Séminaire BOURBAKI Juin 2024
75e année, 2023–2024, no 1225, p. 607 à 641
doi : 10.24033/ast.1242

UN APERÇU DES MODULES DE PERSISTANCE ET DE LEURS APPLICATIONS

par **Grégory Ginot**

Introduction

La *persistance homologique* a en grande partie été développée pour être un des outils principaux de l'étude géométrique des « données », aussi appelée *analyse topologique des données*. Cette dernière est un domaine assez récent et très actif qui trouve de nombreuses applications dans divers domaines scientifiques, notamment via les méthodes issues de la théorie de la persistance homologique, qui a des racines anciennes en géométrie différentielle et topologie algébrique.

D'un point de vue purement mathématique, la théorie de la persistance peut se voir comme l'étude d'*espaces filtrés* $(X_i)_i$ via des invariants algébriques (aussi effectifs que possible) et des propriétés métriques naturelles qu'on peut leur associer (ce dernier point étant l'un des apports récents cruciaux). Et c'est essentiellement ce qui sera développé dans ces notes. Il s'agit le plus souvent d'espaces topologiques, variétés ou complexes simpliciaux (mais cela a du sens dans d'autres cadres).

Il est important de noter qu'un des points clés de la théorie de *la persistance est qu'elle vise à étudier la filtration en elle-même* et pas simplement à la voir comme un outil pour calculer l'espace colimite $\bigcup_{i\in I} X_i$.

Avant de détailler un peu plus la persistance et les modules persistants, on va essayer de préciser pourquoi elle intervient en analyse topologique des données. L'idée de cette branche est d'étudier des données (complexes et possiblement en grande dimension) en essayant d'évaluer leur géométrie et/ou leur topologie, avec comme objectif des tâches de type classification des données, visualisation, régression ou reconnaissance de patterns communs. Souvent, ces données apparaissent comme un espace métrique $(\mathbb{X}, d)$, typiquement un grand sous-ensemble fini d'un espace euclidien — qu'on qualifie souvent de nuage de points.

Par « information topologique » extraite d'un tel ensemble, on entend la forme que prennent ces données dans l'espace ambiant (possiblement de grande dimension). Ce qu'on appelle la *manifold hypothesis* prescrit que (après avoir éventuellement partitionné les données en groupes) ces données s'accumulent sur une sous-variété X

de l'espace ambiant. On peut alors penser les données comme une *discrétisation* de cet objet continu X, qui peut être *a priori* inconnu de l'observateur. L'analyse topologique des données a pour but d'utiliser (et d'approximer donc) la topologie de X pour comprendre ou organiser les données.

La persistance homologique s'appuie sur les méthodes de la topologie algébrique classique *adaptées aux espaces filtrés* pour estimer la topologie des données. Reprenons l'exemple d'un sous-ensemble $\mathbb{X}$ d'un espace disons euclidien (muni de la distance induite), vu comme une approximation d'une sous-variété X. À tout point x de notre ensemble discret $\mathbb{X}$, on peut associer la boule $B(x,r)$ (c'est-à-dire épaissir notre point x) et la réunion $\mathbb{X}(r) = \bigcup_{x\in\mathbb{X}} B(x,r)$ de ces boules. Si $\mathbb{X}$ est une bonne approximation de X, alors cette réunion de boules représente un espace topologique qui, pour un r ni trop petit (sinon on ne voit que les points) ni trop grand, va être un épaississement de X qui lui sera *homotope*. En pratique, on remarque que cette famille $(\mathbb{X}(r))_{r\geqslant 0}$ est filtrée au sens où $r < r'$ implique $\mathbb{X}(r) \subset \mathbb{X}(r')$. On obtient donc ici un espace topologique filtré $\mathbb{X}(r)$ dont la filtration est paramétrée par l'ensemble $\mathbb{R}_+$. En pratique, si $\mathbb{X}$ est fini, il n'y a qu'un nombre fini de valeurs de r où la topologie de $\mathbb{X}(r)$ change, et donc notre espace est en fait équivalent à un espace filtré par un sous-ensemble fini de $\mathbb{R}$ (que l'on ne connaît pas forcément *a priori*). Ceci constitue un premier exemple d'espace filtré apparaissant en analyse topologique de données.

Une autre famille standard (et très proche) d'exemples est donnée par les sous-niveaux d'une fonction $f\colon X \to \mathbb{R}$, c'est-à-dire la collection des $\{x \in X,\ f(x) < t\}_{t\in\mathbb{R}}$ qui est là encore naturellement paramétrée par $\mathbb{R}$.

Un dernier exemple standard est donné par le cas d'un graphe (avec la distance de plus court chemin). On a alors des informations topologiques évidentes comme le nombre de composantes connexes du graphe (qui correspondent à partitionner) et le nombre de cycles dans chaque composante. Notons que ces informations topologiques sont là aussi « à homotopie près », au sens où le nombre de composantes connexes rend un arbre et un point équivalents, et où le nombre de cycles correspond au groupe fondamental (de la composante connexe considérée) du graphe. On peut aussi extraire des informations de plus grande dimension, ce qui revient à construire un complexe simplicial (par exemple le complexe de clique du graphe) qui tient compte des connectivités d'ordre supérieur dans le graphe.

La persistance homologique a pour objet d'étudier les espaces filtrés (par un sous-ensemble de $\mathbb{R}$ dans le cadre originel, mais il y a des développements importants en direction du cadre à plusieurs paramètres, *cf.* 5) via des invariants topologiques. Un principe, du point de vue des applications en analyse topologique des données, est donc que l'on souhaite avoir trois propriétés fondamentales :

- ils doivent être invariants par homotopie (de façon à ne pas être sensibles à de petites déformations, comme dans l'exemple des réunions de boules ci-dessus) ;
- ils doivent être calculables sur machine (de façon à avoir des applications) ;

- ils doivent être comparables pour divers jeux de données, ce qui va se faire par l'intermédiaire de (pseudo-)métriques si possible là aussi implémentables sur machine.

Les invariants qui vont réaliser cela seront encodés par la notion de *module de persistance* associée à un objet filtré dans la section 1. Ces modules sont essentiellement la structure obtenue en passant à l'homologie dans un espace filtré. Cette structure a de très bonnes propriétés dans le cadre algébrique, notamment le fait que l'on peut décomposer les modules de persistance en indécomposables simples caractérisés par un intervalle de $\mathbb{R}$. La donnée de ces intervalles est un objet combinatoire, appelé son *code-barres*, qui permet de le manipuler informatiquement et combinatoirement facilement. Ces propriétés seront le cœur de la partie 2.

Comme évoqué plus haut, une idée clé de la persistance est le fait que les modules de persistance ont une (pseudo)-métrique naturelle, appelée *distance d'entrelacement*, qui a une traduction combinatoire sur le code-barres appelée distance bottleneck, c'est-à-dire du « goulot de bouteille » en français). C'est l'outil essentiel pour comparer des objets persistants et obtenir des estimées intéressantes dans les applications. Ces distances seront définies et comparées dans la partie 3.

Dans les applications, un point important est que les distances que l'on a obtenues pour comparer les modules persistants (ou leurs codes-barres associés à des données) reflètent bien la géométrie des données. Ceci est établi par ce que l'on appelle les théorèmes de stabilité énoncés en partie 4, qui essentiellement garantissent que l'homologie persistante fournit des estimateurs consistants (au sens statistique du terme) de la topologie des données.

Dans la section 5, nous évoquerons des généralisations et raffinements importants de l'homologie persistante, notamment le passage à plusieurs paramètres qui est un enjeu important non seulement dans la théorie mais aussi dans les applications concrètes à d'autres sciences.

Enfin, dans la partie 6, nous parlerons du problème de l'inférence, c'est-à-dire que nous reviendrons au problème initial de déterminer la topologie d'un espace X à partir d'une discrétisation et expliciterons ce qu'apportent les outils introduits dans les sections précédentes dans ce contexte.

Notons que du point de vue de l'étude des espaces filtrés, il n'est évidemment pas nécessaire que ces derniers proviennent de données. La persistance homologique a un intérêt mathématique propre. Elle a de fait trouvé des applications en géométrie symplectique (via les filtrations naturelles dans les complexes d'homologie de Floer : L. Polterovich et al., 2020), a été étudiée en lien avec la théorie de Morse avant même l'analyse topologique des données, et a également des applications en géométrie que nous évoquerons brièvement dans la partie 6.

Notations et conventions

- ▷ Dans toute la suite $\mathbb{F}$ sera un corps. Dans les applications pratiques de la persistance homologique, ce corps est en général un corps fini (notamment pour éviter d'avoir des problèmes d'arrondi dans les calculs sur machine).
- ▷ Si on ne précise rien, les groupes d'homologie seront les groupes d'homologie singulière (pour un espace topologique) ou les groupes d'homologie simpliciale (pour un complexe simplicial) qui calculent précisément l'homologie singulière de l'espace topologique dont le complexe simplicial est une triangulation.
- ▷ Si on ne précise rien, les groupes d'homologie ou espaces vectoriels considérés seront sur le corps $\mathbb{F}$. On notera $\mathbb{F}-\mathrm{Vect}$ la catégorie des $\mathbb{F}$-espaces vectoriels.

On renvoie le lecteur à Barannikov (1994), Botnan et Lesnick (2023), Chazal, Silva et al. (2016), Oudot (2015) et L. Polterovich et al. (2020) pour des références détaillées sur la persistance homologique et ses applications en géométrie symplectique, et des preuves détaillées de résultats des sections 2, 4 notamment.

1. Modules de persistance

Nous introduisons dans cette partie la notion de module de persistance, qui est le miroir algébrique des espaces filtrés. Cette notion peut se généraliser pour des structures autres que des filtrations par des sous-ensembles de $\mathbb{R}$ et pour des structures autres que les espaces vectoriels, comme nous le verrons ci-dessous.

Définition 1.1 (Module de persistance)**.** Un module de persistance est la donnée d'une collection $(M(t))_{t\in\mathbb{R}}$ de $\mathbb{F}$-espaces vectoriels et pour tout $s \leqslant t$, d'applications linéaires $\iota^M_{s\leqslant t}\colon M(s) \to M(t)$, vérifiant :

- ▷ $\iota^M_{s\leqslant s} = \mathrm{id}_{M(s)}$;
- ▷ pour tout $s \leqslant t \leqslant u$, on a $\iota^M_{t<u} \circ \iota^M_{s<t} = \iota^M_{s<u}$.

On notera simplement $M = (M(s), \iota^M_{s\leqslant t})_{s,t}$ cette donnée.

Si M et N sont deux modules de persistance, un morphisme $f\colon M \to N$ est la donnée d'une collection $f_t\colon M(t) \to N(t)$ $(t \in \mathbb{R})$ vérifiant la condition suivante : pour tout $s < t$, le diagramme

$$\begin{array}{ccc} M(s) & \xrightarrow{\iota^M_{s<t}} & M(t) \\ {\scriptstyle f_s}\downarrow & & \downarrow{\scriptstyle f_t} \\ N(s) & \xrightarrow[\iota^N_{s<t}]{} & N(t) \end{array}$$

est commutatif.

La composition des morphismes se fait point par point ($(g \circ f)_t = g_t \circ f_t$) et l'on obtient ainsi une catégorie.

Exemple 1.2 (Module à support sur un intervalle). Soit E un $\mathbb{F}$-espace vectoriel et soit I un intervalle de $\mathbb{R}$. On dispose d'un module de persistance associé défini comme suit : pour tout $t \in \mathbb{R}$, on a

$$E_I(t) = \begin{cases} E & \text{si } t \in I \\ \{0\} & \text{sinon,} \end{cases}$$

et les morphismes structuraux sont donnés, pour tout $s \leqslant t$, par $\iota_{s\leqslant t}^{E_I} = \mathrm{id}_E$ si $t, s \in I$ et (nécessairement) le morphisme nul sinon.

Exemple 1.3. Pour tout réel t, on pose $\widetilde{\mathbb{F}}(t) = \mathbb{F}$ et, pour $s < t$,

$$\iota_{s\leqslant t}^{\widetilde{\mathbb{F}}} = \begin{cases} \mathrm{id}_{\mathbb{F}} & \text{si } t \leqslant 0 \\ 0 & \text{si } s \leqslant 0 \text{ et } t > 0 \\ \mathrm{id}_{\mathbb{F}} & \text{si } s > 0 \end{cases}$$

On vérifie que cette structure définit bien un module de persistance tel que pour tout t, on ait $\widetilde{\mathbb{F}}(t) = \mathbb{F}_{\mathbb{R}}(t)$, mais dont les morphismes structuraux différent. En particulier, $\widetilde{\mathbb{F}}$ n'est pas isomorphe à $\mathbb{F}_{\mathbb{R}}$ comme module de persistance (ce que l'on peut vérifier élémentairement et qui découle aussi de la proposition 2.3 ci-dessous).

La notion de module de persistance que nous avons donnée est un cas particulier associé à la structure d'ensemble ordonné $(\mathbb{R}, \leqslant)$. Plus généralement, soit $(S, <)$ un ensemble partiellement ordonné. On lui associe la catégorie $S^<$ dont les objets sont les éléments de S et les ensembles de morphismes

$$\mathrm{Hom}_{S^<}(s,t) = \begin{cases} \{*\} & \text{si } s = t \text{ ou } s < t\,; \\ \emptyset & \text{sinon.} \end{cases}$$

Autrement dit, il y a un unique morphisme de s vers t si $s \leqslant t$ et aucun sinon. L'unicité garantit que l'on obtient bien une catégorie.

Rappelons que si $\mathscr{C}$ est une petite catégorie et $\mathscr{D}$ une catégorie, la catégorie $\mathrm{Fun}(\mathscr{C}, \mathscr{D})$ est la catégorie dont les objets sont les foncteurs de $\mathscr{C}$ vers $\mathscr{D}$ et les morphismes les transformations naturelles entre foncteurs.

On peut réécrire la notion de module de persistance comme suit.

Lemme 1.4. *La catégorie des modules de persistance de la définition 1.1 est la catégorie* $\mathrm{Fun}(\mathbb{R}^<, \mathbb{F}-\mathrm{Vect})$ *des foncteurs de la catégorie associée à* $(\mathbb{R}, \leqslant)$ *vers celle des* $\mathbb{F}$*-espaces vectoriels.*

On peut également remplacer la catégorie des $\mathbb{F}$-espaces vectoriels par toute catégorie. Ainsi on a la notion générale suivante.