est la somme alternée des nombres des i-simplexes :

$$\chi(T) = \sum_{i=0}^{d} (-1)^i \, |\sigma_i|$$

Proposition 1.1 La caractéristique d'Euler est invariante par homéomorphisme.

1.2.2 Groupes d'homologie

Definition 1.15 Une p-chaine est une somme (formelle) de p-simplexes. On suppose que l'opérateur somme est défini modulo 2.

Ici, la somme est réellement la différence symmétrique (ou la disjonction exclusive) sur $\mathbb{F}_2^{|\sigma_p|}$, et définit le groupe C_p des p-chaînes. Le rang de C_p est $|\sigma_p|$ et son ordre est $2^{|\sigma_p|}$

On peut généraliser à des coefficients plus généraux. Informatiquement, il faut voir la notion de chaîne comme un masque binaire sur l'ensemble des p-simplexes.

Definition 1.16 L'opérateur de bordure ∂ d'un p-simplexe renvoie la $(p-1)$-chaîne des $(p-1)$-faces du simplexe. On l'étend aux p-chaînes comme un morphisme de $C_p \to C_{p-1}$.

Pour un triangle, c'est l'ensemble de ses arêtes. Pour une arête, c'est l'ensemble des extrémités.

Definition 1.17 Un p-cycle est une p-chaîne dont la bordure est vide. On définit Z_p le groupe des p-cycles comme sous-groupe de C_p.

Definition 1.18 Le groupe B_p des p-bordures est l'image de $C_{(p+1)}$ par ∂.

Lemme 1.1 Pour tout $x \in C_p, p \geq 2$, $\partial\partial x = 0$.

Démonstration. Il suffit de vérifier le résultat sur les p-simplexes et d'étendre par somme. Puisque pour tout $(p-2)$-faces τ on a exactement 2 $(p-1)$-co-faces de τ dans un p-simplexe σ, on a le résultat. ■

On obtient directement :

Proposition 1.2 B_p est un sous-groupe de Z_p.

Si on a trois 1-simplexes e_1, e_2, e_3 mais pas leur coface commune $\{e_1, e_2, e_2\}$, on a un exemple d'inclusion stricte. Ceci nous amène à définir la notion de trou, en isolant les cycles :

Definition 1.19 Le p-ème groupe d'homologie H_p est le quotient de Z_p par B_p.

Démonstration. B_p étant un sous-groupe de Z_p, H_p est bien défini. ■

Géométriquement, on peut étendre un p-cycle à un autre p-cycle lorsqu'ils encapsulent le même "trou", c'est-à-dire lorsque qu'on peut "étendre" le premier cycle en encapsulant un $(p+1)$-simplexe. Une classe d'homologie est un élément de H_p, ou plutôt sa classe d'équivalence dans Z_p.

Pour calculer $|H_p|$, on énumère C_p, on élimine les chaînes de bordure non-vide pour calculer Z_p, et on peut ensuite énumérer les classes d'homologie.

Definition 1.20 On définit le p-ème nombre de Betti β_p comme le rang du groupe H_p. Ici, c'est $\log_2 |H_p|$.

La formule logarithmique pour β_p vient du calcul modulo 2 dans notre opération de groupe.

Proposition 1.3 La caractéristique d'Euler d'une triangulation T d'un espace topologique X de dimension d vérifie :

$$\chi(T) = \sum_{i=0}^{d} (-1)^i \beta_i(T)$$

2 Homologie Persistente

3 Fonctions de Morse

4 Inférence Topologique

5 Théorie de Morse Discrète

6 Noyaux et Statistiques