Journées mathématiques X-UPS

Année 2024

Anayse topologique de données

Mathieu CARRIÈRE

Théorie de la persistance (2)

Journées mathématiques X-UPS (2024), p. 47-63.
https://doi.org/10.5802/xups.2024-04

Les Éditions de l'École polytechnique
Route de Saclay
F-91128 PALAISEAU CEDEX
https://www.editions.polytechnique.fr

Centre de mathématiques Laurent Schwartz
CMLS, École polytechnique, CNRS,
Institut polytechnique de Paris
F-91128 PALAISEAU CEDEX
https://portail.polytechnique.edu/cmls/

Publication membre du
Centre Mersenne pour l'édition scientifique ouverte
www.centre-mersenne.org

4. Théorie de la persistance (2/2) : stabilité

Nous avons vu au chapitre 3 que les diagrammes de persistance sont bien définis (dans le cas des modules pdf), et qu'ils sont calculables en temps cubique pour les filtrations simpliciales. Le but de ce chapitre est d'explorer le troisième pilier de la théorie de la persistance, à savoir le théorème de stabilité, qui stipule que les diagrammes de persistance issus des sous-niveaux de deux fonctions proches en norme infinie sont eux-mêmes proches, au sens d'une certaine distance entre diagrammes, que nous allons définir en section 4.1. Nous formulerons ensuite le théorème de stabilité en section 4.2, et nous verrons ses répercussions directes en inférence géométrique et statistique dans les sections 4.3 et 4.4. Enfin, nous clôturerons ce chapitre en abordant la version algébrique et générale du théorème de stabilité au niveau des modules de persistance, via la distance d'entrelacement, en section 4.5.

Remarque 4.1. — *Dans ce chapitre, et dans un souci de simplification, nous nous restreignons à $T = \mathbb{R}$, $\mathbb{K} = \mathbb{Z}/2\mathbb{Z}$, et aux diagrammes de cardinaux finis, et dont les coordonnées des points sont elles-aussi finies. Ces diagrammes apparaissent d'ailleurs très fréquemment en pratique, lorsqu'ils sont calculés sur des jeux de données de taille finie. On dénote par* Dgm *l'ensemble de ces diagrammes. Remarquons pour finir que la plupart des notions définies dans ce chapitre se généralisent sans problème aux diagrammes avec une infinité de points, et/ou des points dans le plan étendu $\mathbb{R}\times]-\infty,+\infty]$.*

4.1. Distances entre diagrammes de persistance. — Une manière naturelle de comparer des diagrammes de persistance consiste à associer les points des deux diagrammes entre eux, et de mesurer le coût d'un tel appariemment en calculant les distances entre les points qui ont été ainsi mis en correspondance. Cette approche, qui sous-tend effectivement la définition des distances usuelles entre diagrammes, nécessite cependant de ne chercher que des correspondances *partielles* entre les points des diagrammes, car on ne peut pas garantir a priori que les cardinaux des diagrammes sont égaux.

Définition 4.1. — *Soient $D, D' \in$ Dgm deux diagrammes de persistance. Une* correspondance partielle *$\mathcal{C}$ entre D et D' est un sous-ensemble du produit cartésien $D \times D'$ tel que les projections*

$$\pi_1 : \begin{cases} \mathcal{C} \to D \\ (p,p') \mapsto p \end{cases} \qquad \textit{et} \qquad \pi_2 : \begin{cases} \mathcal{C} \to D' \\ (p,p') \mapsto p' \end{cases}$$

sont injectives. L'ensemble des correspondances partielles entre D et D' est dénoté $\mathcal{C}(D, D')$.

En outre, le coût d'ordre $q \in \mathbb{N}$ *associé à la correspondance partielle $\mathcal{C}$, et dénoté $c_q(\mathcal{C})$, se calcule ainsi :*

$$c_q(\mathcal{C}) := \left(\sum_{(p,p')\in\mathcal{C}} \|p - p'\|_\infty^q + c_q^1(\mathcal{C}) + c_q^2(\mathcal{C}) \right)^{1/q},$$

où les coûts additionnels $c_q^1(\mathcal{C}), c_q^2(\mathcal{C})$ sont définis par

$$c_q^1(\mathcal{C}) := \sum_{p\in D\setminus \mathrm{im}(\pi_1)} \|p - \pi_\Delta(p)\|_\infty^q,$$
$$c_q^2(\mathcal{C}) := \sum_{p'\in D'\setminus \mathrm{im}(\pi_2)} \|p' - \pi_\Delta(p')\|_\infty^q,$$

où $\pi_\Delta(\cdot)$ dénote la projection sur la diagonale.

Le coût d'une correspondance partielle $\mathcal{C}$ consiste donc à additionner les distances entre les points mis en correspondance par $\mathcal{C}$, ainsi que les distances à la diagonale des points restants. Il est en effet bon de rappeler ici que la distance à la diagonale joue un rôle particulier dans la théorie de la persistance, en cela qu'elle correspond de manière équivalente à la longueur des barres dans le code-barre, et donc à la durée de vie des différentes classes d'homologie persistante. Voir la figure 18 pour un exemple de correspondance.

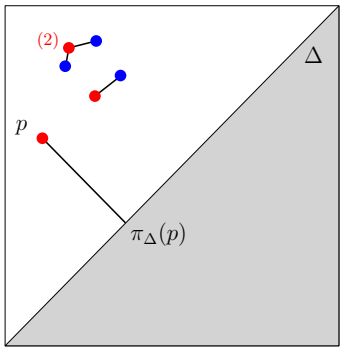


FIGURE 18. Exemple de correspondance partielle entre deux diagrammes. En haut du plan, deux points bleus sont appariés au même point rouge de multiplicité deux.

Remarque 4.2. — *Étant donné deux diagrammes de persistance D et D', la correspondance partielle triviale $\mathcal{C} = \varnothing \in \mathcal{C}(D, D')$ entre D et D' a donc pour coût la somme des distances à la diagonale des points de $D \cup D'$.*

Une fois le coût d'une correspondance bien défini, la distance la plus naturelle entre diagrammes de persistance consiste finalement à calculer le coût du meilleur appariemment entre leurs points.

Définition 4.2. — *Soient D et D' deux diagrammes de persistance. Soit $q \in \mathbb{N}^*$. La* distance de Wasserstein d'ordre q *entre D et D' est définie par :*

$$d_q(D, D') := \inf \{c_q(\mathcal{C}) \,|\, \mathcal{C} \in \mathcal{C}(D, D')\}. \tag{10}$$

Lorsqu'on laisse q tendre vers $+\infty$, la distance de Wasserstein d'ordre q tend vers une distance particulière et importante, appelée la *distance du goulot de bouteille* d_b :

Définition 4.3. — *Soient D et D' deux diagrammes de persistance. La* distance du goulot de bouteille *entre D et D' est définie par :*

$$d_b(D, D') := \inf \{c_\infty(\mathcal{C}) \,|\, \mathcal{C} \in \mathcal{C}(D, D')\}, \tag{11}$$

où $c_\infty(\mathcal{C}) := \sup \{\|p - p'\|_\infty \,|\, (p, p') \in \mathcal{C}\} \cup \{\|p - \pi_\Delta(p)\|_\infty \,|\, p \in D \setminus \mathrm{im}(\pi_1)\} \cup \{\|p' - \pi_\Delta(p')\|_\infty \,|\, p' \in D' \setminus \mathrm{im}(\pi_2)\}$.

4.2. Théorème de stabilité des diagrammes de persistance. — L'intérêt principal de la distance du goulot de bouteille est qu'elle permet d'énoncer simplement un résultat de stabilité, ou de robustesse, associé aux diagrammes de persistance issus des sous-niveaux de fonctions continues à valeurs réelles. Ce théorème est apparu pour la première fois dans Cohen-Steiner et al. (2007), et a ensuite donné lieu à plusieurs versions plus générales. On en trouvera une preuve dans (Chazal et al., 2016, Section 5.6).

Théorème 4.1 **(admis).** — *Soit X un complexe simplicial compact, et soient $f, g : X \to \mathbb{R}$ deux fonctions continues à valeurs réelles définies sur X. Soient $D(f)$ et $D(g)$ les diagrammes de persistance obtenus à partir des filtrations issues des sous-niveaux de f et g. Alors, on a :*

$$d_b(D(f), D(g)) \leq \|f - g\|_\infty, \tag{12}$$

où $\|f - g\|_\infty := \max \{|f(x) - g(x)| \,|\, x \in X\}$.

Il est à noter que les distances de Wasserstein satisfont elles aussi à un théorème de stabilité, mais la borne supérieure est plus difficile à énoncer, voir à ce sujet (Mileyko et al., 2011, Proposition 5). Ce théorème est d'une importance capitale au sein de la théorie de la persistance pour plusieurs raisons.

D'une part, il permet de formaliser la notion d'importance attachée à la longueur des barres d'un code-barre (ou des distances à la diagonale d'un diagramme de persistance) : lorsque les fonctions ne diffèrent que légèrement, le théorème de stabilité impose aux diagrammes d'être proches au sens de la distance du goulot de bouteille, ce qui en retour signifie qu'il existe deux sous-ensembles de points dans ces diagrammes dont les coordonnées sont similaires, et que les points restants sont nécessairement près de la diagonale, de telle manière que leurs coûts ne pèsent pas trop sur la distance du goulot de bouteille. *Les distances à la diagonale faibles caractérisent donc les petites oscillations, tandis que le signal global se lit dans les points éloignés de la diagonale.* Voir la figure 19.

D'autre part, ce théorème a des répercussions directes en inférence géométrique et en statistique, que nous allons maintenant explorer.

4.3. Application aux filtrations de Čech. — Une des applications les plus courantes de la théorie de la persistance est l'inférence géométrique sur des nuages de points, où la tâche consiste à identifier les structures géométriques d'un espace topologique à partir d'un échantillonnage de cet espace uniquement. Cela peut permettre d'identifier que les données observées sont en réalité tirées sur un espace qui a la topologie, par exemple, d'un tore, comme cela arrive couramment lorsque l'on travaille sur des données angulaires (puisqu'un tore n'est en définitive qu'un produit cartésien d'intervalles de la forme $[0, 2\pi]$). Voir par exemple la figure 22. Une telle inférence permet ensuite de raffiner les modèles ou les descripteurs construits sur ces données, comme le montrent les nombreux modèles génératifs et prédictifs spécifiquement optimisés pour traiter des données vivant sur des sphères, des tores, des espaces hyperboliques, etc.

Une technique courante pour caractériser la géométrie et la topologie d'un nuage de points consiste à étudier les sous-niveaux de la fonction distance au nuage, qui ensemble forment une filtration dont la version combinatoire est appelée la filtration de Čech.

Définition 4.4. *— Soit $P \subset \mathbb{R}^d$ un nuage de points fini (plus généralement un sous-espace compact) de $\mathbb{R}^d$. La* fonction distance à P,

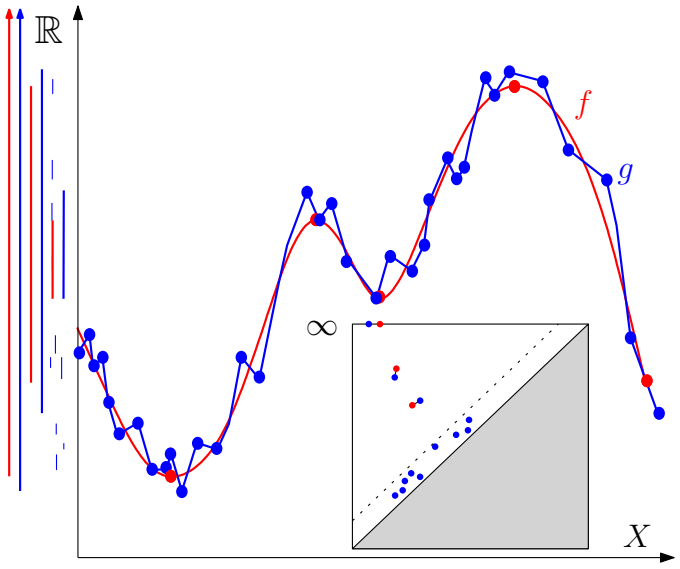


FIGURE 19. Deux fonctions continues définies sur (une triangulation de) la droite réelle, leurs code-barres, et leurs diagrammes de persistance. Comme ces deux fonctions sont proches en norme infinie, leurs diagrammes de persistance sont à distance du goulot de bouteille faible ; en effet, ils ne diffèrent que via des points proches de la diagonale.

dénotée f_P, est définie par :

$$f_P : x \in \mathbb{R}^d \mapsto \min_{p \in P} \|x - p\|_2.$$

La filtration induite par les sous-niveaux de f_P est donc une filtration de $\mathbb{R}^d$ qui consiste à faire grossir des boules centrées sur les points de P. Voir la figure 20. En particulier, cette filtration permet de capturer des structures topologiques et géométriques de P à plusieurs échelles : en effet, les groupes d'homologie calculés sur l'union des boules pour un rayon donné ne sont pas nécessairement isomorphes à ceux calculés pour un rayon plus petit ou plus grand. En un sens, l'homologie persistante évite de devoir choisir et fixer un rayon en capturant les groupes d'homologie à toutes les échelles, ou rayons, possibles, et en encodant les classes d'homologie qui persistent durant certaines plages de rayons.

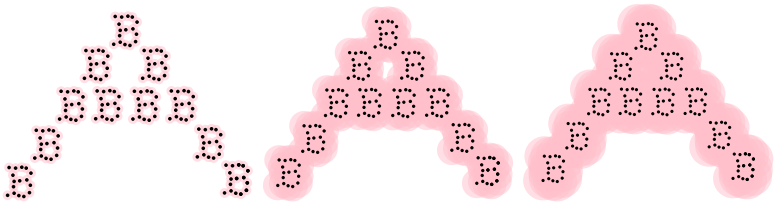


FIGURE 20. Exemple de nuage de points dont la topologie de l'union des boules change en fonction du rayon : pour des rayons faibles, l'union des boules a la même topologie que plusieurs lettres 'B' (gauche), pour des rayons plus grands, la topologie de l'union est celle d'une seule lettre 'A' (milieu), et pour des rayons extrêmes, l'union est formée d'une seule composante connexe sans autre structures topologiques particulières.

Bien que théoriquement satisfaisante, il est évidemment impossible de filtrer complètement l'espace Euclidien $\mathbb{R}^d$ en pratique. Heureusement, on peut démontrer [5] que les sous-niveaux de f_P ont les mêmes groupes d'homologie que les complexes simpliciaux issus d'une filtration calculable qui s'appelle la *filtration de Čech*.

Définition 4.5. — *Soit $P = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ un nuage de points fini de $\mathbb{R}^d$. La filtration de Čech associée à P est la filtration $\{\Delta_n^{P,\mathrm{Cech}}(r) \,|\, r \geq 0\}$ du simplexe complet Δ_n^P à $n = |P|$ sommets identifiés bijectivement aux points de P, définie, pour tout simplexe $\sigma \in \Delta_n^P$, par :*

$$\sigma = \{x_{i_1}, \ldots, x_{i_k}\} \in \Delta_n^{P,\mathrm{Cech}}(r) \Longleftrightarrow \bigcap_{j=1}^{k} B(x_{i_j}, r) \neq \varnothing,$$

où $B(x,r) := \{y \in \mathbb{R}^d \,|\, \|x - y\| \leq r\}$ désigne la boule Euclidienne de rayon r. Le diagramme de persistance associé est dénoté $D_{\mathrm{Cech}}(P)$.

Voir la figure 23. Le diagramme de persistance de Čech permet donc de fournir une estimation de la topologie d'un espace à partir seulement de complexes simpliciaux calculés sur un nuage de points issu de cet espace.

5. La preuve utilise un résultat important appelé le *théorème du nerf*.

De plus, étant donnés deux nuages de points P et P', une application directe du théorème de stabilité permet d'obtenir :

$$d_b(D_{\text{Cech}}(P), D_{\text{Cech}}(P')) = d_b(D(f_P), D(f_{P'})) \leq \|f_P - f_{P'}\|_\infty.$$

La raison pour laquelle ce résultat est important est qu'il permet de relier les diagrammes de Čech à une distance très utile en géométrie, la *distance de Hausdorff*.

Définition 4.6. — *Soient X, Y deux sous-espaces compacts de $\mathbb{R}^d$. La* distance de Hausdorff *entre X et Y est définie par :*

$$d_H(X,Y) := \max\{\max\{f_Y(x) \mid x \in X\}, \max\{f_X(y) \mid y \in Y\}\},$$

où f_X, f_Y sont définies comme en Définition 4.4.

Lemme 4.1. — *Soient X, Y deux sous-espaces compacts de $\mathbb{R}^d$. Alors, on a :*

$$d_H(X,Y) = \|f_X - f_Y\|_\infty.$$

Démonstration. — Montrons d'abord que $d_H(X,Y) \leq \|f_X - f_Y\|_\infty$. Supposons que $d_H(X,Y)$ soit atteinte par un certain $x^* \in X$. Alors, on a :

$$\begin{aligned} d_H(X,Y) &= f_Y(x^*) \\ &= f_Y(x^*) - f_X(x^*) \\ &\leq |f_Y(x^*) - f_X(x^*)| \\ &\leq \|f_Y - f_X\|_\infty. \end{aligned}$$

Le même argument peut être appliqué au cas où $d_H(X,Y)$ est atteinte par un certain $y^* \in Y$, en intervertissant simplement X et Y.

Montrons maintenant que $d_H(X,Y) \geq \|f_X - f_Y\|_\infty$. Soit $z \in \mathbb{R}^d$. Alors, on a :

$$\begin{aligned} f_X(z) - f_Y(z) &= \|z - \pi_X(z)\|_2 - \|z - \pi_Y(z)\|_2 \\ &\leq \|z - \pi_X(\pi_Y(z))\|_2 - \|z - \pi_Y(z)\|_2 \\ &\leq \|\pi_X(\pi_Y(z)) - \pi_Y(z)\|_2 \text{ par l'inégalité triangulaire} \\ &= f_X(\pi_Y(z)) \\ &\leq \sup\{f_X(y) \mid y \in Y\} \\ &\leq d_H(X,Y). \end{aligned}$$

La même séquence d'inégalités peut être appliquée pour borner $f_Y(z) - f_X(z)$, en intervertissant simplement X et Y. Ceci permet d'obtenir $|f_X(z) - f_Y(z)| \leq d_H(X,Y)$ pour tout $z \in \mathbb{R}^d$, et donc $\|f_X - f_Y\|_\infty \leq d_H(X,Y)$. □

Ce lemme permet d'obtenir le résultat suivant, dont nous allons observer les retombées statistiques dans la section suivante.

Corollaire 4.1. — *Soient X un sous-espace compact de $\mathbb{R}^d$, et $\hat{X}_n \subset X$ un échantillonage de X à n points. Alors, on a :*

$$d_b(D(f_X), D_{\text{Cech}}(\hat{X}_n)) \leq d_H(X, \hat{X}_n).$$

4.4. Inférence géométrique et intervalles de confiance. — De manière générale, étant donné un estimateur $\hat{\theta}_n$ d'une quantité cible θ^* associée à une distribution de probabilité (comme, par exemple, sa moyenne), et calculé sur un échantillon de n points issu de cette distribution, une question centrale de la statistique consiste à construire des *intervalles de confiance* pour celui-ci, c'est-à-dire, pour un seuil de confiance $\alpha \in [0, 1]$ donné, de trouver $d_\alpha > 0$ tel que $\mathbb{P}(|\hat{\theta}_n - \theta^*| > d_\alpha) \leq \alpha$. L'inférence géométrique consiste ainsi à calculer des estimateurs pour déterminer (avec un certain niveau de probabilité) la topologie du support de certaines distributions de probabilité. Supposons par exemple que ce support soit un sous-espace compact X de $\mathbb{R}^d$, et que sa topologie soit caractérisée par $D(f_X)$. Un estimateur possible, suggéré par le corollaire 4.1, est donc $D_{\text{Cech}}(\hat{X}_n)$. De plus, ce corollaire permet de transférer le calcul d'intervalles de confiance pour $d_b(D(f_X), D_{\text{Cech}}(\hat{X}_n))$ à celui d'intervalles de confiance pour $d_H(X, \hat{X}_n)$. En effet, en supposant connu un certain d_α associé à un seuil de confiance α pour $d_H(X, \hat{X}_n)$, il s'ensuit trivialement du corollaire 4.1 que

$$\mathbb{P}(d_b(D(f_X), D_{\text{Cech}}(\hat{X}_n)) > d_\alpha) \leq \mathbb{P}(d_H(X, Y) > d_\alpha) \leq \alpha.$$

En outre, la distance de Hausdorff est un objet bien connu en inférence géométrique, car il caractérise à quel point un échantillonage est uniforme—voir sur ce point la figure 21, et ses intervalles de confiance sont faciles à obtenir pour des mesures de probabilité appelées mesures (a, b)-standard.

Dans cette section, nous allons donc présenter quelques résultats d'inférence géométrique via la distance de Hausdorff, et leur utilisation pour les diagrammes de persistance. Ces résultats sont tirés et adaptés de Chazal et al. (2015); Fasy et al. (2014), que l'on pourra consulter pour une exposition et des preuves plus détaillées.

Définition 4.7. — *Soit μ une mesure de probabilité à support compact $X \subset \mathbb{R}^d$. Soient $a, b > 0$ deux réels positifs. La mesure μ est*

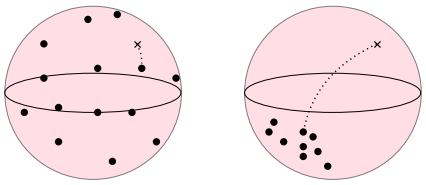


FIGURE 21. Exemple de calcul de la distance de Hausdorff sur deux échantillonnages d'une sphère. Sur l'échantillon de gauche, la distance est faible car n'importe quel point de la sphère (comme celui représenté par une croix) est proche d'au moins un point de l'échantillon. Au contraire, pour l'échantillon de droite, il existe des points de la sphère pour lesquels le point de l'échantillon le plus proche reste quand même très éloigné.

appelée (a,b)*-standard si on a, pour tout* $x \in X$ *et* $r > 0$ *:*

$$\mu(B(x,r)) \leq \min\{1, ar^b\}.$$

L'intérêt des mesures (a,b)-standards est qu'elles permettent de contrôler des quantités aisément reliables à la distance de Hausdorff entre le support et un échantillon de celui-ci, dénommées *nombre de recouvrement* (*covering number* en anglais) et *nombre de stockage* (*packing number* en anglais).

Définition 4.8. — *Soit* X *un sous-espace compact de* $\mathbb{R}^d$ *et* $r > 0$*. Le* nombre de recouvrement *de* X *de rayon* r *est défini par :*

$$n_{\mathrm{co}}(X,r) := \min\{k \in \mathbb{N}^* \mid \exists (x_1,\ldots,x_k) \in X^k \text{ tels que } X \subseteq \bigcup_{i=1}^{k} B(x_i,r)\}.$$

Le nombre de stockage *de* X *est défini par :*

$$n_{\mathrm{st}}(X,r) := \max\{k \in \mathbb{N}^* \mid \exists (x_1,\ldots,x_k) \in X^k \text{ tels que } \forall 1 \leq i \leq k, B(x_i,r) \subseteq X \text{ et } \forall 1 \leq i \neq j \leq k, B(x_i,r) \cap B(x_j,r) = \varnothing\}.$$

Le nombre de recouvrement de rayon r est donc le nombre minimal de boules de rayon r avec lesquelles on peut entièrement couvrir X, tandis que le nombre de stockage de rayon r est le nombre maximal de boules de rayon r que l'on peut arranger dans X sans qu'elles ne s'intersectent. Ces deux quantités sont reliées de la manière suivante :

Lemme 4.2. — *Soit X un sous-espace compact de $\mathbb{R}^d$ et $r > 0$. Alors, on a :*

$$n_{\mathrm{co}}(X, 2r) \leq n_{\mathrm{st}}(X, r).$$

Démonstration. — Soit $k \in \mathbb{N}^*$ un entier réalisant le nombre de stockage, et $c_1, \ldots, c_k$ les centres des boules correspondantes. Supposons par l'absurde que $\bigcup_{i=1}^{k} B(c_i, 2r)$ ne recouvre pas X. Il existe donc un certain $x \in X$ tel que $x \notin \bigcup_{i=1}^{k} B(c_i, 2r)$, c'est-à-dire tel que $\|x - c_i\|_2 > 2r$ pour tout $1 \leq i \leq k$. Il s'ensuit que la boule $B(x, r)$ n'intersecte aucune des boules $B(c_i, r)$ (sinon la distance $\|x - c_i\|_2$ serait inférieure à $2r$ par l'inégalité triangulaire), et donc que l'on peut augmenter le nombre de stockage d'une unité. C'est une contradiction, car ce nombre est maximal. Ainsi, $\bigcup_{i=1}^{k} B(c_i, 2r)$ recouvre entièrement X, d'où l'on peut tirer l'inégalité désirée. □

Pour les mesures (a, b)-standards, on peut facilement borner supérieurement le nombre de stockage et le nombre de recouvrement.

Lemme 4.3. — *Soit μ une mesure (a, b)-standard de support compact $X \subset \mathbb{R}^d$. Alors $n_{\mathrm{co}}(X, 2r) \leq n_{\mathrm{st}}(X, r) \leq \max\{1, 1/(ar^b)\}$.*

Démonstration. — Le résultat étant trivial pour $r \geq a^{-1/b}$, on suppose que $r < a^{-1/b}$. Soit $k \in \mathbb{N}^*$ un entier réalisant le nombre de stockage, et $c_1, \ldots, c_k$ les centres des boules correspondantes. On a donc

$$\begin{aligned} \mu\left(\bigcup_{i=1}^{k} B(c_i, r)\right) &= \sum_{i=1}^{k} \mu\left(B(c_i, r)\right) \text{ car les boules sont disjointes} \\ &\leq k \cdot ar^b \leq 1. \end{aligned}$$

Les inégalités recherchées découlent directement des deux dernières inégalités de la séquence. □

On peut finalement utiliser le nombre de recouvrement pour le calcul des intervalles de confiance de la distance de Hausdorff, et obtenir le résultat suivant :

Proposition 4.1. — *Soit μ une mesure (a, b)-standard de support compact $X \subset \mathbb{R}^d$. Soit $\hat{X}_n$ un échantillonage de X à n points. Soit $d_\alpha > 0$. Alors on a :*

$$\mathbb{P}(d_H(X, \hat{X}_n) > d_\alpha) \leq \min\left\{1, \frac{4^b}{ad_\alpha^b} \exp(-na \cdot (d_\alpha/2)^b)\right\}.$$

Démonstration. — Cette preuve passe par une formulation équivalente de la distance de Hausdorff, que l'on laisse en exercice au lecteur :

Exercice 4.1. — *Montrer que :*

$$d_H(X, \hat{X}_n) = \inf\{\varepsilon > 0 \,|\, X \subseteq \bigcup_{\hat{x}\in\hat{X}_n} B(\hat{x}, \varepsilon)\}.$$

On a donc

$$\mathbb{P}(d_H(X, \hat{X}_n) > d_\alpha) = \mathbb{P}(\sup_{x\in X} \min_{\hat{x}\in\hat{X}_n} \|x - \hat{x}\|_2 > d_\alpha).$$

Soit $k \in \mathbb{N}^*$ un entier réalisant le nombre de recouvrement de rayon $d_\alpha/2$, et $c_1, \ldots, c_k$ les centres des boules correspondantes. De plus, pour tout $x \in X$, soit $c_x \in \{c_1, \ldots, c_k\}$ le centre tel que $x \in B(c_x, d_\alpha/2)$. On a :

$$\begin{aligned}\min_{\hat{x}\in\hat{X}_n} \|x - \hat{x}\|_2 &\leq \|x - \hat{x}_i\|_2 \leq \|x - c_x\|_2 + \|c_x - \hat{x}_i\|_2, 1 \leq i \leq n,\\ &\leq \|x - c_x\|_2 + \max_{1\leq j\leq k} \min_{\hat{x}\in\hat{X}_n} \|c_j - \hat{x}\|_2\\ &\leq d_\alpha/2 + \max_{1\leq j\leq k} \min_{\hat{x}\in\hat{X}_n} \|c_j - \hat{x}\|_2\end{aligned}$$

Ainsi on obtient

$$\sup_{x\in X} \min_{\hat{x}\in\hat{X}_n} \|x - \hat{x}\|_2 \leq d_\alpha/2 + \max_{1\leq j\leq k} \min_{\hat{x}\in\hat{X}_n} \|c_j - \hat{x}\|_2,$$

et donc

$$\mathbb{P}(d_H(X, \hat{X}_n) > d_\alpha) \leq \mathbb{P}(\max_{1\leq j\leq k} \min_{\hat{x}\in\hat{X}_n} \|c_j - \hat{x}\|_2 > d_\alpha/2).$$

La probabilité $\mathbb{P}(\max_{1\leq j\leq k} \min_{\hat{x}\in\hat{X}_n} \|c_j - \hat{x}\|_2 > d_\alpha/2)$ est la probabilité qu'il existe un centre pour lequel tous les points de l'échantillon sont à distance au moins $d_\alpha/2$. Ainsi, si l'on dénote $E_{j,i}$ l'évènement $\{\hat{x}_i \in B(c_j, d_\alpha/2)\}$, on a :

$$\begin{aligned}\mathbb{P}(\max_{1\leq j\leq k} \min_{\hat{x}\in\hat{X}_n} \|c_j - \hat{x}\|_2 > d_\alpha/2) &= \mathbb{P}\left(\bigcup_{j=1}^{k} \bigcap_{i=1}^{n} E_{j,i}^c\right)\\ \leq \sum_{j=1}^{k} \mathbb{P}\left(\bigcap_{i=1}^{n} E_{j,i}^c\right) &= \sum_{j=1}^{k} [1 - \mu(B(c_j, d_\alpha/2))]^n\\ &= k \cdot \left[1 - a(d_\alpha/2)^b\right]^n.\end{aligned}$$

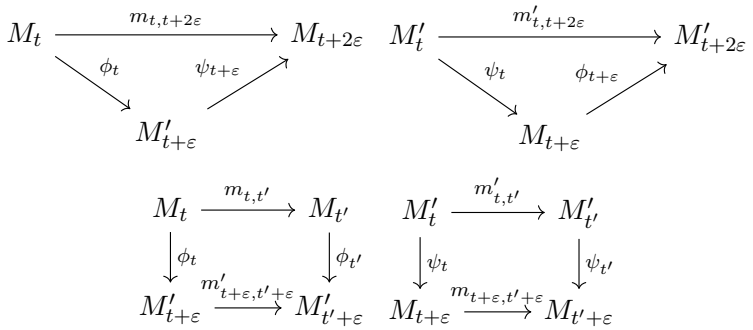
Sachant que $k = n_{\mathrm{co}}(X, d_\alpha/2) \leq \frac{4^b}{a d_\alpha^b}$ d'après le lemme 4.3, et en utilisant l'inégalité $(1-x)^n \leq \exp(-nx)$ pour tout $x \in [0,1]$, on obtient le résultat désiré. □

La proposition 4.1 peut ensuite être utilisée pour le calcul de la *vitesse de convergence*, c'est-à-dire du calcul de l'espérance $\mathbb{E}\left[d_b(D(f_X), D_{\mathrm{Cech}}(\hat{X}_n))\right]$, via l'identité $\mathbb{E}[d] = \int_\alpha \mathbb{P}(d \geq \alpha)\mathrm{d}\alpha$, pour toute variable aléatoire d positive. En outre, il est même possible de démontrer que $D_{\mathrm{Cech}}(\hat{X}_n)$ est un estimateur *optimal*, au sens où il n'existe pas d'estimateur de $D(f_X)$ dont la vitesse de convergence soit meilleure.

La proposition 4.1 est néanmoins limitée par la nécessité de connaître a et b pour pouvoir calculer des intervalles de confiance. Bien que l'on ne connaisse pas forcément leurs valeurs *a priori*, il est toutefois possible de les estimer numériquement. Une fois calculés, les intervalles de confiance peuvent être ensuite visualisés directement sur les diagrammes de persistance, comme montré dans la figure 22.

4.5. Théorème de stabilité des modules de persistance. — La stabilité des diagrammes de persistance, telle que présentée en section 4.2, découle en réalité d'une formulation algébrique de la stabilité, qui, elle, s'énonce au niveau des modules de persistance via la *distance d'entrelacement*.

Définition 4.9. — *Soient* $M = \{\{M_t\}_{t\in\mathbb{R}}, \{m_{t,t'}\}_{t\leq t'\in\mathbb{R}}\}$ *et* $M' = \{\{M'_t\}_{t\in\mathbb{R}}, \{m'_{t,t'}\}_{t\leq t'\in\mathbb{R}}\}$ *deux modules de persistance indexés sur* $\mathbb{R}$. *Soit* $\varepsilon > 0$. *Les modules* M *et* M' *sont* ε-entrelacés *si et seulement si il existe deux familles d'applications linéaires* $\{\phi_t : M_t \to M'_{t+\varepsilon} \,|\, t \in \mathbb{R}\}$ *et* $\{\psi_t : M'_t \to M_{t+\varepsilon} \,|\, t \in \mathbb{R}\}$ *telles que les diagrammes suivants commutent pour tout* $t \leq t' \in \mathbb{R}$ *:*



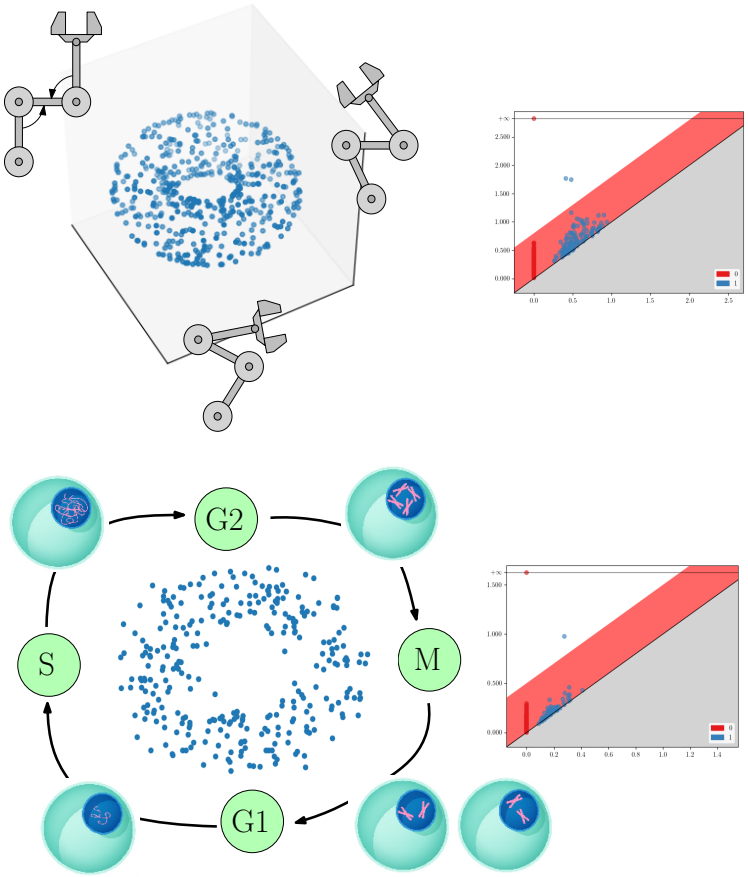


FIGURE 22. Deux exemples d'inférence géométrique sur des données réelles. Dans la première (**haut**), différentes configurations d'un bras robotique sont collectées. Une configuration étant caractérisée par les deux angles aux jointures, les données vivent sur un tore. Dans la deuxième (**bas**), des cellules, ainsi que l'expression de leurs gènes, ont été prélevées à différentes phases du cycle cellulaire, les données vivent donc sur un cercle (plongé dans $\mathbb{R}^{\#\text{gènes}}$). Dans chacun des deux cas, on peut calculer le diagramme de Čech, et, pour un seuil de confiance α donné (dans cette figure, $\alpha = 0.05\%$), visualiser les points correspondants à du bruit (dû par exemple à l'échantillonnage) via une bande de hauteur $2d_\alpha$ au-dessus de la diagonale. On voit que les points caractéristiques de la géométrie des espaces sous-jacents (deux points en dimension 1 et un point en dimension 0 pour le tore, un point en dimension 0 et un point en dimension 1 pour le cercle) sont bien situés en dehors de la bande : ils sont sûrs à 95%.

La distance d'entrelacement *entre M et M' est alors définie comme suit :*

$$d_I(M, M') := \inf \{\varepsilon > 0 \mid M, M' \text{ sont } \varepsilon\text{-entrelacés}\}.$$

Lorsque les modules de persistance sont pdf, et donc que l'on peut calculer leurs diagrammes de persistance, on peut montrer que la distance d'entrelacement est en fait égale à la distance du goulot de bouteille.

***Proposition 4.2* (admise).** — *Soient M, M' deux modules de persistance pdf, et $D(M), D(M')$ les diagrammes de persistance correspondants. Alors, on a :*

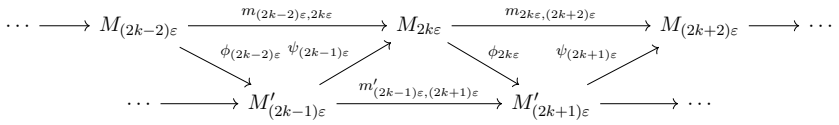
$$d_b(D(M), D(M')) = d_I(M, M').$$

Une preuve de ce résultat est donnée par (Chazal et al., 2016, Section 5.4).

Remarque 4.3. — *Il est relativement simple de prouver une version faible de ce théorème, qui stipule que :*

$$d_b(D(M), D(M')) \leq 8 d_I(M, M').$$

Pour ce faire, il faut d'abord démontrer que l'on peut discrétiser un module continu tout en contrôlant le coût induit sur les diagrammes de persistance. Formellement, pour tout module de persistance pdf $M = \{\{M_t\}_{t\in\mathbb{R}}, \{m_{t,t'}\}_{t\leq t'\in\mathbb{R}}\}$ et pour un pas de discrétisation $\varepsilon > 0$, on peut définir sa discrétisation *$\tilde{M}^{\varepsilon,t_0}$, $t_0 < \varepsilon$, par : $\tilde{M}_{\varepsilon,t_0} = \{\{M_{k\varepsilon+t_0}\}_{k\in\mathbb{Z}}, \{m_{k\varepsilon+t_0,k'\varepsilon+t_0}\}_{k\leq k'\in\mathbb{Z}}\}$, et démontrer que l'on a : $d_b(D(M), D(\tilde{M}_{\varepsilon,t_0})) \leq \varepsilon$ (exercice laissé au lecteur—il suffit de répertorier les changements induits par la discrétisation sur la décomposition de M en modules intervalles). Ensuite, étant donné deux modules pdf M, M', et $\varepsilon > d_I(M, M')$ (en supposant que $d_I(M, M') < \infty$), on peut obtenir le résultat en considérant le diagramme commutatif suivant :*



Le module associé à la ligne supérieure du diagramme est $\tilde{M}_{2\varepsilon,0}$ et celui associé à la ligne inférieure du diagramme est $\tilde{M}'_{2\varepsilon,\varepsilon}$. Appelons M'' le module obtenu en suivant les flèches diagonales du diagramme. On remarque alors que $\tilde{M}_{2\varepsilon,0}$ et $\tilde{M}'_{2\varepsilon,\varepsilon}$ sont aussi des discrétisations

de M'' : $\tilde{M}_{2\varepsilon,0} = \tilde{M}''_{2\varepsilon,0}$ et $\tilde{M}'_{2\varepsilon,\varepsilon} = \tilde{M}''_{2\varepsilon,\varepsilon}$, et on conclut par inégalité triangulaire :

$$\begin{aligned} &d_b(D(M), D(M')) \\ &\leq d_b(D(M), D(\tilde{M}_{2\varepsilon,0})) + d_b(D(\tilde{M}''_{2\varepsilon,0}), D(M'')) \\ &+ d_b(D(M''), D(\tilde{M}''_{2\varepsilon,\varepsilon})) + d_b(D(\tilde{M}'_{2\varepsilon,\varepsilon}), D(M')) \leq 8\varepsilon. \end{aligned}$$

L'intérêt de la proposition 4.2 est qu'elle permet de prouver simplement le théorème 4.1 de stabilité. En effet, si l'on considère deux fonctions $f, g : X \to \mathbb{R}$ définies sur un complexe simplicial compact, et que l'on se donne $\varepsilon > 0$ tel que $\varepsilon > \|f - g\|_\infty$, on obtient aisément

$$f^{-1}((-\infty, t]) \subseteq g^{-1}((-\infty, t + \varepsilon])$$

et

$$g^{-1}((-\infty, t]) \subseteq f^{-1}((-\infty, t + \varepsilon]),$$

pour tout $t \in \mathbb{R}$. La fonctorialité de l'homologie permet ensuite de directement déduire des applications linéaires associées aux inclusions ci-dessus, et donc un ε-entrelacement entre les modules de persistance $M(f)$ et $M(g)$ issus des sous-niveaux de f et g. Ceci permet finalement d'obtenir :

$$d_b(D(f), D(g)) = d_I(M(f), M(g)) \leq \|f - g\|_\infty$$

(ou bien $d_b(D(f), D(g)) \leq 8 d_I(M(f), M(g)) \leq 8\|f-g\|_\infty$ si l'on passe par la version faible présentée en remarque 4.3). Un autre avantage de la distance d'entrelacement est qu'elle permet parfois de contrôler simplement la proximité entre deux modules de persistance issus de filtrations spécifiques, et donc aussi entre leurs diagrammes de persistance (via la proposition 4.2). Une illustration classique de cet avantage concerne les diagrammes de Čech, définis en section 4.3. En effet, en pratique, le calcul effectif de diagrammes de Čech est souvent coûteux (même si possible via l'utilisation de complexes simpliciaux spécifiques appelés complexes alpha). On préfère donc souvent calculer une approximation facilement calculable de la filtration de Čech, appelée *filtration de Vietoris-Rips*.

Définition 4.10. — *Soit $P = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ un nuage de points fini de $\mathbb{R}^d$. La filtration de Vietoris-Rips associée à P est la filtration $\{\Delta_n^{P,\mathrm{Rips}}(r) \,|\, r \geq 0\}$ du simplexe complet Δ_n^P à $n = |P|$ sommets identifiés bijectivement aux points de P, définie, pour tout simplexe $\sigma \in \Delta_n^P$, par :*

$$\sigma = \{x_{i_1}, \ldots, x_{i_k}\} \in \Delta_n^{P,\mathrm{Rips}}(r) \iff \|x_{i_j} - x_{i_{j'}}\|_2 \leq r, \forall 1 \leq j, j' \leq k.$$

Le diagramme de persistance associé est dénoté $D_{\mathrm{Rips}}(P)$.

Lemme 4.4. — *Soit* $r \geq 0$. *Alors, on a les inclusions suivantes entre complexes simpliciaux :*

$$\Delta_n^{P,\mathrm{Rips}}(r) \subseteq \Delta_n^{P,\mathrm{Cech}}(r) \subseteq \Delta_n^{P,\mathrm{Rips}}(2r).$$

Démonstration. — Soit $r \geq 0$. Montrons la première inclusion. Soit $\sigma = \{x_{i_1}, \ldots, x_{i_k}\}$ un k-simplexe de $\Delta_n^{P,\mathrm{Rips}}(r)$. Alors, pour tout $1 \leq j \leq k$, on a $\|x_{i_1} - x_{i_j}\|_2 \leq r$ et donc $x_{i_1} \in B(x_{i_j}, r)$. Ainsi $x_{i_1} \in \bigcap_{j=1}^{k} B(x_{i_j}, r)$ et $\sigma \in \Delta_n^{P,\mathrm{Cech}}(r)$.

Montrons maintenant la deuxième inclusion. Soit $\sigma = \{x_{i_1}, \ldots, x_{i_k}\}$ un k-simplexe de $\Delta_n^{P,\mathrm{Cech}}(r)$. Alors, pour tout $1 \leq j, j' \leq k$, on a $B(x_{i_j}, r) \cap B(x_{i_{j'}}, r) \neq \varnothing$. Soit $z \in B(x_{i_j}, r) \cap B(x_{i_{j'}}, r)$. Alors, par l'inégalité triangulaire, on a :

$$\|x_{i_j} - x_{i_{j'}}\|_2 \leq \|x_{i_j} - z\|_2 + \|z - x_{i_{j'}}\|_2 \leq 2r,$$

et donc $\sigma \in \Delta_n^{P,\mathrm{Rips}}(2r)$. □

Voir la figure 23.

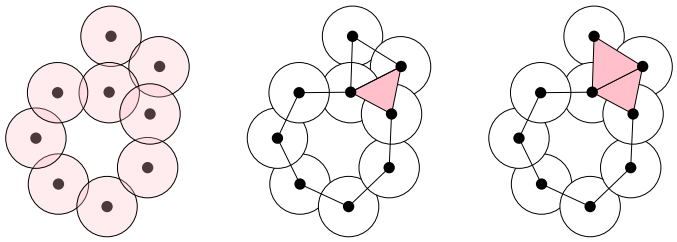


FIGURE 23. Exemple de nuage de points et de boules de rayon r centrées sur les points (**gauche**), son complexe de Čech de rayon r (**centre**), et son complexe de Vietoris-Rips de rayon $2r$ (**droite**). On peut voir que le complexe de Vietoris-Rips contient un simplexe supplémentaire : en effet, il suffit pour les complexes de Vietoris-Rips que toutes les paires de points aient une distance inférieure au rayon pour pouvoir ajouter le simplexe correspondant.

La fonctorialité de l'homologie permet donc d'obtenir directement le corollaire suivant :

Corollaire 4.2. — *Soit $P = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ un nuage de points fini de $\mathbb{R}^d$. Soit $D^{\log}_{\text{Cech}}(P)$ et $D^{\log}_{\text{Rips}}(P)$ les logarithmes*[6] *des diagrammes de Čech et Vietoris-Rips correspondants :*

$$D^{\log}_{\text{Cech}}(P) := \{(\log(p_x), \log(p_y)) \in \mathbb{R}^2 \,|\, (p_x, p_y) \in D_{\text{Cech}}(P)\},$$

$$D^{\log}_{\text{Rips}}(P) := \{(\log(p_x), \log(p_y)) \in \mathbb{R}^2 \,|\, (p_x, p_y) \in D_{\text{Rips}}(P)\}.$$

Alors, on a :

$$d_b(D^{\log}_{\text{Cech}}(P), D^{\log}_{\text{Rips}}(P)) \leq \log(2).$$

Références

Chazal, F., de Silva, V., Glisse, M., and Oudot, S. (2016). *The structure and stability of persistence modules*. SpringerBriefs in Mathematics. Springer-Verlag.

Chazal, F., Glisse, M., Labruère, C., and Michel, B. (2015). Convergence rates for persistence diagram estimation in topological data analysis. *Journal of Machine Learning Research*, 16(110) :3603–3635.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1) :103–120.

Fasy, B., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6) :2301–2339.

Mileyko, Y., Mukherjee, S., and Harer, J. (2011). Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12) :124007.

6. Ces logarithmes sont bien définis car les coordonnées des points des diagrammes de Čech et Vietoris-Rips sont positives par définition.