Journées mathématiques X-UPS

Année 2024

Anayse topologique de données

Steve OUDOT

Introduction à la théorie de la persistance à travers un exemple d'application

Journées mathématiques X-UPS (2024), p. 3-19.
https://doi.org/10.5802/xups.2024-01

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

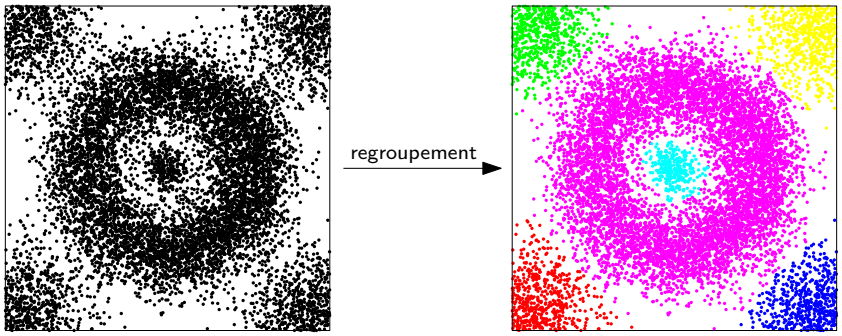

FIGURE 1. Un échantillon de points dans $\mathbb{R}^2$ (à gauche) et un regroupement de ces points (à droite).

1. Introduction à la théorie de la persistance à travers un exemple d'application

Le but de ce chapitre est d'introduire quelques unes des principales idées sur lesquelles repose la théorie de la persistance, qui est le fondement mathématique de l'analyse topologique de données et le sujet central de cet ouvrage. Pour ce faire, nous allons nous placer dans un cadre restreint, celui de la persistance des pics d'une fonction réelle, et nous allons prendre pour prétexte l'application de la théorie à la classification non-supervisée (aussi appelée *regroupement*) en apprentissage machine. L'exposé alternera donc entre des sections de présentation du contexte applicatif et des sections plus formelles introduisant les notions mathématiques. Parmi ces sections, la seule qui soit vraiment utile pour les chapitres suivants est la 1.4, qui introduit la persistance dans notre cadre. Le lecteur qui ne s'intéresserait pas au contexte applicatif peut sans risque limiter sa lecture à cette seule section avant de passer au chapitre suivant.

1.1. Le problème du regroupement. — Soit $P = \{p_1, \cdots, p_n\}$ un ensemble fini de points de l'espace euclidien $\mathbb{R}^d$ – dans le jargon on parle d'*échantillon de points*, de *jeu de données* ou encore de *nuage de points*, selon le domaine scientifique considéré (statistiques, apprentissage machine, ou géométrie). On suppose que les points proviennent de m classes distinctes, que l'on ne connaît pas – m lui-même n'est généralement pas connu non plus. L'objectif est donc (au besoin) de déterminer le nombre m de classes, puis de partitionner le nuage P en m sous-ensembles appelés *clusters*, $P = \bigsqcup_{l=1}^{m} C_l$, de manière à respecter au mieux les m classes sous-jacentes. Voir la figure 1 pour une

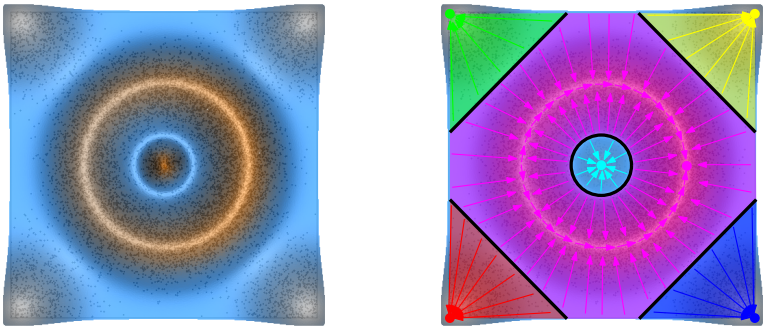

FIGURE 2. À gauche : le nuage de points P de la figure 1 superposé à la densité de probabilité f selon laquelle les points de P ont été échantillonnés iid. Les valeurs de f sont représentées par une palette de couleurs, du bleu (densité faible) à orange (densité forte). À droite : les six pics de f et leurs variétés stables (de la même couleur que le pic correspondant), ainsi qu'une représentation synthétique du flot de gradient de f.

illustration. Cette partition du nuage P s'appelle un *regroupement* des points (*clustering* en anglais).

Tel que formulé, le problème est clairement mal posé puisque, en l'absence d'informations complémentaires sur la manière dont les points sont obtenus à partir des classes sous-jacentes, il n'y a aucun moyen de déterminer si un regroupement du nuage est meilleur qu'un autre. Il faut donc formuler des **hypothèses sur la génération des données**.

Chaque méthode de clustering vient avec son propre jeu d'hypothèses, duquel découle un critère de qualité sur les partitions de P. Ici on va s'intéresser aux approches dites *basées sur la densité*, qui formulent l'hypothèse suivante qu'on adoptera tout au long de ce chapitre – voir la figure 2 (gauche) pour une illustration :

Hypothèse 1.1. — *Les points $p_1, \cdots, p_n$ sont échantillonnés de manière iid selon une mesure de probabilité μ, de densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$. La mesure μ comme sa densité f sont supposées* **inconnues**.

Les classes sont alors définies comme des sous-ensembles deux-à-deux disjoints de $\mathbb{R}^d$ associés aux maxima locaux – aussi appelés *modes* ou *pics* – de la densité f. Plus précisément, elles correspondent

aux *variétés stables* des modes, c'est-à-dire qu'il y a exactement une classe par mode x, et cette classe est définie comme étant le lieu des points de $\mathbb{R}^d$ qui convergent asymptotiquement vers x lorsqu'ils sont poussés continûment le long du flot de gradient de la fonction f – voir la figure 2 (droite) pour une illustration. Nous allons maintenant formaliser cette notion en utilisant un peu de théorie de Morse, dont une référence classique est le livre de Milnor Milnor (1963).

1.2. Définition mathématique des classes à partir de la densité f. — Comme on vient de le dire, pour définir les classes associées aux modes de notre densité f nous allons pousser les points de l'espace $\mathbb{R}^d$ continûment le long du flot de gradient de f. De toute évidence il nous faut faire des hypothèses de régularité sur f afin d'avoir un gradient et un flot bien définis, ainsi qu'un contrôle sur la convergence des trajectoires des points. Voici un jeu simplifié d'hypothèses, qu'on supposera vérifiées dans toute la suite du chapitre :

Hypothèse 1.2. — *On suppose que la densité f :*
(i) est lisse, de classe C^∞ ;
(ii) s'annule à l'infini, c'est-à-dire que $\lim_{\|x\|\to\infty} f(x) = 0$ *;*
(iii) est de type Morse*, c'est-à-dire que ses* points critiques*, i.e. les points x où le gradient $\nabla f(x)$ s'annule, sont en nombre fini et* non-dégénérés*, c'est-à-dire que la matrice hessienne de f en chacun de ces points est inversible.*

L'hypothèse (iii) joue un rôle clé dans la suite. Elle implique en particulier que les points critiques de f sont isolés dans $\mathbb{R}^d$. Bien qu'elle puisse paraître plus restrictive que les autres hypothèses a priori, puisqu'elle impose des conditions sur les quantités différentielles d'ordre 1 et 2 de f, il s'avère que les fonctions de type Morse forment un ouvert dense pour la topologie C^2 dans l'espace des fonctions C^∞, donc on ne perd presque rien en généralité en ajoutant l'hypothèse (iii).

Comme la fonction f est de classe C^∞, son champ de gradient est localement lipschitzien et peut donc être intégré en un flot continu $\Phi\colon \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$. Plus précisément, par le théorème de Cauchy-Lipschitz global, la ligne de flot $\varphi_x\colon t \mapsto \Phi(t,x)$ issue d'un point $x \in \mathbb{R}^d$ est l'unique solution de l'équation différentielle ordinaire suivante :

$$\begin{cases} \dot{\varphi}_x(t) = \nabla f(\varphi_x(t)) \\ \varphi_x(0) = x \end{cases}$$

Cette solution dépend continûment à la fois du paramètre t et de la condition initiale x, d'où la continuité du flot Φ.

On regarde maintenant, pour tout point $x \in \mathbb{R}^d$, si et où converge la ligne de flot φ_x lorsque $t \to +\infty$. Le cas particulier où $f(x) = 0$ est trivial : en tant que densité de probabilité, f est positive ou nulle, donc x est un minimum local de f et, à ce titre, lui-même un point critique de f, stationnaire pour le flot Φ par définition. Pour le cas général $f(x) > 0$, l'hypothèse que f s'annule à l'infini (sur $\mathbb{R}^d$ qui est localement compact) implique que l'ensemble de sur-niveau $\{y \in \mathbb{R}^d \mid f(y) \geq f(x) > 0\}$, dans lequel se trouve l'image de la ligne de flot φ_x, est compact ; il s'ensuit alors que la ligne de flot φ_x converge bien lorsque $t \to +\infty$, et par définition la limite est un point critique de f. Ainsi, à la limite, le flot amène tous les points de $\mathbb{R}^d$ à des points critiques de f.

Définition 1.1. — *La* variété stable *d'un point critique x de f est l'ensemble des points $y \in \mathbb{R}^d$ tels que $x = \lim_{t \to +\infty} \varphi_y(t)$.*

Il découle de ce qui vient d'être dit que les variétés stables des points critiques partitionnent l'espace $\mathbb{R}^d$. Toutefois, tous les points critiques ne sont pas des pics : il y a également les minima locaux, ainsi que les points critiques de type *selle*. Dans la suite on ne retiendra que les variétés stables des pics.

Définition 1.2. — *Les classes correspondant à la densité f sont par définition les variétés stables des pics de f.*

L'idée derrière le fait de ne regarder que les variétés stables des pics est que, en général, identifier les points critiques requiert d'évaluer des quantités différentielles, typiquement le gradient de f (voire la matrice hessienne si on veut déterminer le type des points critiques), dont le calcul est instable numériquement. Les pics, quant à eux, ne nécessitent pas de quantités différentielles pour être caractérisés (definition 1.5), leur calcul en pratique s'avère donc beaucoup plus stable numériquement.

En contrepartie, on peut s'interroger sur la pertinence de laisser de côté les autres types de points critiques, impliquant de fait que l'union des classes ne couvre pas $\mathbb{R}^d$ tout entier. En réalité, la théorie de Morse nous garantit que les classes couvrent presque tout l'espace :

Proposition 1.1. — *Sous l'hypothèse 1.2, le complémentaire de l'union des variétés stables des pics de f est une union finie de sous-variétés différentielles de $\mathbb{R}^d$ de codimensions strictement positives, et donc de mesure de Lebesgue nulle dans $\mathbb{R}^d$.*

En pratique, la probabilité qu'un point p_i de notre échantillon P ne soit pas parmi les classes est donc nulle, puisque les p_i sont échantillonnés selon la mesure μ qui a une densité par rapport à la mesure de Lebesgue. Ainsi, nos hypothèses initiales et notre définition des classes sont génériquement compatibles avec notre problème.

1.3. Calcul des clusters à partir du nuage de points P. — Comme on l'a supposé dans l'hypothèse de départ 1.1, la mesure μ et sa densité f ne sont pas connues en pratique. Il nous va donc falloir trouver un moyen de simuler la montée de gradient à partir des données $p_1, \cdots, p_n$ pour pouvoir former des clusters qui approximent les classes. Il existe tout un éventail de méthodes pour ce faire. Les unes, comme par exemple *mean-shift* Comaniciu and Meer (2002), adoptent une approche purement numérique en tentant d'approximer localement le gradient de f puis de simuler la montée de gradient continue par une montée de gradient approximé discrète dans $\mathbb{R}^d$. Les autres, plus anciennes comme celle que nous allons voir ici Koontz et al. (1976), adoptent une approche combinatoire en remplaçant $\mathbb{R}^d$ par un objet discret appelé *graphe de voisinage*, construit à partir des données, dans lequel le gradient est approximé en chaque sommet par une arête incidente.

Pré-traitement. — En pré-traitement on construit le graphe de voisinage et on approxime la densité aux sommets du graphe.

Définition 1.3. — *Un* prédicat de voisinage *est une fonction symétrique $P \times P \to \{0,1\}$ qui vaut 0 sur la diagonale $\{(p_i, p_i) \mid p_i \in P\}$.*

Définition 1.4. — *Étant donné un prédicat de voisinage $\sigma\colon P \times P \to \{0,1\}$, le* graphe de voisinage *correspondant est le graphe combinatoire $G = (P, E)$ non-orienté dont les sommets sont les points de P et les arêtes forment l'ensemble $E = \{(p_i, p_j) \in P \times P \mid \sigma(p_i, p_j) = 1\}$.*

Exemple 1.1. — *Voici deux constructions classiques de graphes de voisinage, dont la première est illustrée dans la figure 3 (gauche) :*

- *le graphe de r-*voisinage*, pour un réel $r \geq 0$, correspond au prédicat $(p_i, p_j) \mapsto \mathbb{1}_{p_i \neq p_j} \mathbb{1}_{\|p_i - p_j\|_2 \leq r}$ qui teste si deux points donnés sont à distance euclidienne au plus r l'un de l'autre ;*
- *le graphe de k-*voisinage*, pour un entier $k \geq 1$, correspond au prédicat $(p_i, p_j) \mapsto \mathbb{1}_{p_i \neq p_j} \mathbb{1}_{p_j \in ppv_k(p_i)} \mathbb{1}_{p_i \in ppv_k(p_j)}$, où $ppv_k(p_l)$ désigne l'ensemble des k plus proches voisins de p_l parmi les points du nuage P pour la distance euclidienne.*

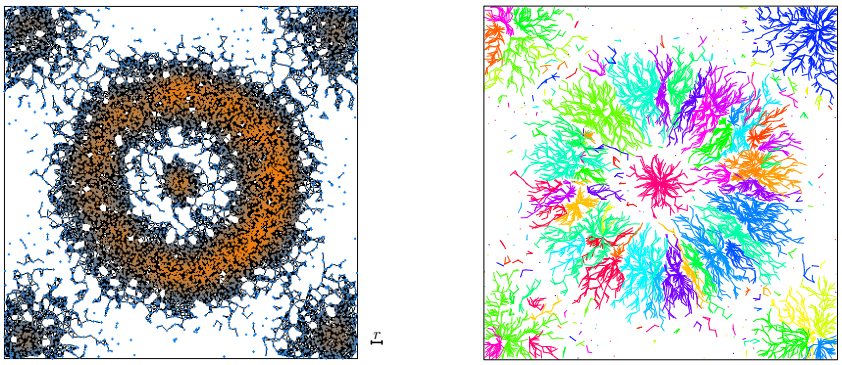

FIGURE 3. À gauche : le graphe G de r-voisinage sur le nuage de points P de la figure 1, pour la valeur de r montrée au centre. Les arêtes du graphe sont en noir, tandis que les valeurs de l'estimateur $\hat{f}$ aux sommets sont représentées comme à la figure 2. À droite : le résultat de l'algroithme sur cette entrée, avec une couleur distincte par cluster. Les arêtes représentées sur le dessin sont celles des arbres de la forêt couvrante de G calculée par l'algorithme.

Une fois un tel graphe de voisinage $G = (P, E)$ construit à partir de P, on calcule une estimation $\hat{f}(p_i)$ de la densité f en chaque point $p_i \in P$. L'estimation de la densité est en soi un sujet à part entière, qui sort du cadre de cet exposé. Notons simplement que le domaine des statistiques nous fournit tout un éventail d'estimateurs ayant chacun des propriétés spécifiques. Ici nous ferons simplement l'hypothèse (relativement forte mais classique) que l'estimateur $\hat{f}$ approxime la densité f en norme sup sur P, c'est-à-dire que l'on peut borner l'erreur de l'estimateur comme suit :

$$\|\hat{f} - f\|_\infty \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} |\hat{f}(p_i) - f(p_i)| \leq \varepsilon(n) \tag{1}$$

où la quantité $\varepsilon(n)$ dépend uniquement de n, pas des points du nuage P, et tend vers 0 lorsque $n \to +\infty$. La borne $\varepsilon(n)$ en elle-même n'a pas d'importance ici, et pour simplifier nous la supposerons vérifiée de manière déterministe et non pas seulement avec forte probabilité comme c'est le cas en pratique. Ainsi nous avons équipé notre graphe de voisinage $G = (P, E)$ d'un champ scalaire $\hat{f} \colon P \to \mathbb{R}^+$ approximant la densité f aux sommets – voir encore la figure 3 (gauche). C'est l'entrée que nous fournissons à l'algorithme.

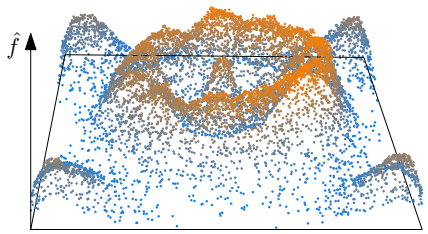

FIGURE 4. Graphe de l'estimateur de densité $\hat{f}$ restreint au nuage de points P de la figure 1.

L'algorithme. — En chaque sommet p_i du graphe nous approximons le gradient de f par l'arête reliant p_i à son voisin p_j dont l'estimée $\hat{f}(p_j)$ est la plus élevée, **à condition que celle-ci soit plus élevée que** $\hat{f}(p_i)$. Dans le cas contraire, p_i est un maximum local de $\hat{f}$ dans le graphe G et on le déclare donc comme étant un pic de la densité.

L'ensemble des arêtes ainsi sélectionnées forme une *forêt couvrante* de G, c'est-à-dire un ensemble de sous-graphes qui sont des arbres (i.e. des sous-graphes connexes sans cycles) et qui couvrent des sous-ensembles deux-à-deux disjoints de sommets dont l'union est le nuage P tout entier[1]. Ces arbres sont les clusters produits par l'algorithme. Chacun contient une unique racine, son sommet dont la valeur de $\hat{f}$ est la plus élevée, qui par construction est un maximum local de $\hat{f}$ dans G et sert donc d'approximation pour un éventuel pic de f. L'arbre en lui-même sert d'approximation pour la variété stable associée à ce pic dans $\mathbb{R}^d$.

Résultat. — La figure 3 (droite) montre un exemple de résultat de l'algorithme, qui, il faut l'avouer, est de piètre qualité. Ce qui frappe immédiatement, c'est la multiplication des clusters (plusieurs dizaines) par rapport au nombre de classes sous-jacentes (six seulement). Ceci s'explique essentiellement par le fait que l'estimateur $\hat{f}$ est beaucoup plus bruité que la densité f, n'étant qu'une approximation en norme sup d'après (1) – voir la figure 4 pour une illustration du bruit dans l'estimateur. Ainsi, en plus de quelques pics "légitimes" associés aux pics de f dans $\mathbb{R}^d$, s'ajoute dans le graphe G tout un tas de pics fallacieux dûs au bruit dans $\hat{f}$. Afin de distinguer parmi

1. Techniquement, certains sommets de G peuvent être isolés, auquel cas on les ajoute à la forêt en tant qu'arbres singletons.

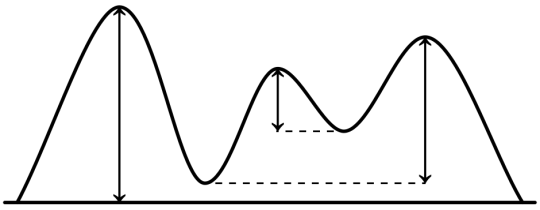

FIGURE 5. Les flèches verticales montrent la proéminence de trois pics sur une île. Une ligne pointillée horizontale relie chaque pic (excepté le plus haut) à son col le plus élevé. Source : *Wikipedia* (`https://en.wikipedia.org/wiki/File:Relative-height.png`, image sous licence CC BY-SA 3.0 Deed).

les pics ceux qui sont légitimes de ceux qui ne le sont pas, nous allons utiliser une notion mathématique qui quantifie l'importance de chaque pic : la **persistance.** Cette notion va également nous fournir une **hiérarchie sur les pics** de $\hat{f}$ dans G, hiérarchie qui va nous permettre de réparer le clustering produit par l'algorithme en fusionnant les clusters associés aux pics fallacieux avec les clusters de leurs parents dans la hiérarchie.

1.4. La persistance des pics d'une fonction réelle. — Alors que le terme *persistance* est utilisé communément en analyse topologique de données, dans le cas particulier des pics il renvoie à un concept plus ancien, la *proéminence*, introduit par les alpinistes et les géographes, que nous allons présenter en premier afin de nourrir l'intuition du lecteur.

1.4.1. *Définition des alpinistes et géographes* [(2)]. — Avec le développement de l'exploration alpine dans la deuxième moitié du XIX$^{\text{ème}}$ siècle, des listes de sommets, conquis ou à conquérir, ont commencé à émerger. Rapidement s'est posée la question de déterminer ce qu'est un sommet. En effet, sans critère restrictif, n'importe quelle antécime, épaule, ou même pierre pourrait être vue comme un sommet en elle-même. Le critère principal retenu pour distinguer les sommets des autres types de protubérances a été celui de la proéminence, qui doit être supérieure à un certain seuil pour que la protubérance puisse être considérée comme un sommet indépendant. Ce seuil varie d'un classement à l'autre : de 30 mètres (la longueur de corde de l'alpinisme

2. Le contenu de cette sous-section est largement repris de l'article correspondant sur *Wikipedia* (`https://fr.wikipedia.org/wiki/Proéminence`).

sommet	continent	altitude (m)	proéminence (m)
Everest	Asie	8849	8849
K2	Asie	8611	4020
Kangchenjunga	Asie	8586	3922
Lhotse	Asie	8516	610
Makalu	Asie	8485	2378
Cho Oyu	Asie	8188	2340
Dhaulagiri I	Asie	8167	3357
Manaslu	Asie	8163	3092
Nanga Parbat	Asie	8126	4608
Annapurna I	Asie	8091	2984

TABLE 1. Liste des 10 sommets les plus hauts du monde, classés par ordre décroissant d'altitude. Le seuil minimal de proéminence requis ici est de 500 mètres.

classique) pour la liste officielle des 82 sommets des Alpes de plus de 4000 mètres, à 1500 mètres pour les sommets dits *ultra-proéminents*.

La proéminence pour les alpinistes et les géographes se définit comme "la différence d'altitude entre un sommet donné et l'ensellement ou le col le plus élevé permettant d'atteindre une cime encore plus haute". Autrement dit, c'est "le dénivelé minimum de la descente à parcourir pour pouvoir remonter sur un sommet plus élevé". Voir la figure 5 pour une illustration. Notons que l'ensellement peut se situer au niveau de la mer mais pas au-dessous. Ainsi, le plus haut pic d'une île a une proéminence égale à sa hauteur. Les tables 1 et 2 donnent les listes des dix sommets les plus hauts du monde d'une part, des dix sommets les plus proéminents du monde d'autre part, et comme on peut le constater elles sont bien différentes.

1.4.2. *Définition mathématique.* — On va maintenant définir formellement la proéminence, ou persistance. Pour cela on fixe un espace topologique X et une fonction $f : X \to \mathbb{R}$, sans plus d'hypothèses pour le moment – des hypothèses sur f et X apparaîtront au cours de l'exposé. Notons dès à présent que l'approche adoptée ici pour définir la persistance transcrit directement les idées de la section 1.4.1 en requérant peu d'outils mathématiques. En contrepartie, elle donne lieu à des énoncés parfois inutilement techniques et se généralise mal – voir à ce propos l'hypothèse 1.3 et les exemples et commentaires qui l'entourent. La bonne approche, fondée sur l'homologie, sera adoptée dans le chapitre 3 – voir en particulier la remarque 3.3.

sommet	continent	altitude (m)	proéminence (m)
Everest	Asie	8849	8849
Aconcagua	Amérique	6960	6960
Denali	Amérique	6191	6155
Kilimanjaro	Afrique	5895	5885
Pico Cristóbal Colón	Amérique	5700	5509
Mont Logan	Amérique	5959	5250
Pico de Orizaba	Amérique	5636	4922
Massif Vinson	Antarctique	4892	4892
Puncak Jaya	Océanie	4884	4884
Mont Elbrouz	Europe	5642	4741

TABLE 2. Liste des 10 sommets les plus proéminents du monde, classés par ordre décroissant de proéminence.

Définition 1.5. — *Un point $x \in X$ est un* maximum local *(ou* pic*) de f s'il existe un voisinage U de x dans X tel que $f(x) = \max_U f$.*

Pour quantifier la proéminence des pics, on regarde l'évolution des composantes connexes par arc dans les *sur-niveaux* de la fonction f alors que le niveau diminue progressivement de $+\infty$ jusqu'à $-\infty$.

Définition 1.6. — *Étant donné un niveau $t \in \mathbb{R}$, l'*(ensemble de) sur-niveau *de f associé est $f^{-1}([t, +\infty))$.*

Définition 1.7. — *Étant donné un pic $x \in X$ de f, pour tout niveau $t \leq f(x)$ on définit $C_t(x)$ comme étant la composante connexe par arc du sur-niveau $f^{-1}([t, +\infty))$ à laquelle appartient x.*

Lorsque t diminue, le sur-niveau de f associé ne fait que croître, ainsi que ses composantes connexes par arc. En conséquence :

Corollaire 1.1. — *Étant donné un pic $x \in X$ de f, pour tous niveaux $t \leq t' \leq f(x)$ on a $C_t(x) \supseteq C_{t'}(x)$.*

Ainsi, informellement, on peut traquer la croissance de la composante connexe par arc contenant notre pic x tandis que l'on abaisse le niveau t, et détecter la première valeur de t à laquelle cette composante fusionne avec celle d'un pic plus élevé : à cette valeur $t = h(x)$ particulière, x émerge comme un pic secondaire d'une montagne plus élevée, et sa persistance est donnée par la différence $f(x) - h(x) \geq 0$. Dans le cas particulier où la composante de x fusionne avec celle d'un

pic de même hauteur, on départage les deux pics en désignant l'un comme étant secondaire de l'autre de manière arbitraire.

Formellement, on suppose donné un ordre total $\preceq$ sur X (ce qui en toute généralité requiert une version faible de l'axiome du choix) et on considère l'ordre lexicographique suivant sur X :

$$y \geq x \iff \begin{cases} f(y) > f(x) & \text{ou} \\ f(y) = f(x) & \text{et} \quad y \preceq x \end{cases} \tag{2}$$

On note $>$ l'ordre total strict associé :

$$y > x \iff y \geq x \quad \text{et} \quad y \neq x$$

Définition 1.8. — *Pour tout pic $x \in X$ de f on définit :*

- *l'instant de* naisssance *de x comme étant la valeur $f(x) \in \mathbb{R}$;*
- *l'instant de* décès *de x par $h(x) = \sup I(x) \in \mathbb{R} \cup \{-\infty\}$, où $I(x) = \{t \leq f(x) \mid \exists y > x$ pic de f tel que $C_t(y) = C_t(x)\}$;*
- *l'*intervalle de persistance *de x comme étant l'intervalle $]h(x), f(x)] \subseteq \mathbb{R}$, que l'on maintient ouvert à gauche par convention car cette extrémité peut être à l'infini ;*
- *la* persistance *(ou* proéminence *en géographie) de x comme étant la différence $f(x) - h(x) \in \mathbb{R}^+ \cup \{+\infty\}$.*

Remarque 1.1. — *Pour les pics x tels que $h(x) = -\infty$, la proéminence telle que définie ici est infinie, tandis que pour les géographes la proéminence de ces pics est égale à leur hauteur (voir la section 1.4.1 et en particulier la figure 5).*

Définition 1.9. — *Le* code-barres *de f est le multiensemble des intervalles de persistance des pics de f. La* multiplicité *d'un intervalle est le nombre (potentiellement infini) de ses occurrences dans le multiensemble.*

Exemple 1.2. — *Considérons la fonction $x \mapsto -\sin(x)\cos(3x)$ sur l'intervalle $X = [-5, 2] \subset \mathbb{R}$ muni de l'ordre usuel sur les réels, dont le graphe est représenté à la figure 6. Cette fonction possède cinq pics, aux abscisses $x = -5$, $x = \arctan\sqrt{2 + \sqrt{\frac{11}{3}}} + k\,\pi$ et $x = -\arctan\sqrt{2 - \sqrt{\frac{11}{3}}} + k\,\pi$ pour $k \in \{-1, 0\}$. Les intervalles de son code-barres sont, de gauche à droite :*

$$]-a, c] \qquad]-b, b] \qquad]-\infty, a] \qquad]-b, b] \qquad]-a, a]$$

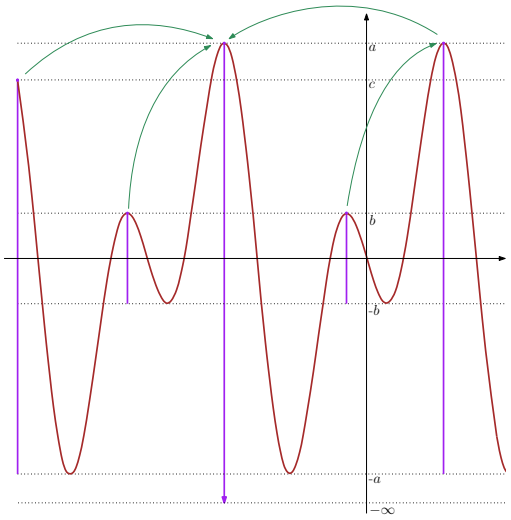

FIGURE 6. Graphe (en brun) et code-barres (en violet) de la fonction $x \mapsto -\sin(x)\cos(3x)$ sur l'intervalle $X = [-5, 2]$ muni de l'ordre usuel sur les réels. Par souci de clarté, les intervalles de persistance sont ancrés à l'aplomb des pics correspondants de la fonction. Les liens de parenté formant la hiérarchie des pics sont matérialisés par des flèches courbes vertes reliant chaque pic à son parent lorsqu'il existe.

où

$$a = \left(3\sqrt{11/3}+5\right)\left(2+\sqrt{11/3}\right)^{1/2}\left(3+\sqrt{11/3}\right)^{-2} \approx 0.88$$
$$b = \left(3\sqrt{11/3}-5\right)\left(2-\sqrt{11/3}\right)^{1/2}\left(3-\sqrt{11/3}\right)^{-2} \approx 0.19$$
$$c = \sin(5)\cos(15) \approx 0.73$$

Les notions d'intervalle de persistance et de code-barres sont définies pour toute fonction $f\colon X \to \mathbb{R}$, toutefois elles n'ont de sens que lorsqu'on fait l'hypothèse que les composantes connexes par arc contenant les pics couvrent l'intégralité des sur-niveaux de la fonction, c'est-à-dire :

Hypothèse 1.3. — $\forall t \in \mathbb{R}, \quad f^{-1}([t,+\infty[) = \bigcup_{\substack{x \text{ pic de } f \\ f(x)\geq t}} C_t(x)$

Dans le cas contraire en effet, des composantes connexes par arc dans les sur-niveaux peuvent être ignorées et donc des barres ne pas apparaître ou bien apparaître avec des extrémités erronées dans le code-barres, comme dans les exemples ci-dessous :

Exemple 1.3. — *La fonction identité sur* $\mathbb{R}$ *ou sur l'intervalle* $]0,1[$ *n'a aucun pic et donc son code-barres est vide. De même pour la fonction* $x \mapsto \begin{cases} 1-|x| \text{ si } x \neq 0 \\ 0 \text{ si } x = 0 \end{cases}$ *sur l'intervalle* $[-1,1]$.

Exemple 1.4. — *La fonction* $f\colon x \mapsto -(x^3+2)\exp(-x)$ *sur* $\mathbb{R}^+$ *a un unique pic en* $x=1$*, dont la proéminence est* $+\infty$ *alors qu'on s'attendrait à ce qu'elle soit finie car* $f(1) = -\frac{3}{e}$ *et* $\lim_{t\to+\infty} f = 0$.

À partir de maintenant nous supposerons donc l'hypothèse 1.3 vérifiée. C'est le cas par exemple lorsque X est compact et f est continue, ou encore lorsque $X = \mathbb{R}^d$ et f est continue, positive ou nulle et s'annule à l'infini comme sous l'hypothèse 1.2.

Hiérarchie des pics. — En plus du code-barres, nous pouvons définir une notion de parent et de là une hiérarchie sur les pics. Pour cela nous faisons l'hypothèse additionnelle suivante, également vérifiée sous l'hypothèse 1.2 :

Hypothèse 1.4. — *Le nombre de pics de la fonction* f *est fini.*

Exercice 1.1. — *Soit* $x \in X$ *un pic de* f *dont la proéminence est finie* ($h(x) > -\infty$)*. Montrer que l'ensemble* $I(x)$ *de la définition 1.8 est alors non-vide, de même que l'ensemble*

$$J(x) \;=\; \{y \text{ pic de } f \mid y > x \text{ et } C_t(y) = C_t(x)\ \forall t \in I(x)\}$$

et que ce dernier admet un maximum pour l'ordre $\geq$ *de l'équation* (2).

Le maximum de $J(x)$ est appelé le *parent* de x. Il est strictement supérieur à x pour l'ordre $\geq$ donc la relation de parenté induit une hiérarchie sur les pics, au sommet de laquelle se trouvent les pics de proéminence infinie. La figure 6 montre cette hiérarchie pour la fonction de l'exemple 1.2.

Diagrammes de persistance et stabilité. — Une autre manière de représenter graphiquement les codes-barres est sous la forme de multiensembles de points dans le plan étendu $\mathbb{R} \times [-\infty, +\infty[$, appelés *diagrammes de persistance*, dans lesquels chaque copie du point (a,b) correspond à une copie de l'intervalle $]b,a]$ dans le code-barres associé,

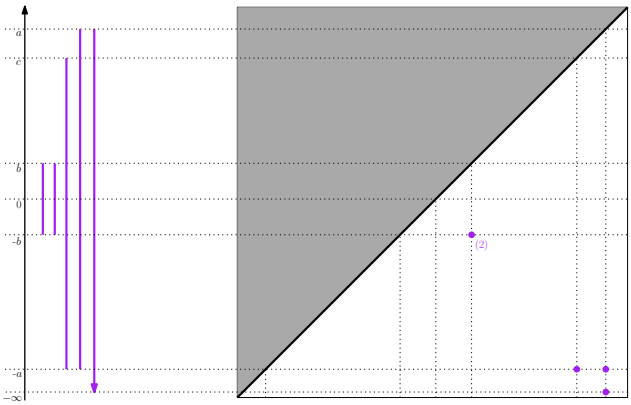

FIGURE 7. Le code-barres issu de la figure 6 (à gauche) et son diagramme de persistance correspondant (à droite). La multiplicité du point $(b, -b)$ dans le diagramme est indiquée entre parenthèses.

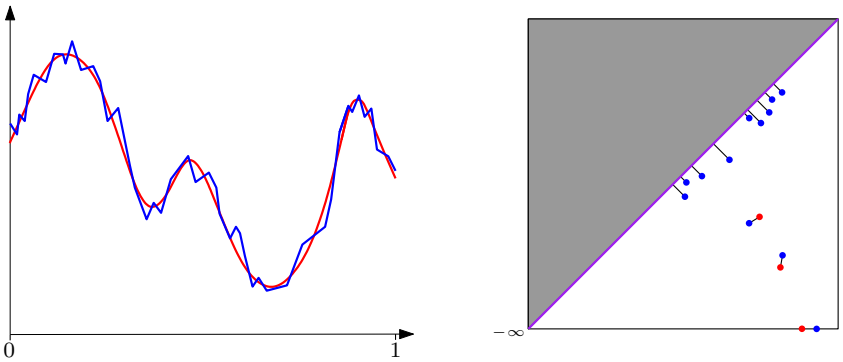

FIGURE 8. Deux fonctions $[0, 1] \to \mathbb{R}$ (à gauche) et leurs diagrammes de persistance (à droite) avec, marqué par des segments en noir, l'appariement optimal entre les points des deux diagrammes pour la distance de transport. Dans cet exemple, la distance de transport est légèrement inférieure à la différence des deux fonctions en norme sup.

avec $a > b \geq -\infty$. Voir la figure 7 pour une illustration. Cette représentation alternative contient la même information mais elle offre l'avantage de montrer les codes-barres comme des nuages de points, ou plutôt comme des mesures empiriques, plus facilement interprétables. Par ailleurs, la métrique naturelle entre codes-barres, appelée *distance*

du goulot de bouteille (*bottleneck distance* en anglais), peut être interprétée comme une distance de transport partiel entre diagrammes de persistance, eux-mêmes vus comme des mesures empiriques. La définition précise de cette distance sera donnée au chapitre 4, mais en attendant nous pouvons déjà la visualiser sur l'exemple de la figure 8. L'interprétation des deux diagrammes dans l'exemple est que chacune des deux fonctions considérées possède trois pics proéminents, le reste des pics (dont les points correspondants dans les diagrammes sont localisés près de la diagonale $y = x$) pouvant être considéré comme du bruit. Le plan de transport optimal associé à la distance du goulot de bouteille entre les diagrammes donne un appariement explicite entre les pics proéminents des deux fonctions, et nous indique par ailleurs d'ignorer les pics non-proéminents (en appariant avec la diagonale leurs points correspondants dans les diagrammes).

Ainsi, les diagrammes de persistance fournissent une représentation synthétique des intervalles de persistance des pics d'une fonction, tandis que la distance de transport entre diagrammes indique comment mettre en correspondance au mieux les pics de fonctions différentes selon leurs intervalles de persistance. Cette observation empirique est appuyée par un résultat fondamental de la théorie de la persistance, appelé le *théorème de stabilité*, qui dit en substance que l'opérateur D, qui associe à toute fonction $X \to \mathbb{R}$ son diagramme de persistance lorsque celui-ci existe, est 1-lipschitzien. Dans notre contexte le résultat s'énonce de la manière suivante – voir encore la figure 8 pour une illustration :

Théorème 1.1. — *Pour toutes fonctions $f, g \colon X \to \mathbb{R}$ vérifiant les hypothèses 1.3 et 1.4, on a l'inégalité suivante, où d_b désigne la distance du goulot de bouteille et $\|\cdot\|_\infty$ désigne la norme sup sur X :*

$$d_b(D(f),\, D(g)) \quad \leq \quad \|f - g\|_\infty$$

1.5. Retour à l'application au clustering. — Revenons maintenant à notre application et utilisons la persistance pour corriger les défauts de l'algorithme de clustering présenté à la section 1.3. Nous n'allons en fait pas modifier l'algorithme en lui-même, mais plutôt ajouter un post-traitement des clusters qu'il produit.

Post-traitement. — Étant donné le regroupement $P = \bigsqcup_{l=1}^{m} C_l$ produit par l'algorithme sur le graphe de voisinage $G = (P, E)$, ainsi que les pics $x_1, \cdots, x_m \in P$ de $\hat{f}$ dans G associés à chacun des clusters $C_1, \cdots, C_m$, on calcule l'intervalle de persistance de chaque pic x_l à l'intérieur du graphe G, ainsi que son parent (s'il existe) parmi les

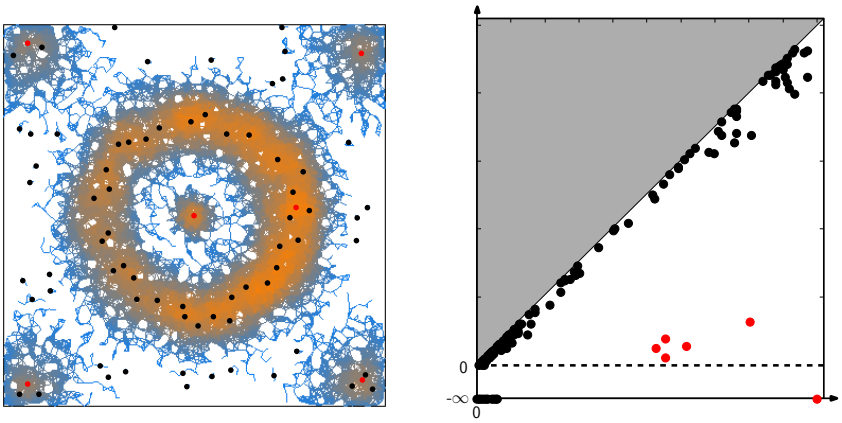

FIGURE 9. À gauche : localisation des pics de l'estimateur de densité $\hat{f}$ dans le graphe de voisinage G de la figure 3. Les six pics les plus proéminents de $\hat{f}$ (en rouge) sont situés près des six pics de la vraie densité f (voir la figure 2), tandis que le reste des pics de $\hat{f}$ (en noir) se répartit dans des zones de faible norme du gradient de f. À droite : le diagramme de persistance de $\hat{f}$, avec, en rouge, les points correspondant aux intervalles de persistance des six pics de $\hat{f}$ les plus proéminents, clairement séparés du reste des points (correspondant au reste des pics de $\hat{f}$) dans le diagramme.

autres pics. Les détails du calcul importent peu, il suffit de savoir qu'il est possible de le faire en un temps quasi-linéaire en la taille du graphe, par un algorithme similaire à celui de Kruskal pour le calcul d'arbre couvrant minimal (Cormen et al., 2009, section 23.2).

Il s'avère que le multiensemble des intervalles de persistance des pics dans G coïncide avec le code-barres de $\hat{f}$, vue comme une fonction réelle sur le graphe G, vu lui-même comme un espace topologique stratifié par ses sommets et ses arêtes, avec interpolation linéaire des valeurs de $\hat{f}$ le long des arêtes. Voir la figure 9 pour une illustration.

Étant donné un choix de seuil $\tau \in \mathbb{R}^+$ sur la persistance, on fusionne ensuite itérativement les clusters associés aux pics de persistance plus petite que τ dans les clusters de leurs parents. Notons que, dans ce cas particulier, la structure stratifiée de l'espace G et la nature linéaire par morceaux de la fonction $\hat{f}$ garantissent que tout pic qui n'est pas un maximum global de $\hat{f}$ sur la composante connexe de G où il se trouve a un parent. Les maxima globaux, quant à eux, ont par définition

une persistance infinie. De ce fait, la procédure fusionne bien tous les clusters associés aux pics de persistance plus petite que τ dans d'autres clusters, et *in fine* dans des clusters associés à des pics de persistance plus grande que τ. C'est ainsi que nous avons obtenu le clustering de la figure 1 à partir de celui de la figure 3, par un choix adéquat de seuil τ.

Choix du seuil de persistance. — En pratique, pour choisir la valeur du seuil τ on peut s'appuyer sur le diagramme de persistance de $\hat{f}$ dans G, noté $D(\hat{f})$, que l'on a calculé lors du post-traitement. Grâce au théorème de stabilité 1.1, on peut montrer que, sous des hypothèses d'échantillonnage adéquates, le diagramme $D(\hat{f})$ exhibe une séparation claire entre, d'une part, les intervalles de persistance des pics de $\hat{f}$ correspondant aux pics de la densité sous-jacente f, et d'autre part, les intervalles de persistance du reste des pics de $\hat{f}$, qui correspondent à du bruit – voir encore la figure 9 pour une illustration. Grâce à cette séparation, l'utilisateur (ou une méthode statistique) peut sélectionner un seuil τ adapté. Les détails de l'analyse théorique et des garanties associées se trouvent dans l'article Chazal et al. (2013) qui a introduit cette approche, appelée *ToMATo* pour *Topological Mode Analysis Tool*.

Références

Chazal, F., Guibas, L. J., Oudot, S., and Skraba, P. (2013). Persistence-based clustering in Riemannian manifolds. *Journal of the ACM*, 60(6) :1–38.

Comaniciu, D. and Meer, P. (2002). Mean shift : A robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5) :603–619.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*. MIT Press, Cambridge, MA, 3rd edition.

Koontz, W. L. G., Narendra, P. M., and Fukunaga, K. (1976). A graph-theoretical approach to non-parametric cluster analysis. *IEEE Trans. Comput.*, C-25 :936–944.

Milnor, J. W. (1963). *Morse theory*. Number 51 in Annals of Mathematics Studies. Princeton university press.