Analyse topologique de données

GT Math/Info M1 — 8 ECTS

Vadim Lebovici*

S2 - 2024/2025

1 Présentation

Ce groupe de travail propose une introduction à la théorie et aux applications de l'homologie persistante, un outil central en analyse topologique de données. L'objectif de cet outil est d'extraire des informations topologiques de données complexes, qu'elles soient représentées sous forme de nuages de points finis ou de graphes, en construisant une famille croissante d'objets géométriques à partir des données, appelée *filtration* ; cf. fig. 1.

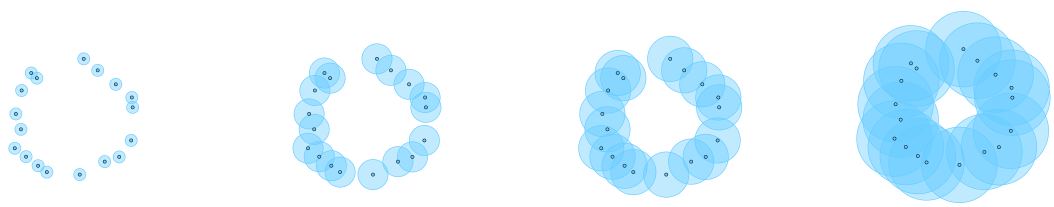


FIGURE 1 – Filtration de Čech d'un nuage de points dans le plan euclidien.

Nous commencerons par introduire les complexes simpliciaux et l'homologie simpliciale, notions fondamentales de la topologie algébrique. L'homologie persistante repose sur l'analyse des groupes d'homologie d'une filtration simpliciale, ce qui permet d'étudier la structure sous-jacente des données à différentes échelles. Nous étudierons ensuite l'homologie persistante d'un point de vue algébrique et sa représentation sous forme de codes-barres, ainsi que les résultats de stabilité associés.

D'un point de vue algorithmique, nous aborderons la construction et le calcul des codes-barres. Nous explorerons ensuite les applications de l'homologie persistante dans divers domaines de l'analyse de données, notamment :

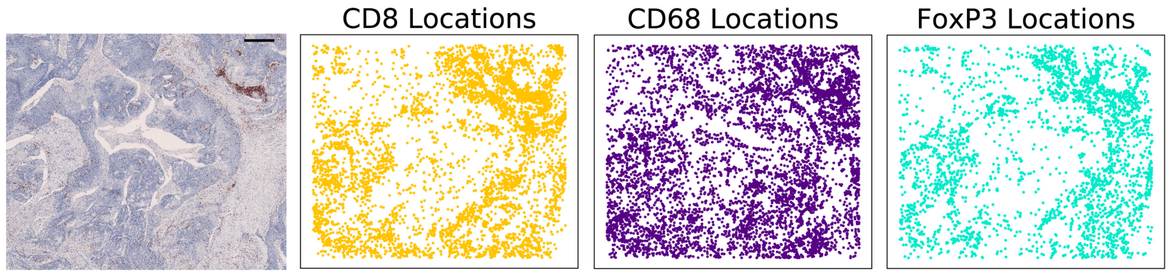
— L'inférence topologique : reconstruire la structure d'un objet topologique inconnu à partir d'échantillons finis de points.

*lebovici@math.univ-paris13.fr

— Le partitionnement (clustering) : détecter des regroupements naturels dans un jeu de données.

— La classification supervisée : prédire l'appartenance à un groupe à partir d'exemples connus.

Des applications concrètes seront présentées, notamment dans les domaines scientifiques et médicaux (cf fig. 2).



(a) (b) (c) (d)

FIGURE 2 – Images numérisées d'immunohistochimie de carcinome épidermoïde (2a) et emplacements des cellules immunitaires de type CD8 (2b), CD68 (2c) et FoxP3 (2d).

Organisation du travail. Le module s'organisera en plusieurs phases :

— *Cours magistraux* pour introduire les concepts fondamentaux.

— *Présentations orales* par les étudiants d'articles clés en analyse topologique des données.

— *Séances d'exercices* pour consolider la compréhension des notions abordées.

— *Travaux pratiques* sur ordinateur afin d'expérimenter les outils étudiés sur des jeux de données réels.

Prérequis. Le contenu des cours et le déroulé des séances seront adaptés à tous·tes· les étudiant·es, de mathématiques comme d'informatique. En fonction du profil, les aspects mathématiques ou algorithmiques des thématiques abordées pourront au choix être approfondies.

Validation. La présence est obligatoire pour valider le cours et la note finale sera la moyenne pondérée d'un contrôle de connaissance (40%) et d'un exposé donné (60%). Pour l'exposé, vous serez évalués sur trois critères : la compréhension et la présentation rigoureuse du contenu de votre exposé, le choix du contenu que vous mettez en valeur dans votre présentation, et enfin la qualité pédagogique de l'exposé.

Préparation de l'exposé. Je vous demande de venir me voir au moins une fois dans la semaine précédant l'exposé, bien à l'avance et en ayant travaillé avant, pour parler de ce que vous traiterez et poser des questions si besoin.

2 Programme

Les séances de 3h seront découpées en 2×1h30. Les trois premières ont un contenu fixé, les autres pourront être modifiées en fonction des aspirations et des goûts des étudiant·es.

1. Mardi 25 mars (cours, TD)
   — Séance introductive, présentation de l'analyse topologique de données
   — Complexe simpliciaux, homologie simpliciale I
2. Mardi 1 avril (cours, TD, TP)
   — Complexe simpliciaux, homologie simpliciale II
   — Approfondissements maths : théorie de l'homologie
   — Approfondissements info : implémentation
3. Mardi 8 avril (cours, TD, TP)
   — Existence et unicité du code-barres (cas finiment présenté) [11]
   — Algorithme de calcul du code-barre [10]
4. Vacances
5. TBA (exposé, TP)
   — Inférence géométrique et filtrations simpliciales (théorie) [10]
   — Inférence géométrique et filtrations simpliciales (pratique) [3, 10]
6. TBA (exposé, TP)
   — Stabilité des codes-barres (théorème d'isométrie) I [4, 11]
   — Sensibilité aux données aberrantes et filtration DTM [2]
7. TBA (exposé, TP)
   — Stabilité des codes-barres (théorème d'isométrie) II [4, 11]
   — Vectorisation des codes-barres, stabilité et performances [1, 9]
8. TBA (exposé, TP)
   — Stabilité Wasserstein pour les fonctions Lipschitz [6]
   — Outils fondés sur la caractéristique d'Euler et multipersistance [8]
9. TBA (exposé, TP)
   — Applications au clustering (ToMaTo) : théorie [10]
   — Applications au clustering (ToMaTo) : pratique [10]
10. TBA (exposé, TP)
    — Différentiation de codes-barres et optimisation topologique [5]
    — Extraction et comparaison de cycles signifiants dans les données [7]

Références

[1] Henry ADAMS et al. “Persistence images : a stable vector representation of persistent homology”. English. In : *J. Mach. Learn. Res.* 18 (2017). Id/No 8, p. 35. ISSN : 1532-4435. URL : `jmlr.csail.mit.edu/papers/v18/16-337.html`.

[2] Hirokazu ANAI et al. “DTM-based filtrations”. English. In : *Topological data analysis. Proceedings of the Abel symposium 2018, Geiranger, Norway, June 4–8, 2018*. Cham : Springer, 2020, p. 33-66. ISBN : 978-3-030-43407-6 ; 978-3-030-43410-6 ; 978-3-030-43408-3. DOI : `10.1007/978-3-030-43408-3_2`.

[3] Ulrich BAUER. “Ripser : efficient computation of Vietoris-Rips persistence barcodes”. In : *J. Appl. Comput. Topol.* 5.3 (2021), p. 391-423. ISSN : 2367-1726. DOI : `10.1007/s41468-021-00071-5`. URL : `https://doi.org/10.1007/s41468-021-00071-5`.

[4] Ulrich BAUER et Michael LESNICK. “Induced matchings and the algebraic stability of persistence barcodes”. English. In : *J. Comput. Geom.* 6.2 (2015), p. 162-191. ISSN : 1920-180X. DOI : `10.20382/jocg.v6i2a9`.

[5] Mathieu CARRIERE et al. “Optimizing persistent homology based functions”. In : *Proceedings of the 38th International Conference on Machine Learning*. Sous la dir. de Marina MEILA et Tong ZHANG. T. 139. Proceedings of Machine Learning Research. PMLR, 2021, p. 1294-1303.

[6] David COHEN-STEINER et al. “Lipschitz functions have L_p-stable persistence”. English. In : *Found. Comput. Math.* 10.2 (2010), p. 127-139. ISSN : 1615-3375. DOI : `10.1007/s10208-010-9060-6`.

[7] Inés GARCÍA-REDONDO, Anthea MONOD et Anna SONG. “Fast topological signal identification and persistent cohomological cycle matching”. English. In : *J. Appl. Comput. Topol.* 8.3 (2024), p. 695-726. ISSN : 2367-1726. DOI : `10.1007/s41468-024-00179-4`.

[8] Olympio HACQUARD et Vadim LEBOVICI. “Euler Characteristic Tools for Topological Data Analysis”. In : *Journal of Machine Learning Research* 25.240 (2024), p. 1-39. URL : `http://jmlr.org/papers/v25/23-0353.html`.

[9] Michael E. Van HUFFEL et al. “Discrete Transforms of Quantized Persistence Diagrams”. In : *2025 Proceedings of the Symposium on Algorithm Engineering and Experiments (ALENEX)*, p. 68-80.

[10] Steve Y. OUDOT. *Persistence theory. From quiver representations to data analysis*. English. T. 209. Math. Surv. Monogr. Providence, RI : American Mathematical Society (AMS), 2015. ISBN : 978-1-4704-2545-6.

[11] Leonid Polterovich et al. *Topological persistence in geometry and analysis*. English. T. 74. Univ. Lect. Ser. Providence, RI : American Mathematical Society (AMS), 2020. isbn : 978-1-4704-5495-1 ; 978-1-4704-5679-5. doi : 10.1090/ulect/074.