Utilisation de la théorie des signatures dans la construction d'hypergraphes dynamiques pour des signaux multi-dimensionnels.

Rémi VAUCHER[1] Stéphane CHRÉTIEN[2] Paul MINCHELLA[3]

[1]Halias Technologies, Meylan, France

[2]Laboratoire ERIC, Université Lumière Lyon 2, Bron, France

[3]Centre Léon Bérard, Lyon, France

Résumé – Depuis quelques années, la détection d'anomalie a pris une place prépondérante en traitement du signal. Ces anomalies peuvent prendre des formes très variées, et sont souvent difficiles à mettre en évidence par des principes généraux. Dans le présent travail, nous nous intéressons à des signaux multidimensionnels et cherchons quelles sont les interdépendances entre ces signaux. Motivés par les récentes avancées en Analyse Topologique des Données, nous proposons une nouvelle méthode pour inférer ces interdépendances à partir de la construction d'un hypergraphe fondés sur le calcul des signatures associées à ces signaux. La théorie des signatures, outil dont on trouve l'origine dans les travaux de K.T. Chen [2], et développé plus récemment par Terry Lyons à partir de 2002 [11], est un domaine actuellement très exploré en Machine Learning [10]. Nous montrons dans le présent travail que les Signatures permettent de construire des Hypergraphes dont la structure caractérise la présence d'une possible anomalie.

Abstract – In recent years, anomaly detection has taken on a prominent role in signal processing. These anomalies can take many different forms and are often difficult to identify using general principles. In the present work, we focus on multidimensional signals and aim to determine the interdependencies between them. Motivated by the recent advances in Topological Data Analysis, we propose a new method to infer these interdependencies based on the construction of a hypergraph derived from the computation of signatures associated with these signals. Signature theory, originally introduced in the work of K.T. Chen [2] and more recently developed by Terry Lyons since 2002 [11], is currently a highly active area of research in Machine Learning [10]. In this work, we show that Signatures make it possible to construct Hypergraphs whose structure characterizes the presence of a potential anomaly.

1 Introduction.

1.1 Analyse Topologique des Données et Hypergraphes intra/inter-signaux.

L'Analyse Topologique des Données (TDA) est un ensemble de techniques très récentes et en pleine expansion [1]. Elle englobe des méthodes consistant à construire une suite croissante de **complexes simpliciaux**, cas particuliers d'**hypergraphes**, sur un ensemble de sommets, voir Section 1.4. Usuellement utilisée pour étudier la forme des nuages de points, des méthodes émergentes permettent maintenant d'utiliser ces outils sur des signaux multi-dimensionnels [1, 12].

Nous présentons dans le présent travail une méthode pour créer ce type de structure. Notre algorithme généralise l'approche que nous avons proposée dans [4] en permettant l'analyse de signaux plus généraux, en particulier

- **les signaux multivariés :** est considéré un ensemble de variables temporelles $X(t) = (X^1(t), ..., X^d(t))$. Un hypergraphe peut être construit sur ces signaux permettant de mettre en évidence et de représenter une structure de multi-dépendances éventuellement évolutives. A titre d'exemple, un signal issu d'un électroencéphalogramme est mesuré par un grand nombre de capteurs monodimensionnels au sein desquels il est scientifiquement important de chercher une organisation en fonction des possibles stimuli appliqués.

  Dans ce cas précis, notre méthode crée un hypergraphe à partir de données temporelles dans $BV([0;T],\mathbb{R})$ (i.e. **chemins à variations bornées** voir Définition , Section 1.3), chaque sommet représentant un $X^i \in BV([0;T],\mathbb{R})$.

- **les signaux multivariés multi-sources :** par "multivariés multi-sources", nous entendons ici des signaux multivariés pour lesquels des liens en hautes dimensions peuvent émerger. Nous avons donc $X = (X_1(t), ..., X_n(t))$ où **chaque** X_i **est multivariée :** pour tout i, $X_i(t) = \left(X_i^1(t), ..., X_i^{d_i}(t)\right)$. Il est à noter que **nous ne requérons aucune homogénéité des dimensions des signaux**.

  Dans ce cas précis, nous créons un hypergraphe à partir de données dans $BV([0,T],\mathbb{R}^d)$ où $d = \max_i d_i$.

- **les images/surfaces 2D :** nous entendons ici un ensemble d'objets dont chacun possède une évolution sur deux **temps** (t_1, t_2) **distincts**. Un exemple immédiat serait un hypergraphe régissant un ensemble de vidéos. Cette déclinaison de notre méthode reste pour l'instant théorique et s'appuie sur [9, 8] pour lesquels un code n'est pas encore disponible, mais qui devrait être implémentés dans un futur proche.

1.2 Exemple introductif.

Comment détecter un début d'épidémie en prenant en compte des **pathologies émergentes** ? Une solution est d'appliquer un algorithme de détection d'anomalies à des séries de comp-

tage représentant les possibles pathologies, par exemples les symptômes des personnes admises aux urgences.

Notre solution est de considérer ces données de comptages réparties **géographiquement**. Chaque région/hôpital est représentée par une série temporelle multidimensionnelle $\text{Symp}_{\text{Zone } i}(t) = (\text{Symptome}_{\text{Zone } i,1}(t), ..., \text{Symptome}_{\text{Zone } i,d}(t))$.

Il est ensuite logique de se dire que les patients vont se déplacer et contaminer d'autres régions. Notre but est d'étudier ces **graphes d'interactions** entre régions, et donc entre séries temporelles multidimensionnelles, pour détecter un changement caractérisant un début d'anomalie.

Pour être plus précis, nous étudions, par le biais des **signatures** comment la **géométrie d'un série temporelle multidimensionnelle** va être en mesure **d'expliquer** la géométrie d'une autre série.

1.3 Signatures de chemins rugueux.

Définition 1 *Soit un **chemin**[1] $X = (X^1, ..., X^d) : [0;T] \to \mathbb{R}^d$. X est dit à **variations bornées** si*

$$V(X) = \sup_{S \in \mathcal{S}} \sum_{i=0}^{n_p - 1} \|X_{t_{i+1}} - X_{t_i}\| < \infty \tag{1}$$

pour $\mathcal{S}$ l'ensemble de toutes les subdivisions de la forme $\{t_1, \dots, t_{n_p}\}$ avec $t_i \in [0,T]$, $t_i < t_{i+1}$. L'ensemble des fonctions à variations bornées de $[0;T]$ dans $\mathbb{R}^d$ est noté $BV([0;T], \mathbb{R}^d)$

Les chemins à variations bornées sont nos objets de base, la majorité des résultats en théorie des Signatures requérant cette hypothèse. De plus, tout signal échantillonné, considéré comme une interpolation linéaire entre ses points d'échantillonnages, est automatiquement à variations bornées. Nous pouvons donc définir la **signature** de ces chemins :

Définition 2 *Soit $X \in BV([0;T], \mathbb{R}^d)$. La **signature** du chemin X est une collection de tenseurs*

$$S_{[0;T]}(X) = (1, S^{(1)}_{[0;T]}(X), \dots, S^{(k)}_{[0;T]}(X), \dots) \tag{2}$$

*dont chaque tenseur $S^{(k)}_{[0;T]}(X) \in \mathbb{R}^{\underbrace{d \times d \times \cdots \times d}_{k \text{ fois}}}$ est créé à partir des **intégrales itérées de** X :*

$$\left(S^{(k)}_{[0;T]}(X)\right)_{i_1 \dots i_k} = \int \cdots \int_{0 < t_1 < \cdots < t_k < T} dX^{i_1}(t_1) \dots dX^{i_k}(t_k). \tag{3}$$

Pour des raisons computationnelles, nous ne considérerons, sauf mention contraire, que la **signature tronquée à l'ordre** k : $S^{\leq k}(X) = \left(1, S^{(1)}_{[0;T]}(X), \dots, S^{(k)}_{[0;T]}(X)\right)$

La signature d'un chemin possède de nombreuses propriétés qui font d'elle un outil très pertinent pour le traitement du signal.

- Invariance par reparamétrisation et translation : peut importe la vitesse ou le lieu de parcourt du chemin, la signature reste la même.
- La signature est caractéristique de la loi du signal : elle contient toutes les informations pour déterminer de manière unique la loi jointe d'un chemin.
- La signature (de dimension infinie) est bijective.

Pour plus de détails, nous renvoyons le lecteur à [3].

Proposition 1 *L'espace des signatures tronquées à l'ordre k de chemins d-dimensionnels, noté $G^k(\mathbb{R}^d)$, est **un groupe de Lie**.*

A ce titre, $G^k(\mathbb{R}^d)$ est munie d'une **topologie** induite et d'une **algèbre de Lie** permettant la linéarisation des problèmes. Pour plus de détails quand à cette structure et sa construction, nous conseillons vivement [6].

De plus, parlant de topologie, l'application signature est un homéomorphisme de $BV([0;T], \mathbb{R}^d)$ dans $G(\mathbb{R}^d)$. Cette propriété motive l'idée d'étudier la topologie des signaux au travers de celle de l'espace des signatures. Pour ce faire, nous mobilisons la notion de **complexe simplicial**.

1.4 Complexes simpliciaux.

Nous considérons ici un ensemble $V = \{v_1, ..., v_N\}$ de sommets.

Définition 3 *Un **simplexe (abstrait) de dimension** n σ_n est la donnée d'un ensemble de $n+1$ sommets $\{v_{i_1}, ..., v_{i_{n+1}}\}$ et de tous ses sous ensembles possibles.*

*Lorsque c'est possible, σ_n est représenté par **l'enveloppe convexe** de $n+1$ sommets $\{v_{i_1}, ..., v_{i_{n+1}}\}$ linéairement indépendants. On dit alors que σ_n est une **réalisation géométrique** du n-simplexe.*

*Si σ est un n-simplexe (simplexe de dimension n) composé des sommets $v_{i_1}, v_{i_2}, \dots, v_{i_n}$, nous noterons généralement $[v_{i_1} \dots v_{i_n}]$. Cette notation est générique et ne **définit pas une orientation sur** σ.*

Cette distinction est très importante : la version géométrique permet de visualiser graphiquement le principe des n-simplexes, comme illustré dans les Figures 1 et 2. Notons que les sommets sont de dimension infinie. Notons de plus que nous ne considérons pour l'instant aucune notion d'orientation sur nos simplexes. Cette feature additionnelle sera abordée en détail dans un travail ultérieur.

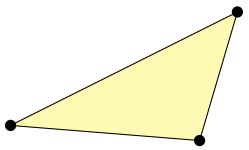

FIGURE 1 : 2-simplex

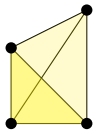

FIGURE 2 : 3-simplex

[1] Un chemin est juste une fonction de $[0;T]$ dans $\mathbb{R}^d$ pour un certain d.

Remarque : Un n-simplexe est l'équivalent en TDA d'une **hyper-arête** en théorie des hypergraphes. De fait, nous utiliserons indifféremment ces deux appellations au cours de cette présentation.

Définition 4 *Un **complexe simplicial** $\mathcal{C}$ sur un ensemble fini de sommets V est une collection $\{\sigma_k\}$ de simplexes tels que :*

- $v \in \mathcal{C}$ *si* $v \in V$
- $\tau \subseteq \sigma \in \mathcal{C} \Rightarrow \tau \in \mathcal{C}$

*La **dimension** d de $\mathcal{C}$ est le plus grand entier k tel qu'il existe un k-simplexe dans $\mathcal{C}$.*

En surfant sur les deux théories (hypergraphes et TDA), il est possible d'utiliser une large batterie d'outils d'analyse :

- La théorie des hypergraphes nous apporte des **distances** entre hypergraphes. Il est alors possible de quantifier avec précision la différence simplexe à simplexe. Nous renvoyons le lecteur à [13] pour plus de détails.
- La topologie des données nous apporte des **caractérisations topologiques** à déformations continues près. Il est alors possible de quantifier la différence de forme entre nos complexes simpliciaux. Nous renvoyons le lecteur à [1] pour plus de détails.

2 Création de complexes simpliciaux dans l'espace des signatures.

Avant tout, nous définissons la notion de **link**.

Définition 5 *Soit σ un n-simplexe dans un complexe simplicial $\mathcal{C}$. Son **link** dans $\mathcal{C}$ est l'ensemble* $\mathrm{Link}(\sigma)$ *de tous les simplexes $\tau \in \mathcal{C}$ tels que :*

- $\sigma \cap \tau = \varnothing$
- *Le simplexe défini par $\sigma \cup \tau$ appartient à $\mathcal{C}$*

Exemples :

- Soit un sommet v_i dans $\mathcal{C}$. Un sommet $v_j \in \mathcal{C}$, $j \neq i$ est dans le link de v_i si et seulement si l'arête $[v_i v_j]$ est dans $\mathcal{C}$.
- On considère une arête $[v_i v_j]$. Un triangle $[v_k v_l v_m]$ est dans le link de $[v_i v_j]$ si et seulement si le 5-simplexe $[v_i v_j v_k v_l v_m]$ appartient à $\mathcal{C}$.

En complément, on définit le **link k-dimensionnel** comme le sous-ensemble des k-simplexes de $\mathrm{Link}(\sigma)$. Notre méthode **construit itérativement** les links du complexe recherché comme une union des links k-dimensionnel. Ce complexe recherché est un **complexe d'explicabilité** : *deux sommets sont liés si l'un est explicable par l'autre avec un modèle linéaire*.

Dans ce qui suit, les sommets v_j sont :

- Soit des chemins bidimensionnels $(t, X^i(t))$ dans le cas où l'on crée le complexe simplicial régissant le signal $X = (X^1, ..., X^d)$.
- Soit des chemins multidimensionnel dans le cas où l'on crée le complexe simplicial régissant les signaux $(X_1, ..., X_N)$ où $X_i = (X_i^1, ..., X_i^{d_i})$ pour tout $1 \leq i \leq N$.

Algorithme 1 : Prédiction du Link k-dimensionnel de v_i.

Données : Un ensemble de sommets $V = \{v_1, ..., v_N\}$; un indice i ; une dimension k ; un intervalle $[a; b] \subset [0; T]$.
Résultat : Le link $(k-1)$-dimensionnel de v_i.

1 Calcul de la signature S_i du sommet v_i sur $[a; b]$ (éventuellement **augmenté** pour la cohérence des dimensions)
2 Pour tout sous-ensemble $v_{i_1}, ..., v_{i_k}$ de longueur k, calculer la signature $S_{i_1 \dots i_k}$ du chemin défini par le multi-sommet $\{v_{i_1}, \dots, v_{i_k}\}$ sur $[a; b]$.
3 Par un algorithme de sélection de variable, déterminer les coefficients de la régression

$$S_i = \sum_{\{v_{i_1}, \dots, v_{i_k}\} \subset V \setminus \{v_i\}} \beta_{i_1 \dots i_k} S_{i_1 \dots i_k}. \quad (4)$$

pour $\{v_{i_1}, ..., v_{i_k}\} \subset V \setminus \{v_i\}$ **faire**
4 **si** $\beta_{i_1 \dots i_k} \neq 0$ **alors**
5 Inclure le $(k-1)$-simplexe $[v_{i_1} \dots v_{i_k}]$ dans le link $(k-1)$-dimensionnel de v_i.
6 **fin**
7 **fin**

Le principe fondamental de notre méthode repose sur le fait qu'un complexe simplicial sur un ensemble de sommet V **est l'union des link de tous ses sommets**, et le link d'un sommet est l'union de tous ses link k-dimensionnels. Nous obtenons donc au final l'Algorithme 2

Algorithme 2 : Construction du complexe simplicial d'explicabilité sur V

Données : Un ensemble de sommets $V = \{v_1, \dots, v_k\}$; une dimension K ; un intervalle $[a; b] \subset [0; T]$.
Résultat : Le complexe simplicial d'explicabilité de l'ensemble V.

1 **pour** $v_i \in V$ **faire**
2 **pour** $k \leq K - 1$ **faire**
3 Prédire le Link k-dimensionnel de v_i grâce à l'algorithme 1
4 **fin**
5 **pour** $\sigma = [v_{i_1} \dots v_{i_k}]$ *dans le link k-dimensionnel de v_i* **faire**
6 Inclure le simplexe $[v_i v_{i_1} \dots v_{i_k}]$ dans $\mathcal{C}$.
7 **fin**
8 **fin**

3 Expérimentations.

Pour des expérimentations sur données réelles sur signaux multivariés, nous renvoyons le lecteur à [4, 5].

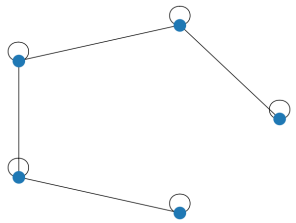

FIGURE 3 : Graphe généré. Chaque sommet est un signal bivarié Y_i.

Dans cette expérimentations, nous tentons de prédire le graphe d'explicabilité dans un système d'échanges entre plusieurs signaux multivariés.

Nous générons 5 signaux bivariés $(Y_1, ..., Y_5)$, avec $Y_i = (Y_i^1, Y_i^2)$ selon le schéma suivant :

- $Y_1(0) \sim \mathcal{N}\left((0,0), \Sigma^2_{start}\right)$,
- $Y_i(0) = (0,0)$ pour $i \neq 1$,
- $$\begin{cases} Y_1(t) & = AY_1(t-1) + CY_2(t-1) + \epsilon_1(t) \\ Y_i(t) & = AY_i(t-1) + BY_{i-1}(t-1) + CY_{i+1}(t-1) \\ & \quad +\epsilon_i(t) \quad \text{si} \quad i \in \{2,3,4\} \\ Y_5(t) & = AY_5(t-1) + BY_4(t-1) + \epsilon_5(t) \end{cases},$$
  $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ de sorte d'obtenir le graphe présent en Figure 3.
- La matrice d'adjacence M (théorique) attendue entre les sommets $(Y_i)_{i\in\{1,...,5\}}$ est

$$M = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

3.1 Résultats.

Dans le cas normalisé, les résultats sont convaincants : nous avons appliqué notre méthode sur un échantillon de 100 signaux $(Y_1, ..., Y_5)$. La moyenne des matrices d'adjacences estimées est :

$$\begin{pmatrix} 0 & 0.8 & 0.27 & 0.29 & 0.29 \\ 0.82 & 0 & 0.66 & 0.23 & 0.42 \\ 0.32 & 0.66 & 0 & 0.69 & 0.34 \\ 0.37 & 0.26 & 0.72 & 0 & 0.75 \\ 0.36 & 0.33 & 0.31 & 0.75 & 0 \end{pmatrix} \tag{5}$$

Ces résultats montrent de manière convaincante une propension de notre méthode à récupérer les bons liens de corrélation.

4 Conclusion et perspectives.

Dans cette contribution, nous généralisons la méthode proposée dans [4] avec l'optique de la rendre applicable à des familles de signaux plus riches et permettant d'accéder à des applications plus complexes.

Dans une contribution future, nous prévoyons d'extraire une orientation sur les faces des simplexes reflétant une relation de causalité au sens de [7] et de tester la cohérence globale des orientations pour comprendre les relations de dépendances entre les signaux multidimensionnels.

Références

[1] Frédéric CHAZAL et Bertrand MICHEL : An introduction to topological data analysis : fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:108, 2021.

[2] Kuo-Tsai CHEN : Integration of paths, geometric invariants and a generalized baker-hausdorff formula. *Annals of Mathematics*, 65(1):163–178, 1957.

[3] Ilya CHEVYREV et Andrey KORMILITZIN : A primer on the signature method in machine learning. *arXiv preprint arXiv :1603.03788*, 2016.

[4] Stéphane CHRÉTIEN, Ben GAO, Astrid THÉBAULT GUIOCHON et Rémi VAUCHER : Leveraging the power of signatures for the construction of topological complexes for the analysis of multivariate complex dynamics. *In International Conference on Complex Networks and Their Applications*, pages 283–294. Springer, 2023.

[5] Stéphane CHRÉTIEN, Ben GAO, Astrid THEBAULT-GUIOCHON et Rémi VAUCHER : Time topological analysis of eeg using signature theory. *arXiv preprint arXiv :2404.15328*, 2024.

[6] Peter K FRIZ et Nicolas B VICTOIR : *Multidimensional stochastic processes as rough paths : theory and applications*, volume 120. Cambridge University Press, 2010.

[7] Chad GIUSTI et Darrick LEE : Iterated integrals and population time series analysis. *In Topological Data Analysis : The Abel Symposium 2018*, pages 219–246. Springer, 2020.

[8] Darrick LEE : The surface signature and rough surfaces. *arXiv preprint arXiv :2406.16857*, 2024.

[9] Darrick LEE et Harald OBERHAUSER : Random surfaces and higher algebra. *arXiv preprint arXiv :2311.08366*, 2023.

[10] Terry LYONS et Andrew D MCLEOD : Signature methods in machine learning. *arXiv preprint arXiv :2206.14674*, 2022.

[11] Terry LYONS et Zhongmin QIAN : *System control and rough paths*. Oxford University Press, 2002.

[12] Andrea SANTORO, Federico BATTISTON, Giovanni PETRI et Enrico AMICO : Unveiling the higher-order organization of multivariate time series. *Nature Physics*, 19(2):221–229, 2023.

[13] Amit SURANA, Can CHEN et Indika RAJAPAKSE : Hypergraph similarity measures. *IEEE Transactions on Network Science and Engineering*, 10(2):658–674, 2022.