Approches topologiques pour l'analyse de signaux physiologiques

Alexandre BOIS[1], Brian TERVIL[1], Laurent OUDRE[1],

[1]Université Paris Saclay, Université Paris Cité, ENS Paris Saclay, CNRS, SSA, INSERM, Centre Borelli, F-91190, Gif-sur-Yvette, France
alexandre.bois@ens-paris-saclay.fr, brian.tervil@ens-paris-saclay.fr
laurent.oudre@ens-paris-saclay.fr

Résumé – L'analyse topologique des données (TDA) est un ensemble de techniques dérivées de la topologie algébrique, qui consiste à étudier l'évolution de la structure des données à travers différentes échelles. Cet article décrit une méthode non paramétrique basée sur la TDA pour l'étude de bases de données de signaux physiologiques. Nous avons appliqué cette méthode à l'étude de la marche chez des patients atteints de sclérose en plaques.

Abstract – Topological data analysis (TDA) is a set of techniques derived from algebraic topology that consists in studying the evolution of the structure of data through different scales. This article describes a non-parametric TDA-based method for the study of databases of signals. We applied this method to the study of locomotion of patients suffering from multiple sclerosis.

1 Introduction

L'étude de signaux physiologiques est utile dans de nombreuses applications biomédicales notamment dans des tâches de prévention, d'aide au diagnostic ou de suivi de patients. Dans certains signaux physiologiques, tels que les électrocardiogrammes ou les signaux d'accélérométrie enregistrés durant l'activité de marche, une structure particulière apparaît, sous forme de motifs répétitifs. Ces motifs se répètent généralement dans le temps de façon régulière mais peuvent être perturbés par des anomalies, des oscillations ou du bruit. Savoir comparer les signaux physiologiques et relier les différences entre les signaux aux différences dans les phénomènes biomécaniques qu'ils mesurent permet de les quantifier et donc de produire des analyses plus fines de ces phénomènes.

Dans cet article, nous utilisons des techniques d'analyse topologique de données (TDA) pour développer une méthode non paramétrique d'analyse de signaux physiologiques. La TDA regroupe un ensemble de techniques issues de la topologie algébrique, consistant à étudier l'évolution de la structure des données à différentes échelles. Elle bénéficie de fondements théoriques [1] qui garantissent une forte robustesse au bruit et à certaines transformations des données. La TDA a déjà été utilisée dans la littérature pour analyser diverses séries temporelles dont des séries temporelles médicales [2, 3], notamment pour étudier le mouvement humain et la marche [4]. Les caractéristiques topologiques extraites grâce à la TDA sont, dans la plupart des cas, intégrées dans des algorithmes de *machine learning* (pour de la classification, par exemple), et permettent souvent d'en améliorer les performances. En revanche, leur interprétation propre est mise de côté, au profit de la tâche d'apprentissage.

Notre approche est différente, car elle vise justement à utiliser les outils de la TDA pour analyser et interpréter les signaux physiologiques, mais également pour fournir des métriques entre signaux. Notre méthode consiste à représenter chaque signal d'une base de données par un objet issu de la TDA appelé code-barres de persistance (ou simplement code-barres), puis à les comparer en utilisant une notion de distance entre code-barres. Un algorithme de visualisation est ensuite utilisé et des caractéristiques peuvent être calculées pour comparer différent sous-ensembles de signaux. Notre idée principale est la suivante : les codes-barres résument les oscillations du signal qu'ils représentent, la distance entre deux codes-barres traduit donc une différence au niveau de ces oscillations et est donc pertinente pour comparer des signaux. De plus, l'utilisation de la TDA nous a permis d'obtenir une méthode non paramétrique et ne nécessitant que des signaux bruts, sans étape de segmentation préalable.

Notre méthode est générale et peut s'appliquer à n'importe quel type de signaux. Dans cet article, nous l'avons appliquée à des signaux de marche (vitesses angulaires) de sujets sains et de patients atteints de la sclérose en plaques (SEP) [5]. Cette application servira donc de fil rouge lors de la description mathématique de la méthode mais des extensions sont actuellement étudiées sur d'autres terrains expérimentaux.

2 Protocole et données

Notre base de données est composé de signaux de marche mesurés sur 10 sujets sains et 22 patients atteints de SEP. Le protocole consiste à marcher 12 mètres (avec demi tour) avec un capteur inertiel Xsens® fixé sur chaque pied et mesurant la vitesse angulaire autour de l'axe de rotation de la cheville de l'avant vers l'arrière. Chaque sujet effectue l'aller-retour deux

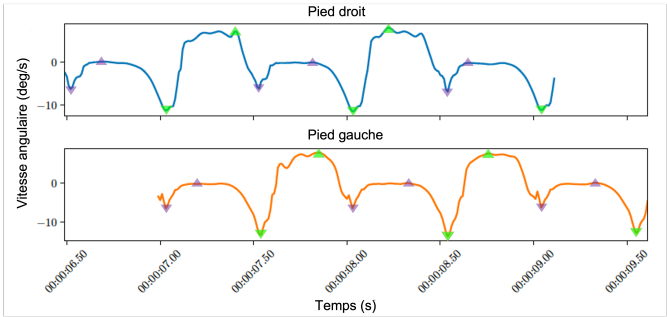

FIGURE 1 – Deux signaux de marche issus d'un sujet sain. Les triangles marquent les extrema locaux correspondant aux plus grandes barres des codes-barres de persistance.

fois lors d'une première session (M0) puis deux nouvelles fois six mois plus tard (M6). Les allers et retours sont traités séparément et les demi-tours sont ignorés. Pour chaque sujet, il y a donc au total 8 signaux à M0 (2 allers, 2 retours, pour chaque pied) et 8 à M6 (voir figure 1 pour des exemples typiques de signaux de marche). La sévérité de la SEP chez les patients est évaluée à chaque session à l'aide de l'EDSS (Expanded Disability Status Scale [6]), qui est une échelle allant de 0 (bilan neurologique normal) à 10 (décès) par incréments de 0,5. Les patients étudiés ici ont des EDSS compris entre 2 et 6,5 (un patient avec un EDSS de 7 ou plus est incapable de réaliser le protocole).

3 Méthode

Notre méthode est une procédure en trois étapes. Tout d'abord, on calcule pour chaque signal un *code-barres de persistance*, outil issu de la TDA, permettant de représenter tous les phénomènes saillants du signal sous forme de barres. Ensuite, on calcule les distances deux à deux entre tous les signaux grâce à une distance adaptée entre codes-barres, dite *distance de bottleneck*. Enfin, ces distances sont projetées dans un espace 2D afin de pouvoir visualiser les distances relatives entre les différents signaux et ainsi fournir un outil pertinent d'analyse.

3.1 Calcul des codes-barres de persistance

On considère un signal représenté par une fonction continue $f : \mathbb{R} \to \mathbb{R}$. Notre méthode repose sur l'étude des ensembles de sous-niveaux de f. Pour un seuil $\alpha \in \mathbb{R}$, le sous-niveau F_α de f est défini par :

$$F_\alpha = f^{-1}(]-\infty, \alpha]). \quad (1)$$

Le but de la TDA est d'étudier l'évolution des données à différentes échelles. Ici, on regarde l'apparition et la disparition des composantes connexes des sous-niveaux F_α lorsque α augmente. La naissance et la mort de composantes connexes sont définies comme suit : pour un α donné, si F_α possède une composante connexe dont aucun point n'appartient à un F_β pour $\beta < \alpha$, alors on dit que cette composante est née en α. Si deux composantes d'un F_β tel que $\beta < \alpha$ ont fusionné dans F_α alors on dit que la plus jeune des deux est morte en α. Le code-barres de persistance est l'ensemble des paires (date de naissance, date de mort) des composantes connexes des F_α, pour α variant de $-\infty$ à $+\infty$.

En pratique, le code-barres correspondant à un signal peut être construit en appariant des minima et des maxima locaux selon l'algorithme suivant (illustré en figure 2) :

1. Marquer le niveau (sur l'axe Y) de tous les extrema locaux du signal. Le premier et dernier point peuvent être ignorés s'ils sont des maxima locaux.
2. Commencer à faire grandir une barre vers le haut en partant du minimum global.
3. À chaque fois que les barres atteignent le niveau d'un autre **minimum local**, commencer une nouvelle barre à ce point. Ensuite, faire grandir toutes les barres jusqu'au niveau du prochain extremum.
4. À chaque fois que les barres atteignent le niveau d'un **maximum local**, si ce point a une barre à sa gauche et une à sa droite, alors la plus courte des deux s'arrête de grandir. Ensuite, faire grandir toutes les barres jusqu'au niveau du prochain extremum.
5. Lorsque les barres atteignent le maximum global, arrêter : la dernière barre va grandir jusqu'à l'infini (cette barre représente l'unique composante connexe correspondant au signal tout entier, qui ne meurt pas).
6. Le code-barres est constitué de toutes les paires de coordonnées verticales (début, fin) des barres (on ignore l'axe du temps), avec une barre qui grandit jusqu'à $+\infty$. On représente généralement les barres horizontalement comme à la fin de la figure 2.

Intuitivement, le code-barre de persistance représente donc de façon structurée tous les événements saillants (vallées, crêtes, oscillations...) présents dans un signal. Chaque barre représente une variation particulière observée dans le signal : plus la barre est longue, et plus la variation est importante. On a représenté sur la figure 1 deux signaux de marche. Sur chacun des signaux, un pas correspond à quatre phénomènes successifs qui se manifestent sous forme de vallées (marquées par les ▼ et ▼ sur chaque signal de la figure 1) et crêtes (marquées par les ▲ et ▲ sur chaque signal de la figure 1) sur la forme d'onde. Ces phénomènes correspondent à des événements biomécaniques particuliers (pose du talon, talon à plat, levé de l'orteil puis du talon). Il y a donc, pour chaque pied, deux pas complets et un pas incomplet. En observant les code-barres associés sur la figure 3, ces phénomènes sont visibles immédiatement : on observe trois grandes barres (correspondant aux deux premières paires (▼,▲), et à la paire formée par le troisième ▼ et l'infini), et trois petites barres (correspondant aux trois paires (▼, ▲)). Les autres barres correspondent à des oscillations, des irrégularités dans la marche ou du bruit.

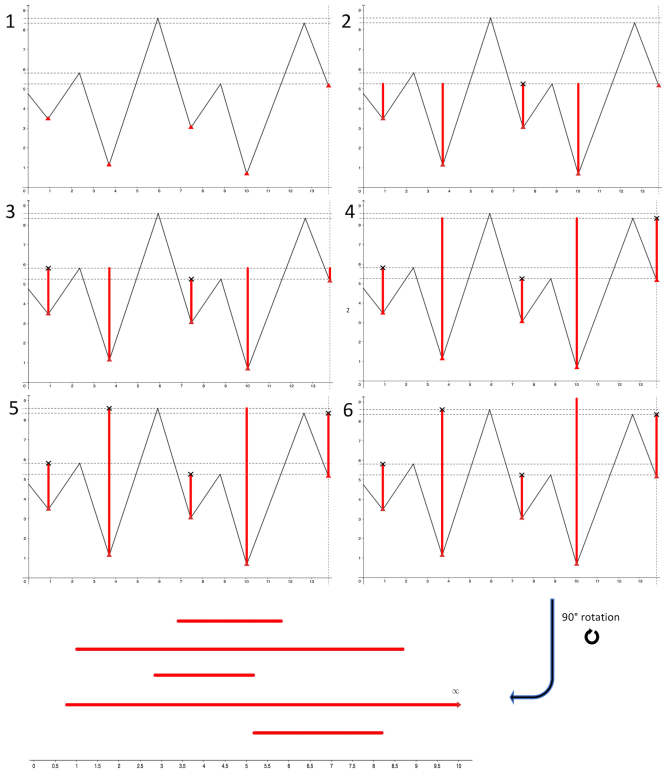

FIGURE 2 – Construction d'un code-barres de persistance.

3.2 Distance de bottleneck entre les diagrammes de persistance

Il est possible de définir une distance sur l'espace des codes-barres : la *distance de bottleneck*. Un code-barres est un ensemble de paires (x, y) qui sont les coordonnées de début et de fin des barres. La même paire peut apparaître plusieurs fois, et y peut valoir $+\infty$ (cela arrive exactement une fois pour une fonction continue définie sur un intervalle). La distance de bottleneck est basée sur des idées venant du transport optimal. Soient B et B' deux codes-barres. On considère l'ensemble $\Gamma(B, B')$ des bijections de B vers B'. Notons que si deux ensembles finis ont un cardinal différent, alors il n'y a pas de bijection entre les deux. C'est pourquoi on considère que chaque code-barres contient aussi chaque barre (x, x) de longueur nulle une infinité de fois. Pour tout $\gamma \in \Gamma(B, B')$ et toute barre $b = (x, y) \in B$ telle que $\gamma(b) = b' = (x', y')$, on définit :

$$||b - b'||_\infty = \begin{cases} |x - x'| & \text{si } y = y' = \infty \\ \max(|x - x'|, |y - y'|) & \text{sinon.} \end{cases} \tag{2}$$

La distance de bottleneck entre B et B' est alors :

$$d_{Bot}(B, B') = \inf_{\gamma \in \Gamma(B,B')} \sup_{b \in B} ||b - \gamma(b)||_\infty. \tag{3}$$

Le point clé de cette approche réside dans ses théorèmes de stabilité [1], qui montrent que, sous des hypothèses assez générales, les codes-barres correspondant à des signaux similaires sont proches en termes de distance de bottleneck.

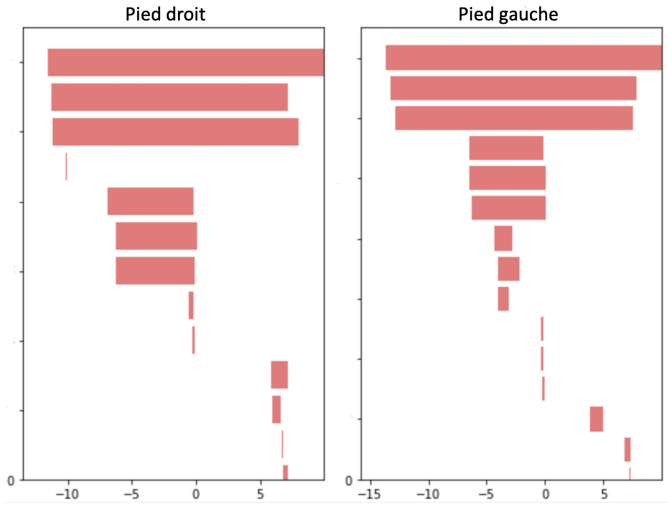

FIGURE 3 – code-barres correspondant aux signaux de la figure 1.

Comme expliqué ci-dessus, le nombre de grandes barres dans un code-barres de signal de marche correspond au nombre de pas. La distance de bottleneck entre deux codes-barres correspondant à deux exercices de marche avec un nombre de pas différent sera donc grande. En effet, l'un des deux aura plus de grandes barres que l'autre et la bijection optimale associera donc une grande barre à une barre plus petite. La distance de bottleneck est donc très sensible au nombre de pas, hors ce nombre peut varier à cause de nombreux facteurs comme la taille, l'âge ou le pied qui fait le premier pas. Pour réduire cette sensibilité et se concentrer davantage sur les oscillations, nous proposons de compter le nombre k de pas de chaque signal (en utilisant une fonction d'autocorrélation par exemple) et de retirer les k plus grandes barres de son code-barre. La matrice de distance est ensuite calculée pour l'ensemble des données.

3.3 Visualisation et analyse

Les codes-barres (et les signaux qu'ils représentent) peuvent être vus comme des points dans un espace métrique équipé de la distance de bottleneck. Nous utilisons l'algorithme UMAP [7] pour les visualiser dans le plan euclidien de telle sorte que la distance dans le plan 2D respecte la structure induite par la matrice de distance de bottleneck sur nos codes-barres. Ainsi, deux codes-barres proches (relativement aux autres) selon la distance de bottleneck seront représentés par deux points proches par UMAP. Nous utilisons UMAP pour projeter en 2D, avec comme paramètres $\texttt{n\_neighbors} = 45$, $\texttt{min\_dist} = 0.3$ (ces paramètres ont peu d'influence sur la qualité des résultats).

Pour analyser la base de données, nous proposons de partitionner le nuage de points retourné par UMAP en différents groupes (par exemple : sujets sains/malades ou partition selon l'EDSS) et d'analyser ces groupes pour savoir s'ils ont facilement séparables et denses. Pour cela, nous calculons trois caractéristiques pour une partition donnée : le score de silhouette

de chaque paire de groupe, la distance au carrée moyenne de chaque groupe et le diamètre au carré de chaque groupe. Le score de silhouette est défini comme suit : Soit $X = (x_i)_{1 \leq i \leq n}$ un nuage de points partitionné en groupes $(C_i)_{i \in I}$. On note $|C|$ le cardinal d'un ensemble C, $||.||_2$ la norme euclidienne et on considère deux indices distincts $i, j \in I$. Le score de silhouette d'un point $x \in C_i$ par rapport à C_j est :

$$Sil(x, C_j) = \frac{b - a}{\max(a, b)}, \tag{4}$$

où $a = \frac{1}{|C_i|-1} \sum_{y \in C_i, y \neq x} ||x - y||_2$ est la distance moyenne entre x et les points de son groupe, et $b = \frac{1}{|C_j|} \sum_{y \in C_j} ||x-y||_2$ est la distance moyenne entre x et les points du groupe j. Le score de silhouette de C_i par rapport à C_j est alors :

$$Sil(C_i, C_j) = \frac{1}{|C_i|} \sum_{x \in C_i} Sil(x, C_j). \tag{5}$$

Le score de silhouette est utilisé pour juger la qualité d'une partition en termes de *partitionnement*. Il a une valeur entre -1 et 1. Une valeur proche de 1 signifie que les deux groupes sont facilement séparables et denses, une valeur proches de 0 signifie qu'ils se mélangent, une valeur proche de -1 indique un mauvais partitionnement.

La distance au carré moyenne et le diamètre au carré sont deux mesures complémentaires de la densité des groupes. En effet, un groupe peu dense aura un grand diamètre (relativement aux autres) et une (relativement) grande distance moyenne au carré, alors qu'un groupe dense avec un point anormal situé loin des autres aura aussi un grand diamètre mais une plus petite distance au carré moyenne.

4 Résultats

Nous avons appliqué notre méthode à la base de données de signaux de marche décrite dans la section 2 avec plusieurs partitions. La figure 4 représente une partition selon le score EDSS (avec un score de 0 pour les sujets sains) dont le but est d'étudier l'impact de la sévérité de la SEP sur la marche.

On observe sur la figure 4 une évolution globale de la sévérité de la maladie de la gauche vers la droite. Pour quantifier cela, nous avons calculé les scores de silhouette du groupe des points d'EDSS inférieur ou égal à i par rapport au groupe des points d'EDSS supérieur à i, pour tout i. Ces scores montrent qu'il y a toujours une bonne séparation, ce qui confirme l'observation ci-dessus. Les différences d'oscillations détectées par notre distance de bottleneck entre codes-barres reflète donc les différents niveaux de sévérité de la SEP. Par ailleurs, on observe des variations chez des sujets qui ont les mêmes EDSS, ce qui indique que notre méthode pourrait affiner cette échelle clinique semi-quantitative et opérateur-dépendante.

Nous avons aussi pu détecter d'autres phénomènes en partitionnant nos données selon les différentes sessions (M0 ou M6) de chaque individu. Les mesures de densité ont permis de détecter une valeur anormale chez les sujets sains et de découvrir une erreur lors de l'acquisition de ce signal. Certains

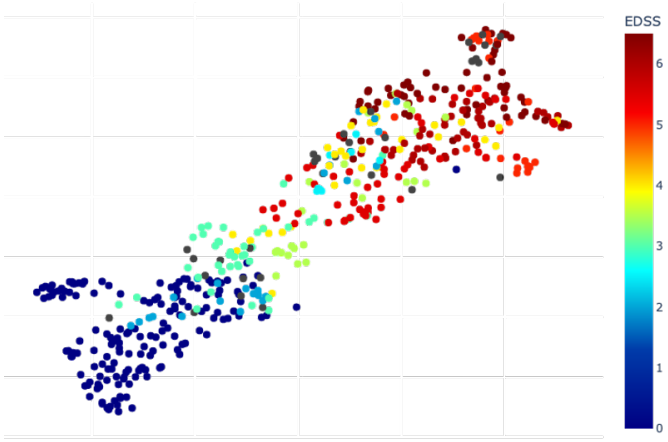

FIGURE 4 – Nuage de points représentant nos signaux de marche, partitionnés selon l'EDSS.

patients ont des groupes correspondants à leurs sessions M0 et M6 très nettement séparables (selon les scores de silhouette), ce que l'on interprète comme une évolution de la maladie : ces patients ont en effet vu leur EDSS augmenter entre les deux sessions. Enfin, en comparant les signaux de pied gauche et droit de certains sujets, nous avons pu détecter de l'asymétrie entre les deux pieds.

Références

[1] H. Edelsbrunner and J. Harer, *Computational topology : an introduction*. American Mathematical Soc., 2010.

[2] N. Ravishanker and R. Chen, "Topological data analysis (tda) for time series," *arXiv preprint arXiv :1909.10604*, 2019.

[3] M. Dindin, Y. Umeda, and F. Chazal, "Topological data analysis for arrhythmia detection through modular neural networks," in *Canadian Conference on Artificial Intelligence*, pp. 177–188, Springer, 2020.

[4] Y. Yan, Y.-S. Liu, C.-D. Li, J.-H. Wang, L. Ma, J. Xiong, X.-X. Zhao, and L. Wang, "Topological descriptors of gait nonlinear dynamics toward freezing-of-gait episodes recognition in parkinson's disease," *IEEE Sensors Journal*, 2022.

[5] A. Vienne, R. P. Barrois, S. Buffat, D. Ricard, and P.-P. Vidal, "Inertial sensors to assess gait quality in patients with neurological disorders : a systematic review of technical and analytical challenges," *Frontiers in psychology*, vol. 8, p. 817, 2017.

[6] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis : an expanded disability status scale (edss)," *Neurology*, vol. 33, no. 11, pp. 1444–1444, 1983.

[7] L. McInnes, J. Healy, and J. Melville, "Umap : Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv :1802.03426*, 2018.