

N° National de Thèse : XXX



THÈSE

en vue de l'obtention du grade de

Docteur de l'Université de LYON
délivré par Université Lumière Lyon 2

Discipline : Mathématiques Appliquées

Laboratoire ERIC

École Doctorale Infomaths

Présentée et soutenue publiquement le 9 janvier 2025
par **Rémi VAUCHER**

Détection d'anomalies sur signaux temporels par la méthode des signatures et l'analyse topologique des données.

Directeur de Thèse : M. Stéphane CHRÉTIEN Encadrant industriel : M. Laurent TESTARD

Devant la commission d'examen formée de :

M.	Stéphane CHRÉTIEN	<i>Université Lumière Lyon 2</i>	Directeur
M.	Julien JACQUES	<i>Université Lumière Lyon 2</i>	Président
M.	Frédéric CHAZAL	<i>INRIA</i>	Examinateur
Mme.	Gabriela CIUPERCA	<i>Université Claude Bernard Lyon 1</i>	Examinateuse
Mme.	Marianne CLAUSEL	<i>Université de Lorraine</i>	Rapportrice
M.	Badhi GATTAS	<i>Université d'Aix-Marseille</i>	Rapporteur
Mme.	Lyudmila GRIGORYEVA	<i>University of St Gallen</i>	Examinateuse
M.	Laurent TESTARD	<i>Halias Technologies</i>	Examinateur

Laboratoire ERIC
5 av. Pierre Mendès France
69500 Bron

École doctorale 512
INSA de Lyon - Batiment Blaise
Pascal
Campus LyonTech - La DOua 7 av.
Jean Capelle
69621 Villeurbanne CEDEX

Remerciements

Merci à tout plein de gens.

Table des matières

1	Introduction	3
1.1	Motivation(s)	3
1.1.1	Un peu de contexte	3
1.1.2	Plan de la thèse	4
1.2	La détection d'anomalies.	5
1.2.1	Définition informelle	5
1.2.2	Revue de littérature (non exhaustive) sur la détection d'anomalie.	7
1.3	La signature d'un chemin à variations bornées.	8
1.3.1	Chemins à variations bornées	8
1.3.2	Signature d'un chemin à variations bornées	10
1.3.3	L'espace $\mathcal{G}(\mathbb{R}^d)$	17
1.3.4	Calcul effectif des signatures	19
1.3.5	Injectivité de l'application signature.	21
1.3.6	Propriétés analytiques	24
1.4	Analyse topologique des données	24
1.4.1	Complexe simplicial, de Čech, de Vietoris Rips	25
1.4.2	Homologie persistante.	31
2	Analyse topologique de plusieurs signaux temporels dans l'espace des signatures.	45
2.1	Introduction	46
2.2	Background on Signatures and topology	47
2.2.1	Recalls on the theory of Signatures	47
2.2.2	Recalls on topology	48
2.3	Continuity, differentiability for signature transform	49
2.6	Building simplicial complexes between time series	53
2.6.1	Presentation of the method	53
2.6.2	Our greedy order stratified algorithm :	53
2.6.3	Known results about the LASSO and the construction of simplicial complexes	56
2.7	Numerical experiments and Applications	57
2.7.1	Practical choices and hyper parameters	57
2.7.2	Modelling interactions in Functional MRI datasets	58
2.8	Discussion and future work	59
3	Sélection de variable en régression entière : Algorithme de Lagarias Odlyzko étendu.	61

TABLE DES MATIÈRES

3.1	Introduction	61
3.2	Background on high dimensional regression and variable selection	62
3.2.1	High-Dimensional Regression and support recovery for the LASSO estimator	62
3.3	Estimating integer relations in the regression vector : LLL and friends	65
3.3.1	Gram-Schmidt orthogonalisation	65
3.3.2	Lattices	67
3.3.3	Lovász reduced bases and the LLL algorithm	68
3.4	Extended Lagarias-Odlyzko Algorithm	70
3.4.1	Historical background : the sumset problem	70
3.4.2	The extended Lagarias Odlyzko algorithm	70
3.4.3	Theoretical results.	72
3.5	Proof of Theorem 3.10	73
3.5.1	Preliminary result	73
3.5.2	Main proof	74
4	Applications en oncologie, neuroscience et acoustique	81
4.1	Détection de dynamiques malignes en oncologie.	82
4.1.1	Introduction	82
4.1.2	A stochastic model for the dynamics of tumorous biomarkers	83
4.1.3	Details on the biomarker dynamics	84
4.1.4	The mathematical tools : The Signature transform, Principal Geodesics and Topological Data Analysis	87
4.1.5	Our statistical testing procedure	94
4.1.6	Visualisation of the data and analysis of the delineability of the benign/malignant classes	97
4.1.7	Empirical performance of our testing procedure	98
4.1.8	Future works	99
4.2	Time topological analysis of EEG using signature theory	102
4.2.1	Introduction	102
4.2.2	Empirical study	105
4.3	Etude de données acoustiques.	110
4.3.1	Séparation des anomalies	110
4.3.2	Conclusion	110
5	Générer des signaux multivariés sous une loi donnée à partir des signatures : le problème de l'inversion.	115
5.1	Introduction	115
5.2	Signature estimation and signal recovery using Median of Means	116
5.2.1	Signature of paths from their coefficients in a dictionary	116
5.2.2	Anti-concentration for the 3-Signature coefficients	117
5.2.3	Estimation of the signal decomposition using Median of Means (MoM)	119
5.3	Inversion de la signature : pistes vers une amélioration de l'algorithme Pfeffer-Seigal-Sturmfelds (PSS).	122
5.3.1	Ajuster la fonction objectif à la structure de groupe de $G^N(\mathbb{R}^d)$	123
5.3.2	S'affranchir de la restriction au plus court chemin	124
5.4	Expériences	125

5.5 Perspectives	127
Conclusion	131
Bibliographie	133
A "Rudiments" sur les groupes de Lie	141
A.1 Rappels rapides pour non algébristes.	141
A.2 Groupes de Lie	142
A.3 Algèbre de Lie	143
A.3.1 Action d'un groupe (de Lie) sur une variété : algèbre de Lie associée à un groupe de Lie.	145
A.3.2 Application exp et log	147

Travaux publiés ou en vue de soumission :

- Une méthode de détection d'anomalie sur image radar conçue pour la société HALIAS a été acceptée aux Journées de la Statistique 2024 qui s'est tenue à Bordeaux [4], ainsi qu'à la conférence CAP 2024 qui s'est tenue à Lille.
- Le Chapitre 2 est une révision de [26] publié dans les proceedings de la conférence internationale Complex Networks 2023 à Menton, France.
- Le Chapitre 3 fera l'objet d'une soumission à la conférence internationale EUSIPCO 2025.
- Le Chapitre 4 contient un travail accepté aux Journées de la Statistiques 2024 [27] et un travail à soumettre à la conférence internationale IDA 2025 qui se tiendra à Konstanz.
- Le Chapitre 5 contient une publication parue dans les proceedings de la conference internationale GSI 2023 qui s'est tenue à Saint Malo [28].
- La Conclusion du manuscrit mentionne le travail original (en construction) présenté à l'École d'automne en statistique bayésienne 30 Octobre-3 Novembre, 2023 [29].

Chapitre 1

Introduction

Liminaire, pas luminaire !

Pierre Desproges

1.1 Motivation(s)

1.1.1 Un peu de contexte

Cette thèse CIFRE est le fruit d'une collaboration étroite entre le laboratoire ERIC de l'Université Lumière Lyon 2 et Halias Technologies. Plutôt que de me focaliser sur un seul algorithme dédié à une tâche précise, Halias m'a laissé la chance de travailler majoritairement de manière exploratoire. C'est pourquoi, si vous ne regardez que la table des matières, cette thèse vous paraîtra peut être décousue¹. Il est donc important de lire ce paragraphe pour saisir le fil rouge de ce manuscrit.

Mon travail s'est construit au fil du temps autour des axes suivants :

- ⇒ **La détection d'anomalies** : Tout d'abord sur des images radar, puis finalement sur des signaux temporels multidimensionnels (ce que l'on appellerait en statistiques des séries temporelles multivariées), l'un des axes principaux de cette thèse était de développer des algorithmes indiquant à l'utilisateur : "*Ici, ça n'est pas comme d'habitude/partout ailleurs.*" Ces algorithmes avaient, et ont toujours, pour vocation de s'intégrer dans une plateforme de monitoring à plus grande échelle.
- ⇒ **De l'explicabilité** : Cet axe s'est imposé naturellement à moi de par la présence de Ben Gao en alternance à Halias. Plus compétent que moi sur tout ce qui touche aux réseaux de neurones, je lui ai laissé cette partie pour me concentrer sur des algorithmes impliquant un maximum d'outils mathématiques dont les résultats pouvaient s'expliquer de manière explicite.
- ⇒ **De belles mathématiques** : Toutes les mathématiques sont belles. Pour autant, chaque mathématicien possède sa propre sensibilité, encore plus arrivé à haut niveau. La mienne se situait à l'intersection de la géométrie (étude quantitative des formes)

1. Ce que Boulet dans [8] appellera "le syndrôme de Frankenstein"

et de la topologie (étude qualitative des formes)². Elle se positionne maintenant dans une zone encore plus précise : l'intersection des deux domaines précédents et celui des statistiques ou/et probabilités.

⇒ **De l'application aux données réelles :** Tout d'abord, dans une thèse CIFRE, on ne peut se soustraire aux applications réelles. Aucune entreprise ne souhaite d'algorithmes qui ne sont efficaces que sur des processus stochastiques parfaitement synthétiques. Mais plus que le côté industriel, j'aimerais pouvoir apporter ma propre réponse à la question :

"Hey m'sieur m'sieur ! En vrai de vrai, ça sert à quoi les maths dans la vraie vie ?"³

Un grand nombre d'apports de cette thèse sont motivés par des sujets de recherches en médecine (notamment grâce à mes comparses neuroscientifiques). Il faut bien garder à l'esprit que les données traitées au sein d'Halias sont très souvent propriétés intellectuelles d'autres entreprises : leur exploitation dans des articles de recherche est donc exclue. Dans certains cas, nous avons pu compenser par des jeux de données publics. C'est la raison pour laquelle les études menées pour Halias ne seront pas toutes présentes dans cet écrit.

1.1.2 Plan de la thèse

Pour démarrer, à la suite de cette sous-partie, nous allons introduire le contexte global sur la détection d'anomalies, sur les signatures, et sur la topologie des données.

Le premier chapitre non-liminaire introduit la méthode principale de cette thèse : un algorithme de création de complexe simplicial sur une série multivariée. Cet algorithme propose deux innovations : l'introduction de l'espace des signatures et l'application d'un algorithme de sélection de variables. Ces deux outils permettront de créer une structure topologico-statistique naturelle en mobilisant tous les leviers que fournit la théorie des signatures. Ce travail sera l'occasion d'étudier une nouvelle propriété théorique des signatures : *L'application est un C^1 -homéomorphisme de l'espace des chemins de variations bornées (à diverse relation d'équivalence près) sur son image.*

De manière assez logique, l'algorithme SigTopo nécessite une procédure de sélection de variable. Dans le chapitre 3 l'utilisation de ce que nous pourrions appeler de manière simpliste un "LASSO à coefficient entier" : l'algorithme Extended Lagarias-Odlyzko. Nous prouvons une amélioration des hypothèses de cet algorithme dans le théorème ? grâce à un résultat provenant de la théorie autour du problème de Littlewood-Offord.

Après l'introduction des méthodologies théoriques des chapitres précédents, les applications pratiques sont abordées dans le Chapitre 4. Nous y mettons en place des approches efficaces pour des challenges en analyse de données de Science Cognitive, en Oncologie et en Acoustique. Nous établirons aussi de quelles manières nous pouvons utiliser la théorie des Signatures et/ou la topologie des données dans un contexte de détection d'anomalie

2. Ces deux parenthèses sont des interprétations très personnelles et sont sujet à débat.

3. comme le pensait une proportion de mes anciens élèves du lycée Robert Doisneau de Vaulx-en-Velin.

(phénomène hors distribution ou pathologie).

Enfin, dans un dernier chapitre, nous aborderons quelques sujets en cours ou mis de côté mais représentant un intérêt certain pour la détection d'anomalies sur des séries multivariées. Ces travaux annexes portent sur la génération aléatoire de données temporelles, par l'utilisation de signatures. Ces travaux impliquent notamment une réflexion sur un problème encore difficile : l'inversion de l'application signature.



Remarque(s)

Les expériences de chaque chapitre sont accessibles sur <https://remivaucher.github.io/manuscrit>

1.2 La détection d'anomalies.

1.2.1 Définition informelle

Commençons par le commencement : qu'est ce qu'une anomalie ? Mathématiquement, c'est très difficile à caractériser. Pour citer [15] : *déetecter une anomalie, c'est détecter un schéma qui ne correspond pas à celui attendu.*

Mathématiquement, il est assez intuitif de se dire qu'une anomalie dans un échantillon est une observation dont la loi est différente de celle de l'échantillon. Or, plusieurs phénomènes viennent contredire cette intuition :

- ⇒ Si une observation est de probabilité très faible mais **non nulle**, peut-on vraiment parler d'anomalie ? Cela reviendrait à qualifier toute observation rare d'anomalie.
Exemple : Lors de l'élection présidentielle de 2017, seul 0.18% des votants (et donc 0.14% de la population française ayant la possibilité de voter) ont voté pour Jacques Cheminade. Si nous choisissons un français au hasard, la probabilité qu'il ait voté pour J. Cheminade est donc < 0.14%. Si toutefois cela arriverait, serait-ce une anomalie ?
- ⇒ Si maintenant il est clair que l'observation en cause est de probabilité nulle *au vu de l'échantillon*, il n'en reste pas moins que la loi **théorique** de l'échantillon (et non des observations) n'est peut-être plus la même. On parlerait alors de **dérive de distribution**.
Exemple : Les épisodes climatiques forment un exemple de dérive de distribution. Le réchauffement global de la planète entraîne des changements dans la loi de probabilité régissant les événements climatiques (tornade, cyclone, tempête, canicules, etc.) Une canicule en Russie risque, avec le temps, de passer du statut d'anomalie à celui d'événement rare.
- ⇒ Selon la problématique métier, la notion d'anomalie n'est pas la même. En changeant juste de paradigme, un comportement normal peut devenir anormal, et inversement.

Voyons plusieurs exemples d'anomalies dans divers domaines d'application.

Les fraudes bancaires : Plusieurs jeux de données proposent des observations de retrait/paiement bancaire. L'objectif ici est de déterminer les transactions traduisant un vol de carte/identité. De nombreuses solutions ont été publiées qu'elles soient basées sur un réseau de neurones [1] ou du clustering [6].

Pour autant, comment différencier une transaction frauduleuse d'une transaction rare (on pourra penser à une dépense pour vacances/déménagement) ? Dans ce cas précis, le problème ne se pose pas : la plupart des jeux de données libres sont annotés. Toutefois, la question vient à se poser si l'on vient à envisager un algorithme non-supervisé.

Les fuites en pleine eau : Ce domaine est beaucoup plus complexe à traiter, et il fut le point de départ de cette thèse. La difficulté majeure est la disponibilité des données anormales. En effet, il existe des images radar disponibles en libre accès. Pour autant, les images radar de fuites de pétrole et/ou gaz sont rarement partagées. De plus, il est intéressant de se dire qu'il existe des évènements rares normaux : les fuites biologiques 1.1.

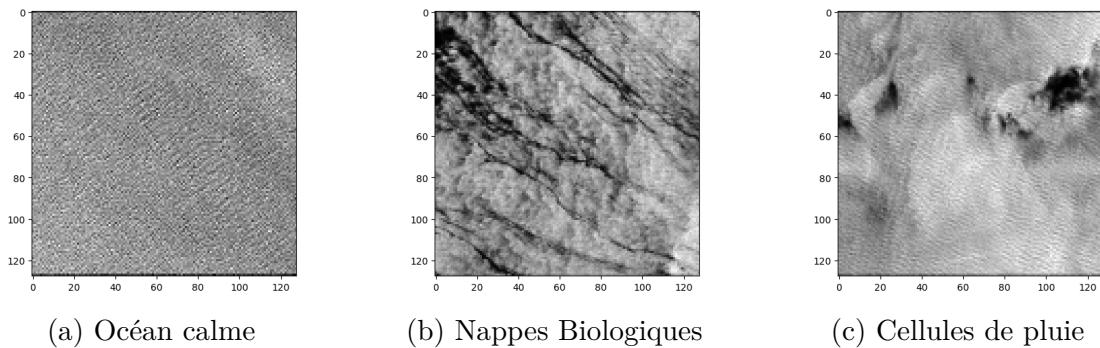


FIGURE 1.1 – Différentes situation en imagerie radar.

Le second aspect intéressant réside dans le type de données. En effet, la détection d'anomalies sur image représente un domaine très vaste et en constante évolution. Dans la section 1.2.2, nous donnerons deux exemples d'algorithmes capables de gérer ce type de problème.

Le niveau sonore de chantier : Pour assurer un niveau sonore raisonnable en ville, les municipalités disposent du droit de stopper tout chantier dont le niveau sonore dépasserait un certain seuil. Pour cela, des microphones sont installés à proximité et enregistrent le niveau sonore produit par le chantier.

Il est toutefois nécessaire de faire la différence entre :

- Une anomalie due au chantier qui a augmenté son niveau sonore : le chantier **doit** être arrêté.
- Un défaut du microphone qui produit alors un bruit blanc très intense : le chantier **ne doit pas** être arrêté et le microphone doit être changé/réparé.
- Un chien allemand énervé qui aboie à moins d'un mètre du microphone : le chantier **ne doit pas** être arrêté et le microphone **ne doit pas** être changé.

Ce problème de détection d'anomalies est intéressant dans le sens où plusieurs types d'anomalies peuvent apparaître. Pour cela, nous disposons généralement d'une décomposition fréquentielle du signal sonore (donc un signal temporel multivarié) accompagné d'un niveau sonore moyen sur une période donnée (LAeq)

1.2.2 Revue de littérature (non exhaustive) sur la détection d'anomalie.

Nous allons voir ici plusieurs grands principes en détection d'anomalies sans forcément en présenter tous les algorithmes.

Au même titre que d'autres tâches en apprentissage automatique, ces algorithmes peuvent être séparés en deux catégories : l'apprentissage supervisé ou non-supervisé. De même, si l'on considère leurs sorties on retrouve les scores ou les labels (on pourra faire un parallèle avec régression/classification). Voyons maintenant une liste des grands mouvements présents en détection d'anomalie.⁴

Les algorithmes de classification. Le principe est très simple et très intuitif : ces algorithmes doivent classifier les données en 0 (normal) ou 1 (anomalie). Pour ce faire, [91] apprends des frontières autour des données classées comme normales, puis classe toute données en dehors de ces frontières comme anomalie.

Les algorithmes de clustering : Nous regroupons ici les techniques de clustering usuelles et les k plus proches voisins. Ces algorithmes utilisent pour la plupart la notion de distance pour détecter les anomalies. On retiendra dans cet exemple l'Isolation Forest [71] et son extension [56]. Le concept est là aussi intuitif : lorsque l'on sépare itérativement et aléatoirement l'espace (par une forêt aléatoire), une anomalie se retrouve très vite isolée. On attribue un score plus ou moins élevé selon la rapidité d'isolation de chaque donnée.

Les algorithmes statistiques : Ce sont des algorithmes pour lesquels toute observation de faible probabilité est une anomalie. L'un des tests les plus basiques s'applique à des données sous hypothèses Gaussienne. Considérant, pour $X \sim \mathcal{N}(\mu, \sigma^2)$, que

$$\mathcal{P}[\mu - 3\sigma < X < \mu + 3\sigma] \simeq 0.997$$

alors toute donnée en dehors de cet intervalle $[\mu \pm 3\sigma]$ est une anomalie.

Les algorithmes basés sur la théorie de l'information : Le principe de cette méthode est qu'une anomalie change significativement l'information de tout un jeu de données (l'entropie par exemple) [67].

Les algorithmes basés sur une technique de projection/réduction de dimension : L'hypothèse ici est qu'une donnée anormale se détachera encore plus des autres si on plonge les données dans un espace bien choisi. Un exemple se retrouve dans [37] où une anomalie est une donnée qui se détache trop de la structure "principale" de corrélation.

Les algorithmes de reconstructions : Ces méthodes sont très certainement les plus récentes, car c'est avec les méthodes de génération aléatoire guidée que les résultats

4. Cette "classification" s'inspire majoritairement de [15]

sont les plus épataints. Si l'on prend AnoDDPM [103], l'algorithme apprend à bruitter/débruiter une image avec un algorithme de Stable Diffusion. L'entraînement est fait sur des données totalement normales, dans le but de "déconstruire" l'anomalie pour reconstruire une image normale. La différence entre l'image d'origine et sa reconstruction donnera une carte thermique indiquant l'anomalie. Un travail conjoint avec Julien Bastian, Stéphane Chrétien et Ben Gao a été présenté à ce sujet aux Journées de la Statistique 2024 et CAP 2024.

Les méthodes que nous allons présenter se situent :

- ⇒ Pour le chapitre 4 section 4.1 dans la catégorie des méthodes statistiques pour la première partie et dans la catégorie des méthodes de plongement/réduction de dimension pour la seconde partie.
- ⇒ Pour le chapitre 4 section 4.2 et 4.3 dans la catégorie des algorithmes basés sur la théorie de l'information.

1.3 La signature d'un chemin à variations bornées.

C'est dans les années 60, par le biais de 3 articles [19–21] que Kuo-Tsai Chen définit l'objet signature pour un chemin lisse. Cet objet, issu de la géométrie théorique, fut ramené au devant de la scène dans les années 90 par Terry Lyons pour l'analyse des chemins rugueux [43, 46, 47, 75–77]. Utilisée comme une représentation⁵ d'une trajectoire temporelle de tout type, la signature trouve parfaitement sa place dans le contexte de Machine Learning. Par ailleurs, l'intégration des signatures dans l'apprentissage automatique a permis très régulièrement d'atteindre l'état de l'art. Au lieu d'une longue liste de références incomplète, le lecteur intéressé pourra se reporter au compendium établi dans [74].



Remarque(s)

- Dans cette partie introductory, nous nous contenterons des définitions et propriétés indispensables à la compréhension des signatures. Certaines propriétés plus avancées seront amenées plus tard. De même, pour les démonstrations, nous renvoyons le lecteur à [47, 77]. Les notions de groupe et d'algèbre de Lie sont abordées en annexe A avec leur application aux signatures.
- Pour les calculs de signature, nous utiliserons majoritairement les packages python Signatory [60] et iisignature [89]

1.3.1 Chemins à variations bornées

Avant de définir les signatures, il est important de comprendre sur quels objets ces dernières s'appliquent.

5. Non, ce terme n'a pas été choisi au débotté.



Definition 1.3.1

Soit $d \in \mathbb{N}^*$. Un chemin d -dimensionnel est une application

$$X : I \subset \mathbb{R} \rightarrow \mathbb{R}^d$$

$$t \rightarrow X_t = (X_t^1, \dots, X_t^d)$$

avec $X^i : I \subset \mathbb{R} \rightarrow \mathbb{R}$ pour tout $1 \leq i \leq d$ et la convention d'écriture $X_t = X(t)$.

De manière générale, dans les chapitres qui suivront, nous ferons l'hypothèse que $I = [0; T]$ et, après pré-traitement $I = [0; 1]$. Très ponctuellement, il nous arrivera de considérer $I = \mathbb{R}$. Conformément à la littérature, nous nous intéresserons à $d > 1$, le cas $d = 1$ étant particulier, voir [41] Exemple 1.4.

En géométrie, et notamment dans le travail de KT Chen, la notion de chemin implique le plus régulièrement une notion de lissité, voire lissité presque-partout dans les situations les plus extrêmes. Or dans notre cas, nous allons être amenés à étudier des chemins stochastiques, donc potentiellement lisses nulle part (au sens usuel du terme). Nous allons donc se fixer une hypothèse bien moins forte :



Definition 1.3.2 Chemin à variations bornées

On considère $X : I \subset \mathbb{R} \rightarrow \mathbb{R}^d$ continu. On définit, pour $I = [0, T]$ et $[s, t] \subset I$:

$$D(I) = \{(t_i) = (t_0, t_1, \dots, t_n) \in I^n \mid s = t_0 < t_1 < \dots < t_n = t\}$$

La **1-variation** de X sur $[s, t]$ est définie par

$$\|x\|_{1-var, [s, t]} = \sup_{(t_i) \in D([s, t])} \sum_{i=0}^n \|X_{t_{i+1}} - X_{t_i}\|$$

Pour être plus précis, nous allons tout simplement demander que X soit à **variations bornées** :



Definition 1.3.3

On se place dans le contexte des hypothèses précédentes. On dit que X est à **variations bornées** si

$$\|X\|_{1-var} := \|X\|_{1-var, I} < \infty$$



Remarque(s)

- Nous ne définissons ici que ce qui est nécessaire. La 1-variation est un cas particulier de la p -variation, très liée aux contrôles. Nous renvoyons encore au Friz-Victoir pour plus de théorie autour du contrôle.
- $\|\cdot\|_{1-var}$ n'est pas une norme (de manière assez évidente, 0 à pour antécédent toute fonction constante).
- Toute fois, l'espace vectoriel $BV(I, \mathbb{R}^d) = \{X \in C^0(I, \mathbb{R}^d) \mid \|X\|_{1-var} < \infty\}$ (parfois noté $C^{1-var}(I, \mathbb{R}^d)$ dépendamment des auteurs) peut être muni de la norme

$$\|X\|_{BV} := \|X\|_{1-var} + \sup_{t \in I} \|X_t\|$$

Par ailleurs, cette norme permet de munir l'espace $BV(I, \mathbb{R}^d)$ d'une structure de Banach.



Remarque(s)

- On pourrait interpréter cette norme (en réalité semi-norme) d'une manière géométrique. En effet, si l'on considère $X \in C^2(I, \mathbb{R}^d)$ une variété unidimensionnelle lisse, alors, en posant $h_i = t_{i+1} - t_i > 0$

$$\begin{aligned}\|X\|_{1-var} &= \sup_{(t_i) \in D(I)} \sum_{i=0}^n \|X_{t_{i+1}} - X_{t_i}\| \\ &= \sup_{(t_i) \in D(I)} \sum_{i=0}^n h_i \left\| \frac{X_{t_{i+1}} - X_{t_i}}{h_i} \right\| \\ &= \int_I \|\dot{X}_t\| dt\end{aligned}$$

on retrouve que la 1-variation correspond à la longueur du chemin X .

1.3.2 Signature d'un chemin à variations bornées

Maintenant que nous avons défini les objets que nous allons étudier par le truchement des signatures, il est temps de définir ces dernières. Il existe plusieurs manières de construire/voir les signatures⁶ :

⇒ **Comme un catalogue organisé de quantités décrivant une forme** : Sans être la plus élégante, ni même la plus complète, cette manière de voir les signatures a le mérite d'être accessible au plus grand nombre. C'est donc pour cela que j'ai choisi cette voie.

6. Il arrivera (notamment en annexe) que nous passions d'un point de vue à l'autre pour faciliter certaine preuve (cf. [14] remarque 1.3.8)

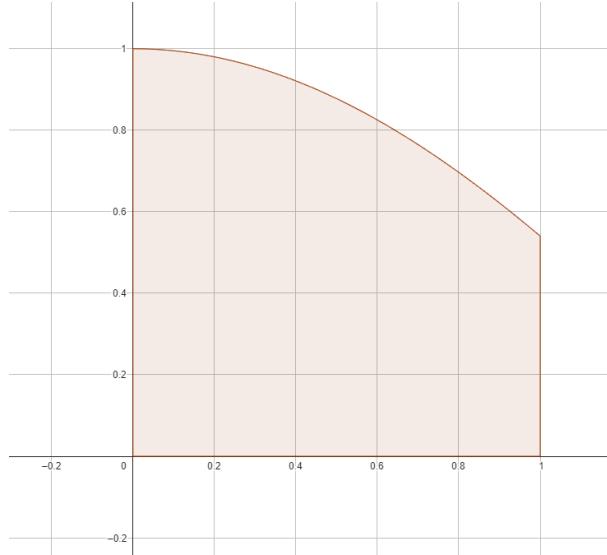


FIGURE 1.2 – L'aire dessinée ici correspond à $\int_0^1 \cos(t)dt$

- ⇒ **Avec des formes différentielles, des complexes de De Rham, des algèbres extérieures, et j'en passe :** C'est à mon sens la plus distinguée. La plus difficile aussi peut être. Il faut un niveau très conséquent en géométrie algébrique. Il est à noter que c'est en passant par cette batterie d'outils que K.T. Chen définit pour la première fois les signatures. Nous renvoyons toutefois le lecteur avide de géométrie de haute voltige à [50, 51].
- ⇒ **Avec des équations différentielles contrôlées :** Je considère cette définition comme la plus naturelle, à défaut d'être l'originelle. Toutefois, ce point de vue ne nous sera pas utile au regard de l'utilisation faite des signatures dans ce manuscrit.

Démarrons par un des outils les plus connus de l'analyse : l'intégrale de Riemann. Lorsque l'on calcule une intégrale de Riemann, on calcule

$$I = \int_a^b f(t)dt$$

Il est bien connu que cette expression désigne l'aire sous la courbe de f entre a et b .

Sur la figure 1.2 on voit l'aire sous la courbe définie par $f(t) = \cos(t)$ entre 0 et 1. Comme on l'apprend en seconde, la courbe du plan définie par une fonction f est l'ensemble des points $(t, f(t))$. Ce qui est intéressant, c'est que sans s'en rendre compte nous venons de créer notre premier chemin à variations bornées de dimension 2. Et nous avons aussi récupéré une information géométrique quantitative sur cette courbe.

Pour autant, avons nous bien tout saisi ce que nous avons fait ? Le signe intégrale donne la notion d'aire, la fonction f est bien visible sur notre graphique... mais à quoi donc peut correspondre ce dt ?

Et bien pour comprendre cela, revenons à la définition même de l'intégrale de Riemann : c'est la limite de l'aire sous la courbe d'une fonction en escalier.

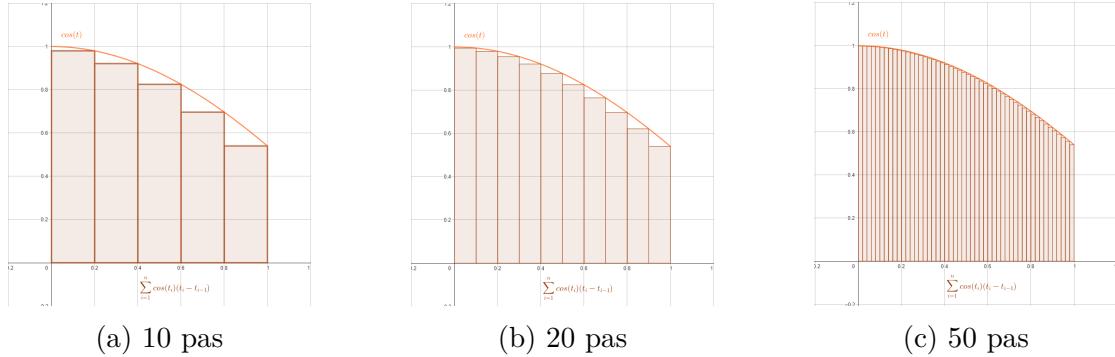


FIGURE 1.3 – Trois étapes différentes de la convergence vers $\int_0^1 \cos(t)dt$

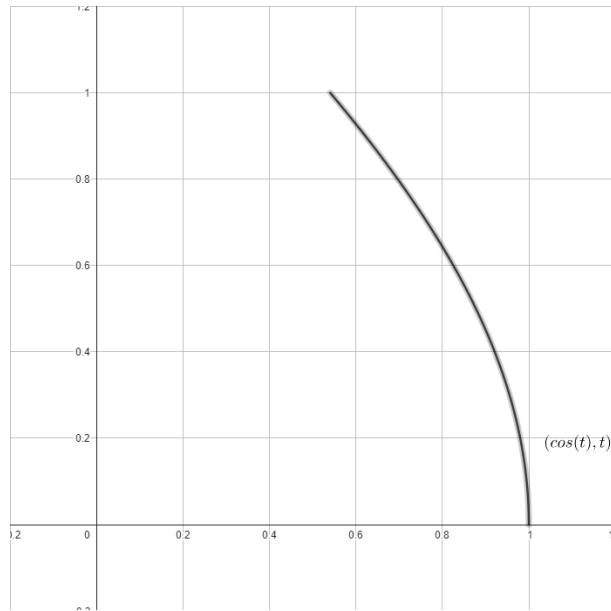


FIGURE 1.4 – Chemin $(\cos(t), t)$ pour $0 < t < 1$

Comme nous le savons et/ou comme nous pouvons le voir, à chaque étape, on somme des aires de rectangles dont la hauteur est $H_i = \cos(t_i)$ et dont la largeur est donnée par un intervalle plus ou moins grand avant t_i : $L_i = t_i - t_{i-1}$. Comme ce qui nous intéresse est le résultat asymptotique, nous pouvons assumer que tous les L_i sont égaux à une valeur que nous appellerons Δt

Le passage à la limite se fait justement sur Δt et c'est cette limite que nous appellerons dt : Nous regardons l'aire sous la courbe pour des petits pas de temps.

Oui, mais qu'arriverait-il si j'avais inversé les deux dimensions ? Autrement dit : si je considère le chemin bidimensionnel $(\cos(t), t)$ que se passe-t-il ? Déjà, regardons ce que cela donne graphiquement :

Nous pouvons déjà voir sur la Figure 1.4, que l'on a plus la même chose. On peut aussi remarquer que la courbe a subi des transformations simples : symétrie axiale et rotation d'angle $\frac{\pi}{2}$. De fait, l'aire I calculée plus tôt n'est plus l'aire délimitée par la courbe et les

droites d'équations $y = 0$, $x = 0$ et $x = 1$, mais celle délimitée par la courbe et les droites $x = 0$, $y = 0$ et $y = 1$.

Qu'en est il alors de l'aire sous la courbe dans le sens où nous avons l'habitude de le voir ?

Et bien, si l'on suit la construction précédente de l'intégrale de Riemann, il suffit de prendre la limite de la fonction en escalier suivante :

$$\sum_{i=1}^n t_i (\cos(t_i) - \cos(t_{i-1})) = \sum_{i=1}^n t_i \delta \cos(t)$$

Là aussi, on peut se dire que l'on va prendre $\delta \cos(t) = \cos(t_i) - \cos(t_{i-1})$ constant (ce qui implique que $t_i - t_{i-1}$ ne l'est plus). Et bien nous avons une deuxième information géométrique quantitative que l'on définit en passant à la limite par :

$$I' = \int_0^1 t d\cos(t)$$

Si l'on passe à des notations cohérentes avec notre définition d'un chemin ($X_t^1 = t$ et $X_t^2 = \cos(t)$), alors nous avons créé deux quantités géométriques pour décrire notre chemin :

$$I = \int_0^1 X_t^2 dX_t^1 \quad \text{et} \quad I' = \int_0^1 X_t^1 dX_t^2$$

Ces quantités ainsi créées sont ce que l'on appelle des **intégrales le long d'un chemin** :



Definition 1.3.4 Intégrale le long d'un chemin uni-dimensionnel

On considère un chemin $X : [a; b] \rightarrow \mathbb{R}$ et un chemin $Y : [a; b] \rightarrow \mathbb{R}$. Alors l'intégrale du chemin Y le long du chemin X est :

$$\int_a^b Y_t dX_t \tag{1.1}$$

Ce type d'intégrale généralise la version de Riemann et est appelée **intégrale de Stieltjes**

Proposition 1.3.1

- Si Y et X sont des chemins au moins C^1 , alors :

$$\int_a^b Y_t dX_t = \int_a^b Y_t \dot{X}_t dt$$

- Sinon, on définit une subdivision $t_0 = a < t_1 < \dots < t_{n-1} < t_n = b$, une suite $(s_i)_{i \in \{1, \dots, n\}}$ telle que $s_i \in [t_{i-1}; t_i]$ et

$$\int_a^b Y_t dX_t = \lim_{\sup\{|t_i - t_{i-1}|, i=1, \dots, n\} \rightarrow 0} \sum_{i=1}^n Y_{s_i} (X_{t_i} - X_{t_{i-1}})$$

Remarque(s)

- D'aucuns reconnaîtront dans le cas C^1 un morceau de la dérivée d'un produit :

$$(fg)' = f dg + g df$$

C'est tout à fait normal : si **au moins l'une des deux intégrales de Stieltjes existe**, alors la règle de Leibniz implique automatiquement l'existence de l'autre.

- On pourrait être poussé à croire que les deux intégrales sont égales. C'est généralement faux.

Continuons notre processus. Toujours avec la même ligne directrice, nous pouvons, à partir du chemin X , créer les deux intégrales de Stieltjes suivantes :

$$\int_a^b X_t^1 dX_t^1 \quad \text{et} \quad \int_a^b X_t^2 dX_t^2$$

De plus, il est possible de les organiser en une matrice :

$$M = \begin{pmatrix} \int_a^b X_t^1 dX_t^1 & \int_a^b X_t^1 dX_t^2 \\ \int_a^b X_t^2 dX_t^1 & \int_a^b X_t^2 dX_t^2 \end{pmatrix}$$

Cette simple matrice résume les mesures d'aires possibles à l'aide du chemin (X^1, X^2) .

Retournons à un degré inférieur avec le vecteur suivant :

$$V_1(t) = \begin{pmatrix} \int_a^t dX_s^1 \\ \int_a^t dX_s^2 \end{pmatrix} = \begin{pmatrix} X_t^1 - X_a^1 \\ X_t^2 - X_a^2 \end{pmatrix}$$

Quitte à considérer $\tilde{X}_t = X_t - X_a$ (nous aborderons ce point plus tard), on peut considérer $X_a = 0$. V devient donc :

$$V^1(t) = \begin{pmatrix} X_t^1 \\ X_t^2 \end{pmatrix}$$

Nous pouvons alors remarquer que :

$$\forall 1 \leq i, j \leq 2 \quad M_{i,j} = \int_a^b V_i^1(s) dX_s^j$$

Notons cette matrice V^2 , on peut définir :

$$\forall 1 \leq i, j \leq 2 \quad (V^2)_{i,j}(t) = \int_a^b V_i^1(s) dX_s^j$$

Si l'on récapitule :

- V^1 donne les écarts absolus entre X_a et X_b . C'est donc logique d'obtenir un objet de même dimension que X .
- V^2 donne des mesures d'aires définies par X . Une aire étant un produit de deux longueurs, il est cohérent que V^2 vive dans un espace de matrices de taille $d \times d$.

De manière logique, nous sommes amenés à étendre ces objets pour obtenir des volumes de plus haute dimension. Pour cela, on a besoin des définitions suivantes :

Definition 1.3.5

On pose $(\mathbb{R}^d)^{\otimes k} = \underbrace{\mathbb{R}^{d \times d \times \dots \times d}}_{k \text{ fois}}$ l'espace vectoriel des tenseurs à k dimensions sur \mathbb{R}^d . Soit $Y : \mathbb{R} \mapsto (\mathbb{R}^d)^{\otimes k}$ et $X : \mathbb{R} \mapsto \mathbb{R}^d$. Alors **l'intégrale du chemin tensoriel Y contre le chemin X** (notée ici $T(t)$) vérifie

$$T(t) = \int_a^t Y_s dX_s \in (\mathbb{R}^d)^{\otimes k+1}$$

et, pour $(i_1, \dots, i_{k+1}) \in \llbracket 1, d \rrbracket^{k+1}$

$$(T(t))_{i_1, \dots, i_{k+1}} = \int_a^t Y_{i_1, \dots, i_k}(s) dX_s^{k+1}$$

Muni de cet outil, on peut tout naturellement créer une suite de tenseurs $S_I^{(k)}(X)$ telle que :

\Rightarrow Pour $I = [a; b]$, $X : I \subset \mathbb{R} \rightarrow \mathbb{R}^d$ est un chemin à variation bornée.

$\Rightarrow S_I^{(0)}(X) = 1$

\Rightarrow Pour tout $k \geq 1$ on définit $S_I^{(k+1)}(X) = \int_I S_{[a,t]}^{(k)}(X) dX_t$



Remarque(s)

- Chaque tenseur $S^{(k)}$ représente une collection de volumes de dimension k créés à partir du chemin X . Ces tenseurs sont appelés **signatures d'ordre k** (ou de niveau k).
- **Attention** : dans beaucoup d'articles de la littérature, les auteurs alternent allègrement entre les espaces d'appartenance du chemin X selon les besoins de la théorie. Il est évident que l'intégrale de Riemann-Stieltjes **n'est pas correctement définie sur tous les espaces fonctionnels**. Pour autant, et heureusement, plusieurs généralisations ^a permettent de calculer ces quantités selon l'espace fonctionnel considéré.

a. Nous introduisons les intégrales d'Ito et de Stratonovich dans une prochaine annexe



Definition 1.3.6 Signature

- Soit un chemin $X : I \subset \mathbb{R} \rightarrow \mathbb{R}^d$ à variation bornée. Sa **signature** sur $J \subset I$ est la suite de tenseurs $S_J(X) = (S_J^{(k)}(X))_{k \geq 0}$
- On appelle **signature tronquée à l'ordre N** la suite finie

$$S_J^N(X) = (S_J^{(k)}(X))_{0 \leq k \leq N}$$

Cette définition, agréable dans sa faible complexité, n'en reste pas moins incommodante lorsqu'il s'agira d'effectuer des opérations. Pour pallier ces difficultés, nous allons apporter un peu plus de structures à ces objets, ainsi qu'à leur ensemble dans la section suivante. Avant cela, finissons sur un exemple fondamental :

Exemple(s)

Signature d'un chemin linéaire : On considère $X : [0; T] \rightarrow \mathbb{R}^d$ tel que $X_t = X_0 + (X_T - X_0)t$. Alors, pour tout multi-indice $K = (i_1, \dots, i_n) \in \{1, \dots, d\}^d$:

$$\begin{aligned} S_{[0;T]}^I(X) &= (X_T^{i_1} - X_0^{i_1}) \dots (X_T^{i_k} - X_0^{i_k}) \\ &= \frac{1}{n!} \prod_{j=1}^k (X_T^{i_j} - X_0^{i_j}) \end{aligned}$$

Il est intéressant de voir que, sur ce simple cas, les différents degrés k de la signature donnent exactement des volumes k -dimensionnels, et, pour être plus exact, des volumes de simplexes k -dimensionnels dont les k côtés s'intersectant en $O_{\mathbb{R}^d}$ sont de longueurs $X_T^{i_j} - X_0^{i_j}$.

1.3.3 L'espace $\mathcal{G}(\mathbb{R}^d)$

Définition 1.3.7

On appelle **algèbre tensorielle de \mathbb{R}^d** et on note $T(\mathbb{R}^d)$ l'algèbre associative

$$T(\mathbb{R}^d) = \bigoplus_{k=0}^{+\infty} (\mathbb{R}^d)^{\otimes k}$$

Pour (e_1, \dots, e_d) une base de \mathbb{R}^d , un élément $x \in T(\mathbb{R}^d)$ peut être vu comme une série formelle

$$\sum_{k=0}^{+\infty} \sum_{(i_1, \dots, i_k) \in \llbracket 1, d \rrbracket^k} \lambda_{i_1, \dots, i_k} e_{i_1} \otimes \dots \otimes e_{i_k}$$

où $\{e_{i_1} \otimes \dots \otimes e_{i_k} \mid k \geq 1\}$ est une base de $T(\mathbb{R}^d)$

Remarque(s)

- ⇒ On peut définir de manière équivalente $T^N(\mathbb{R}^d) = \bigoplus_{k=0}^N (\mathbb{R}^d)^{\otimes k}$ qui est aussi une algèbre associative (non-commutative).
- ⇒ Pour les personnes ayant une appétence pour les séries formelles non commutatives (les polynômes à d variables non commutatives), il est possible de regarder $T(\mathbb{R}^d)$ par leur prisme en identifiant e_i à X_i . On a alors $\lambda e_{i_1} \otimes \dots \otimes e_{i_n} \simeq \lambda X_{i_1} \dots X_{i_n}$

En tant qu'algèbre, l'espace $T(\mathbb{R}^d)$ est stable par somme, par multiplication scalaire et par le produit tensoriel défini ci dessous :

Definition 1.3.8

On considère $x, y \in T(\mathbb{R}^d)$. Alors $x \otimes y$ est défini par :

$$\begin{aligned} x \otimes y &= \left(\sum_{k=0}^{\infty} \sum_{(i_1, i_2, \dots, i_k) \in \llbracket 1, d \rrbracket} \lambda_{i_1, \dots, i_k} e_{i_1} \otimes \cdots \otimes e_{i_k} \right) \otimes \left(\sum_{k=0}^{\infty} \sum_{(i_1, i_2, \dots, i_k) \in \llbracket 1, d \rrbracket} \mu_{i_1, \dots, i_k} e_{i_1} \otimes \cdots \otimes e_{i_k} \right) \\ &= \lambda_0 \mu_0 + \sum_{i=1}^d (\lambda_0 \mu_i + \lambda_i \mu_0) e_i + \sum_{i_1, i_2=1}^d (\lambda_0 \mu_{i_1, i_2} + \lambda_{i_1} \mu_{i_2} + \lambda_{i_1, i_2} \mu_0) e_{i_1} \otimes e_{i_2} + \dots \end{aligned}$$

Une manière plus conforme à la définition originelle du produit tensoriel serait la suivante : Pour tout $m, n \in \mathbb{N}$ (éventuellement $\in \llbracket 1, N \rrbracket$ dans le cas de $T^N(\mathbb{R}^d)$),

- L'application $\otimes : (\mathbb{R}^d)^{\otimes m} \times (\mathbb{R}^d)^{\otimes n} \rightarrow (\mathbb{R}^d)^{\otimes(m+n)}$ est bilinéaire non-commutative. (Pour s'économiser des lourdeurs d'écriture, on utilise la même notation pour tous les couples (m, n) ainsi que pour \otimes étendu à $T(\mathbb{R}^d)$).
- Pour tout $E_{i_1, 0, \dots, i_m, 0} = (\mathbf{1}_{i_1, 0, \dots, i_m, 0}(i_1, \dots, i_m))_{i_1, \dots, i_m}$ (resp. $F_{j_1, 0, \dots, j_n, 0} = (\mathbf{1}_{j_1, 0, \dots, j_n, 0}(j_1, \dots, j_n))_{j_1, \dots, j_n}$) vecteurs de la base canonique de $(\mathbb{R}^d)^{\otimes m}$ (resp. $(\mathbb{R}^d)^{\otimes n}$), on a :

$$E_{i_1, 0, \dots, i_m, 0} \otimes F_{j_1, 0, \dots, j_n, 0} = (\mathbf{1}_{i_1, 0, \dots, i_m, 0, j_1, 0, \dots, j_n, 0}(i_1, \dots, i_m, j_1, \dots, j_n))_{i_1, \dots, i_m, j_1, \dots, j_n}$$

- Pour tout $e_{i_1} \otimes \cdots \otimes e_{i_m}$ (resp. $f_{j_1} \otimes \dots \otimes f_{j_n}$) vecteur de la base canonique de $(\mathbb{R}^d)^{\otimes m}$ (resp. $(\mathbb{R}^d)^{\otimes n}$) :

$$(e_{i_1} \otimes \cdots \otimes e_{i_m}) \otimes (f_{j_1} \otimes \dots \otimes f_{j_n}) = e_{i_1} \otimes \cdots \otimes e_{i_m} \otimes f_{j_1} \otimes \dots \otimes f_{j_n}$$

Muni de cet autre point de vue, on peut redéfinir le produit tensoriel $x \otimes y$ pour $x = (x_0, x_1, \dots, x_n, \dots)$ et $y = (y_0, y_1, \dots, y_n, \dots)$ et pour tout $i \in \mathbb{N}$ $x_i, y_i \in (\mathbb{R}^d)^{\otimes i}$:

$$x \otimes y = \sum_{k=0}^{\infty} \left(\sum_{j=0}^k x_j \otimes y_{k-j} \right)$$

Retour aux signatures : La signature $S(X)$ d'un chemin peut être identifiée par un élément de $T(\mathbb{R}^d)$ (et donc par une série formelle) grâce à l'identification

$$\lambda_{i_1, \dots, i_k} = S^{i_1 \dots i_k}(X)$$

ce qui implique la possibilité d'écrire $S(X) = \sum_{k=0}^{\infty} \sum_{i_1, \dots, i_k \in \llbracket 1, d \rrbracket} S^{i_1, \dots, i_k}(X) e_{i_1} \otimes \cdots \otimes e_{i_k}$. On note $\mathcal{G}(\mathbb{R}^d)$ (resp. $\mathcal{G}^N(\mathbb{R}^d)$) l'image des chemins à variations bornées dans l'espace $T(\mathbb{R}^d)$ (resp. $T^N(\mathbb{R}^d)$) par l'application signature. Un élément de $T(\mathbb{R}^d)$ est un élément de $\mathcal{G}(\mathbb{R}^d)$ si et seulement si il vérifie la propriété induite par le théorème du Shuffle product :

**Definition 1.3.9** *Shuffle product*

On définit un (n, m) -mélange comme une permutation de $\llbracket 1, d \rrbracket$ vérifiant :

$$\sigma^{-1}(1) < \dots < \sigma^{-1}(n) \quad \text{et} \quad \sigma^{-1}(n+1) < \dots < \sigma^{-1}(n+m)$$

On considère maintenant deux multi-indices $I = (i_1, \dots, i_n)$ et $J = (j_1, \dots, j_m)$ de $\llbracket 1, d \rrbracket$. Alors, le **Shuffle product** de I et J est un ensemble fini de multi-indices défini par :

$$I \sqcup\!\sqcup J = \left\{ (r_{\sigma(1)}, \dots, r_{\sigma(n+m)}) \mid \sigma \in \text{Mélange}(n, m) \right\}$$

avec $(r_1, \dots, r_n, r_{n+1}, \dots, r_{n+m}) = (i_1, \dots, i_n, j_1, \dots, j_m)$

**Theorem 1.1** *Identité du Shuffle product*

On considère $X \in BV(\mathbb{R}^d)$ et deux multi-indices I et J de $\llbracket 1, d \rrbracket$. Alors :

$$S^I(X) \cdot S^J(X) = \sum_{K \in I \sqcup\!\sqcup J} S^K(X)$$

Ce principe du shuffle product est assez difficile à visualiser de manière générale, mais il exprime tout simplement des notions de géométrie élémentaire que l'on constate sur de simples exemples :

**Exemple(s)**

On considère $X = (X^1, \dots, X^d)$ un chemin d -dimensionnel sur $I = [0; 1]$. Rappelons que, pour tout $1 \leq i \leq d$: $S_I^i(X) = X_1^i - X_0^i$. Pour simplifier l'écriture, on notera $s_{i_1 i_2 \dots i_k} = S_I^{i_1, \dots, i_k}(X) /$

- Premier niveau du shuffle product : pour tout $1 \leq i, j \leq d$, $s_i s_j = s_{ij} + s_{ji}$. Considérant que s_{ij} et s_{ji} sont des aires, il est plutôt logique de les exprimer comme un produit de deux longueurs, voir Fig 1.5a
- Deuxième niveau du shuffle product : pour tout $1 \leq i, j, k \leq d$, $s_{ij} s_k = s_{ijk} + s_{ikj} + s_{kij}$. De nouveau, on obtient qu'un volume est le produit d'une longueur, voir Fig 1.5b

1.3.4 Calcul effectif des signatures

Comment calculer une suite infinie d'intégrales sur des chemins tels que les séries temporelles ? Il est évident qu'il nous faut une technique autre que l'intégrale : approximer numériquement une intégrale est déjà complexe, et ici nous devons faire la même chose pour une intégrale de Stieltjes. Heureusement, un théorème va rendre le calcul des signatures (numériquement) très simple. Pour commencer, définissons ce qu'est la concaténation de deux chemins :

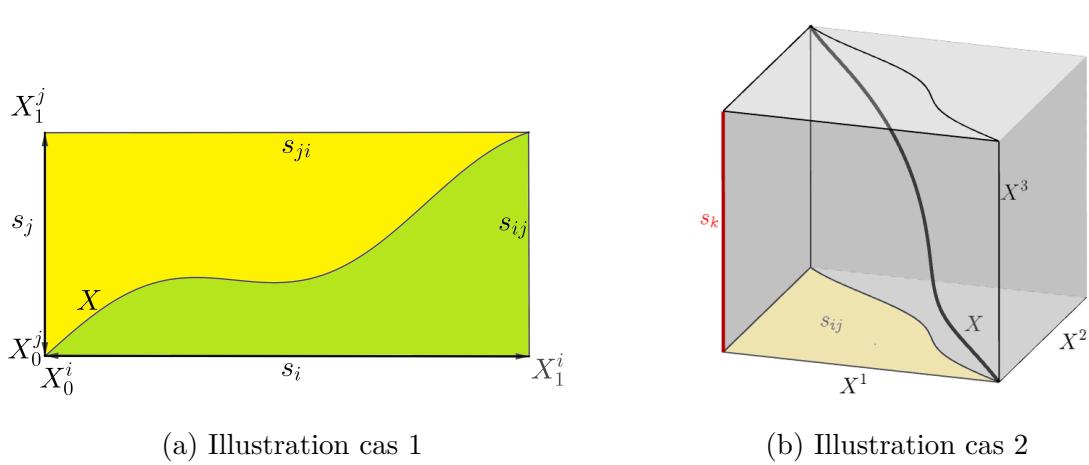


FIGURE 1.5

Definition 1.3.10

Soit $X : [a, b] \rightarrow \mathbb{R}^d$ et $Y : [b, c] \rightarrow \mathbb{R}^d$. Leur **concaténation** est le chemin $X \star Y : [a; c] \rightarrow \mathbb{R}^d$ tel que :

$$(X \star Y)_t = \begin{cases} X_t & \text{si } t \in [a, b] \\ Y_t - Y_b + X_b & \text{si } t \in [b, c] \end{cases}$$

On a alors :

Theorem 1.2 Identité de Chen

Soit $X : [a, b] \rightarrow \mathbb{R}^d$ et $Y : [b, c] \rightarrow \mathbb{R}^d$. Alors :

$$S_{[a;c]}(X \star Y) = S_{[a;b]}(X) \otimes S_{[b;c]}(Y)$$

C'est grâce à cette identité que nous allons pouvoir calculer les signatures. On considère ici une série temporelle multivariée $X = (X^1, \dots, X^d)$ échantillonnée sur (t_0, \dots, t_n) . Nous pouvons considérer une interpolation linéaire de X :

$$\forall 1 \leq i \leq n, \forall t \in [t_{i-1}, t_i] \quad X_t = \phi_i(t) = \frac{t - t_{i-1}}{t_i - t_{i-1}} X_{t_i} + \left(1 - \frac{t - t_{i-1}}{t_i - t_{i-1}}\right) X_{t_{i-1}}$$

On a donc que $X = \phi_0 \star \dots \star \phi_{n-1}$. Grâce à l'identité de Chen :

$$S_{[t_0, t_n]}(X) = \bigotimes_{i=1}^n S_{[t_{i-1}, t_i]}(\phi_i)$$

Or, comme vu précédemment, la signature de chaque ϕ_i possède une formule explicite. Le calcul de $S(X)$ revient donc à un produit tensoriel de plusieurs signatures "pré-calculées".

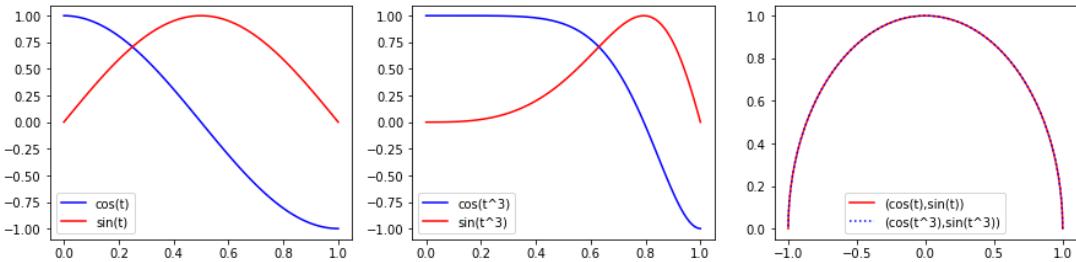


FIGURE 1.6 – Illustration de l'invariance par reparamétrisation

1.3.5 Injectivité de l'application signature.

Pour s'assurer de l'injectivité de la signature, nous allons étudier toutes ses propriétés d'invariances. Les propriétés qui suivent vont drastiquement influencer l'espace de départ de l'application signature. Pour l'instant, nous considérons $BV(I, \mathbb{R}^d) = \{X : I \subset \mathbb{R}^d, \|X\|_{1-\text{var}} < +\infty\}$ comme notre espace de départ.
:

Invariance par translation : La première invariance est immédiate : si l'on considère une constante $a \in \mathbb{R}$ et $\tilde{X}_t = X_t + a$, alors $d\tilde{X}_t = dX_t$, et donc la signature est invariante par translation. De fait, nous serons amenés, dans un cadre statistique, à étudier $\tilde{X} = X_t - \mathbb{E}_{t \sim \mathcal{U}(I)}[X_t]$. Ceci implique que la signature d'un chemin **centré en sa moyenne sur tous les temps est la même que le chemin originel**. On établit la relation d'équivalence : $X \sim_\lambda Y$ si et seulement si il existe $\lambda \in \mathbb{R}^d$ tel que $Y = X + \lambda$. On en déduit une restriction de la signature à $BV(I, \mathbb{R}^d) / \sim_\lambda$ (qui pourrait s'apparenter à un projectif fonctionnel).

Invariance par reparamétrisation : La seconde invariance qui nous intéresse est plus subtile : la signature extrayant des quantifications géométriques (longueurs, aires, volumes...) on s'attend à ce qu'elle ne soit pas affectée par la **vitesse de parcours du chemin**. Et en effet, on a le résultat suivant :

 **Theorem 1.3** *Invariance par reparamétrisation*

On considère un chemin $X \in BV(I, \mathbb{R}^d)$ et une fonction $\phi : I \rightarrow I$ continûment dérivable, croissante et surjective. On considère aussi $\tilde{X}_t = X_{\phi(t)}$ pour tout $t \in I$. (Naturellement $\tilde{X} \in BV(I, \mathbb{R}^d)$). Alors :

$$S_I(X) = S_I(\tilde{X})$$

Dans la figure 1.6, on voit sur les deux premiers graphiques la différence de vitesse de parcours, mais la stricte superposition des chemins dans le 3ème graphique.

En considérant, pour toute reparamétrisation ϕ de I , les applications δ_ϕ telles que

$$\delta_\phi(X) = X_{\phi(t)}$$

on peut définir la relation d'équivalence $X \sim_\phi Y$ si et seulement s'il existe une reparamétrisation ϕ telle que $Y = \delta_\phi(X)$. Cela nous amène à restreindre la signature à $BV(I, \mathbb{R}^d) / \sim_\phi$.

Invariance par "Tree-like" équivalence. On commence par définir l'équivalence en question :

☀️ Definition 1.3.11 Version analytique

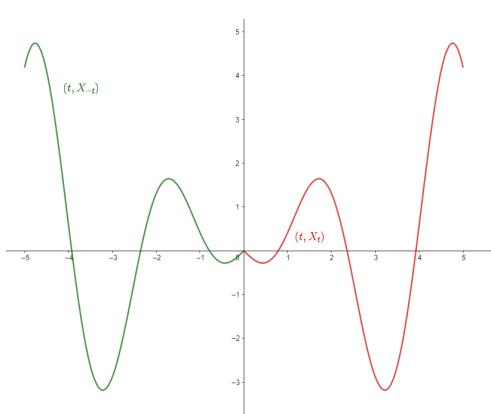
Un chemin $X : [0, T] \rightarrow \mathbb{R}^d$ est considéré "tree-like" ^a s'il existe une fonction continue $h : [0; T] \rightarrow \mathbb{R}_+$ telle que $h(0) = h(T) = 0$, et pour tout $s, t \in [0; T]$, avec $s \leq t$:

$$\|X_s - X_t\| \leq h(s) + h(t) - 2 \inf_{u \in [s, t]} h(u)$$

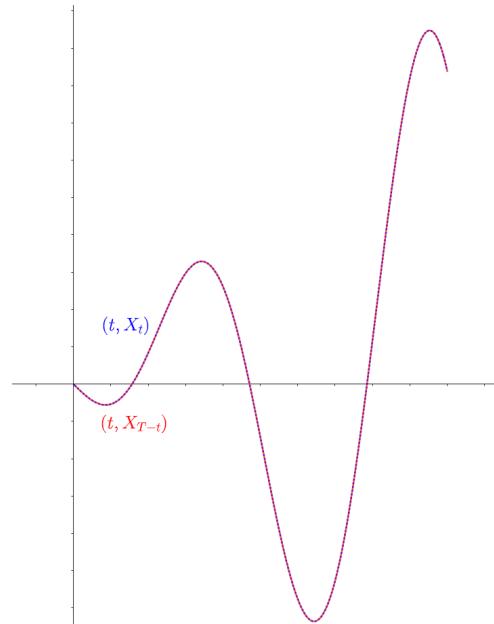
Cette fonction h est appelée **fonction de hauteur**.

a. Je n'ai aucune traduction correcte, et ChatGPT est pire.

Un aller-retour suivant la même courbe est un exemple plus explicite d'un chemin tree-like.



(a) La partie verte et rouge représente le même chemin parcouru en temps renversé.



(b) La concaténation de X_t et de X_{T-t} (sur $[0; T]$) est un chemin **tree-like**

FIGURE 1.7

Une autre manière plus formelle de définir ce genre de chemin est de dire qu'un chemin X est tree-like si et seulement si sa signature $S(X)$ est celle d'un chemin

constant :

$$S(X) = S(a) = (1, 0, 0, 0, \dots, 0, \dots) = \mathbf{1}$$

On en déduit le théorème

Theorem 1.4

Soit $X : [0; T] \rightarrow \mathbb{R}^d$ un chemin à variation bornée. Alors :

1. $S(X) = \mathbf{1}$ si et seulement si X est tree-like.
2. Soit $Y : [0; T] \rightarrow \mathbb{R}^d$ un autre chemin à variation bornée. On note $\overleftarrow{Y} = Y_{T-t}$. Alors $X \sim Y$ si et seulement si $X \star \overleftarrow{Y} = \mathbf{1}$
3. Entre autre, $S(X) \otimes S(\overleftarrow{X}) = \mathbf{1}$.

Au final, on obtient les trois résultats suivants :

Theorem 1.5 Unicité [54]

Pour tout $X, Y \in BV(\mathbb{R}^d)$, $S(X) = S(Y)$ si et seulement si $X \sim Y$.

Lemma 1.3.1

Soit $X \in BV(\mathbb{R}^d)$ un chemin avec **au moins une coordonnée strictement monotone**. Alors $S(X)$ détermine X de manière unique, à translation et reparamétrisation près.

Dans la suite, nous considérerons les chemins comme *augmentés* par le temps : $\tilde{X}_t = (t, X_t^1, X_t^2, \dots, X_t^d)$. De même, nous allons considérer $\tilde{X}_t = X_T - \mathbb{E}_{t \sim \mathcal{U}[0;T]}$. Sauf s'il y a un risque d'ambiguïté, nous ne noterons plus I , et nous considérerons $BV(\mathbb{R}^d)$ comme l'ensemble des chemins réduit par translation, reparamétrisation et par tree-like équivalence.

Le dernier résultat est plus algébrique et sera exploité dans les chapitre 2 et 5 :

Proposition 1.3.2

Soit $X : [0, T] \rightarrow \mathbb{R}^d$ un chemin à variations bornées. Alors :

1. $S(X) \otimes \mathbf{1} = S(X)$
2. $S(\overleftarrow{X})$ est un **inverse** de $S(X)$ pour le produit tensoriel.

En combinant ces deux propriétés avec l'identité de Chen, $\mathcal{G}(\mathbb{R}^d)$ est **un groupe** pour le produit tensoriel.

Pour finir, un théorème que nous explicitons dans l'annexe A :

Theorem 1.6

 Le groupe $(\mathcal{G}^N(\mathbb{R}^d), \otimes)$ est muni d'une structure de groupe de Lie.

1.3.6 Propriétés analytiques

Dans cette dernière et courte section, on donne la continuité des applications *signature* et *signature tronquée à l'ordre N*

$$S_I : BV(\mathbb{R}^d) \rightarrow T(\mathbb{R}^d) \quad \text{et} \quad \pi_n \circ S : BV(\mathbb{R}^d) \rightarrow T^N(\mathbb{R}^d)$$

pour $I = [0; T]$. On a d'abord le résultat suivant :

Proposition 1.3.3

 Soit $X \in BV(\mathbb{R}^d)$. Alors :

$$\|S^{(k)}(X)\|_F \leq \frac{1}{n!} \|x\|_{1-var}^n < \infty$$

 On en déduit rapidement

$$\|S(X)\|_F \leq \exp(\|x\|_{1-var}^n) < \infty$$

... ce qui amène naturellement

Corollary 1.3.1

 Les applications $S : BV(\mathbb{R}^d) \rightarrow T(\mathbb{R}^d)$ et $\pi_n \circ S : BV(\mathbb{R}^d) \rightarrow T^N(\mathbb{R}^d)$ sont continues.

1.4 Analyse topologique des données

Cette introduction à l'analyse topologique des données suit majoritairement [17, 18, 95]. Nous passerons toutefois sous silence nombre d'objets de topologie théorique⁷. Nous invitons les lecteurs avides de théorie à aller comprendre ce qu'est un module, une relation d'homotopie, un faisceau ou bien un fibré vectoriel⁸.

7. intéressants, importants à une connaissance profonde, mais accessoires à une compréhension correcte.

8. Note aux statisticiens : c'est aussi terrible qu'on le pense, mais d'une satisfaction intense à la compréhension.

L'incursion de la géométrie dans le domaine des statistiques n'est pas si récent : l'ACP n'est finalement qu'un précurseur algébrique de l'apprentissage de variété (manifold learning). Par la suite, des algorithmes tels que tSNE, Spectral Embedding ou Locally Linear Embedding (LLE) ont apporté une solution plus précise à ce problème. Toutefois, ces algorithmes ont pour but de trouver une représentation en "faible" dimension de nos données. De fait, une partie de l'information géométrique est perdue au profit d'une dimension inférieure. L'objectif de la topologie des données est de récupérer des informations de contextes sur la variété topologique portant les données sans pour autant en amener une construction visualisable.

C'est avec les travaux préliminaires de [39, 106] que l'analyse topologique des données (TDA) pris son essor. L'idée derrière ces techniques est assez simple : les données ont une forme, et celle-ci contient des informations structurelles importantes. Dans cette courte introduction à la TDA, nous allons dans un premier temps aborder les outils permettant de reconstruire la forme en question. Dans un second temps, nous verrons des outils mathématiques permettant d'analyser quantitativement l'objet mathématique obtenu.

Remarque(s)

Les packages python utilisés dans ces travaux sont majoritairement GUDHI [78] et giotto-tda [99] pour l'algorithme Mapper

1.4.1 Complexe simplicial, de Čech, de Vietoris Rips

On considère $V = \{v_1, \dots, v_N\}$ un ensemble de sommets sur lequel on veut construire une structure topologique. De manière générale, ces sommets sont des points dans \mathbb{R}^n , ou plus généralement dans un espace métrique.



Definition 1.4.12 k -simplexe

Un k -simplexe de V est un ensemble σ_k formé d'un sous-ensemble $V_k = \{v_{i_0}, \dots, v_{i_k}\} \subset V$ de taille $k + 1$ et de tous ses sous-ensembles :

$$\sigma_n = \{V_k\} \cup \{V_i \mid V_i \subset V_k\}$$

On dira aussi que σ_k est un simplexe de dimension k . Les éléments de V sont appelés **sommets**, et tout simplexe $\sigma_i \subset \sigma_k$ est appelé une **face**.

Dans la suite, on notera $\sigma_{i_0 i_1 \dots i_k}$ le k -simplexe dont les $k + 1$ sommets sont $\{v_{i_0}, v_{i_1}, \dots, v_{i_k}\}$.

 **Exemple(s)**

On considère un ensemble $V = \{a, b, c, d\}$ de 4 points dans \mathbb{R}^n .

- $\{\{a\}\}$ est un 0-simplexe (et tout élément de V , autrement dit tout sommet, est un 0-simplexe).
- $\{\{a, b\}, \{a\}, \{b\}\}$ est un 1-simplexe.
- $\{\{a, b\}\}$ n'est pas un simplexe.

 **Remarque(s)**

La définition précédente est celle d'un k -simplexe **abstrait**. Dans la plupart des cas, on peut avoir une réalisation géométrique, qui correspond d'ailleurs plus à ce que l'on veut obtenir (la notion de simplexe abstrait est discrète, tandis que sa réalisation géométrique amène une notion de continuité).

 **Definition 1.4.13**

Soit σ_k un k -simplexe abstrait de sommets $V_k = \{v_1, \dots, v_k, v_{k+1}\}$. Une **réalisation géométrique** de σ_k est l'enveloppe convexe C des $k + 1$ points de V_k telle que C est k dimensionnelle.

 **Remarque(s)**

Lorsqu'il y a une notion d'indépendance linéaire, alors il faut et il suffit que V_k soit un ensemble de $k + 1$ points linéairement indépendants.

Ces simplexes représentent les outils de bases pour reconstruire une approximation de la structure topologique sous-jacente à V . Il est temps maintenant de définir la notion principale :

 **Definition 1.4.14 Complexe Simplicial**

Un **complexe simplicial** \mathcal{C} sur V est une collection de simplexes de V telle que :

- Toute face de $\sigma \subset \mathcal{C}$ est un simplexe de \mathcal{C}
- Soit σ_1, σ_2 deux simplexes de \mathcal{C} . Leur intersection est soit vide, soit une face commune aux deux.

La **dimension** de \mathcal{C} est le plus haut k pour lequel \mathcal{C} admet un k -simplexe.

L'union des simplexes de \mathcal{C} est un sous-ensemble de \mathbb{R}^n héritant de sa topologie. Cet ensemble est appelé **espace sous-jacent** et peut être vu comme un espace topologique à part entière.

Remarque(s)

- ⇒ Comme nous le verrons avec les complexes de Čech ou de Vietoris-Rips, la dimension d'un complexe simplicial sur $V \subset \mathbb{R}^n$ peut être de dimension $N \leq n$, et donc ne pas admettre de réalisation géométrique dans \mathbb{R}^n . La topologie dont hérite \mathcal{C} est donc celle de \mathbb{R}^N .
- ⇒ Bien que ce soit anecdotique sur \mathbb{R}^n , il faut bien garder à l'esprit qu'un complexe simplicial peut hériter de topologies différentes selon celle considérée sur l'espace de départ (on pourra penser aux espaces fonctionnels). De même, deux complexes simpliciaux construits sous la même topologie peuvent l'exprimer de manière différente selon la méthode utilisée (ce sera le cas pour les complexes de Čech et Vietoris-Rips).

Maintenant que nous nous sommes munis d'un outil créant une structure topologique sur un nuage de point (vu comme un ensemble de sommets), il nous faut une méthode de construction. Nous allons en voir deux dans cette introduction. On se place, dans la suite dans un espace métrique (E, d) .



Definition 1.4.15 Complexe de Čech

On considère un ensemble $V = \{v_1, \dots, v_N\}$ de points de (E, d) tel que N est non nécessairement égal à $\dim(E)$. Alors le **complexe de Čech au niveau α** est l'ensemble $\text{Cech}_\alpha(V)$ des simplexes tels que :

$$\sigma_{i_0 \dots i_k} \in \text{Cech}_\alpha(V) \Leftrightarrow \bigcap_{j=0}^k \overline{B}(v_{i_j}, \frac{\alpha}{2}) \neq \emptyset$$

(On rappellera que $\overline{B}(v, \alpha) = \{x \in \mathbb{R}^n, d(v, x) \leq \alpha\}$ et dénote ici les boules **fermées** de \mathbb{R}^n).



Definition 1.4.16 Complexe de Vietoris Rips

On considère un ensemble $V = \{v_1, \dots, v_N\}$ de points de (E, d) tel que N est non nécessairement égal à $\dim(E)$. Alors le **complexe de Vietoris-Rips au niveau α** est l'ensemble $\text{Rips}_\alpha(V)$ des simplexes tels que :

$$\sigma_{i_0 \dots i_k} \in \text{Rips}_\alpha(V) \Leftrightarrow d(v_{i_j}, c_{i_l}) \leq \alpha, \quad \forall j, l \in \{0, \dots, k\}$$

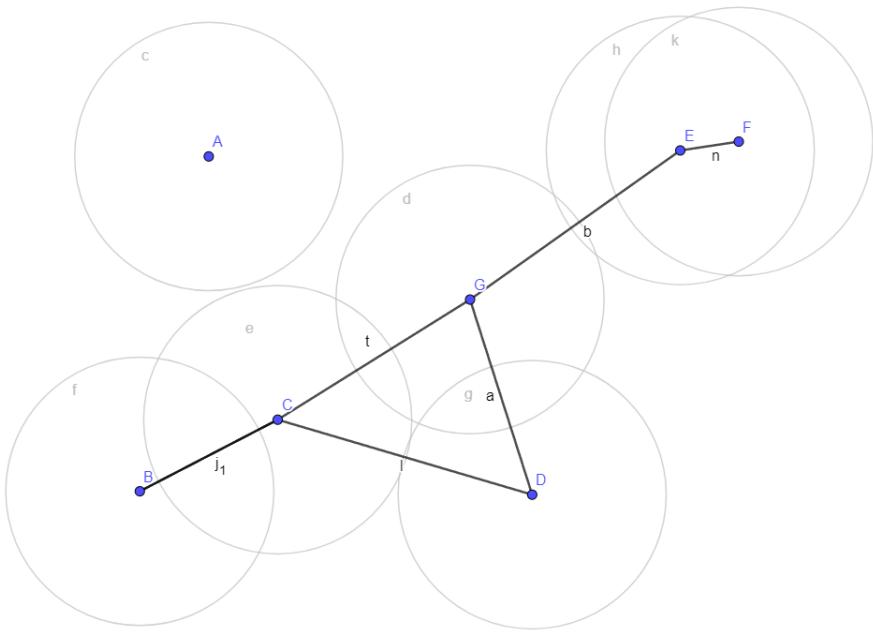


FIGURE 1.8 – Complexe de Čech au niveau α

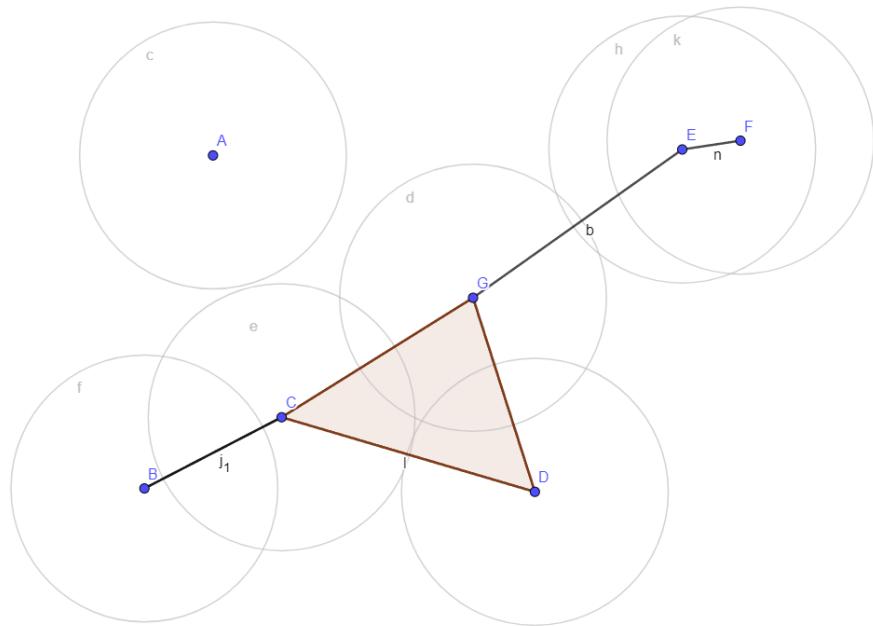


FIGURE 1.9 – Complexe de Vietoris Rips associé au niveau α

On a la suite d'inclusion suivante :

$$\text{Rips}_\alpha \subseteq \text{Cech}_{\sqrt{2}\alpha} \subseteq \text{Rips}_{\sqrt{2}\alpha}$$

Dans leur construction, le complexe de Vietoris Rips est plus simple : il suffit de considérer les distances par paires. Tout n -uplets dont les distances 2 à 2 sont inférieures à α forment un $(n - 1)$ -simplexe. Pour Čech, il faut, en plus de cette propriété, vérifier que tout point dans l'enveloppe convexe du n -uplet appartient à toutes les boules fermées $\bar{B}(x_i, \alpha)$. Un exemple non trivial de différence entre ces deux complexes se trouvent dans la figure 1.10

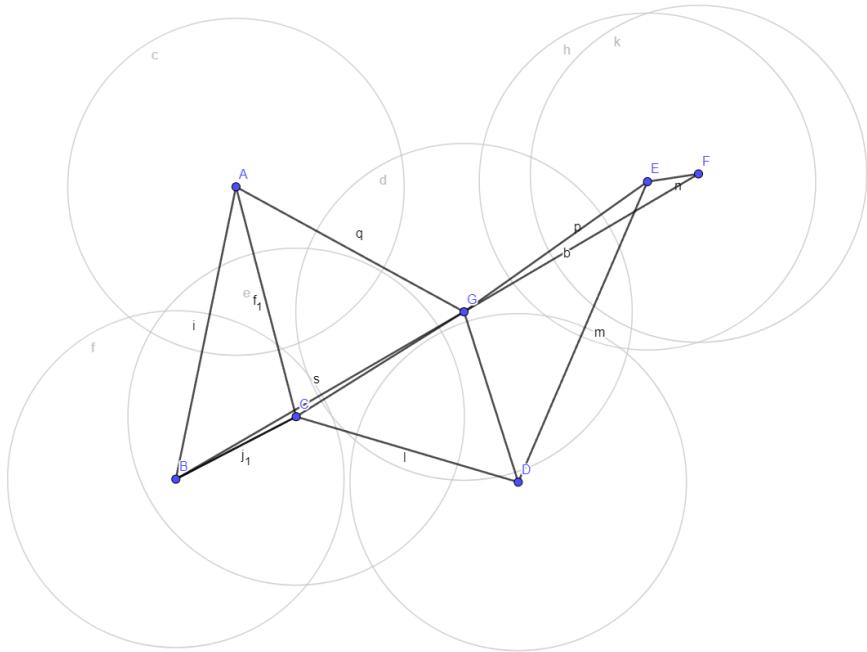


FIGURE 1.10 – Sur cette construction où nous n'avons dessiné que le squelette 1-D, les points A, B, C et G sont liés par paires et par triplets autant dans le complexe de Rips que dans celui de Čech. Toutefois, il n'existe pas d'intersection non vide pour les 4 boules fermées centrées sur ces 4 points et de rayon α . Le 3-simplexe formé de A, B, C, G est donc dans le complexe de Rips **mais pas** dans celui de Čech

Le complexe de Čech est toutefois un cas particulier de la construction suivante :



Definition 1.4.17 Nerf d'un recouvrement

On considère un espace topologique \mathcal{M} muni d'un recouvrement d'ouvert (U_i) (i.e. $\bigcup U_i = \mathcal{M}$). Le **nerf** de (U_i) est un complexe simplicial abstrait \mathcal{C} tel que :

- L'ensemble de ses sommets est l'ensemble des U_i .
- Le simplexe dont les sommets sont les $\{U_{i_0}, U_{i_1}, \dots, U_{i_k}\}$ est inclus dans \mathcal{C} si et seulement si $\bigcap U_i \neq \emptyset$

Exemple(s)

Regardons le nerf d'un recouvrement de la sphère. Pour cela, on va choisir un recouvrement en 4 ouverts de la sphère (on verra plus tard pourquoi ce recouvrement est pertinent). On définit donc :

$$\mathcal{U} = \{U_1, U_2, U_3, U_4\}$$

dont on épargnera la paramétrisation, mais que l'on représente ci-dessous dans la figure 1.11

Ces 4 ouverts sont créés de manière à n'avoir aucune intersection commune aux 4, mais une intersection pour chaque triplet d'ouverts. Le nerf associé est visible dans la figure 1.12

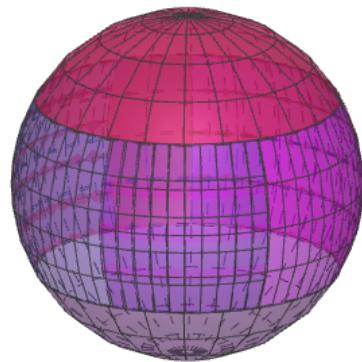


FIGURE 1.11 – Recouvrement de la sphère par 4 ouverts

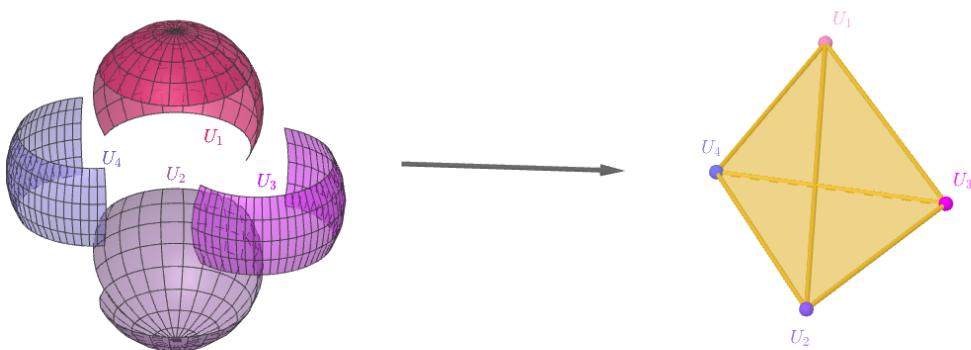


FIGURE 1.12 – Nerf du recouvrement

On peut alors voir le complexe de Čech comme le nerf d'un recouvrement $\{B(v_i, \alpha)\}$, chaque $B(v_i, \alpha)$ étant identifié par v_i . Maintenant, la question qui se pose est : comment utiliser la notion de **nerf d'un recouvrement** sur un ensemble de données avec une notion plus générale que le complexe de Čech ?

Dans cette partie, nous allons noter $\mathcal{X} = \{x_1, \dots, x_N\} \subset \mathbb{R}^n$ (bien que ce soit aussi valide sur des espaces de dimension infinie). Nous voulons créer un complexe simplicial sur \mathcal{X} en utilisant le nerf d'un recouvrement.

⇒ **Première étape :** Nous devons disposer d'un recouvrement de \mathcal{X} . Pour ceci, nous allons créer une collection d'ouvert dans le sens suivant :

*Deux éléments x_i et x_j sont dans le même ouvert V_i s'ils sont proches **au sens des données**.*

Bien sur, il est naturel de penser à la notion de distance, mais comment l'exploiter ?

Une solution, au sens statistique du terme, est d'utiliser un algorithme de clustering pour délimiter des sous-ensembles de données proches qui constitueront **une base** pour nos ouverts. Attention toutefois : il convient ici de laisser une certaine marge de manœuvre à l'algorithme pour que les clusters aient la possibilité d'une intersection non vide. En effet, suivant la définition du nerf d'un recouvrement, si les ouverts sont tous disjoints deux à deux, alors les seuls simplexes existant dans le complexe obtenu seront les V_i et nous aurons donc un complexe de dimension 0, ce qui n'amène aucune information supplémentaire par rapport au clustering simple.

⇒ On considère alors l'ensemble de clusters $V = \{V_1, \dots, V_i\}$ comme un ensemble de sommets sur lequel construire notre complexe grâce à la relation donnée par les intersections non-vides.

Cette méthode de construction (initié par [96]) donnera lieu à l'algorithme Mapper.

1.4.2 Homologie persistante.

Avant de démarrer cette sous-partie, nous renvoyons le lecteur à [95] pour une explication théorique des groupes d'homologie et de leur invariance par relation d'homotopie. Si le temps m'est laissé, une annexe verra peut être le jour à ce sujet.

Résumons ce que nous avons pour l'instant. Sur un ensemble de données $\mathcal{V} = \{v_1, \dots, v_N\} \subset \mathbb{R}^n$, nous avons construit une structure \mathcal{C} . Cette structure reflète-t-elle la topologie de \mathcal{V} ? Le résultat suivant nous donne cette garantie dans le cas où \mathcal{C} est construit comme le nerf d'un recouvrement :



Theorem 1.7 Nerve theorem

On considère $\mathcal{V} = (V_i)_{i \in I}$ un recouvrement d'un espace topologique \mathcal{M} . On considère aussi que \mathcal{V} vérifie :

- V_i est ouvert pour tout i .
- Pour tout $I_k \subset I$, $\cap_{i \in I_k} V_i$ est soit vide soit **contractile** (i.e. homotopiquement équivalent à un point).

Alors \mathcal{M} et le nerf de \mathcal{V} sont homotopiquement équivalents.

La relation d'homotopie induit une conservation de certaines propriétés topologiques : la **connexité** (simple ou par arcs), les **groupes d'homotopies**, les **groupes de co-homologies** et, cette dernière étant la principale propriété qui va nous intéresser, les **groupes d'homologies**.

Sans rentrer ici dans les détails de cette notion, nous allons présenter plusieurs outils permettant de quantifier les groupes d'homologies d'un complexe simplicial. En premier lieu, parlons des quantificateurs de l'homologie d'un complexe simplicial fixé.

i - Groupes d'homologies, nombres de Betti.



Remarque(s)

Comme les groupes d'homologies ne sont interprétables qu'aux travers des nombres de Betti et autres outils présentés après, ils sont accessoires à la compréhension des résultats

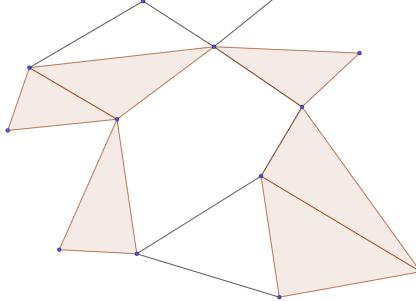
Procédons en deux temps :

- **Les nombres de Betti pour les non-courageux :** Les **nombres de Betti** d'un espace topologique sont une suite (b_0, b_1, \dots) représentant les nombres de cavités $k+1$ dimensionnelles. b_0 représente le nombre de composantes connexes, b_1 représente le nombre de cavités 2-dimensionnelles (donc dans le plan), b_2 le nombre de cavités 3-dimensionnelles, etc. Si l'espace topologique est de dimension k , alors $b_i = 0$ pour tout $i \geq k$.

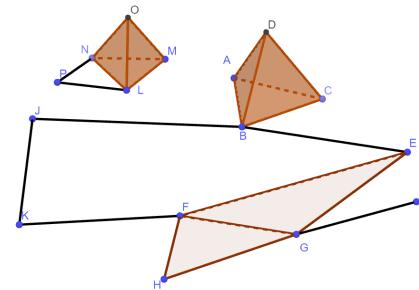


Exemple(s)

- Un cercle possède une composante connexe et une cavité 2-dimensionnelle. On a donc la suite de Betti suivante : $(1, 1, 0, 0, \dots)$
- Une sphère possède une composante connexe et une cavité 3-dimensionnelle. On a donc la suite de Betti suivante : $(1, 0, 1, 0, 0, \dots)$
- Un tore possède une composante connexe, une cavité 3-dimensionnelle, et 2 cavités 2-dimensionnelles (pour s'en convaincre : combien de fois peut on couper le tore en conservant $b_0 = 1$?). On obtient donc la suite de Betti suivante : $(1, 2, 1, 0, 0, \dots)$



(a) Ce complexe possède 1 composante connexe, 3 cavités 2-dimensionnelles et est de dimension 2. Sa suite de Betti est donc : (1,3,0,0,...)



(b) Ce complexe possède 2 composantes connexes, 2 cavités 2-dimensionnelles et 2 cavités 3-dimensionnelles ($OMLN$ et $ABCD$). Pour finir, ce complexe est de dimension 3. Sa suite de Betti est donc (2, 2, 2, 0, 0, ...)

FIGURE 1.13 – 2 exemples de complexes ainsi que leur suite de Betti.

Ce qu'il faut comprendre, c'est qu'en topologie deux objets sont considérés comme les mêmes si on peut les déformer de manière continue pour passer de l'un à l'autre. Et "*de manière continue*" implique la conservation des diverses cavités. La suite de Betti pour un complexe simplicial nous donne donc une caractérisation à déformation continue près de notre espace topologique.

- **Les groupes d'homologies :** On considère un complexe simplicial \mathcal{C} sur $\mathcal{V} = \{v_1, \dots, v_N\}$. On définit l'**espace des k -chaînes** sur \mathcal{C} comme l'ensemble des séries formelles de k -simplexes sur \mathbb{Z}_2 ⁹. Comme on se place sur \mathbb{Z}_2 , on a juste à exprimer la somme de deux complexes simpliciaux σ_1 et σ_2 comme la **différence symétrique** :

$$\sigma_1 + \sigma_2 = \sigma_1 \Delta \sigma_2 = (\sigma_1 \setminus \sigma_2) \cup (\sigma_2 \setminus \sigma_1)$$

En notant $\{\sigma_1, \dots, \sigma_p\}$ l'ensemble des k -simplexes de \mathcal{C} , l'espace des k -chaînes sur \mathcal{C} est donc :

$$C_k(\mathcal{C}) = \{c = \sum_{i=0}^p \epsilon_i \sigma_i, \quad \epsilon_i \in \mathbb{Z}_2\}$$

$C_k(\mathcal{C})$ est un \mathbb{Z}_2 espace vectoriel dont une base est l'ensemble des k -simplexes de \mathcal{C} . De manière intuitive :

9. On rappelle que $\mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z} = \{0, 1\}$ avec $1 + 1 = 0$



Definition 1.4.18 Opérateur de bord

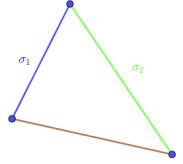
Soit $\sigma_{0,...,k}$ un k -simplexe dont les sommets sont $\{v_{i_1}, ..., v_{i_k}\}$. Le **bord** de $\sigma_{0,...,k}$ est la $(k - 1)$ -chaîne :

$$\partial_k(\sigma_{0,...,k}) = \sum_{i=0}^k (-1)^i \sigma_{0,...,i-1,i+1,...,k}$$

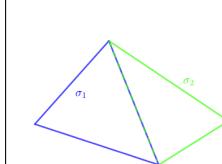
On définit alors :

$$\partial_k : C_k(\mathcal{C}) \rightarrow C_{k-1}(\mathcal{C})$$

comme l'**opérateur de bord de degré k** .



(a) Somme de deux arêtes.



(b) Somme de deux 2-simplexes.

Considérant que \mathcal{C} est construit sur un ensemble fini de sommets, l'ensemble $C_k(\mathcal{C})$ est de dimension $\leq \binom{\text{Card}(V)}{k}$. L'application ∂_k étant linéaire, nous pouvons regarder son image et son noyau.



Definition 1.4.19

On définit :

- $Z_k(\mathcal{C}) = \{c \in C_k(\mathcal{C}) : \partial_k(c) = 0\} \subseteq C_k(\mathcal{C})$. $Z_k(\mathcal{C})$ est le noyau de ∂_k et est appelé **l'espace des k -cycles de \mathcal{C}**
- $B_k(\mathcal{C}) = \{c \in C_k(\mathcal{C}), \exists c' \in C_{k+1}(\mathcal{C}), c = \partial_k(c')\}$. $B_k(\mathcal{C})$ est l'image de ∂_k et est appelé **l'espace des k -frontières de \mathcal{C}**

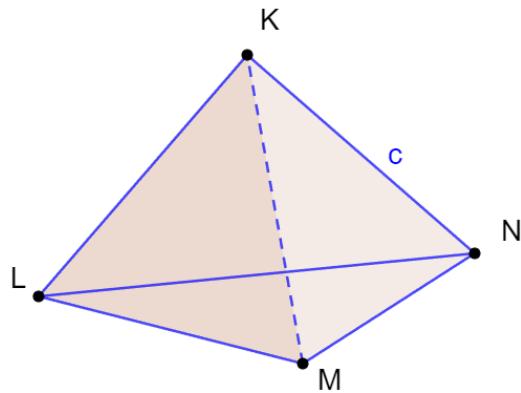
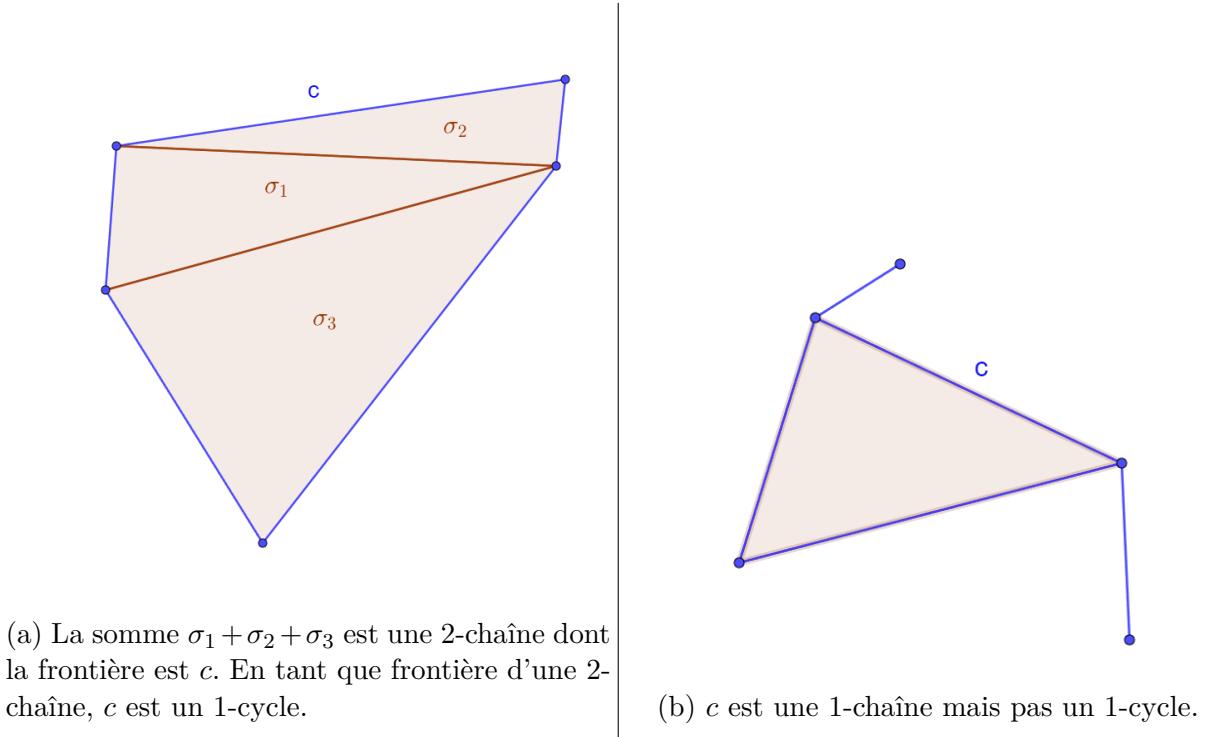
De manière intuitive, une k -frontière est un k -cycle. On a donc la suite d'inclusion suivante :

$$B_k(\mathcal{C}) \subseteq Z_k(\mathcal{C}) \subseteq C_k(\mathcal{C})$$

Cette suite d'inclusion a pour conséquence direct

$$\partial_{k-1} \circ \partial_k \equiv 0$$

Ces notions nous amènent à définir



(c) Ici, le 2-simplexe sur $\{K, L, M, N\}$ forme une 2-chaîne ET un 2-cycle.

FIGURE 1.15 – Plusieurs exemples de chaînes et cycles.

Definition 1.4.20 Groupe d'homologie simplicial et nombres de Betti

On considère \mathcal{C} un complexe simplicial. Le k -ème groupe d'homologie de \mathcal{C} est l'espace vectoriel quotient :

$$H_k(\mathcal{C}) = Z_k(\mathcal{C}) / B_k(\mathcal{C})$$

Le nombre $\beta_k(\mathcal{C}) = \dim(H_k(\mathcal{C}))$ est appelé **k-ème nombre de Betti**.

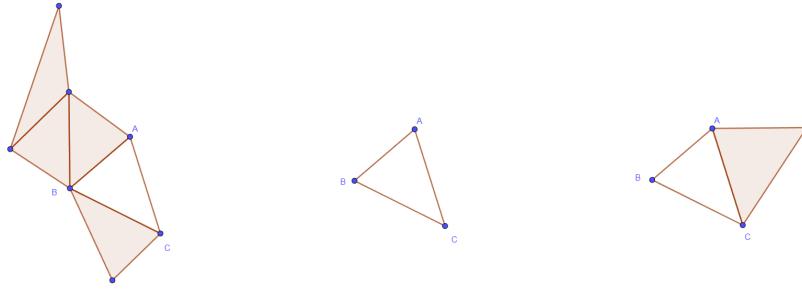


FIGURE 1.16 – Ces trois structures extraites du même complexe ont toutes un 2-cycle commun : ABC . Ce 2-cycle n'est toutefois pas une frontière (le 2-simplexe σ_{ABC} n'existe pas). Elles diffèrent toutes d'un certains nombres de frontières.

Analysons un peu en quoi consiste $H_k(\mathcal{C})$. Pour cela, on rappelle la définition d'un espace vectoriel quotient :



Definition 1.4.21

Soit E un espace vectoriel et F un sous-espace vectoriel de E . On définit E/F comme l'ensemble des classes d'équivalence pour la relation suivante :

Pour tout v, w dans E , v est en relation avec w si et seulement si $v - w \in F$

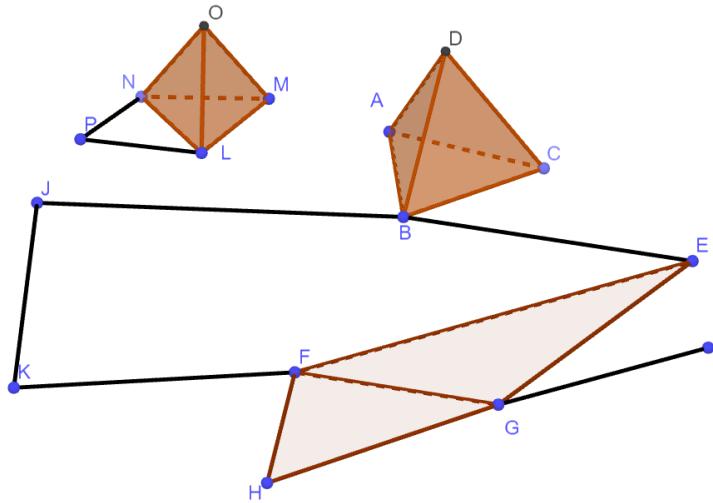
Autrement dit, toute classe d'équivalence $[v]$ se définit comme $[v] = v + F$.

Soit c et c' deux éléments de $Z_k(\mathcal{C})$. c et c' sont donc deux k -cycles de $C_k(\mathcal{C})$. On définit $[c] \in H_k(\mathcal{C})$ la classe d'équivalence de c . Alors c et c' sont en relation si

$$\begin{aligned} c - c' \in B_k(\mathcal{C}) &\Leftrightarrow \exists b \in B_k(\mathcal{C}), c' = c + b \\ &\Leftrightarrow \exists d \in C_{k+1}(\mathcal{C}), c' = c + \partial_{k+1}(d) \end{aligned}$$

Deux chaines sont donc en relation si elles diffèrent d'une frontière existante (voir fig 1.16).

On dira que les classes d'équivalence de $H_k(\mathcal{C})$ sont les **classes d'homologies**.


 FIGURE 1.17 – Le complexe \mathcal{C}

Exemple(s)

Calculons les groupes d'homologies du complexe de la figure 1.17. Puisqu'il n'y a pas d'ambiguité, nous ne préciserons pas que le complexe en question est \mathcal{C} . De même, on notera $\sigma_{i_1 i_2 \dots i_k}$ (resp $\sigma_{AB\dots K}$) le k -simplexe sur $\{v_{i_1}, \dots, v_{i_k}\}$ (resp. $\{A, B, \dots, K\}$) :

⇒ **Le cas $k = 0$** : La première étape avant toute chose est de regarder en quoi consiste C_0 . Il existe seize 0-simplexes (les sommets) :

$$C_0 = \left\{ \sum_{i=0}^{16} \epsilon_i \sigma_i \quad \epsilon_i \in \mathbb{Z}_2 \right\}$$

L'ensemble des frontières est constitué des sommes $v_i + v_j$ telles que σ_{ij} est une arête du complexe.

D'un autre côté, l'ensemble des cycles est constitué des sommes $v_i + v_j$ telles qu'il existe une suite d'arête $(\sigma_{i_m j_m})_m$ avec $j_m = i_{m-1}$ ^a.

On en déduit que pour trouver H_0 , il suffit de trouver tous les points ne pouvant être connectés ensemble. Il s'ensuit que :

$$H_0 = \{[A], [O]\}$$

avec $[A] = \{A, B, C, D, E, F, G, H, I, J, K\}$ et $[O] = \{O, M, N, L, P\}$. H_0 est donc de dimension 2, et donc $\beta_0 = 2$.

a. Sous entendu, v_i et v_j sont connectés par un chemin de dimension 1

Exemple(s)

⇒ **Le cas $k = 1$** : C_1 est l'ensemble des 1-chaines du complexe.

L'espace Z_1 contient les cycles de dimension 1 : ce sont les chaines fermées (dans le sens "lacet fermé"). Par exemple, les chaines $\sigma_{FG} + \sigma_{GH} + \sigma_{HF}$ et $\sigma_{LP} + \sigma_{PN} + \sigma_{NO} + \sigma_{OM} + \sigma_{ML}$ sont des cycles.

L'espace B_1 contient les frontières de dimension 1 : ce sont les cycles dont l'intérieur est "plein". Par exemple $\sigma_{FG} + \sigma_{GH} + \sigma_{HF}$ et $\sigma_{AD} + \sigma_{DC} + \sigma_{CB} + \sigma_{BA}$ sont des frontières. Le cycle $\sigma_{LP} + \sigma_{PN} + \sigma_{NO} + \sigma_{OM} + \sigma_{ML}$ quant à lui n'est pas une frontière. On en déduit qu' H_1 contient autant de classes que de cycles de *taille minimale*^a.

$$H_1 = \{[c_1], [c_2]\}$$

avec $c_1 = \sigma_{LP} + \sigma_{PN} + \sigma_{NL}$ et $c_2 = \sigma_{JB} + \sigma_{BE} + \sigma_{EF} + \sigma_{FK} + \sigma_{KJ}$. Il s'ensuit que $\dim(H_1) = 2$, et donc $\beta_1 = 2$.

⇒ **Le cas $k = 2$** : Ce cas est le plus simple. En effet, le complexe \mathcal{C} étant de dimension 2, il ne contient aucun simplexe de dimension 3. De fait, aucune 2-chaine ne peut être une frontière. Il s'ensuit :

$$B_2 = \emptyset \quad \Rightarrow H_2 = Z_2$$

Il suffit donc de trouver les 2-cycles de \mathcal{C} . On a donc :

$$H_2 = \{[\sigma_{LMON}], [\sigma_{ABCD}]\}$$

On conclut finalement que $\dim(H_2) = \beta_2 = 2$

a. Dans le sens où il n'existe pas 3 sommets de la chaîne formant un 1-simplexe de \mathcal{C} .

ii - Filtrations

L'homologie est un outil qui permet de caractériser un complexe simplicial sur un nuage de point. Une question naturelle est :

Ce complexe simplicial est-il unique ?

Il est évident que non : il suffit, pour s'en convaincre, de considérer les complexes de Čech/Rips à deux niveaux différents. Une nouvelle question naturelle vient ensuite :

Comment quantifier la topologie d'un nuage de point en incluant la variabilité des complexes simpliciaux construits dessus ?

Comme toute question naturelle en mathématiques, il est difficile d'apporter une réponse précise. Toutefois, il est possible d'apporter une réponse à **méthode de construction fixée**¹⁰.

10. En pratique tout du moins : le principe de filtration est indépendant d'une méthode de construction.

**Definition 1.4.22**

Soit \mathcal{C} un complexe simplicial sur un ensemble $\mathcal{V} = \{v_1, \dots, v_N\}$ de sommets dans \mathbb{R}^n . **Une filtration de \mathcal{C}** est une collection de complexe $\mathcal{F} = (\mathcal{C}_r)_{r \in [0;T]}$ telle que :

- Pour tout $r \leq r'$ dans $[0;T]$, alors $\mathcal{C}_r \subseteq \mathcal{C}_{r'}$
- $\bigcup_{r \in [0;T]} \mathcal{C}_r = \mathcal{C}$

**Exemple(s)**

Les ensembles $(\text{Cech}_\alpha)_{\alpha \in [0;T]}$ (voir figure 1.18) et $(\text{Rips}_\alpha)_{\alpha \in [0;T]}$ forment des filtrations pour, respectivement, Cech_T et Rips_T .

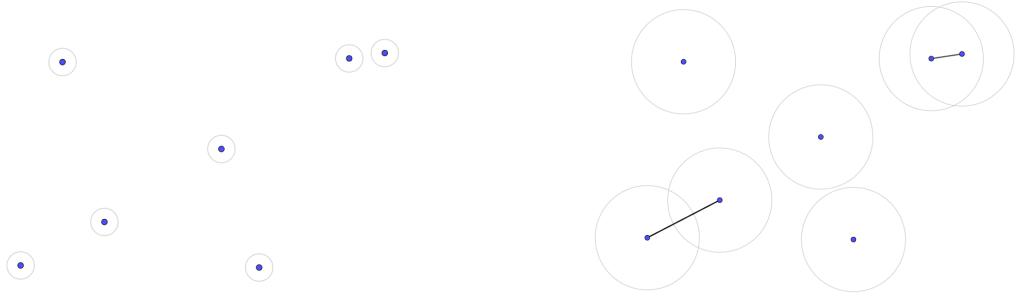
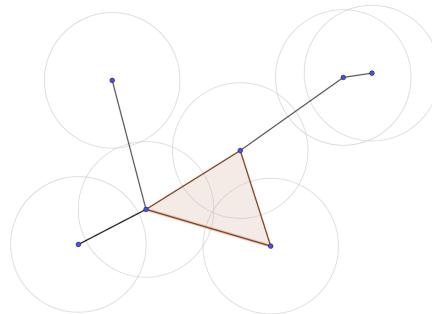
(a) Le complexe de Čech pour $\alpha \simeq 0.5$ (b) Le complexe de Čech pour $\alpha \simeq 2.2$ (c) Le complexe de Čech pour $\alpha \simeq 3.25$

FIGURE 1.18 – Des éléments de la filtration de Čech sur un nuage de point.

Il est assez intuitif que, si l'ensemble de sommets \mathcal{V} est fini, alors la collection (\mathcal{C}_r) le sera aussi. De même, en choisissant T assez grand, alors le complexe simplicial $\mathcal{C} = \bigcup \mathcal{C}_r$ est juste un simplexe de dimension $\text{card}(V) - 1$ (pour Rips, toutes les paires de points vont finir par se connecter, pour Čech, toutes les $B(x_i, r)$ vont finir par avoir une intersection commune).

Remarque(s)

Le paragraphe ci-dessous définit les espaces d'homologie persistante. Cette définition n'est pas nécessaire pour une compréhension purement statistique. Elle permet toutefois de comprendre les fondements théoriques des diagrammes de persistance.

Cette suite croissante (pour l'inclusion) de complexes simpliciaux induit une suite pour chaque groupe d'homologie : pour k fixé, nous pouvons étudier la suite $(H_k(\mathcal{C}_r))_r$. Ainsi, à k fixé, l'inclusion $\mathcal{C}_s \subset \mathcal{C}_t$ induit une application linéaire $H_k(\mathcal{C}_s) \rightarrow H_k(\mathcal{C}_t)$. On va définir un espace d'homologie incluant des informations sur \mathcal{C}_s et \mathcal{C}_t .

Definition 1.4.23 Espace d'homologie persistante

Soit \mathcal{C} un complexe simplicial et $\mathcal{F} = (\mathcal{C}_r)$ une filtration sur \mathcal{C} . Les **espaces d'homologie persistante pour la dimension k** sont définis, pour $s \leq t$, par :

$$H_k^{s,t}(\mathcal{C}) = Z_k(\mathcal{C}_s) / (B_k(\mathcal{C}_t) \cap Z_k(\mathcal{C}_s))$$

Les éléments de $H_k^{s,t}(\mathcal{C})$ sont appelés classes d'homologie persistante.

Remarque(s)

Pour comprendre ce que contient $H_k^{s,t}$, regardons ce que contient $B_k(\mathcal{C}_s) \cap Z_k(\mathcal{C}_t)$.

- $B_k(\mathcal{C}_t)$ contient les frontière de dimension k appartenant à \mathcal{C}_t .
 - $Z_k(\mathcal{C}_s)$ contient les cycles de dimension k appartenant à \mathcal{C}_s
 - Un k -cycle dans \mathcal{C}_s reste un k -cycle dans \mathcal{C}_t . On a donc l'inclusion $Z_k(\mathcal{C}_s) \subseteq Z_k(\mathcal{C}_t)$.
 - De même, on a l'inclusion $B_k(\mathcal{C}_t) \subseteq Z_k(\mathcal{C}_t)$.
- ⇒ L'intersection $B_k(\mathcal{C}_t) \cap Z_k(\mathcal{C}_s)$ définit les cycles existant au temps s qui sont une frontière au temps t .
- ⇒ Les éléments de $H_k^{s,t}$ sont donc des classes d'homologie de \mathcal{C}_s qui **persistent** dans \mathcal{C}_t .

En comparant l'indice temporel d'apparition (*naissance*) d'une classe d'homologie à sa date de disparition (*mort*), nous pouvons obtenir une *durée de vie* pour les classes d'homologies (et donc avoir en chaque temps la dimension de H_k pour tout k). Ce sont ces durées de vie que nous allons étudier au travers des **diagramme de persistance**.

Remarque(s)

Le terme de *mort/disparition* d'une classe d'homologie ne reflète pas exactement ce qu'il se passe : une classe d'homologie disparaît au moment où elle ne contient plus d'élément qui n'est pas une frontière de dimension supérieure. A ce moment là, les cycles la composant viennent se greffer aux autres classes par somme de chaines.

ii - Code barres, diagramme de persistance.

Avec le principe de filtration, nous abandonnons l'idée de complexe simplicial unique pour étudier une collection imbriquée de complexe simpliciaux. Grâce à ce concept, nous allons incorporer une certaine variabilité de forme. Nous allons ici essayer de quantifier les structures **persistentes**.

En effet, en évoluant le long d'une filtration $\mathcal{F} = (\mathcal{C}_r)$, des structures vont apparaître et disparaître. Un exemple simple est le passage du complexe initial, où il n'existe aucune connexion entre les sommets, au premier complexe où une connexion existe. En considérant r comme une indexation *temporelle*, la filtration consiste en un chemin (dans notre cas discret) dans l'espace des complexes simpliciaux sur l'ensemble de points considéré. Ce que nous allons faire, c'est quantifier le temps d'apparition d'une k -chaîne de \mathcal{C} et le temps de disparition de cette même chaîne. Ce concept s'appelle **homologie persistante** et est résumé par un **Code barre**¹¹ (voir Figure 1.19).



Definition 1.4.24 Code barre

On considère une filtration $\mathcal{F} = (\mathcal{C}_r)_r$. Pour chaque classe d'homologie persistante $[c_i]$, on définit b_i, d_i respectivement les temps de naissance et de mort de la classe $[c_i]$.

Le **code barre** de \mathcal{C} est défini par l'ensemble des intervalles de vie des classes d'homologie persistantes :

$$BC(\mathcal{C}) = \bigcup_{[c_i]} \{[b_i, d_i]\}$$



Remarque(s)

Pour que la dernière définition soit consistante, on aura les conventions suivantes :

- Si $[c_i]$ n'apparaît pas, alors $\{[b_i, d_i]\} = \emptyset$
- Si $[c_i]$ ne meurt pas, alors $d_i = +\infty$.

De ce code barre découle la notion de **diagramme de persistance**. (voir Figure 1.20)



Definition 1.4.25

Le **diagramme de persistance** est le nuage de points de \mathbb{R}^2 dont chaque élément est la donnée (b_i, d_i) .



Remarque(s)

Dans le chapitre 4, nous définirons au moins une distance sur les diagrammes de persistance. Cette distance permettra de calculer la dissimilarité entre 2 filtrations.

11. Une animation explicative est disponible sur <https://remivaucher.github.io/manuscrit>



FIGURE 1.19 – La filtration induite d'un complexe de Vietoris Rips et de son code barre associé.

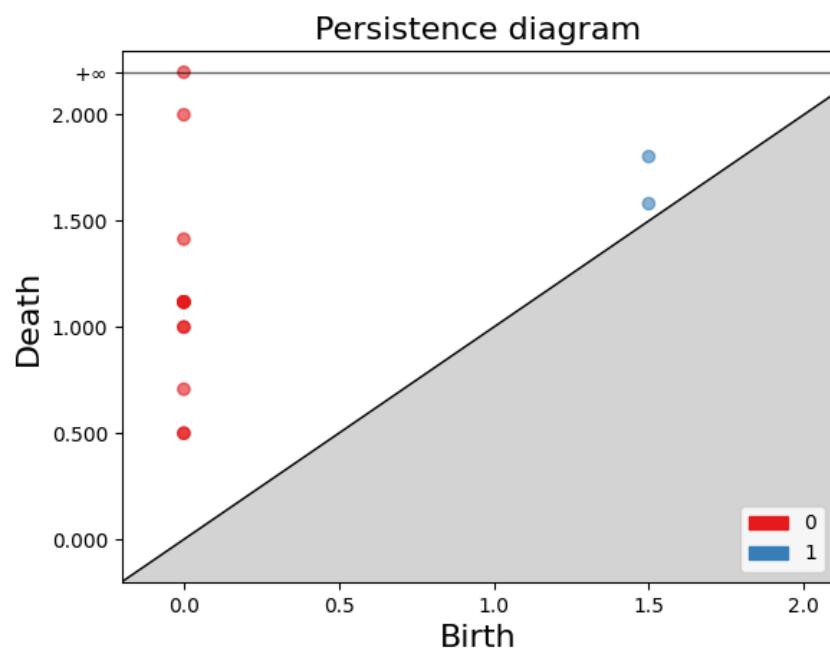


FIGURE 1.20 – Diagramme de persistance pour la filtration de la figure 1.19

Chapitre 2

Analyse topologique de plusieurs signaux temporels dans l'espace des signatures.

Pour rencontrer monsieur Lee, vaut mieux une bonne couverture !

OSS 117

Ce chapitre est une version allongée et retravaillé d'un travail commun avec Stéphane Chrétien, Ben Gao et Astrid Thebault Guiochon, travail soumis et accepté à la conférence *Complex Network 2023*. Dans cette version, nous amenons :

- Une deuxième version de l'algorithme dont le but est de s'appliquer à un échantillon géographique de signaux multivariés. Si $Y = (X^1, \dots, X^d)$ est une série temporelle multivariée, l'objectif est de construire un complexe simplicial sur un échantillon (Y_1, \dots, Y_n) . Par ailleurs, cette méthode est présentée dans sa version la plus générale qui permet de découvrir une structure topologique reflétant l'explicabilité au sein d'un ensemble de séries multivariées **de dimension différentes**.
- Un résultat de dérivabilité pour l'application signature. Bien que la structure topologique soit créé dans l'espace des signatures, l'objectif est que notre résultat soit transposable dans l'espace des signaux de variations bornées. Pour cela, il faut s'assurer que l'application signature transporte bien la topologie. Nous montrons donc que la signature est un homéomorphisme de classe C^1 .
- Des limites pour cet algorithme et des pistes de travail pour une amélioration. En effet, nous allons voir que cet algorithme peut encore être amélioré en étant plus attentif à la structure de l'espace des signatures. Nous allons exhiber certains difficultés et réfléchir à des éléments de solutions possibles.

2.1 Introduction

Topological Data Analysis (TDA) is a new field with a wide range of application in fields such as finance, neuroscience, medicine, etc. TDA addresses the problem of accounting for groupwise interactions in the data and therefore opens very promising prospects to better apprehend complex phenomena than models relying on pairwise interactions only. Several tools from algebraic topology, such as homology groups, homotopy groups, Betti numbers, etc can be put to work in building a set of relevant features that can capture the intricate nature of dependencies.

Some compelling examples of the benefits of using topological features appear regularly in the literature. In [85] the homological features of brain functional networks are shown to take different values in two states depending on the absorption of some drug. More generally, it is shown in [97] that homological cycles in structural brain networks finds connections between regions of early and late evolutionary origin. TDA can also be efficiently used in dynamical settings as well. One very intriguing example is change detection as illustrated in the study of functional brain networks conditioned in different tasks [98]. In [61] it is investigated how speech-related brain regions connectivity changes in different scenarios of speech perception.

In the present paper, we focus on the analysis of dynamical high dimensional phenomena and on the problem of constructing associated relevant topological structures with the aim of proposing news computational tools for deepening our understanding of the higher order structures hidden in time series data. Our main contribution is a new approach to building statistically informed simplicial complexes and possibly more general structures.

Our two main tools will be the basic objects of TDA and Signature theory. Signatures were recently proposed as a very powerful feature map for time series and dynamical systems in [22, 40, 74]. The introduction of Signatures for building topological structures for high dimensional dynamical phenomena is new and appears as a key and very natural ingredient that can accurately account for orientations of the various simplices in the complex at hand while capturing the main shape features from the dynamics. Using Signature in a statistical/machine learning context is an approach which is adopted in a growing number of applications nowadays [40] and our work is also intended to illustrate the relevance of Signature theory combined with statistics/machine learning for building a higher order interaction modelling framework.

In mathematical terms, our proposal is based on the assumption that k -simplices are simply sets of k nodes with there time series attached to them, with an orientation prescribed by the ordering of the nodes in computing their associated k -Signature. Recall that the orientation encoded in the computation of the associated signature carries potentially very interesting interpretation about the causal dependencies of the times series [50]. In the next step, the relevance of incorporating a simplex into our simplicial complex is assessed using a purely statistical procedure : each oriented k -simplex is associated with a corresponding k -Signature that is included into a set of multivariate features that is used to predict the Signatures of all the other potential simplices. More precisely, our construction is a generalisation of the approach developed by Meinshausen and Bühlmann in [79] for Gaussian Graphical Models. Simplices that are selected from the set of potential Simplices whose Signature can predict the Signature of a target simplex in

terms of confidently regressing or predicting¹ the Signature associated with this target are included as candidates for being considered as adjacent to this target. Using this procedure, we obtain a construction of a simplicial complex that accurately incorporates the statistical relationships between all the simplices in terms of regression or prediction, while keeping track of the inherent orientations of the simplices.

The plan of the paper is as follows. In Section 2.2, we recall the necessary background on topological data analysis and Signature theory. In section 2.3 we study how signature transform ensure to retrieve the topology on signals from the signature one. In Section 2.6, we present our method for constructing the simplicial complex using the Signatures of the simplices and the LASSO algorithm. In Section 2.7, we present our numerical experiments on real datasets. A conclusion section completes the paper.

2.2 Background on Signatures and topology

In this section, we summarise the mathematical prerequisites from topology and the theory of Signatures.

2.2.1 Recalls on the theory of Signatures

The theory of Signatures is a new topic of growing interest that emerged as a sub-branch of the theory of rough paths [46, 47, 75] which has a long history in mathematics and control that may have started with the work of Chen [19]. Rough paths provide a new framework for the analysis of stochastic processes and permitted to resolve various open problems, including the existence of a solution to the KPZ equation in mathematical physics, a result for which Martin Heirer was awarded the Fields medal [38]. Recently, this theory developed as a new tool for the analysis of signals in the area of Machine Learning [22, 40, 74] where remarkable performance was achieved for a series of difficult practical problems including the analysis of financial data, medical data and textual data [63, 74]. Lately, an intriguing relationship with recurrent neural networks was exhibited using the viewpoint of control theory [44].

Let us now turn to the definition and some interesting properties of Signatures. Consider a d -dimensional path $X = (X^1, X^2, \dots, X^d) : \mathbb{R} \rightarrow \mathbb{R}^d$. Then, the (truncated)² **signature** of X on $[a, b]$ is an object in $\mathcal{T}(\mathbb{R}^d) = \mathbb{R}^d \oplus \mathbb{R}^{d \times d} \oplus \mathbb{R}^{d \times d \times d} \oplus \dots$, defined, for $j \in \mathbb{N}^*$ by

$$\begin{aligned} (S_{[a,b]}(X))_{i_1, i_2, \dots, i_j} &:= S_{[a,b]}^{i_1 i_2 \dots i_j}(X) \\ &= \int_{a \leq s_1 \leq s_2 \leq \dots \leq s_j \leq b} dX_{s_1}^{i_1} dX_{s_2}^{i_2} \dots dX_{s_j}^{i_j} \end{aligned} \quad (2.1)$$

which lies in $\overbrace{\mathbb{R}^d \times \mathbb{R}^d \times \dots \times \mathbb{R}^d}^j$.

Chen's identity is a very useful result that allows to compute the Signature recursively based on linear interpolation of observed values of a trajectory.

1. for time dependent Signatures

2. The k -truncated version of the signature is $S^{(1)}(X) \oplus S^{(2)}(X) \oplus \dots \oplus S^{(k)}(X)$

Theorem 2.2.1 (Chen's identity). *Let $X : [a, b] \rightarrow \mathbb{R}^d$ and $Y : [b, c] \rightarrow \mathbb{R}^d$. Consider the concatenation of X and Y (noted $X * Y$) defined by :*

$$(X * Y) : [a, c] \rightarrow \mathbb{R}^d$$

$$t \mapsto \begin{cases} X(t) & , t \in [a, b] \\ X(b) - Y(b) + Y(t) & , t \in [b, c]. \end{cases}$$

Then :

$$S_{[a,c]}(X * Y) = S_{[a,b]}(X) \otimes S_{[b,c]}(Y)$$

Augmentation of a path

The Signature defines X uniquely on $[a, b]$ close to *tree-like equivalence* (i.e. there exist $I, J \subset [a, b]$, such that $X|_I(t) = X|_J(b-t)$).

Proposition 2.2.1. *$S_{[a,b]}(X)$ define X uniquely on $[a, b]$ if there exists $1 \leq i \leq d$ such that X^i is strictly monotonic on $[a, b]$.*

This result leads to consider the **time-augmented path** \tilde{X} associated with X , defined as $\tilde{X} = (t, X^1, X^2, \dots, X^d)$ in order to ensure the unicity of $S(X)$. Another augmentation will be useful for our work, namely the Lead-Lag augmentation.

Definition 2.2.1. *Consider a d -dimensional path X with $T+1$ timesteps. The **lead-lag augmentation** of X is a $2d$ -dimensional path $X_{lead,lag} = (X^{Lead}, X^{Lag})$ with $2T+1$ time steps such that :*

$$X^{Lead} = \{X(0), X(1), X(1), X(2), \dots, X(T), X(T)\}$$

$$X^{Lag} = \{X(0), X(0), X(1), X(1), \dots, X(T-1), X(T)\}$$

The Lead-Lag augmentation was found to play an important role in many machine learning applications [40].

2.2.2 Recalls on topology

We now turn to some useful definitions from topology. Consider a set of n vertices $V = \{v_1, \dots, v_n\}$.

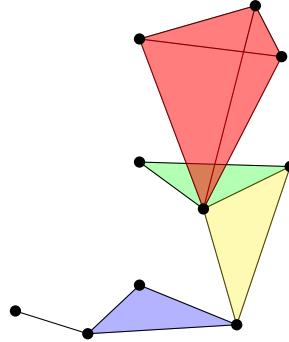
Definition 2.2.2. *For $k < n$, a k -simplex σ_k of V is the collection of a subset V_k of length $k+1$ and all its subsets. The **geometric realization** of a k -simplex is the convex hull C of $k+1$ points, such that $\dim(C) = k$. A face (of dimension l) σ_k is a collection of set in σ_k that form a l -simplex ($l \leq k$).*

Definition 2.2.3. *A **simplicial complex** \mathcal{C} on V is a collection of simplices (of V) such that for every $\sigma^i \in \mathcal{C}$, there exists j with $\sigma^i \cap \sigma^j$ a sub-simplex of both σ^i and σ^j .*

Definition 2.2.4. *The **dimension** of \mathcal{C} is the dimension of the highest simplex in \mathcal{C} (i.e. the highest k such that there exist a k -simplex in \mathcal{C}).*

Definition 2.2.5. *An **orientation** on \mathcal{C} is a order of the vertices for every simplices in \mathcal{C} . We use the notation $[v_1, v_2, v_3]$ to denote an oriented 3-simplex.*

From a geometric point of view, \mathcal{C} is constructed by attaching a group of simplices to each other by binding them with a shared face.



Definition 2.2.6. Consider $k \in N^*$ and σ a k -simplex in a simplicial complex \mathcal{C} . Its **link** $Lk(\sigma, \mathcal{C})$ in \mathcal{C} is the set of all faces $\tau \subset \mathcal{C}$ such that :

- $\sigma \cap \tau = \emptyset$
- $\sigma \cup \tau$ is a face of \mathcal{C}

Remarks :

- The link of a vertex is called **neighborhood** in graph theory.
- In the following algorithm, for a vertex v and a fixed k , we build the subset of all the k -simplex in $Lk(v, \mathcal{C})$ that we call k -dimensional Link (for simplicity). The k -dimensional Link is denoted $Lk_k(v, \mathcal{C})$

The goal of our work is to build a simplicial complex encoding the groupwise interactions explaining the dependencies inside high dimensional time series. The main ingredient in our approach is to make use of relevance measures in predicting *simplex indexed*-groups of time series using other *simplex indexed*-groups of time series as a criterion for selecting potential higher and higher dimensional simplices for (or for the sake of mitigating the computational complexity, sequential greedy) aggregations. We turn to the description of our approach in Section 2.6.

2.3 Continuity, differentiability for signature transform

We prove the following result :

Theorem 2.3.1. Consider that :

- X is a bounded variation path.
- Even if it means considering an augmented version of X , we consider $X_0 = 0$ and $X_t^1 = \text{len}(X|_{[0,t]})$ (or at least X is tree-like reduced).
- X is piecewise linear.

Then, the signature is a C^1 -homeomorphism from the space of X defined as above to its image $G(\mathbb{R}^d)$.

Proof. As there is no doubt that each paths and each signatures are taken on the same interval, we will not precise it.

We begin by a continuity result.

Proposition 2.3.1. Let $X \in BV(\mathbb{R}^d)$. Then, for any $n \geq 0$,

$$\|S^{(k)}(X)\|_{T(\mathbb{R}^d)} \leq \frac{1}{k!} \|X\|_{1-\text{var}} < \infty \quad (2.2)$$

and

$$\|S(X)\|_{T(\mathbb{R}^d)} \leq \exp(\|X\|_{1-\text{var}}) < \infty$$

Now, recall a result from [54] (the reader can refer to this for tree-like equivalence)

Theorem 2.3.2. Consider $X, Y \in BV(\mathbb{R}^d)$. Then, $S(X) = S(Y)$ iff X and Y are tree-like-equivalent. Moreover, there exists a unique path (up to reparametrization and translation) of minimal length in its equivalence class. This path \tilde{X} is called tree-like reduced.

Without giving any more details, it is intuitive that the augmented path $\tilde{X} = (t, X_t)$ is **always** tree-like reduced.³ We use the notation $BV_{\text{red}}(\mathbb{R}^d)$ for the space of tree-like reduced bounded variations paths. At this time, under our hypothesis, signature transform is a continuous bijection. To be an homeomorphism, we need the continuity of the inverse application : $S^{-1} : G(\mathbb{R}^d) \rightarrow BV_{\text{red}}(\mathbb{R}^d)$.

Lemma 2.4. $S^{-1} : G(\mathbb{R}^d) \rightarrow BV(\mathbb{R}^d)$ is continuous.

We use that $G(\mathbb{R}^d)$ is a closed Lie group. Then, if $(Y_n)_{n \in \mathbb{N}}$ is a sequence in $G(\mathbb{R}^d)$, then $\lim Y_n = Y_\infty \in G(\mathbb{R}^d)$. We can deduce from Chow's Theorem [24] and Theorem 2.3.2 that there exists unique piece wise linear paths X_n and X_∞ such that

- $S(X_n) = Y_n$
- $S(X_\infty) = Y_\infty$

Now, we show that $X_n \rightarrow X$ in $BV(\mathbb{R}^d)$. As a matter of fact, we start from the converging sequence Y_n . As a closed Lie group $\subset T(\mathbb{R}^d)$, $G(\mathbb{R}^d)$ inherit its norm $\|\cdot\|_{T(\mathbb{R}^d)}$. Now, it is easy to prove that, (re)defining $\mathbf{1} = (1, 0, 0, 0, \dots)$ as the neutral element of $G(\mathbb{R}^d)$:

$$\begin{aligned} Y_n \rightarrow Y &\Leftrightarrow Y_n^{-1} \otimes Y \rightarrow \mathbf{1} \\ &\Leftrightarrow \|Y_n^{-1} \otimes Y - \mathbf{1}\|_{T(\mathbb{R}^d)} \rightarrow 0 \\ &\Leftrightarrow \|S(X_n)^{-1} \otimes S(X_\infty) - \mathbf{1}\|_{T(\mathbb{R}^d)} \rightarrow 0 \\ &\Leftrightarrow \|\overleftarrow{S(X_n \star X_\infty)} - \mathbf{1}\|_{T(\mathbb{R}^d)} \rightarrow 0 \end{aligned} \quad (2.3)$$

Using that $X_k, k \in \mathbb{N} \cup \{\infty\}$ are piecewise linear and [47] corollary 7.29, then we deduce that :

$$\|\overleftarrow{X_n \star X_\infty}\|_{1-\text{var}} \rightarrow 0 \Leftrightarrow \|X_\infty - X_n\|_{1-\text{var}} \rightarrow 0$$

And so, S^{-1} is continuous. \square

3. As a matter of fact, this is true if \tilde{X} contains any strictly increasing dimension.

Lemma 2.5. *$S(X)$ is differentiable for all $X \in BV(\mathbb{R}^d)$ and, for all $W \in BV(\mathbb{R}^d)$*

$$(d_X S)(W) = \bigoplus_{k=1}^{\infty} \sum_{i=1}^k \mathcal{I}_{i,X} k - 1(W)$$

For this proof, we adopt the following fact (see [86]) : as X is piecewise linear, it exists at least one base $\Psi = (\psi_1, \psi_2, \dots, \psi_m)$ such that

$$X = A\Psi = \left(\sum_{j=1}^m a_{1j} \Psi_j, \sum_{j=1}^m a_{2j} \Psi_j, \dots, \sum_{j=1}^m a_{dj} \Psi_j \right)$$

with $A \in \mathbb{R}^{d \times m}$. In such a configuration, Ψ is a m -dimensional path, and for a level k , using k -mode tensor product notations from [62]

$$S^{(k)}(X) = [\![S^{(k)}(\Psi); \underbrace{A, A, \dots, A}_{k \text{ times}}]\!]$$

In addition, note that $[\![C; A_1, A_2, \dots, A_k]\!]$ is k -linear but non-commutative. To avoid heavy notations, we use the following ones :

- $P_k(C, A) = [\![C; \underbrace{A, A, \dots, A}_{k \text{ times}}]\!]$. e.g. $P_3(C, A) = [\![C; A, A, A]\!]$
- $P_{k,i}(C, A, H) = [\![C; \underbrace{A, A, \dots, A}_{i-1 \text{ times}}, H, \underbrace{A, A, \dots, A}_{k-i+1 \text{ times}}]\!]$ e.g. $P_{4,3}(C, A, H) = [\![C; A, A, H, A]\!]$
- By convention, if $k = 0$, $P_k(C, A) = P_{k,i}(C, A, H) = 1$

From this fact, one can deduce that

$$S(X) = \bigoplus_{k=0}^{\infty} P_k(S^{(k)}(\Psi), A)$$

Consider $W \in BV(\mathbb{R}^d)$ piecewise linear such that $W = H\Psi$. Then, from last equation

$$\begin{aligned} (d_X S)(W) &= \bigoplus_{k=0}^{\infty} (d_X S^{(k)})(W) \\ &= \bigoplus_{k=0}^{\infty} d_A P_k(S^{(k)}(\Psi), .)(H) \end{aligned}$$

and then using the k -linearity

$$(d_X S)(W) = 0 \oplus \bigoplus_{k=1}^{\infty} \sum_{i=1}^k P_{k,i}(S^{(k)}(\Psi), A, H)$$

Using the facts that $S^{(k)}(X) = \int S^{(k-1)} dX$, we have

$$\begin{aligned} P_{k,i}(C, A, H) &= [\![C; A, \dots, A, H, A, \dots, A]\!] \\ &= \int \dots \int dX_{t_1} \dots dX_{t_{i-1}} dH_{t_i} dX_{t_{i+1}} \dots dX_{t_k} \\ &= \mathcal{I}_{i,X}^{k-1}(dW) \end{aligned}$$

where \mathcal{I} is the **insertion map** defined in [16, 42]. We also get a Lipschitz result for $d_X S$. First, we get :

$$\begin{aligned} \|\mathcal{I}_{i,X}^{k-1}(dW_1) - \mathcal{I}_{i,X}^{k-1}(dW_2)\| &= \|\mathcal{I}_{i,X}^{k-1}(d(W_1 - W_2))\| \\ &= \left\| \int \cdots \int dX_{t_1} \dots dX_{t_{i-1}} d(W_1 - W_2)_{t_i} dX_{t_{i+1}} \dots dX_{t_k} \right\| \\ &\leq \left\| \int \cdots \int dX_{t_1} \dots dX_{t_{i-1}} dX_{t_{i+1}} \dots dX_{t_k} \right\| \|W_1 - W_2\|_{1-\text{var}} \\ &\leq \|S^{(k-1)}(X)\| \|W_1 - W_2\|_{1-\text{var}} \end{aligned}$$

Taking the sum over i gives us

$$\begin{aligned} \|(d_X S^{(k)})(W_1) - (d_X S^{(k)})(W_2)\| &= \sum_{i=1}^k \|\mathcal{I}_{i,X}^k(dW_1) - \mathcal{I}_{i,X}^k(dW_2)\| \\ &\leq k \|S^{(k-1)}(X)\| \|W_1 - W_2\|_{1-\text{var}} \end{aligned}$$

Now, we use the dilation property : for any $\lambda \in \mathbb{R}$, $S^{(k)}(\lambda X) = \lambda^k S^{(k)}(X)$:

$$\begin{aligned} \|(d_X S^{(k)})(W_1) - (d_X S^{(k)})(W_2)\| &\leq k \left\| S^{(k-1)}\left(\frac{2}{2}X\right) \right\| \|W_1 - W_2\|_{1-\text{var}} \\ &\leq \frac{k}{2^{k-1}} \|S^{(k-1)}(2X)\| \|W_1 - W_2\|_{1-\text{var}} \\ &\leq \|S^{(k-1)}(2X)\| \|W_1 - W_2\|_{1-\text{var}} \end{aligned}$$

One of the last steps for $d_X S(W)$ use (2.2) :

$$\begin{aligned} \|(d_X S^{(k)})(W_1) - (d_X S^{(k)})(W_2)\| &\leq \|S^{(k-1)}(2X)\| \|W_1 - W_2\|_{1-\text{var}} \\ &\leq \frac{1}{(k-1)!} \|X\|_{1-\text{var}}^{k-1} \|W_1 - W_2\| \end{aligned}$$

Finally, we conclude on $(d_X S)(W)$:

$$\begin{aligned} \|(d_X S)(W_1) - (d_X S)(W_2)\| &= \left\| \bigoplus_{k=1}^{\infty} (d_X S^{(k)})(W_1) - \bigoplus_{k=1}^{\infty} (d_X S^{(k)})(W_2) \right\| \\ &= \left\| \bigoplus_{k=1}^{\infty} [(d_X S^{(k)})(W_1) - (d_X S^{(k)})(W_2)] \right\| \\ &\leq \sum_{k=1}^{\infty} \|S^{(k-1)}(2X)\| \|W_1 - W_2\|_{1-\text{var}} \\ &\leq \sum_{k=0}^{\infty} \frac{1}{k!} \|X\|_{1-\text{var}}^k \|W_1 - W_2\|_{1-\text{var}} \\ &\leq \exp(\|X\|_{1-\text{var}}) \|W_1 - W_2\|_{1-\text{var}} \end{aligned}$$

□

From this result, we deduce an analogous result for the special case of $G^N(\mathbb{R}^d)$, if N is at least the number of linear piece of $X \in BV(\mathbb{R}^d)$ (with the same dissection of $[0; T]$). Then $S^N(\mathbb{R}^d)$ is a bijection.

This results gives us an explicit formula for the deviation of $S^{(k)}(X)$ for a small variation of X in $BV(\mathbb{R}^d)$.

But more importantly, this results ensures that some topological features (like simplicial homology) is invariant by the signature transform.

2.6 Building simplicial complexes between time series

2.6.1 Presentation of the method

In the same spirit as for Gaussian graphical models [79], our goal is to infer a higher order model from data using a model selection based numerical procedure. One brute force method is to use sparse solutions of LASSO type estimators that can be employed to predict all sub-groups of times series based on all other groups of time series.

In the present paper, our goal is to propose a better structured solution to the problem of capturing interesting structures in the dependency relationship among the various components of high dimensional time series. Interesting types of structure often come from Topological Data Analysis (TDA) as presented in e.g. [18]. Inferring such structures is however extremely computationally extensive, and even more so if we account for the necessity of using Cross-Validation types of hyper parameter calibration procedures.

Our proposal is to adopt a principled sequential approach to simplicial complexes estimation. In our approach, the simplices that will be incorporated into the simplicial complex are chosen among completions of existing simplices into one order higher simplices, hence allowing to stratify the construction by the complexity of the interactions. The selection is performed using the LASSO, complemented by inspection of the R^2 criterion. One key ingredient of our construction is the use of the Signature transform as features for prediction. Signatures bring an essential information to the construction of our simplicial complex, namely orientation, since, Signatures encode the order in which the times series are integrated in ((2.1)). Using the orientation can be instrumental for the interpretation of the interactions among the various components of high dimensional time series, and noticing any difference in the prediction capabilities of two different orderings might be extremely useful in practice.

We now turn to the details of the implementation.

2.6.2 Our greedy order stratified algorithm :

Consider a group of d (augmented) time series $\mathcal{G} = \{X^1, X^2, \dots, X^d\}$ with $T + 1$ timesteps each of the form $X^i = [(0, X^i(0)), (t_1, X^i(t_1)), \dots, (t_T, X^i(t_T))]$.

In this section, we introduce our main contribution, namely the construction of a simplicial complex that encodes the multiway dependencies of the various dimensions

in a high-dimensional time series. We also discuss a greedy algorithm for sequentially building the sought for simplicial complex that mitigates the computational complexity. As mentioned earlier in the introduction, the main principle of the our algorithm is to build consistent simplices within groups of time series using regression or prediction error measures. In our implementation, we chose to present the *Signature prediction* version, consisting of predicting the Signature at a simplex as a linear function of the signatures at other simplices of various orders.

In order to keep control on the computational complexity of the method, we now present a lighter greedy algorithm. For any $t \in \llbracket 0, T - L \rrbracket$ with L to be specified.

Closure of a simplicial complex.

In our framework, we need to introduce the following definition.

Definition 2.6.1. Let \mathcal{C} denote a simplicial complex. We denote by $\overline{\mathcal{C}}$ the simplicial complex consisting in appending all k -simplices whose **single** incorporation results in creating a simplex of order $k + 1$ using the simplices already present in \mathcal{C} only.

The sequential algorithm.

We now define the steps of our sequential greedy method as follows.

Algorithm 1 Lasso explicability Simplicial Complex algorithm

Require: Set $\ell = 1$ and set $\mathcal{C}^{(1)} = \{1, \dots, d\}$.

Ensure: The time series interaction simplicial complex

while No more simplex is selected **do**

 Select an (augmented) k -subset of nodes $C^J = \{X^{j_1}, X^{j_2}, \dots, X^{j_k}\}$,
 with $J = \{j_1, \dots, j_k\}$ in $\mathcal{C}^{(\ell)}$, and compute $S_{[t,t+L]}(C^J)$.

for (augmented) k' -combination $C^{J'} = \{X^{j'_1}, X^{j'_2}, \dots, X^{j'_{k'}}\}$ with $J' = \{j'_1, \dots, j'_{k'}\}$
 in $\overline{\mathcal{C}^{(\ell)}}$ **do**

 Compute $S_{[t,t+L]}(C^{J'})$

end for

 Predict $S_{[t,t+L]}(C^J)$ from $(S_{[t,t+L]}(C^{J'}))_{J' \cap J = \emptyset}$ with LASSO

 Compute $S^w(\tilde{X})$

if $R^2 > 0, 67$ **then**

 Select all non-zero β_j LASSO coefficients

else

 Set $\beta_j = 0$ for all j .

end if

end while

Each non-zero β_j coefficient represents a k -simplex whose vertices are

$$\{X_i, X_{j_1}, X_{j_2}, \dots, X_{j_{k+1}}\}.$$

This simplex comes with a natural weight w . As this k -simplex can be produced multiple time (by predicting X_i with $\{X_{j_1}, \dots, X_{j_{k+1}}\}$ or X_{j_l} by $\{X_i, X_{j_1}, \dots, X_{j_{k+1}}\}_{k \neq l}$, this weight is computed as the sum of all the non-zero LASSO coefficients obtained for each prediction.

By iterating the procedure for every possible simplices whose signature is a statistically interesting quantity to predict, and every dimension of simplex $k \leq K$ (for a chosen k), one can produce a simplicial complex among \mathcal{G} .

Remarks

Let us now address some technical question that arise from the proposed construction.

Time dependancy : The algorithm is applied to evolving time series for which the computation of the Signatures is updated incrementally and prediction is performed using these updated Signatures as time increases.

Orientation : This algorithm gives a natural orientation on every simplex, as $S(X_i, X_j) \neq S(X_j, X_i)$ which is often of great potential use for interpretability.

Extension to simplicial complex over multidimensional functional data.

The upper version of this algorithm can be extended to work with multiple multidimensional time series.

- In the simplest configuration, each node represents a d -dimensional trajectory $X_i = (X_{i1}, \dots, X_{id})$, and (X_1, \dots, X_n) is a sample from X . In this case, the algorithm does not change, except for the dimension of each node.

As an example, one can imagine geographical interactions between multidimensional meteorological signals such as (Temp, Pressure, Hygrometry, ...) in each town. This algorithm allows us to discover a correlation structure between cities.

- In another configuration, each node has its own dimension. Our dataset is then :

$$X = (X_1, \dots, X_n)$$

with $X_i \in BV([0, T], \mathbb{R}^{d_i})$. In addition, each X_i does not necessarily describe the same phenomenon.

A clear example comes from medicine : One might want to measure the interactions between different areas of the body, whose measurements are provided by different tests. EEG can have up to 256 channels, ECG up to 12, and EMG can have 2 channels for each area (and can be applied to multiple areas).

This design is a bit tricky, as each node's signature has its own dimension. To overcome this difficulty, we apply path augmentation to all signals except for the one with the highest dimension.

In each case, signature bring a natural feature space for constructing a correlation structure. Even more than in the very first case, signature of each node compress in a unique vector geometrical interaction and statistical feature of each channel.

2.6.3 Known results about the LASSO and the construction of simplicial complexes

In this section, we adapt [12] for the signatures. First, recall that there exists a canonical inner product for $T(\mathbb{R}^d)$: first, we have a canonical inner product $\langle x, y \rangle_{(\mathbb{R}^d)^{\otimes k}}$ on $(\mathbb{R}^d)^{\otimes k}$

$$\langle x, y \rangle_{(\mathbb{R}^d)^{\otimes k}} = \sum_{1 \leq i_1, \dots, i_k \leq d} x_{i_1, \dots, i_k} y_{i_1, \dots, i_k}.$$

Using that every $\mathbf{x} \in T(\mathbb{R}^d)$ can be decomposed as $\mathbf{x} = \bigoplus_{k \geq 0} x^{(k)}$ with $x^{(k)} \in (\mathbb{R}^d)^{\otimes k}$ the projection of \mathbf{x} over $(\mathbb{R}^d)^{\otimes k}$, we get

$$\langle \mathbf{x}, \mathbf{y} \rangle_{T(\mathbb{R}^d)} = \sum_{k \geq 0} \langle x^{(k)}, y^{(k)} \rangle_{(\mathbb{R}^d)^{\otimes k}}$$

Our complex is constructed sequentially dimension after dimension. In the following, we fix a step (and hence, a dimension) k . The objective is to build the k -dimensional Link of each node $X^j, 1 \leq j \leq d$ (see Definition 2.2.6). In order to proceed, we estimate $\beta = (\beta_1, \dots, \beta_p), p = \binom{k-1}{d}$ such that :

$$S(\tilde{X}^j) = \sum_{I_i = \{i_1, \dots, i_{k-1}\} \subset \{1, \dots, j-1, j+1, \dots, d\}^{k-1}} S(X^{I_i}) \beta_i \quad (2.4)$$

where $X^I = (X^{i_1}, \dots, X^{i_{k-1}})$. Going back to the usual framework for the study of the LASSO, our goal is to find $\beta = (\beta_1, \dots, \beta_p)$ such that :

$$S(X^j) = A\beta \quad (2.5)$$

with A being a matrix such that each column is defined by $S(X^I)$ for $I \subset \{1, \dots, j-1, j+1, \dots, d\}^{k-1}$.

Definition 2.6.2 (Coherence property). *For a matrix $A \in \mathbb{R}^{n \times p}$, define the **coherence** as $\mu(A) = \sup_{1 \leq i, j \leq p} |\langle A_i, A_j \rangle|$. The matrix A is said to satisfy the **coherence property** if*

$$\mu(A) \leq C_0(\log(p))^{-1},$$

where $C_0 \in \mathbb{R}_+$ is an absolute constant.

Under the assumption that each column of A is centered, the coherence $\mu(A)$ measures the highest correlation between any two columns of A , and therefore, any matrix A satisfying the coherence property has the property that it contains no highly correlated pair of columns.

Recall that our goal is to recover $\text{Lk}_k(\sigma)$, where σ is a pre-existing simplex of \mathcal{C} . We use the notation of $\hat{\text{Lk}}_k(\sigma)$ for its estimation. We want to maximize $\mathbb{P}[\hat{\text{Lk}}_k(\sigma) = \text{Lk}_k(\sigma)]$. A simplex σ_k is called **eligible** if $\sigma_k \cap \sigma = \emptyset$. For all σ_k eligible, we note X^{σ_k} the path of vertices from σ_k , and $\text{supp}(\beta) = \{i, \beta_i \neq 0\}$. Then,

$$\begin{aligned} \hat{\text{Lk}}_k(\sigma) &= \{\sigma_k \text{ eligible}, \hat{\beta}_{\sigma_k} \neq 0 \text{ in (2.4)}\} \\ &\simeq \text{supp}(\hat{\beta}), \quad \beta \text{ solution estimation of (2.4)} \end{aligned}$$

In the same way, one can deduce $\text{Lk}_k(\sigma) \simeq \text{supp}(\beta)$, β solution of (2.4). Then,

$$\mathbb{P}[\hat{\text{Lk}}_k(\sigma) = \text{Lk}_k(\sigma)] = \mathbb{P}[\text{supp}(\hat{\beta}) = \text{supp}(\beta)]$$

Thus, using [12, Theorem 1.3], we obtain



Theorem 2.1

Let $\text{Lk}_k(X^i) = \text{supp}(\beta)$ be the k dimensional link of X^i . Suppose that

1. A defined from (2.5) obeys the coherence property.
2. $ne_k(X^i)$ is taken from the generic $|\text{supp}(\beta)|$ -sparse model.
3. $N \leq \frac{c_0 p}{\|A\|^2 \log(p)}$ for some positive constant c_0 .
4. $\min_{i \in \text{supp}(\beta)} |\beta_i| > 8\sigma \sqrt{2 \log(p)}$

then

$$\mathbb{P}[\text{supp}(\beta) = \text{supp}(\hat{\beta})] \geq 1 - \frac{2}{p} \left((2\pi \log(p))^{-1/2} + \frac{|\text{supp}(\beta)|}{p} \right) - O(p^{-2\log(2)})$$

2.7 Numerical experiments and Applications

Multivariate times series are omnipresent and high order correlation occurred frequently in e.g. the domains of finance and neuroscience. We evaluated our method on two public data sets analysed by [93] : the fMRI resting-state data from the HCP <https://www.humanconnectome.org/>.

2.7.1 Practical choices and hyper parameters

We consider only simplices up to triangles $k \leq 2$ for now. Due to the concerns about complexity, the depth of signature is set to $depth = 2$ for the construction of both 1-simplex and 2-simplex. More precisely, we use time-augmented path to calculate 2-truncated signatures of 0-simplices, in order to construct 1-simplices. As a design choice, for 2-simplices the predictors (1-simplices composed of two times series) are not augmented when applying LASSO regression. Clearly, many different choices could be imposed on how we model the dependencies between subgroups of time series.

The regularisation term of LASSO is crucial for our method since it directly controls the sparsity in prediction using the linear models on signature features. Recall that the selected groups of time series will immediately be translated into new simplices in our sequentially growing simplicial complex.

In the present numerical experiments, we show that coarsely selected values for these hyperparameters already provide interesting results on a real dataset. In practice, $\lambda_{1\text{-simplex}}$ and $\lambda_{2\text{-simplex}}$ have been empirically tuned to the values $\lambda_{1\text{-simplex}} = 1000$ and $\lambda_{2\text{-simplex}} = 3$ for the fMRI dataset. More experiments based on extensive comparisons over a refined grid will be tested in an extended version of the present paper. The latest version of our implementation is available on our GitHub page :

<https://github.com/remivaucher/thesis - Chapter 2>

2.7.2 Modelling interactions in Functional MRI datasets

Functional connectivity is a neuroscience approach aimed at understanding the organization of the human brain based not solely on spatial proximity and structural factors, but rather on its functionality, i.e. its connectivity patterns between different brain regions and networks. For instance, even seemingly routine tasks such as paying attention during a lecture have been found to activate regions like the pulvinar (within the thalamus), the superior colliculus (in the midbrain), and the posterior parietal cortex [87]. In this perspective, and given that functional brain imaging data can be regarded as time series, the theory of Signatures could prove to be particularly useful.

In the absence of specific tasks or external stimulation (resting, meditating, sleeping, etc.), the brain enters what is known as resting-state. The Default Mode Network (DMN) becomes prominently active during this resting state. This neural network includes key regions such as the medial prefrontal cortex (mPFC), the posterior parietal lobe (PTL), the posterior cingulate cortex (PCC), and the precuneus [9].

We tested our method on resting-state fMRI(rs-fMRI) data⁴ preprocessed by the same pipeline in [93]. The dataset contains 100 cortical (Schaefer100 [94]) and 19 subcortical ROIs (Regions of Interest). In order to evaluate the quality of identified interactions, we have selected a subset of 15 ROIs of which functional connectivities during resting-state are well known. We constructed simplicial complex on all 1200 timesteps, and analysed the top 10 1-simplices and 2-simplices that are the most persistent, i.e. that occurred on most time-steps. Besides, we observed that the life duration distribution of 1-simplices is centred and symmetric, whereas the distribution of 2-simplices is positively skewed.

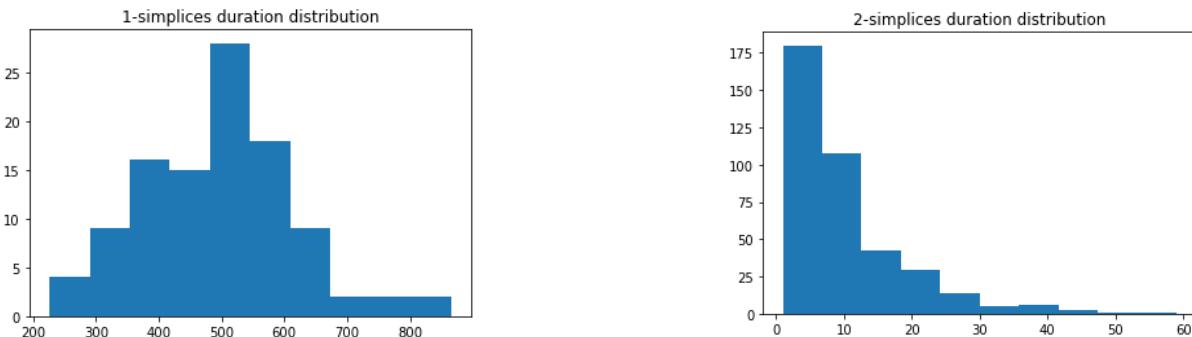


FIGURE 2.1 – Histograms of the observed duration for the discovered 1 and 2-simplices

The 1-simplex representing the interaction between *7Networks LH Vis 9* and *7Networks LH SomMot 4* occurred on most occasions (865 time steps). The most persistent 2-simplex (59 time steps) represents the interaction among *LH Cont Par 1*, *RH Default PFCdPFCm 2* and *RH Default pCunPCC 2*.

Most of the persistent 1-simplices involve the prefrontal cortex, the parietal lobe and the precuneus, which is consistent with literature as all three regions are active during

4. HCP, <http://www.humanconnectome.org/>

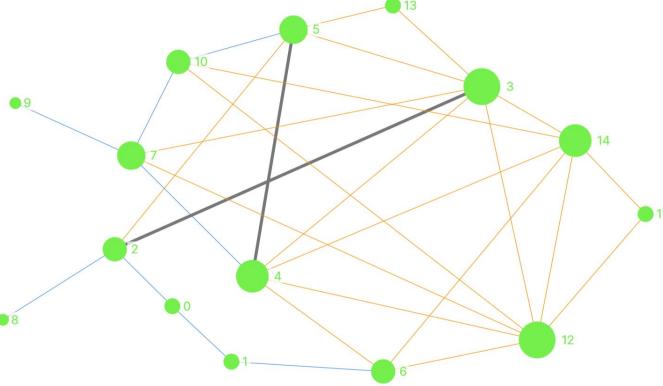


FIGURE 2.2 – Simplicial complex constructed with persistent simplicies. 1-simplices are blue, 2-simplices are defined by their orange 1-simplex faces. The gray signifies the coincidence of a 1-simplex and one face of a 2-simplex. The selected ROI and numbered 0-simplices are matched by the following dictionary : 0 - *LH Vis* 9, 1 - *LH SomMot* 4, 2 - *LH DorsAttn Post* 4, 3 - *Cont Par* 1, 4 - *LH Cont pCun* 1, 5 - *LH Default pCunPCC* 1, 6 - *LH Default pCunPCC* 2, 7 - *RH Cont Par* 1, 8 - *RH Cont PFCl* 1, 9 - *RH Cont pCun* 1, 10 - *RH Default Par* 1, 11 - *RH Default PFCdPFCm* 1, 12 - *RH Default PFCdPFCm* 2, 13 - *RH Default PFCdPFCm* 3, 14 - *RH Default pCunPCC* 2.

resting-state [9]. The most persistent interaction include subregions of the left hemisphere's visual and somatomotor networks. Component *LH SomMot* 4 has previously been associated with components of the left (*LH Default PFC*) and right (*RH Default PFCv* 2) PFC in a study proposing an age prediction pipeline Ayu using rs-fMRI data [7]. In the same study, several visual areas (*RH Vis* 1, 3 and 4) are linked to the PFC areas during rs-fMRI (*LH Default PFC* 1, 2, and 3), although they are located in the right hemisphere.

The top 10 interactions that occurred the most include components from the same three recurrent brain areas that are the PFC, parietal regions and the precuneus, with the latter taking part in all 10 of them. This aligns with previous work [9] as all three are indeed involved in the Default Mode Network which is active during rs-fMRI.

The most persistent simplicies and matching ROIs are represented in Figure 2.3. In particular, the interactions discovered using our approach show excellent coherence with well identified spatial activity zones.

2.8 Discussion and future work

The qualification of high-order interaction of time series is a relatively new research area. Previous work, such as [93], tried to estimate the higher order interactions in high dimensional signals. Our method, based on the theory of Signatures that captures higher dimensional interactions together with what can be encoded as a simplex orientations, is able to leverage much more refined information about the mutual behaviour of the observed phenomena.

From a theoretical point of view, although the orientation of the various simplices discovered in the sequential construction was not used properly, it could be fruitfully exploited

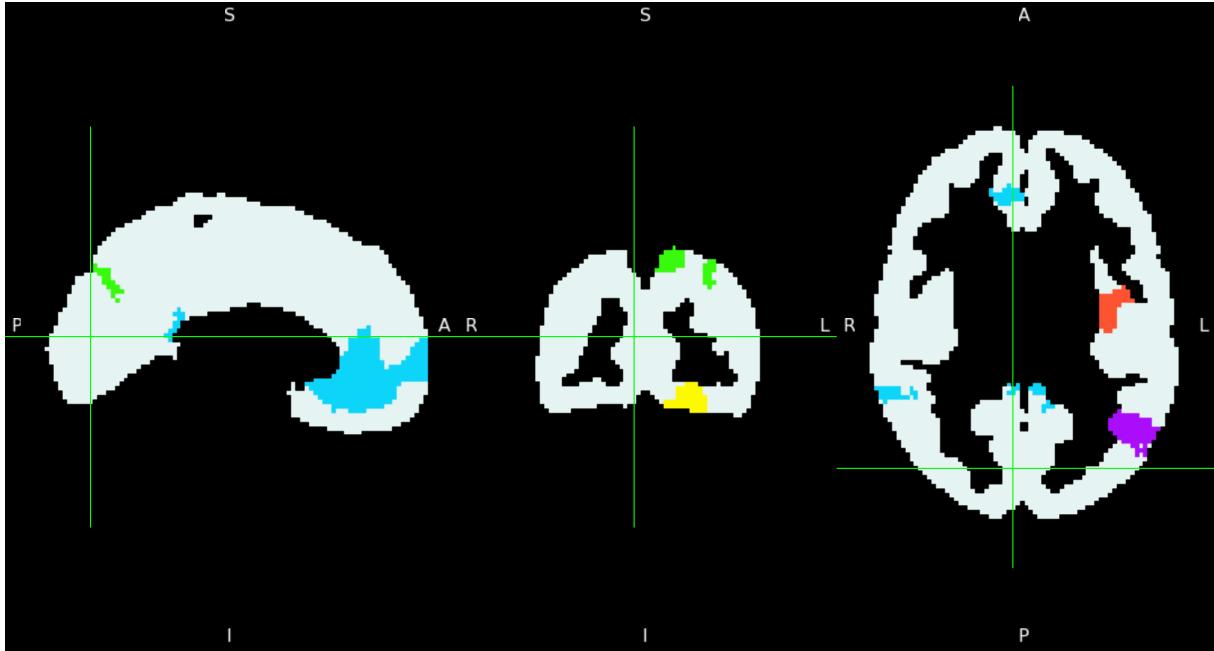


FIGURE 2.3 – Anatomical representation of the 10 most persistent 1-simplices and 2-simplices matched to their corresponding network in the 7-network parcellation by [104]. Only parcels that are part of the most persistent simplices are colored. Each color corresponds to a network as per the following color legend : yellow - *Visual*, red - *Somatomotor*, purple - *Dorsal Attention*, green - *Control*, blue - *Default*.

in the future. Secondly, the problem of dimension consistency could be appropriately tackled using group LASSO types of techniques or SLOPE-based approaches.

To conclude, we mention that the method we just presented could also be handily put to work on nonstationary problems for e.g. change point detection such as in early detection of epidemics, using human in the loop validation steps. The simplicial complex could be sequentially updated as a function of time as well, leading to a dynamical topological structure whose characteristics and abrupt potential changes could help extract valuable information about the emergence of certain interesting phenomena.

Chapitre 3

Sélection de variable en régression entière : Algorithme de Lagarias Odlyzko étendu.

*Jake : Feels like that deserved an
audible gasp !*
*Charles : **Audible Gasp***

Brooklyn 99

3.1 Introduction

This paper addresses the problem of estimating the integer regression vector $\beta^* \in \mathbb{Z}^p$ in the linear model

$$Y = X\beta^* + \epsilon \quad (3.1)$$

with $Y \in \mathbb{R}^n$ is the noisy observation vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, and $\epsilon \in \mathbb{R}^n$ is the noise vector. The rows of the design matrix X are called the features. In the case where the number of features is larger than the number of observations, the problem is usually categorized as belonging to the field of high-dimensional statistics [10, 102], also called the $p \gg n$ regime. This problem is of important interest in statistics and signal processing. Our interest in this problem stems from the possibility of selecting the support of a regression vector based on noisy observations using hard constraints implicitly enforced in a penalisation framework as in Bayesian Information Criterion (BIC)-based approaches [57], [10], [90], or explicitly expressed in the estimation problem [31], with the additional constraint of achieving the computation of the estimator in reasonable computational time. One other direction for solving this problem is the one adopted in the line of research that started with the LASSO estimator [101]. In particular, the LASSO has long been considered an efficient approach to the problem of reconstructing the support and sign of the true vector in linear models in the case where a sparsity assumption on β^* is realistic and the features are sufficiently orthogonal [12]. A version of this type of result for the case where the variance of the observation noise is not known is proposed in, e.g., [25]. On the other hand, the LASSO estimator is not known to work

correctly at recovering a vector with the same support as β^* for n smaller than a certain threshold, namely when n is smaller than the support size $\|\beta^*\|_0$ multiplied by $\log(p)$.

Recently, new results were obtained in [48, 49] for the case where the matrix X and the vector β^* have integer coordinates, or more generally rational coordinates with the same integer denominator. The algorithm proposed is inspired by the Lagarias and Odlyzko Algorithm [45, 64], an algorithm based on the celebrated Lenstra-Lenstra-Lovasz algorithm for finding the shortest vector in a lattice, and which solves a version of the so-called Subset-Sum problem :

Given $p \in \mathbb{Z}_{>0}$ and $y, x_1, \dots, x_p \in \mathbb{Z}_{>0}$, find $S \subset \{1, \dots, p\}$, $S \neq \emptyset$ such that

$$y = \sum_{j \in S} x_j,$$

when at least one such set S exists.

It is proved in [49] that the Extended Lagarias-Odlyzko Algorithm solves the noiseless high dimensional linear regression problem for binary β^* and X generated by i.i.d. elements chosen uniformly from $[2^{\frac{1}{2}(1+\epsilon)p^2}]$ with only one observation, with high probability when $p \rightarrow \infty$.

Our goal in the present work is to extend such results to more general setups, including setting where the columns of X are not independent, for integer and rational possible vectors β^* .

3.2 Background on high dimensional regression and variable selection

Linear regression often considers models of the type (3.1) where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$ and $\epsilon = [\epsilon_1, \dots, \epsilon_n]^t$ is sub-Gaussian with variance proxy σ^2 and $\mathbb{E}[\epsilon] = 0$.

3.2.1 High-Dimensional Regression and support recovery for the LASSO estimator

Selection based on the BIC criterion

Definition 3.2.1

Fix $\tau > 0$ and assume the linear regression model $Y = X\beta^* + \epsilon$. The **BIC estimator** of β^* is defined by any $\hat{\beta}^{BIC}$ such that

$$\hat{\beta}^{BIC} \in \arg \min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|Y - X\beta\|_2^2 + \tau^2 \|\beta\|_0 \right\}$$



Theorem 3.1

Assume that the linear regression model $Y = X\beta^* + \epsilon$ holds, where $W \sim \text{sub}G_n(\sigma^2)$ and that $\|\beta^*\|_0 \geq 1$. Then the BIC estimator $\hat{\beta}^{BIC}$ with regularization parameter

$$\tau^2 = 16 \log(6) \frac{\sigma^2}{n} + 32 \frac{\sigma^2 \log(ed)}{n}$$

satisfies

$$MSE(X\hat{\beta}^{BIC}) = \frac{1}{n} \|X\hat{\beta}^{BIC} - X\beta^*\|_2^2 \leq \|\beta^*\|_0 \sigma^2 \frac{\log(\frac{ed}{\delta})}{n}$$

with probability at least $1 - \delta$.

The Least Absolute Shrinkage and Selection Operator (LASSO)



Definition 3.2.2

Fix $\tau > 0$ and assume the linear regression model $Y = X\beta^* + \epsilon$. The **LASSO estimator** of β^* is defined by any $\hat{\beta}^L$ such that

$$\hat{\beta}^L \in \arg \min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \tau \|\beta\|_1 \right\}$$

The following result is not about support recovery but is important to remember when doubts arise about the performance of the LASSO when the columns are not sufficiently non-collinear.



Theorem 3.2 Slow rate

Assume that the linear regression model $Y = X\beta^* + W$ holds, where $W \sim \text{sub}G_n(\sigma^2)$. Moreover, assume that the columns of X are normalized in such a way that $\max_j \|X_j\|_2 \leq \sqrt{n}$. Then the LASSO estimator $\hat{\beta}^L$ with regularization parameter

$$2\tau = 2\sigma \sqrt{\frac{2 \log(2d)}{n}} + 2\sigma \sqrt{\frac{2 \log(\frac{1}{\delta})}{n}}$$

satisfies

$$MSE(X\hat{\beta}^L) = \frac{1}{n} \|X\hat{\beta}^L - X\beta^*\|_2^2 \leq 4\|\beta^*\|_1 \sigma \sqrt{\frac{2 \log(2d)}{n}} + 4\|\beta^*\|_1 \sigma \sqrt{\frac{2 \log(\frac{1}{\delta})}{n}}$$

with probability at least $1 - \delta$.

Definition 3.2.3 *Incoherence : INC(k)*

We say that the design matrix X has **incoherence** k for some integer $k > 0$ if

$$\left\| \frac{X^t X}{n} - I_d \right\|_\infty \leq \frac{c}{k}$$

for some universal constant c , where the $\|A\|_\infty$ denotes the largest element of A in absolute value.

The following theorem proves that incoherence helps in prediction.

Theorem 3.3 *Fast rate*

Fix $n \geq 2$. Assume that the linear regression model $Y = X\beta^* + \epsilon$ holds, where $\epsilon \sim \text{sub}G_n(\sigma^2)$ Moreover, assume that $\|\beta^*\|_0 \leq k$ and that X satisfies assumption INC(k). Then the LASSO estimator $\hat{\beta}^L$ with regularization parameter defined by

$$2\tau = 8\sigma\sqrt{\frac{2\log(2d)}{n}} + 8\sigma\sqrt{\frac{2\log(\frac{1}{\delta})}{n}}$$

satisfies

$$MSE(X\hat{\beta}^L) = \frac{1}{n}\|X\hat{\beta}^L - X\beta^*\|_2^2 \leq \sigma^2 k \frac{\log(\frac{2d}{\delta})}{n}$$

and

$$\|\hat{\beta}^L - \beta^*\|_2^2 \leq \sigma^2 k \frac{\log(\frac{2d}{\delta})}{n}$$

with probability at least $1 - \delta$.

The next Theorem proves that exact support and sign pattern recovery can be achieved under more stringent incoherence assumptions.


Theorem 3.4 *Support recovery*

Let I be the support of β with cardinality

$$k \leq c_0 p / (\|X\|^2 \log p)$$

for some positive numerical constant c_0 , and suppose that the signal amplitude is above the noise level, i.e.

$$\min_{i \in I} |\beta_i^*| > 8\sigma \sqrt{2 \log p}$$

Then, under the assumptions that the matrix X has all its columns ℓ_2 normalised and satisfies the incoherence property $INC(\log(p))$, the lasso estimate with $\tau = 2\sqrt{2 \log p}$ obeys

$$\text{support}(\hat{\beta}) = \text{support}(\beta^*) \quad \text{and} \quad \text{sign}(\hat{\beta}_i) = \text{sign}(\beta_i^*) \quad \text{for all } i \in I, \quad (3.2)$$

with probability at least $1 - 2p^{-1} \left((2\pi \log p)^{-1/2} + |I|p^{-1} \right) - O(p^{-2 \log 2})$.

3.3 Estimating integer relations in the regression vector : LLL and friends

We will use the following notations. The real vector space $\mathbb{R}^n, n \geq 1$ is endowed with its Euclidean structure and the Lebesgue measure, denoted by μ , on the σ -algebra of Lebesgue measurable sets. The canonical basis of \mathbb{R}^n is denoted by (e_1, e_2, \dots, e_n) . The scalar product of $v, u \in \mathbb{R}^n$, and the Euclidean norm of u are respectively denoted by $v \cdot u$, and $\|u\| = (u \cdot u)^{1/2}$. To a subset $E \subseteq \mathbb{R}^n$, we can associate the real vector space generated by E , which will be denoted by $\langle E \rangle$.

The set of matrices with n rows and m columns, with coefficients in a set \mathbb{S} (in practice \mathbb{R} , \mathbb{Q} or \mathbb{Z}) is denoted $\mathbb{S}^{n \times m}$. The transpose of a given matrix M is denoted by M^\top , its determinant is denoted by $\det M$ and its inverse matrix (when it exists) by M^{-1} .

The Gram-Schmidt orthogonalisation plays a central role for the problem of finding a short vector in a lattice, which is our main focus in the present paper. In the following section, we summarise the main concepts and results about the Gram-Schmidt process.

3.3.1 Gram-Schmidt orthogonalisation

To a system $B = (b_1, \dots, b_p)$ of p vectors of \mathbb{R}^n , we associate the matrix whose rows are the vectors b_i expressed in the canonical basis (e_1, e_2, \dots, e_n) of \mathbb{R}^n . This matrix will be called the row matrix of (b_1, \dots, b_p) and will also be, with a slight abuse of language, designated by B . The sets $\llbracket a, b \rrbracket$ defined by $\llbracket a, b \rrbracket := [a, b] \cap \mathbb{Z}$ will be called integer intervals. The open (resp. closed) ball of radius ρ centered at a is denoted by $B(a, \rho)$ (resp. $\bar{B}(a, \rho)$) and is as usual defined by

$$B(a, \rho) = \{x \in \mathbb{R}^n \mid \|x - a\| < \rho\}, \quad \bar{B}(a, \rho) = \{x \in \mathbb{R}^n \mid \|x - a\| \leq \rho\}.$$

For the sake of completeness, we now give the definition of the well known Gram-Schmidt orthogonalisation.



Definition 3.3.4 Gram-Schmidt orthogonalisation.

Let $B = (b_1, \dots, b_p)$ be a family of linearly independent vectors of \mathbb{R}^n . We denote by B_i the starting family, $B_i := (b_1, \dots, b_i)$ and by H_i the \mathbb{R} -vector space generated by B_i . The **Gram-Schmidt orthogonalized family** is the orthogonal family $B^* = (b_1^*, \dots, b_p^*)$ formed by the vectors b_i^* , where b_i^* is the orthogonal projection of b_i onto the orthogonal of H_{i-1} . More precisely

$$\begin{cases} b_1^* = b_1 \\ b_i^* = b_i - \sum_{k=1}^{i-1} m_{i,k} b_k^*, \quad \text{with} \quad m_{i,j} = \frac{b_i \cdot b_j^*}{\|b_j^*\|^2} \quad \text{for} \quad 1 \leq j < i \leq p \end{cases}$$

We also pose $m_{i,i} = 1$ for $1 \leq i \leq p$ and $m_{i,j} = 0$ for $1 \leq i < j \leq p$. This Gram-Schmidt orthogonalization procedure also constructs the matrix $\mathcal{P} \in \mathbb{R}^{p \times p}$ whose entry $m_{i,j}$ is defined above.

For simplicity, we will denote by B the matrix $\mathbb{R}^{p \times n}$ whose i^{th} row is the vector b_i . If B^* is the matrix $\mathbb{R}^{p \times n}$ whose i^{th} row is the vector b_i^* , then we have $B = \mathcal{P}B^*$.



Definition 3.3.5

Let $B = (b_1, \dots, b_p)$ and let $B^* = (b_1^*, \dots, b_p^*)$ be the family obtained after the Gram-Schmidt orthogonalisation process. The ℓ_2 -norm of b_i^* , denoted by ℓ_i , is called the i^{th} Siegel length.



Definition 3.3.6 Gram matrix.

The Gram matrix of a system $B = (b_1, \dots, b_p)$ of p vectors of \mathbb{R}^n , denoted by $G(b_1, \dots, b_p)$ or $G(B)$, is the matrix of $\mathbb{R}^{p \times p}$ defined by

$$G_{ij} = b_i \cdot b_j \quad \forall i, j \in \llbracket 1, p \rrbracket.$$

If the rows of the matrix B are the vectors of the system (b_1, \dots, b_p) , then the Gram matrix is written $G(B) = B \cdot B^\top$.

Then, we have

$$\det G(B) = \prod_{i=1}^p \|b_i^*\|^2 = \prod_{i=1}^p \ell_i^2.$$

We now turn to the heart of the matter, namely the concepts of lattice and basis reduction.

3.3.2 Lattices

A **Euclidean lattice** of \mathbb{R}^n is the set of linear combinations with integer coefficients of a family $\{b_1, \dots, b_p\}$ of p linearly independent vectors of \mathbb{R}^n , called the lattice basis.

The following result is a basic rephrasing of the notion of lattice in terms of groups.

Proposition 3.3.1

The following assertions are equivalent :

- (i) \mathcal{L} is a discrete additive subgroup of \mathbb{R}^n , which generates a vector subspace of dimension p .
- (ii) There exists a system $B = \{b_1, \dots, b_p\}$ of p linearly independent vectors of \mathbb{R}^n , for which

$$\mathcal{L} = \left\{ \sum_{i=1}^d x_i b_i \mid x_i \in \mathbb{Z} \quad \forall i \in \llbracket 1, p \rrbracket \right\}.$$

The next result is fundamental for justifying the approach adopted in the LLL algorithm to use orthogonalised bases.

Proposition 3.3.2

Let \mathcal{L} be a lattice generated by a basis $B = (b_1, \dots, b_p)$. We denote by B^* the orthogonalized basis of B . Then, for all $w \in \mathcal{L} \setminus \{0\}$ we have

$$\|w\| \geq \min \{\|b_i^*\| ; i \in \llbracket 1, p \rrbracket\}.$$

We now define key quantities about lattices that relate to the length of the vectors in the lattice.

Definition 3.3.7

The first minimum of the network \mathcal{L} , denoted by $\lambda_1(\mathcal{L})$, is the norm of a shortest nonzero vector of \mathcal{L} . More generally, the i -minimum of the \mathcal{L} lattice, denoted by $\lambda_i(\mathcal{L})$, is the smallest positive real number ρ for which the closed ball of radius ρ centered at the origin contains at least i vectors linearly of the \mathcal{L} lattice,

$$\lambda_i(\mathcal{L}) := \min\{\rho > 0 \mid \dim(\bar{B}(0, \rho) \cap \mathcal{L}) \geq i\}.$$

The following theorem by Minkowski is of great relevance, since a condition is given on the minimum size of a set to contain at least one point in the network.

Theorem 3.5 *Minkowski*.

Consider a Euclidean lattice \mathcal{L} of dimension p . We denote by μ the p -dimensional Lebesgue measure, and consider a subset C of the vector subspace generated by \mathcal{L} , μ -measurable, which is convex, symmetric about the origin, and verifies $\mu(C) > 2^p \det \mathcal{L}$. Then C contains at least one point of \mathcal{L} .

Basis reduction is key to the construction of the LLL algorithm.



Definition 3.3.8 Informal definition of the notion of reduced basis.

A basis $B = (b_1, \dots, b_p)$ formed of p vectors of \mathbb{R}^n is a reduced basis if it is formed of short enough and orthogonal enough vectors. These criteria are measured quantitatively by a markup of the orthogonality defect $\rho(B)$ and the length defects $\theta_i(B)$ defined respectively by

$$\rho(B) = \frac{1}{\det \mathcal{L}(B)} \prod_{i=1}^p \|b_i\| = \prod_{i=1}^p \frac{\|b_i\|}{\|b_i^*\|}, \quad \theta_i(B) = \frac{\|b_i\|}{\lambda_i(\mathcal{L}(B))}$$

The orthogonality defect $\rho(B)$ is at least 1, with equality only when the basis B is orthogonal. As a lattice does not have an orthogonal basis, we have in general $\rho(B) > 1$. The basis is "fairly" orthogonal when its orthogonality defect $\rho(B)$ is increased. The length defects are also at least equal to 1. The equalities $\theta_i(B) = 1$ can only occur simultaneously when the basis is minimal, i.e., formed by vectors realizing successive minima. The existence of a minimal basis is not always guaranteed, as soon as the dimension p verifies $p \geq 5$.



Definition 3.3.9 Reduction.

Given a basis B , reducing B consists in finding an equivalent reduced basis.

A good notion of reduction must establish a compromise between the quality of the reduced basis and the complexity of the reduction algorithm. Such a compromise is achieved by the LLL algorithm, invented by Lenstra, Lenstra and Lovász in 1982, presented in the next subsection.

3.3.3 Lovász reduced bases and the LLL algorithm

A basis (a_1, \dots, a_n) of a lattice L of rank n in \mathbb{R}^p is called Lovász-reduced if the lengths l_i of the vectors \hat{a}_i , $i = 1, \dots, n$ and the entries $m_{i,j}$ of the matrix \mathcal{P} defined in definition 5, satisfy the following two conditions :

— Properness :

$$|m_{i,j}| = \frac{|a_i \cdot \hat{a}_j^*|}{\|\hat{a}_j^*\|^2} \leq \frac{1}{2} \text{ for } 1 \leq j < i \leq n.$$

— Lovász condition :

$$l_i^2 \leq s^2 (l_{i+1}^2 + m_{i+1,i}^2 l_i^2) \text{ for } 1 \leq i \leq n-1, \quad (3.3)$$

where s is a real parameter in the interval $]1, 2[$, that is usually chosen to be equal to $t > 2/\sqrt{3}$. The LLL algorithm is defined below in Algorithm 2 and makes use of the integral part of the Gram-Schmidt coefficients.

The main result about the orthogonality defect of the LLL algorithm is the following theorem.

Algorithm 2 The LLL algorithm

Require: A lattice L given by a basis of vectors b_1, \dots, b_p .

Ensure: A Lovász-reduced basis b of the lattice \mathcal{L} .

Compute the system \hat{b} and the matrix \mathcal{P}

Set $i := 1$;

while $i < n$ **do**

$b_{i+1} := b_{i+1} - \lceil m_{i+1,i} \rceil b_i$

if Lovász condition (3.3) **is true then**

set $i := i + 1$;

else if Lovász condition (3.3) **is false then**

swap b_i and b_{i+1}

update \hat{b} and m

if $i \neq 1$ then $i := i - 1$

end if

end while



Theorem 3.6

Consider a real $s > 2/\sqrt{3}$. The LLL algorithm constructs from a basis $B := (b_1, \dots, b_p)$ whose size $\tau(B)$ is defined as

$$\tau(B) = \Theta(pn) \cdot \log M \quad \text{with} \quad M = \max \{B_{i,j}; \quad i \in \llbracket 1, p \rrbracket, j \in \llbracket 1, n \rrbracket\},$$

a basis \hat{B} with the following characteristics :

- (i) the \hat{B} basis is obtained in polynomial time in the size $\tau(B)$ of the B matrix,
- (ii) the orthogonality defect $\rho(\hat{B})$ and the length defects $\theta_i(\hat{B})$ of the basis \hat{B} satisfy

$$\rho(\hat{B}) \leq s^{p(p-1)/2} \quad \theta_i(\hat{B}) \leq s^{p-1}.$$

The main results about the computational complexity of the LLL algorithm are the following.



Theorem 3.7 Lenstra-Lenstra-Lovàsz

Let $\mathcal{L} \subset \mathbb{Z}^n$ a network given by a basis b_1, b_2, \dots, b_p , and let

$$B := \max \{|b_i|^2, \quad i \in \llbracket 1, p \rrbracket\}. \quad (3.4)$$

Then, the number of arithmetic operations performed by the algorithm $LLL(t)$ is in $O(p^3 n \log_t B)$, and the operands are integers whose binary length is in $O(p \log B)$.

Finally, the following theorem proves that the LLL algorithm naturally provides an approximation algorithm for SVP, with an exponential approximation factor in the dimension.



Theorem 3.8 Approximation algorithm for the Shortest Vector Problem.

In a network of dimension $p \geq 2$, the algorithm LLL(t) is a polynomial approximation algorithm for computing the shortest vector, with an approximation factor in s^{p-1} , where s and t are related by the relation $s^2 = 4t^2/(4 - t^2)$.

Given these results, we turn to the question of devising a more efficient algorithm, with better scalability than the LLL algorithm.

3.4 Extended Lagarias-Odlyzko Algorithm

3.4.1 Historical background : the sumset problem

The sumset problem is defined as follow : suppose x_1, \dots, x_p are p random positive integers. Define y by

$$y = \sum_{j=1}^p x_j \beta_j^*, \quad \beta_j^* \in \{0, 1\}, \quad i = 1, \dots, p.$$

The sumset problem is the one of recovering the β_j^* 's given y only. More than 30 years ago, this problem received a lot of attention in the field of cryptography, based on the belief that the problem would be hard to solve in many "real" instances. This would have implied that the already built public key cryptosystems could be considered safe from attacks. Nevertheless this belief was proven to be wrong in the early 80's. In 1986, Frieze [45] used a short argument to prove the following fact : if x_1, \dots, x_p follow an i.i.d uniform distribution on $[2^{\frac{1}{2}(1+\varepsilon)p^2}]$ for some $\varepsilon > 0$ then there exists a polynomial-in- p time algorithm which solves the sumset problem with high probability as $p \rightarrow \infty$. This efficient algorithm is based on the LLL algorithm and is called **extended Lagarias Odlyzko algorithm**.

3.4.2 The extended Lagarias Odlyzko algorithm

We now consider a multidimensional extention of the sumset problem and which will allow us to address the integer regression problem. Let $n, p, R \in \mathbb{Z}_{>0}$. We will consider the model

$$Y = X\beta^* + \epsilon, \tag{3.5}$$

where $Y \in \mathbb{Z}^n, X \in \mathbb{Z}^{n \times p}, \beta^* \in (\mathbb{Z} \cap [-R, R])^p$ and $\epsilon \in \mathbb{Z}^n$. Given Y, X , the regression problem we want to address is the exact estimation problem consisting in recovering the vector β^* . For this purpose, we will use the extended Lagarias-Odlyzko algorithm which takes $Y \in \mathbb{Z}^n, X \in \mathbb{Z}^{n \times p}, \hat{R} \in \mathbb{Z}_{>0}$ and $\hat{\epsilon} \in \mathbb{Z}_{>0}$ as input and outputs $\hat{\beta}$ an estimation of β^* . In reality, the values of R and $\|\epsilon\|_\infty$ are not known, and for this reason, we will use \hat{R} , an estimated upper bound in absolute value for the entries of β^* , and $\hat{\epsilon}$, an estimated upper bound in absolute value for the entries of ϵ . The extended Lagarias-Odlyzko algorithm is a method that leverages the integer structure of the true parameter for solving the problem of estimating β^* .

Algorithm 3 The algorithm

Require: $(Y, X, \hat{R}, \hat{\epsilon})$, where $Y \in \mathbb{Z}^n$, $X \in \mathbb{Z}^{n \times p}$, $\hat{R}, \hat{\epsilon} \in \mathbb{Z}_{>0}$

Ensure: $\hat{\beta}$, an estimate of β^*

- 1: Generate a random vector $Z \in \{\hat{R}+1, \hat{R}+2, \dots, 2\hat{R}+\log p\}^p$ with iid entries uniform in $\{\hat{R}+1, \hat{R}+2, \dots, 2\hat{R}+\log p\}$
- 2: Set $\tilde{Y} = Y + XZ$
- 3: **for** $i = 1$ to n **do**
- 4: **if** $|(\tilde{Y}_2)_i| < 3$ **then**
- 5: Set $(\tilde{Y}_2)_i = 3$
- 6: **else**
- 7: Set $(\tilde{Y}_2)_i = (\tilde{Y}_1)_i$
- 8: **end if**
- 9: **end for**
- 10: Set $m = 2^{n+1+\lceil \frac{1}{3}+p \rceil} (\hat{R}\lceil \sqrt{p} \rceil + \hat{\epsilon}\lceil \sqrt{n} \rceil)$
- 11: Construct matrix A_m as :

$$A_m := \begin{bmatrix} mX & -m\text{Diag}_{n \times n}(Y) & mI_{n \times p} \\ I_{p \times p} & 0_{p \times n} & 0_{p \times p} \\ 0_{n \times p} & 0_{n \times n} & I_{n \times n} \end{bmatrix}$$

- 12: Obtain $\hat{z} \in \mathbb{Z}^{2n+p}$ by running the LLL basis reduction algorithm on the lattice generated by the columns of A_m
- 13: Compute $g = \gcd(\hat{z}_{n+1}, \hat{z}_{n+2}, \dots, \hat{z}_{n+p})$ using Euclid's algorithm
- 14: **if** $g \neq 0$ **then**
- 15: Output $\beta^* = \frac{1}{g}(\hat{z}_{n+1}, \hat{z}_{n+2}, \dots, \hat{z}_{n+p})$
- 16: **else**
- 17: Output $\beta^* = 0_{p \times 1}$
- 18: **end if**

3.4.3 Theoretical results.

In [49], the authors gives the following result :



Theorem 3.9

Suppose

1. $X \in \mathbb{Z}^{n \times p}$ is a matrix with iid entries generated according to a distribution \mathcal{D} on \mathbb{Z} which, for some $N \in \mathbb{Z} > 0$ and constants C and $c > 0$ assigns at most $\frac{c}{2^N}$ probability to each element of \mathbb{Z} and satisfies $\mathbb{E}[|V|] \leq C2^N$, for $V \sim \mathcal{D}$;
2. $\beta^* \in (\mathbb{Z} \cap [-R, R])^p$, $W \in \mathbb{Z}^n$;
3. $Y = X\beta^* + W$.

Suppose furthermore that $\hat{R} \geq R$ and

$$N \geq \frac{1}{2n}(2n+p) \left[2n + p + 10 \log \left(\hat{R}\sqrt{p} + (\|W\|_\infty + 1)\sqrt{n} \right) \right] + 6 \log((1+c)np).$$

Then, for any $\hat{W} \leq \|W\|_\infty$, the algorithm ELO with input (Y, X, \hat{R}, \hat{W}) outputs exactly β^* with probability $1 - \left(\frac{1}{np}\right)$, with high probability as $p \rightarrow +\infty$, and terminates in time at most polynomial in $n, p, N, \log(\hat{R})$ and $\log(\hat{W})$.

Our contribution is to extend this result to the case where X has correlated columns :



Theorem 3.10

Assume that

- (i) $X \in \mathbb{Z}^{n \times p}$ is a matrix with iid rows generated according to a distribution \mathcal{D} on \mathbb{Z}^p , with $X_{\max} = \max\{k \in \mathbb{R}^+, \mathbb{P}[X = k] > 0\}$ such that $\exists N \in \mathbb{N}, X_{\max} = \mathcal{O}(2^N)$
- (ii) $\beta^* \in (\mathbb{Z} \cap [-R, R])^p$, and $W \in \mathbb{Z}^n$.
- (iii) For $i = 1, \dots, n$, and for all $\gamma \in \mathbb{Z}^p$ with $\|\gamma\|_1 = p$, the probability generating function of $\sum_{j=1}^p X_{i,j} \gamma_j \in \{0, 1, \dots, S_\gamma\}$, with $p_0 p_{S_\gamma} > 0$, has only zeros ζ satisfying $|\arg(\zeta)| \geq \alpha > 0$ and $\rho^{-1} \leq |\zeta| \leq \rho$, for some $\rho \geq 1$.
- (iv) $Y = X\beta^* + W$.

Suppose furthermore that $\hat{R} \geq R$ and

$$\begin{aligned} N \geq & \frac{1}{2n} (2n + p) \left[2n + p + 10 \log \left(\hat{R} \sqrt{p} + (\|W\|_\infty + 1) \sqrt{n} \right) \right] \\ & + 6 \log((1+c)np) - \log(\alpha) + \frac{\pi}{\alpha} \log(\rho) - \frac{1}{2} \log(p). \end{aligned} \quad (3.6)$$

For any $\hat{W} \geq \|W\|_\infty$, the algorithm ELO outputs **exactly** β^* with probability $1 - \mathcal{O}\left(\frac{1}{np}\right)$ and terminates in time at most polynomial in $n, p, N, \log(\hat{R})$ and $\log(\hat{W})$.

Proof. See Section 3.5. □

3.5 Proof of Theorem 3.10

3.5.1 Preliminary result

Beginning with (3.6) :

$$\begin{aligned} N &\geq 10 \log \left(\sqrt{p} + \sqrt{n} (\|W\|_\infty + 1) \right) \\ &\geq 5 \log \left(\sqrt{p} \sqrt{n} (\|W\|_\infty + 1) \right) \\ &\geq 2 \log \left(pn (\|W\|_\infty + 1) \right) \end{aligned}$$

We get $2^N \geq \left(pn (\|W\|_\infty + 1) \right)^2$ which implies :

$$\frac{\|W\|_\infty}{2^N} \leq \frac{1}{n^2 p^2}.$$

3.5.2 Main proof

Why ELO use the columns of $A_m \in \mathbb{Z}^{(2n+p) \times (2n+p)}$ as the lattice which will serve as an input for the LLL algorithm ? The first reason comes from the definition of the lattice : every z in the lattice can be written as $z = A_m x$ for a certain $x \in \mathbb{Z}^{2n+p}$.

The second reason come from computing z_0 defined as :

$$z_0 = \begin{bmatrix} mX & -m\text{Diag}_{n \times n}(Y) & mI_{n \times p} \\ I_{p \times p} & 0_{p \times n} & 0_{p \times p} \\ 0_{n \times p} & 0_{n \times n} & I_{n \times n} \end{bmatrix} \begin{bmatrix} \beta \\ 1_n \\ W \end{bmatrix} = \begin{bmatrix} 0_{n \times 1} \\ \beta \\ W \end{bmatrix} \quad (3.7)$$

Step 1 ($\|\hat{z}\|_2$ is controlled by LLL's bound) : The smallest ℓ_2 -norm vector, say z^* , in the lattice has smaller norm than $\|\beta\|_2 + \|W\|_2$. We thus have

$$\|z^*\|_2 \leq \|\beta\|_\infty \sqrt{p} + \|W\|_\infty \sqrt{n}.$$

We use hypothesis ii) in order to get :

$$\|z^*\|_2 \leq (3\hat{R} + \log(p)) \sqrt{p}^+ (\|W\|_\infty + 1) \sqrt{n}$$

Then, the LLL finds a $\hat{z} \in \mathbb{Z}^{2n+p}$ such that

$$\|\hat{z}\|_2 \leq 2^{\frac{2n+p}{2}} \|z^*\|_2 \leq 2^{\frac{2n+p}{2} + 2} p (\hat{R} \sqrt{p} + (\|W\|_\infty + 1) \sqrt{n}) := m_0 < m \quad (3.8)$$

where Theorem 3.8 gives us a scaling by s^{d-1} with $d = 2n + p$ and $s \in]1, 2[$ (see (3.3)) implying that

$$\|\hat{Z}\| < \|s^{2n+p-1} z^*\| < \|2^{2n+p-1} z^*\| < \|2^{2n+p} z^*\| = 2^{\frac{2n+p}{2}} \|z^*\|$$

so long as m is an upper bound to m_0 .

Step 2 (Similarly to (3.7), \hat{z} also satisfies $[\hat{z}_1, \dots, \hat{z}_n] = 0$ when m is sufficiently large) : For $z \in \mathbb{L}_m$ such that $z_{n+1}, \dots, z_{n+p} \neq \lambda\beta$, with $\lambda \in \mathbb{Z}^*$, we have $\|z\|_2 \leq m_0$ implies $z_1, \dots, z_n = 0$ as soon as $m > m_0$ since if non-zero, at least one component is larger than m .

Step 3 (LLL solution is non-zero)

For any non-zero $z \in \mathbb{L}_m$ with $\|z\|_2 \leq m_0$, necessarily

$$z_{n+1}, \dots, z_{n+p} \neq 0 \quad (3.9)$$

with probability $1 - \mathcal{O}(\frac{1}{np})$. That is why the case $\lambda = 0$ is not possible with probability $1 - \mathcal{O}(\frac{1}{np})$. So it suffices to show that there is no triplet $x = (0, x_2, x_3)^t$, $x_2, x_3 \in \mathbb{Z}^n$ for which the vector $z = A_m x \in \mathbb{L}_m$ is non-zero and $\|z\|_2 \leq m_0$, with probability $1 - \mathcal{O}(\frac{1}{np})$. We have

$$z = \begin{bmatrix} m(\text{Diag}_{n \times n}(Y)x_2 + x_3) \\ 0 \\ x_3 \end{bmatrix}.$$

Since $\|z\|_2 \leq m_0$, we have

$$\text{Diag}_{n \times n}(Y)x_2 = x_3$$

or row-wise

$$\forall i \in [n], Y_i(x_2)_i = (x_3)_i.$$

We also have $\forall i = 1, \dots, n, \|Y_i\| \geq \frac{3}{2} \frac{2^N}{n^2 p^2}$ with probability at least $1 - \mathcal{O}(\frac{1}{np})$, simply because

$$(Y_2)_i \geq \left| \sum_{j=1}^p X_{ij} \beta_j \right| - \|W_1\|_\infty \geq \left| \sum_{j=1}^p X_{ij} \beta_j \right| - \|W\|_\infty - 1. \quad (3.10)$$

Using Lemma 3.5.1 gives us

$$\left| \sum_{j=1}^p X_{ij} \beta_j \right| \geq 3 \frac{2^N}{n^2 p^2} \quad (3.11)$$

with probability at least $1 - \mathcal{O}(\frac{1}{np})$. Moreover, we use (3.6) to obtain :

$$\|W\|_\infty \leq \frac{2^N}{n^2 p^2}$$

Combining the latter equations with (3.5.1) gives us the desired result. Therefore we obtain :

$$\forall i \in [n], \frac{3}{2} \frac{2^N}{n^2 p^2} \|(x_2)_i\| \leq \|(x_3)_i\|$$

with probability $1 - \mathcal{O}(\frac{1}{np})$. Since $z = A_m x = (0, 0, x_3)$ and we have supposed that z is non-zero, there exists $i \in [n]$ for which $(x_3)_i \neq 0$. Using that $\forall i \in [n], Y_i(x_2)_i = (x_3)_i$, we have that $(x_2)_i \neq 0$ as well. Since x_2 is integer-valued and $\|z = (0, 0, x_3)\|_2 \leq m_0$, it must be simultaneously true that $\|(x_2)_i\| \geq 1$ and $\|(x_3)_i\| \leq m_0$ for this value of i . With these inequalities, we obtain that

$$\frac{3}{2} \frac{2^N}{n^2 p^2} \leq m_0$$

Thus, we obtain that $\frac{2^N}{n^2 p^2} \leq m_0$, which is equivalent to

$$N \leq 2 \log(np) + \log m_0$$

After replacing m_0 by its value, we have

$$N \leq 2 \log(np) + \frac{2n+p}{2} + \log p + \log(\hat{R}\sqrt{p} + (\|W\|_\infty + 1)\sqrt{n}).$$

On the other hand, we know that

$$N \geq \frac{1}{2n}(2n+p)[2n+p+10\log(\hat{R}\sqrt{p} + (\|W\|_\infty + 1)\sqrt{n})] + 6\log((1+c)np).$$

We conclude that there is no non-zero vector in $\mathbb{L}_m \setminus \{z \in \mathbb{L}_m \mid z_{n+1}, \dots, z_{n+p} = \lambda\beta, \lambda \in \mathbb{Z}\}$ with ℓ_2 norm less than m_0 .

 **Lemma 3.5.1**

We have

$$\mathbb{P} \left(\min_{i=1}^n \left| \sum_{j=1}^p X_{ij} \beta_j \right| \leq K \right) \leq n \times (2\lfloor K \rfloor + 1) c\alpha^{-1} \rho^{\pi/\alpha} S_\beta^{-\frac{1}{2}} \quad (3.12)$$

Proof. We have

$$\mathbb{P} \left(\min_{i=1}^n \left| \sum_{j=1}^p X_{ij} \beta_j \right| \leq K \right) = \mathbb{P} \left(\bigcup_{i=1}^n \left\{ \left| \sum_{j=1}^p X_{ij} \beta_j \right| \leq K \right\} \right) \quad (3.13)$$

$$\leq \sum_{i=1}^n \mathbb{P} \left(\left| \sum_{j=1}^p X_{ij} \beta_j \right| < K \right) \quad (3.14)$$

$$= \sum_{i=1}^n \sum_{k \in \mathbb{Z} \cap [-K, K]} \mathbb{P} \left(\sum_{j=1}^p X_{ij} \beta_j = k \right) \quad (3.15)$$

$$\leq \sum_{i=1}^n (2\lfloor K \rfloor + 1) \max_{k \in \mathbb{Z} \cap [-K, K]} \mathbb{P} \left(\sum_{j=1}^p X_{ij} \beta_j = k \right) \quad (3.16)$$

$$\leq n \times (2\lfloor K \rfloor + 1) \max_{k \in \mathbb{Z} \cap [-K, K]} \mathbb{P} \left(\sum_{j=1}^p X_{ij} \beta_j = k \right). \quad (3.17)$$

Now, using [81, Corollary 4] we have

$$\max_{k \in \mathbb{Z} \cap [-K, K]} \mathbb{P} \left(\sum_{j=1}^p X_{ij} \beta_j = k \right) \leq c\alpha^{-1} \rho^{\pi/\alpha} S_\beta^{-\frac{1}{2}}, \quad (3.18)$$

which implies

$$\mathbb{P} \left(\min_{i=1}^n \left| \sum_{j=1}^p X_{ij} \beta_j \right| \leq K \right) \leq n \times (2\lfloor K \rfloor + 1) c\alpha^{-1} \rho^{\pi/\alpha} S_\beta^{-\frac{1}{2}}. \quad (3.19)$$

□

Step 4 ($\exists \lambda \in \mathbb{Z}_*$, such that $\hat{x}_1 = \lambda\beta$) : Let $\hat{x} = (\hat{x}_1, \hat{x}_2, \hat{x}_3)$ be such that $\hat{z} = A_m \hat{x}$. Since $\hat{z}_2 = \hat{x}_1$ and $\hat{z}_3 = \hat{x}_3$ and since $\|\hat{z}\|_2 \leq m_0$, we necessarily have $\|\hat{x}_1\|_\infty \leq \|\hat{x}_1\|_2 \leq m_0$ and $\|\hat{x}_3\|_\infty \leq \|\hat{x}_3\|_2 \leq m_0$. It can now easily be shown that the set of all $(x_1, x_2, x_3) \in \mathbb{Z}^p \times \mathbb{Z}^n \times \mathbb{Z}^n$ such that :

$$\begin{cases} \|x_1\|_\infty \leq m_0, \|x_3\|_\infty \leq m_0, \\ Xx_1 - \text{Diag}_{n \times n}(Y)x_2 - x_3 = 0, \\ \forall \lambda \in \mathbb{Z} \ x_1 \neq \lambda\beta, \end{cases} \quad (3.20)$$

has very small probability for random matrices X . Indeed, under these assumptions, the following result holds.

Lemma 3.5.2

For any X random matrix satisfying hypothesis i) and iii), for any $(x_1, x_2, x_3) \in \mathbb{Z}^p \times \mathbb{Z}^n \times \mathbb{Z}^n$ satisfying (3.20), we have

$$\mathbb{P}(Xx_1 - \text{Diag}_{n \times n}(Y)x_2 - x_3 = 0) \leq \frac{c^n R^{n\pi/\alpha}}{\alpha^n p^{n/2} X_{\max}^{n/2}}.$$

Proof. We can rewrite

$$Xx_1 - \text{Diag}_{n \times n}(Y)x_2 - x_3 = 0$$

row-wise, for $i \in [n]$ and X_i the i -th row of X , we have

$$\langle X_i, x_1 \rangle - Y_i(x_2)_i = (x_3)_i.$$

By plugging $Y = \langle X_i, \beta \rangle + \epsilon_i$, we obtain

$$\langle X_i, x_1 - (x_2)_i \beta \rangle = (x_3)_i - (x_2)_i \epsilon_i.$$

Using independence between the rows of X , we have

$$\mathbb{P}(Xx_1 - \text{Diag}_{n \times n}(Y)x_2 - x_3 = 0) = \prod_{i=1}^n \mathbb{P}(\langle X_i, x_1 - (x_2)_i \beta \rangle = (x_3)_i - (x_2)_i \epsilon_i).$$

Since,

$$x_1 \neq \lambda \beta, \quad \forall \lambda \in \mathbb{Z},$$

we have

$$x_1 - (x_2)_i \beta \neq 0 \tag{3.21}$$

and at least one of the entries of X_i gets a nonzero multiplicative coefficient in the expression

$$\langle X_i, x_1 - (x_2)_i \beta \rangle = (x_3)_i - (x_2)_i \epsilon_i \tag{3.22}$$

Then,

$$\begin{aligned} \mathbb{P}[\langle X_i, x_1 - (x_2)_i \beta \rangle = (x_3)_i - (x_2)_i \epsilon_i] &= \mathbb{P}[q_i \langle X_i, x_1 - (x_2)_i \beta \rangle = q_i ((x_3)_i - (x_2)_i \epsilon_i)] \\ &\leq \max_y \mathbb{P}[q_i \langle X_i, x_1 - (x_2)_i \beta \rangle = y] \end{aligned} \tag{3.23}$$

with $q_i = \frac{p}{\|x_1 - (x_2)_i \beta\|_1}$. Now, using Corollary 4 from [81], following hypothesis i) and iii) from Theorem 3.10

$$\mathbb{P}[\langle X_i, x_1 - (x_2)_i \beta \rangle = (x_3)_i - (x_2)_i \epsilon_i] \leq c \alpha^{-1} \rho^{\pi/\alpha} S_{q_i(x_1 - (x_2)_i \beta)}^{-1/2} \tag{3.24}$$

As a result, and using that $S_{a\gamma} = aS_\gamma$

$$\prod_{i=1}^n \mathbb{P}(\langle X_i, x_1 - (x_2)_i \beta \rangle = (x_3)_i - (x_2)_i \epsilon_i) \leq \frac{c^n R^{n\pi/\alpha}}{\alpha^n \prod_{i=1}^n q_i S_{x_1 - (x_2)_i \beta}^{1/2}} \quad (3.25)$$

which immediately gives

$$\mathbb{P}(Xx_1 - \text{Diag}_{n \times n}(Y)x_2 - x_3 = 0) \leq \frac{c^n R^{n\pi/\alpha}}{\alpha^n \prod_{i=1}^n q_i S_{x_1 - (x_2)_i \beta}^{1/2}}. \quad (3.26)$$

Let us now focus on $\prod_{i=1}^n q_i S_{x_1 - (x_2)_i \beta}^{1/2}$, beginning with q_i . We have that

$$q_i = \frac{p}{\|x_1 - (x_2)_i \beta\|_\infty}$$

and

$$S_{q_i(x_1 - (x_2)_i \beta)} = \max \left\{ k \in \mathbb{R}^+, \quad \mathbb{P}[\langle X_i, q_i(x_1 - (x_2)_i \beta) \rangle = k] > 0 \right\} \quad (3.27)$$

On the other hand, $q_i \geq \frac{p}{\|x_1 - (x_2)_i \beta\|_\infty}$, and thus

$$\left\| q_i(x_1 - (x_2)_i \beta) \right\|_\infty \geq p. \quad (3.28)$$

Using the notations $(a_i)_j = (q_i(x_1 - (x_2)_i \beta))_j$, we get

$$\langle X_i, q_i(x_1 - (x_2)_i \beta) \rangle = \sum_{j=1}^p X_{ij} (a_i)_j. \quad (3.29)$$

Consider $J \in \{1, \dots, p\}$ such that $\|q_i(x_1 - (x_2)_i \beta)\|_\infty = |a_J|$, and $e_J = (0, \dots, 0, \underbrace{1}_J, 0, \dots, 0)$.

Using (3.27), it follows, as $\mathbb{P}[q_i(x_1 - (x_2)_i \beta) = e_J] > 0$

$$\begin{aligned} S_{q_i(x_1 - (x_2)_i \beta)} &= \max \left\{ k \in \mathbb{R}^+, \quad \mathbb{P}[\langle X_i, q_i(x_1 - (x_2)_i \beta) \rangle = k] > 0 \right\} \\ &\geq \max \left\{ k \in \mathbb{R}^+, \quad \mathbb{P}[\langle X_i, p e_J \rangle = k] > 0 \right\} \\ &= \max \left\{ k \in \mathbb{R}^+, \quad \mathbb{P}[p X_{ij} = k] > 0 \right\} \\ &= p X_{\max}. \end{aligned}$$

Combining this last bound with (3.26) gives

$$\mathbb{P}(Xx_1 - \text{Diag}_{n \times n}(Y)x_2 - x_3 = 0) \leq \frac{c^n R^{n\pi/\alpha}}{\alpha^n p^{n/2} X_{\max}^{n/2}}.$$

□

 **Lemma 3.5.3**

For any $(x_1, x_2, x_3) \in \mathbb{Z}^p \times \mathbb{Z}^n \times \mathbb{Z}^n$ satisfying (3.20), we have $\|x_2\|_\infty = \mathcal{O}(m_0 n^2 p^3)$ with probability $1 - \mathcal{O}\left(\frac{1}{np}\right)$.

Proof. See [49], p.17-18] □

 **Theorem 3.11**

 For any X random matrix satisfying the assumptions of Theorem 3.10, the probability that $\exists \lambda \in \mathbb{Z}_*$, such that $\hat{x}_1 = \lambda \beta$ is at least $1 - \mathcal{O}\left(\frac{1}{np}\right)$.

Proof. For any $r > 0$, there are at most $\mathcal{O}(r^n)$ vectors in \mathbb{Z}^n with ℓ_∞ -norm at most r . Combining this observation with the fact that, from Lemma 3.5.2

$$\mathbb{P}(Xx_1 - \text{Diag}_{n \times n}(Y)x_2 - x_3 = 0) \leq \frac{c^n R^{n\pi/\alpha}}{\alpha^n p^{n/2} X_{\max}^{n/2}}, \quad (3.30)$$

together with the union bound over all the integer vectors (x_1, x_2, x_3) with $\|x_2\|_\infty = \mathcal{O}(m_0 n^4 p^5)$, $\|x_1\|_\infty \leq m_0$ and $\|x_3\|_\infty \leq m_0$, we conclude that the probability that there exist a triplet (x_1, x_2, x_3) satisfying (3.20) is at most of the order

$$\left(m_0 n^2 p^3\right)^n m_0^{n+p} \left(\frac{c^n R^{n\pi/\alpha}}{\alpha^n p^{n/2} X_{\max}^{n/2}}\right) \quad (3.31)$$

which is small as soon as X_{\max} is sufficiently large. Since, using **Step 1**, $\|x_1\|_\infty \leq m_0$ and $\|x_3\|_\infty \leq m_0$ and since, using **Step 2**, \hat{x} satisfies (3.5.2), one deduces that $\exists \lambda \in \mathbb{Z}_*$ such that $\hat{x}_1 = \lambda \beta$ has probability at least

$$1 - \mathcal{O}\left(\left(m_0 n^2 p^3\right)^n m_0^{n+p} \left(\frac{c^n R^{n\pi/\alpha}}{\alpha^n p^{n/2} X_{\max}^{n/2}}\right)\right).$$

Plugging in the value of m_0 and using (3.6), we deduce that the claimed result holds with probability at least $1 - \mathcal{O}\left(\frac{1}{np}\right)$. □

Chapitre 4

Applications en oncologie, neuroscience et acoustique

Nul ne peut se prétendre médecin s'il ne connaît les bases de l'astrologie

Hippocrate

Dans ce chapitre nous appliquons l'algorithme de création de complexe simplicial sur deux types de données réelle dans le but d'étudier son utilisation dans le contexte de détection d'anomalies.

- Dans la première partie, nous étudions un problème en détection précoce de développement de tumeur cancéreuse. Ce travail est motivé par des discussions avec Loïc Verlingue du Centre Léon Bérard qui nous a introduit à ces nouveaux défis et les projets de l'entreprise GRAIL dans le domaine. La problématique est de détecter les dynamiques malignes en estimant un coefficient de régression exponentielle. L'idée d'utiliser la théorie des Signatures nous a vite paru très pertinente, les données (s'agissant de prélèvements sanguins) étant en faible nombre et à échantillonnage irrégulier. Nous proposons un test avec correction, fondé sur une étude théorique des premiers coefficients de signatures pour des processus de morts avec immigration, motivée par le modèle utilisé par GRAIL, ainsi qu'une étude par la topologie des données et la structure de groupes des Signatures des-dits processus.
- Dans une seconde partie, nous développons un travail présenté aux Journées de la Statistique 2024. L'objectif ici est de considérer des données avec une organisation spatiale intrinsèque : des données d'electroencéphalogramme. Pour être plus précis, nous étudions les variations de la structure topologique obtenue grâce à l'algorithme du chapitre 2. Dans cette optique, nous étudions l'évolution temporelle des nombres de Betti sur des patients présentant des crises d'épilepsie.

Pour pouvoir obtenir ces nombres de Betti, nous introduisons une première filtration adaptée au complexe créé, sujet qui n'a pas été discuté dans le chapitre 2.

- Dans une seconde partie, nous étudions l'application de l'algorithme sur des données

acoustiques étiquetées pour la détection d'anomalies. Outre la présence d'étiquetage de données anormales, ces données ne sont pas interprétables spatialement. Ce jeu de données étant à but industriel, il va permettre de voir si notre complexe, et sa filtration, permettent de bien distinguer données anormales de données normales. Nous réinvestissons la méthode exploratoire utilisée dans la section 4.1.

Dans le même temps, nous en profiterons pour étudier la faisabilité d'une seconde filtration possible basée sur le paramètre de parcimonie du LASSO.

4.1 Détection de dynamiques malignes en oncologie.

Nous abordons dans cette section l'étude des trajectoires des taux de ctDNA dans les prélèvements sanguins de patients afin de détecter une explosion précoce et d'alarmer sur la présence d'une dynamique cancéreuse. Nous montrons que la théorie des Signatures est très pertinente pour ce problème et proposons un test statistique fondé sur l'approximation de la loi du processus qui modélise ces dynamiques sous l'hypothèse nulle.

4.1.1 Introduction

The early detection of aggressive cancerous tumors through blood samples has been the subject of recent research and is a potential game changer in cancer detection, with high expected societal impact. Traditional diagnostic methods have relied heavily on imaging techniques and biopsy procedures, which, despite their efficacy, come with limitations such as invasiveness, high costs, and the risk of missing early-stage tumors. Recent advances in molecular biology and bioinformatics have opened up new avenues for non-invasive cancer detection, leveraging the analysis of biomarkers in blood samples. In particular, an important advance in the diagnostic and surveillance toolbox for oncologists is circulating tumor DNA (ctDNA). This recent technology was observed to permit efficient detection of microscopic levels of cancer tissue which can be exploited for efficient patient monitoring. In practice, accurately detecting the potential signs of malignancy poses a formidable machine learning challenge due to the scarcity of the observation data.

The present paper proposes a new approach to early cancer detection from few blood samples using the theory of signatures for extracting and amplifying the distinctive features of the specific birth and death processes that govern the shedding of ctDNA biomarkers in the blood stream. We will leverage the power of Signature Theory, a recent tool in Machine Learning with a long history in Control Theory and Stochastic Analysis, which was shown to achieve state of the art prediction capabilities without training in the feature extraction phase. Signatures extract meaningful nonlinear information about the signals, and seamlessly adapts to irregular sampling, a key advantage in our set up where blood samples arrive at irregular sampling times. We will complement Signature based feature extraction results with some results obtained using Topological Data Analysis, a fast growing field of statistics that leverages topological concepts to unveil persistent structures using a statistical viewpoint.

The main contribution of the present work is to introduce recent techniques from Machine Learning such as Signature Theory and Topological Data Analysis to address the challenge of earliest detection. More precisely, based on a birth and death process

model for the generation of biomarkers by cancer cells and the elimination from the body, we show in particular how Signature Theory allows to come up with an efficient statistical testing procedure by computing an efficient approximation of second order Signatures for relevant approximations using Poisson processes. We illustrate our findings using simulation experiments that confirm a good detection power for the malignant cases, relying only on an extremely small number of observations.

This paper is organized as follows :

- A quick explanation on birth and death processes that allow us to model tumor growth.
- A brief reminder on signatures and topological data analysis (TDA)
- The testing framework
- An exploratory section using TDA and Principal Geodesic Analysis
- Results

4.1.2 A stochastic model for the dynamics of tumorous biomarkers

A cell can divide at a birth rate b or it can die at a death rate equal to d , leading to a Birth and Death process. When a cell dies by apoptosis, it releases ctDNA into the bloodstream at a rate q_d . Moreover, Finally, each day, an ϵ proportion of ctDNA is eliminated by organism.

The dynamics of tumor growth has been a topic of extensive research in various sub-fields of applied mathematics. Stochastic models are in particular playing an important rôle for capturing the variability of the signal [36]. Based on such stochastic models, the dynamics of CtDNA biomarkers in the blood as a function of time has been extensively studied in [2].

Background on point processes, birth and death processes and Cox models

In this section, we recall the basic models that underpin the dynamics of the observed ctDNA levels.

Definition 4.1.1. *A **Point process** over \mathbb{R}_+ is an increasing sequence of random variables*

$$0 < T_1 < T_2 < \dots < T_n < \dots$$

defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with values in \mathbb{R}_+ and which satisfies $T_n \xrightarrow[n \rightarrow +\infty]{\text{P.s.}} +\infty$. These random times T_n , $n \in \mathbb{N}$ are used to model the times at which an event occurs. The successive time intervals between two successive events are defined as $S_1 = T_1$ and $S_n = T_n - T_{n-1}$, $n \in \mathbb{N}^$.*

One can define the counting process $(N_t)_{t \geq 0}$ associated with the point process $\{T_n, n \in \mathbb{N}\}$ defined by :

$$N_t = \sup\{n \in \mathbb{N}, T_n \leq t\} = \sum_{j \geq 1} \mathbb{1}_{\{T_j \leq t\}}$$

This process is time continuous and describes the number of events that have occurred before a time $t \in \mathbb{R}_+$.

Birth and death processes :

Poisson processes are a special case of a more general type of process call Birth and Death process (BDP). These processes are defined formally as follows.

Definition 4.1.2. A Birth and Death process is a point process $(X_t)_{t \geq 0}$ with state space $S = \mathbb{N}$ whose transitions from k to $k + 1$ occur at rate λ_k for $k \in \mathbb{N}$, denoted the birth rates, and whose transitions from k to $k - 1$ occur at rate μ_k for $k \in \mathbb{N}_*$, denoted the death rate.

To such processes, we associate the transition matrix

$$P_{ij}(t) = \mathbb{P}(X_t = j \mid X_0 = i) \quad (4.1)$$

with $P_{ij}(0) = \delta_{ij}$ (Kronecker symbol) which satisfies the differential equation

$$\frac{dP_{i0}(t)}{dt} = \mu_1 P_{i1}(t) - \lambda_0 P_{i0}(t) \quad (4.2)$$

and

$$\frac{dP_{ij}(t)}{dt} = \lambda_{j-1} P_{i,j-1}(t) + \mu_{j+1} P_{i,j+1}(t) - (\lambda_j + \mu_j) P_{ij}(t). \quad (4.3)$$

The Cox process :

The doubly stochastic Poisson process (DSPP) is a generalization of the Poisson process when the intensity of the occurrence of the points is influenced by an external process called information process such that the intensity becomes a random process. This process was introduced by Cox [34]. In our study, the Cox process which models the ctDNA level in the blood is a Poisson process whose intensity is proportional to the Birth and Death process modelling the tumor size as a function of time.

Simulation :

We used the **BirDepy** python package to simulate trajectories, considering that cells number of a tumor can be modelled by a birth and death process [58].

4.1.3 Details on the biomarker dynamics

When tumorous cells are present in the body, their death through apoptosis have a probability to release specific biomarker, called circulating DNA, or ctDNA, in the bloodstream.

We now present the tumor growth model and the resulting ctDNA biomarker shedding process. We will distinguish between different situations : the malignant tumor model, the benign tumor model and the biomarker shedding model.

Malignant tumors :

The primary tumor grows stochastically from a single malignant cell at time $t = 0$. The tumor size at time t , denoted A_t and expressed in number of cells, is modeled by a supercritical branching birth death process with net growth rate

$$r = b - d > 0,$$

where b and d denote the birth and death rate per day, respectively.

This process has a probability of $\delta = \min\{1, d/b\}$ to go extinct. See e.g. Durrett. Normal or benign tumor cells also shed cell-free DNA (cfDNA) into the bloodstream. Cells in benign tumors and expanded subclones often harbor the same cancer-associated mutations as cancer cells. Hence, ctDNA shed from benign tumors can be difficult to distinguish from ctDNA shed from malignant tumors.

Benign tumors :

Normal or benign tumor cells also shed cell-free DNA (cfDNA) into the bloodstream. Let B_t denotes the size of the benign population of cells at time t that shed the biomarker in the bloodstream, and denote their shedding rate as λ_{bn} . We assume that benign lesions roughly replicate at a constant size. Hence, benign cells divide and die at the same rate

$$b_{bn} = d_{bn},$$

for simplicity, we assume that their population size B_t remains constant over time, i.e. ($B_t = B_0$ for all t). Normal or benign tumor cells also shed cell-free DNA (cfDNA) into the bloodstream. Cells in benign tumors and expanded subclones often harbor the same cancer-associated mutations as cancer cells. Hence, ctDNA shed from benign tumors can be difficult to distinguish from ctDNA shed from malignant tumors. Considering again that the biomarker is exclusively shed by cells undergoing apoptosis, the shedding rate of benign cells is

$$\lambda_{bn} = d_{bn} \cdot q_{d, bn}$$

per day.

Biomarker shedding :

Normal or benign tumor cells also shed cell-free DNA (cfDNA) into the bloodstream. Although apoptotic benign cells shed ctDNA with the same probability as malignant cells ($q_d = q_{d,bn}$), we will assume that the shedding rate λ_{bn} of benign cells is lower than the shedding rate λ of malignant cells because benign cells typically replicate at a lower rate than cancer cells. The circulating biomarker is eliminated from the bloodstream at an elimination rate ε which can be calculated from the biomarker half-life time $t_{1/2}$ as $\varepsilon = \log(2)/t_{1/2}$. We denote as C_t^A and C_t^B the amount of biomarker (i.e., number of hGE) circulating in the bloodstream at time t shed by malignant and benign cells, respectively. The total amount of the biomarker circulating at time t is thus $C_t = C_t^A + C_t^B$. Since malignant and benign cells shed the biomarker independently from each other,

the processes (A_t, C_t^A) and (B_t, C_t^B) can be studied separately. The stochastic process (A_t, C_t^A) is a two-type branching process governed by the following transitions

$$\begin{array}{lll} A \rightarrow AA & \text{rate} & b \\ A \rightarrow C & \text{rate} & d \cdot q_d \\ A \rightarrow \emptyset & \text{rate} & d(1 - q_d) \\ C \rightarrow \emptyset & \text{rate} & \varepsilon \end{array}$$

as illustrated in Figure 4.1 below.

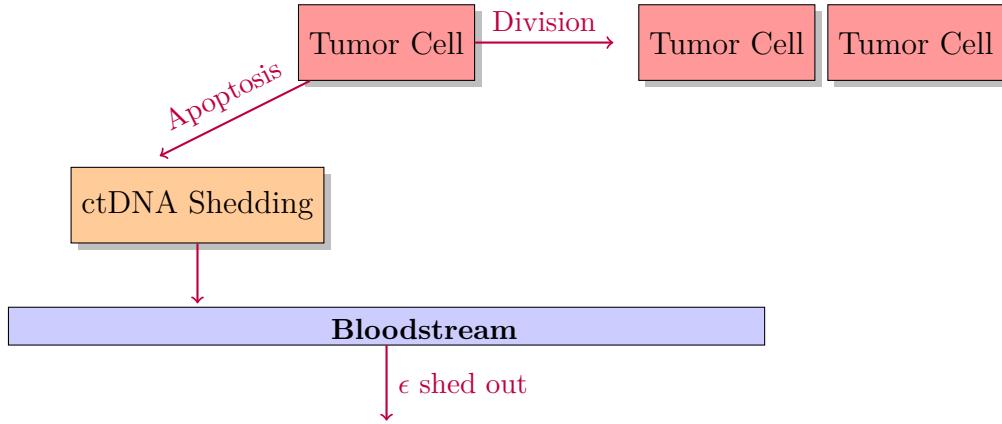


FIGURE 4.1 – The shedding process in the case of apoptosis

The process is initialized at time $t = 0$ with a single cancer cell and no circulating biomarkers, that is $(A_0, C_0^A) = (1, 0)$. Since we assume that the benign cell population B_t remains constant over time and biomarker units are eliminated from the bloodstream independently at the same rate, the process C_t^B is a branching pure-death process with constant immigration. We assume that C_t^B is at equilibrium at time $t = 0$.

The C_t^A process :

The biomarker is shed by cancer cells as a Cox process C_t^A (or doubly stochastic Poisson process with intensity $(\lambda \cdot A_t)_{t \geq 0}$). Furthermore, for large times

$$\lim_{t \rightarrow \infty} A_t = W e^{rt}$$

$$\text{with } W \stackrel{d}{=} \sum_{i=1}^{A(0)} \chi_i \psi_i$$

where $\chi_i \sim \text{Bern}(r/b)$ and $\psi_i \sim \text{Exp}(r/b)$ are independent. $W = 0$ if and only if the process A_t goes extinct. After some generating functionalology, we get for large times that

$$C_t^A \stackrel{\mathcal{D}}{\approx} \text{Geom} \left(\frac{r(\epsilon + r)}{b\lambda e^{rt} + r(\epsilon + r)} \right).$$

The C_t^B process :

The C_t^B process is a branching pure-death process with immigration, with death rate ε and immigration rate

$$B_0 \cdot \lambda_{\text{bn}} = B_0 \cdot d_{\text{bn}} \cdot q_d,$$

because the population size B_t is constant over time. The probability generating function for such a process is given by

$$\mathcal{C}^B(y, t) = \left(1 + (y - 1)e^{-\varepsilon t}\right)^{C_0^B} e^{\frac{\lambda_{\text{bn}}}{\varepsilon} B_0(y-1)(1-e^{-\varepsilon t})}.$$

When t is large,

$$\lim_{t \rightarrow \infty} \mathcal{C}^B(y, t) = e^{\frac{\lambda_{\text{bn}}}{\varepsilon} B_0(y-1)}.$$

independently of the initial biomarker level. The right hand side is the probability generating function of a Poisson random variable with mean $B_0 \cdot \lambda_{\text{bn}}/\varepsilon$, and so for large t we have

$$C_t^B \sim \text{Poisson} \left(\frac{B_0 \cdot \lambda_{\text{bn}}}{\varepsilon} \right).$$

As a matter of fact, one can easily assume that $\lambda_{\text{bn}} = \varepsilon$ and $B_0 = \mathbb{E}_{t \in \mathbb{R}_+}[C_t^B]$.

The C_t process and sampling :

By combining the previous results, we find the asymptotic limit for the distribution of the total ctDNA biomarker amount present in the bloodstream when the primary tumor is made of a large number of cells is the sum of a Cox process and a Death process with immigration, with unknown parameters.

4.1.4 The mathematical tools : The Signature transform, Principal Geodesics and Topological Data Analysis

This section introduces the technical background about Signature Theory, Principal Geodesic Analysis and Topological Data Analysis in order to make the reader comfortable when using these tools in the next section where they will be put to work in the application to Birth and Death Processes for addressing the .

Signature Theory

Signature Theory is a set of tools for analysing multidimensional signals, multivariate time series, of the form

$$X(t) = (X(t)^1, \dots, X(t)^d),$$

$t \in [0, T]$, which may not be regularly sampled in time and for which the sampling times are not necessarily synchronized across its d dimensions. Signatures are multiple integrals

of the product of the derivatives of each component, each with its own variable, and the variables of integration are constrained by a certain given order, which ensures that the speed of certain component is multiplied by the speed of another component at past times only or at future times only. The Signatures come naturally from the expansion of solutions to forced ODE's using the Picard iterations, but in the 60's, very interesting structures were discovered by Chen and others about Signatures, making them a very interesting tool for mathematicians, data scientists and engineers as well. Several very good references on Signature Theory are [22], [74], [14], as well as the first chapter from [41]. The formal definition of Signatures is given below.

Definition 4.1.3 (k^{th} -order Signature). *The signature of order k , denoted by*

$$S_{[0,t]}^{(k)}(X) \in \overbrace{\mathbb{R}^{d \times d \times \dots \times d}}^{k \text{ times}} = (\mathbb{R}^d)^{\otimes k} \quad (4.4)$$

is defined for every word $i_1 i_2 \dots i_k$ from $\{1, \dots, d\}$ by

$$S(X)_{0,t}^{i_1, \dots, i_k} = \int_{0 < t_k < t} \dots \int_{0 < t_1 < t_2} \frac{dX^{i_1}}{dt_1}(t_1) \dots \frac{dX^{i_k}}{dt_k}(t_k). \quad (4.5)$$

When the signal speeds $dX^{i_j}/dt_j(t_j)$, $j = 1, \dots, k$ do not exist, more general definitions allow to circumvent this purely technical difficulty [22]. An interesting class of paths to consider for which Signatures can be defined is the space of bounded variation paths from $[0, T] \subset \mathbb{R}$ in \mathbb{R}^d , denoted by $\mathcal{BV}(\mathbb{R}^d)$, i.e. all path X such that

$$\|X\|_{\mathcal{BV}} := \sup_{D \subset J} \sum_{\substack{t_i \in D \\ i \neq 0}} \|X_{t_i} - X_{t_{i-1}}\|_2 < \infty \quad (4.6)$$

Signatures of order 2 are matrices and signatures of order 3 and larger belong to the type of mathematical objects called tensors [70], [35].

Given the definition of all k^{th} -order Signatures, Signatures are now defined as the list of all signatures of all orders, arranged by increasing order, associated with a given multidimensional signal.

Definition 4.1.4 (Signature). *The **signature** of a path $X : \mathbb{R} \rightarrow \mathbb{R}^d$ over $I \subset \mathbb{R}$ is an infinite sequence of tensors defined by*

$$S_I(X) = \bigoplus_{i=0}^{\infty} S_I^{(i)}(X) \in \bigoplus_{d=1}^{\infty} R^{\otimes d} =: T(\mathbb{R}^d)$$

Signatures have very useful properties that make them ideal for feature extraction. Signatures would be of poor interest if not coming with very interesting properties allowing the space of Signatures to be endowed with very powerfull structures.

— The first property is **Invariance to Reparametrisation**, i.e.

$$S_I(\tilde{X}) = S_I(X) \quad (4.7)$$

for all $t \in [0, T]$ with $\tilde{X}(s) = X_{\phi(s)}$ for ϕ any surjective, increasing differentiable function $\phi : [0, T] \mapsto [0, T]$. This property is very simple and intuitive as it means that only the trajectory as a curve is preserved but not the speed at which it is traversed. For instance, the digit "3" is a curve in a 2-dimensional space, but Signatures forget how this digit actually came to be written.

- Another very important property is the **Shuffle Product Identity**. Consider a path $X : [0, T] \mapsto \mathbb{R}^d$ and two multi-indexes $I = (i_1, \dots, i_k)$ and $J = (j_1, \dots, j_l)$ with $i_1, \dots, i_k, j_1, \dots, j_l \in \{1, \dots, d\}$ with possible multiple repetitions. Notice that $S(X)_{0,T}^I$ (resp. $S(X)_{0,T}^J$) is a real number that is the corresponding components of the k^{th} -order Signature tensor (resp. l^{th} -order Signature tensor). The Shuffle Product of two index sets I and J , denoted by $I \sqcup J$, can be easily defined after considering the following examples :

$$\begin{aligned}\{1, 2\} \sqcup \{3\} &= \{\{1, 2, 3\}, \{1, 3, 2\}, \{3, 1, 2\}\} \\ \{1, 2\} \sqcup \{3, 4\} &= \{\{1, 2, 3, 4\}, \{1, 3, 2, 4\}, \{1, 3, 4, 2\}, \{3, 1, 2, 4\}, \{3, 1, 4, 2\}, \{3, 4, 1, 2\}\} \\ \{1, 2\} \sqcup \{2, 1\} &= \{\{1, 2, 2, 1\}, \{1, 2, 2, 1\}, \{1, 2, 1, 2\}, \{2, 1, 2, 1\}, \{2, 1, 1, 2\}, \{2, 1, 1, 2\}\}.\end{aligned}$$

The Shuffle Product is more generally defined as all the ways of merging two index sets while still preserving the order of appearance for each initial set in the result set. More rigorously, any shuffle product can be obtained recursively from the two following basic properties

$$I \sqcup \emptyset = I \tag{4.8}$$

$$Ia \sqcup Jb = \{(I \sqcup Jb)a, (Ia \sqcup J)b\} \tag{4.9}$$

where Ia is the concatenation of the index set I with the singleton $\{a\}$ and Jb is the concatenation of the index set J with the singleton $\{b\}$. Using these definitions, the Shuffle Product Identity is the following identity

$$S(X)_{0,T}^I \cdot S(X)_{0,T}^J = \sum_{K \in I \sqcup J} S(X)_{0,T}^K. \tag{4.10}$$

- Finally, **Chen's Property** completes the presentation of the most elementary properties of Signatures. Let $X : [0, T] \mapsto \mathbb{R}^d$ and $Y : [T, T'] \mapsto \mathbb{R}^d$ be two paths. Define the concatenation of X and Y by

$$X * Y(t) = \begin{cases} X(t) & \text{for } t \in [0, T] \\ Y(t) & \text{for } t \in [T, T']. \end{cases} \tag{4.11}$$

Then, for the concatenation of X and Y , we have

$$S(X * Y)_{0,T'} = S(X)_{0,T} \otimes S(Y)_{T,T'}. \tag{4.12}$$

Chen's identity is often used when modelling the signal using piecewise affine. Given a path $X : [0, T] \mapsto \mathbb{R}^d$ and a set of timestamps $T_\nu = \{t_1, \dots, t_\nu\}$, one defines X_{T_ν} as the piecewise linear path obtained by linearly interpolating the observed values of X at times t_1, \dots, t_ν . Then it is quite easy to compute the Signature of the path using Chen's identify :

$$S(X_{t_\nu}) = \bigotimes_{i=1}^{\nu-1} S(X)_{t_{i+1}, t_i}$$

In our work, we will employ the coefficients of first and second degree of the signature, as well as the **Levy Area**.

Definition 4.1.5. In the case $d = 2$, the Levy Area is given by :

$$A_I(X) = \frac{1}{2} (S_I^{12}(X) - S_I^{21}(X))$$

Geometrically, the Levy area computes the area enclosed by a path X on each side of the $[X(0)X(T)]$ segment as illustrated in Figure 4.2. The Signatures of order k and the

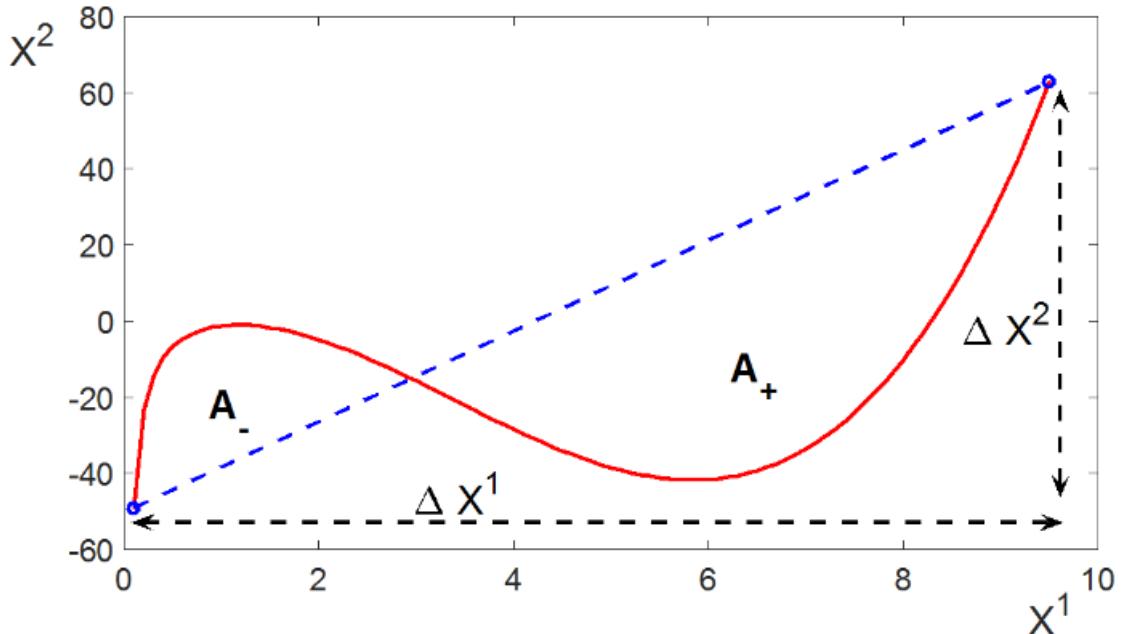


FIGURE 4.2 – Levy Area is computed as $A_- + A_+$

Levy Area will be instrumental for building the detection scheme.

Principal Geodesic Analysis : leveraging Lie group structure in dimension reduction.

Recall that Signatures can be defined on quite general spaces of paths such as the BV -class $\mathcal{BV}(\mathbb{R}^d)$. Consider $S : \mathcal{BV}(\mathbb{R}^d) \rightarrow T(\mathbb{R}^d)$ the application that send a path X of bounded variation to its signature.

A very important fact is that the set $S(\mathcal{BV}(\mathbb{R}^d)) = G(\mathbb{R}^d) \subset T(\mathbb{R}^d)$ has a **Lie group** structure for the tensor product \otimes . The k -level truncated signature space $G^k(\mathbb{R}^d)$ inherits this structure in finite dimension. Further details on this can be found in [47].

Since the truncated signatures are high dimensional objects even for moderate orders, it is important to employ appropriate method for mapping them to lower dimensional spaces. Dimensionality reduction is moreover often seen as a preliminary step before trying to find a relevant clustering technique, adapted to the shape of the clusters. It can be also tailored to the specific structure of the space where the data leaves. In the present case, using vanilla PCA is precluded by the fact that this does not account for the Lie-group structure, since in Lie-groups "the smallest path between A and B " is not a straight line in the euclidean space.

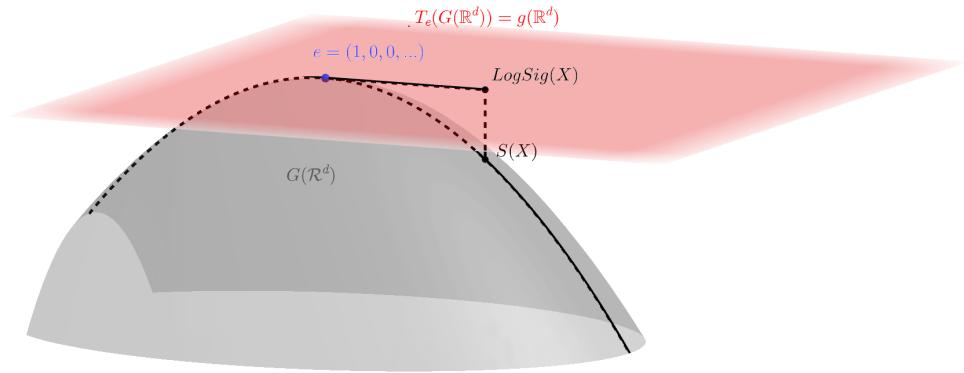


FIGURE 4.3 – A simplified view of Signature space $G(\mathbb{R}^d)$ and its Lie Algebra (\mathbb{R}^d)

One convenient way of building a dimensionality reduction for Lie groups is to project the data onto the associated Lie algebra; see Fig 4.3. This can be done through the log application. The next step is to perform PCA for the projected data, as $\log(G(\mathbb{R}^d))$ is the tangent space to $G(\mathbb{R}^d)$ at the unit element, and therefore, possesses a vector space structure. Another, more refined, option is to resort to Principal Geodesic Analysis (PGA) that finds principal component fitting the Lie group shape; we refer to [82] for further details on using this technique for Signatures. Rather than minimizing the distance to a direction (as a straight line in \mathbb{R}^d), PGA try to minimize the distance to a direction inside the Lie group where the distance is considered along geodesic curves on the Lie Group.

In our setting, we seek such objects in $G(\mathbb{R}^d)$ as the shortest path between a point (to be determined) and the identity signature $(1, 0, 0, \dots)$, see Fig 4.4. An important result is the existence of such geodesics. It is true if one considers the Carnot-Caratheodory norm; see [47] Prop 7.42.

Background on Topological Data Analysis.

In order to unveil the hidden structures and shapes in our data set, Topological Data Analysis (TDA) can become a very relevant tool. We start with n vertices $V = \{v_1, \dots, v_n\}$, each vertex representing a signal dimension.

Definition 4.1.6. *For $k < n$, an abstract k -simplex σ of V is a subset of cardinality $k + 1$ of vertices from V , together with all its subsets. The **geometric realization** of a k -simplex is the convex hull C of $k + 1$ points, such that $\dim(C) = k$ when considered as a affine manifold with boundaries. Any subset of σ forming an l -simplex, for $l \leq k$, will be called an l -dimensional face of σ .*

The role of Topological Data Analysis in our study will become clear if we think in terms of shapes : a malignant trajectory cannot have the same shape as a benign. In a more global point of view, one can think that the shape of malignant space is different from the space of benign one.

In the first case, we create vertices from a single trajectory, using the time-delay embedding, and we compare each created structure with a notion of distance. In the second

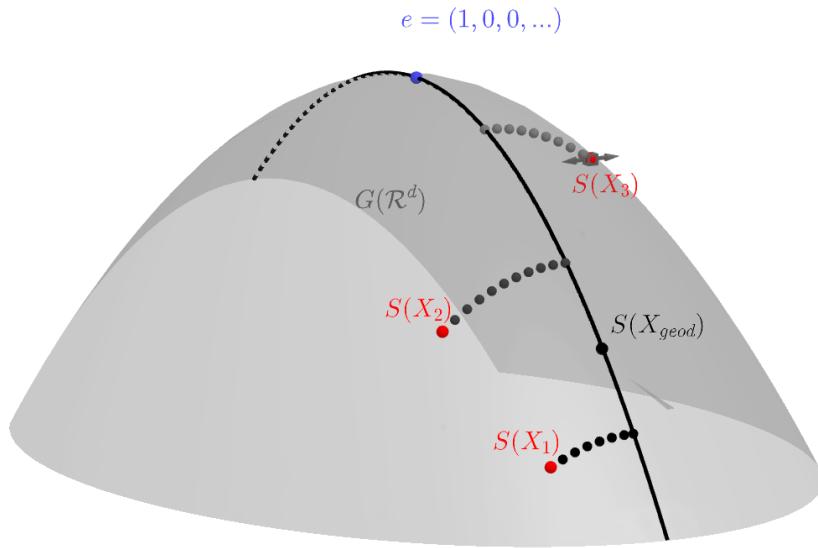


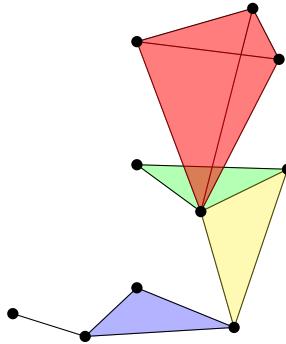
FIGURE 4.4 – Simplification of PGA principle : One has to find X_{geod} such that all dotted curves from points to geodesic linking e to $S(X_{geod})$ are minimal.

case, we consider that each vertex is a whole trajectory, and we create the intrinsic structure carrying either benign or malignant trajectories.

Let us now jump to the definition of simplicial complexes.

Definition 4.1.7. An abstract **simplicial complex** \mathcal{C} on V is a collection of simplices of V such that for every $\sigma^i \in \mathcal{C}$, there exists j with $\sigma^i \cap \sigma^j$ a sub-simplex of both σ^i and σ^j . The **dimension** of \mathcal{C} is the dimension of the highest simplex in \mathcal{C} (i.e. the highest k such that there exist a k -simplex in \mathcal{C}).

From a geometric point of view, \mathcal{C} is constructed by attaching a group of simplices to each other by binding them with a shared face.



A possible algorithm for building a simplicial complex over V (usually in \mathbb{R}^d), and the specific one used in our work, is **Mapper**; see [96]. The principal idea is to unveil a *manifold structure* by an approximation of a **cover**.

Definition 4.1.8. Let \mathbb{M} be a manifold. A **cover** of \mathbb{M} is a family of open sets $\mathcal{U} = (U_i)_{i \in I}$ such that $\mathbb{M} = \bigcup_{i \in I} U_i$

The **nerve** of this cover is a simplicial complex \mathcal{M} over the vertex set $\mathcal{U} = \{U_i, \quad i \in I\}$ with the following rule :

$$\sigma = [U_{i_0}, \dots, U_{i_k}] \in \mathcal{M} \iff \bigcap_{j=0}^k U_{i_j} \neq \emptyset$$

One possible way to construct such a cover, is to chose a relevant **continuous** real valued function f over V , often called *height function*. As a continuous function, $f^{-1}(I)$ for $I =]a; b[$ is an open set of \mathbb{R}^d . For a well-suited number of open intervals $(I_j)_{j \in J}$ with $\bigcup_{j \in J} I_j = \mathbb{R}$, then $(f^{-1}(I_j))_{j \in J}$ is a cover of \mathbb{M} . We can then interpret the nerve as a skeleton for the manifold \mathbb{M} . The main problem faced with using this construction of a simplicial complex is the difficulty of choosing an appropriate height function. Another difficulty is induced by the somewhat biased nature of the proposed construct when based on a single arbitrary height function. A better approach is to compute **the persistent homology** [39] which will be based on spanning different simplicial complexes using a single **scale parameter** for building sequence of increasing neighborhoods for each vertex. One way of achieving this is to build a rigourous mathematical framework based on the notion of **filtration** for simplicial complexes.

Definition 4.1.9. Consider a simplicial complex \mathcal{C} over $V = \{v_1, \dots, v_n\}$. A **filtration** over \mathcal{C} is a (finite) sequence of simplicial complex $(\mathcal{C}_r)_{r \in I}$ such that :

$$V = \mathcal{C}_0 \subset \mathcal{C}_1 \subset \dots \subset \mathcal{C}_r = \mathcal{C}$$

For theoretical consideration about persistent homology, especially **persistent homological class**, see [17, 95, 106]. In the present work, we will leverage the idea that filtrations generate persistence diagrams, and, moreover, two persitence diagrams can be compared using specific distances.

Definition 4.1.10. The **persistence diagram** $PD(\mathcal{F})$ of the filtration $\mathcal{F} = (\mathcal{C}_r)_{r \in I}$ is the set of all (b_i, d_i) with :

- $\#PD(\mathcal{F})$ is the number of persistent homological class \bar{h}_i along the filtration.
- b_i is the first r where \bar{h}_i appear in \mathcal{C}_r .
- d_i is the last r where \bar{h}_i is not merged with another class.

An advantage of persistence diagrams is that we can endow the set of persistence diagrams with several metrics. Let us introduce the appropriate definition of two possible instances.

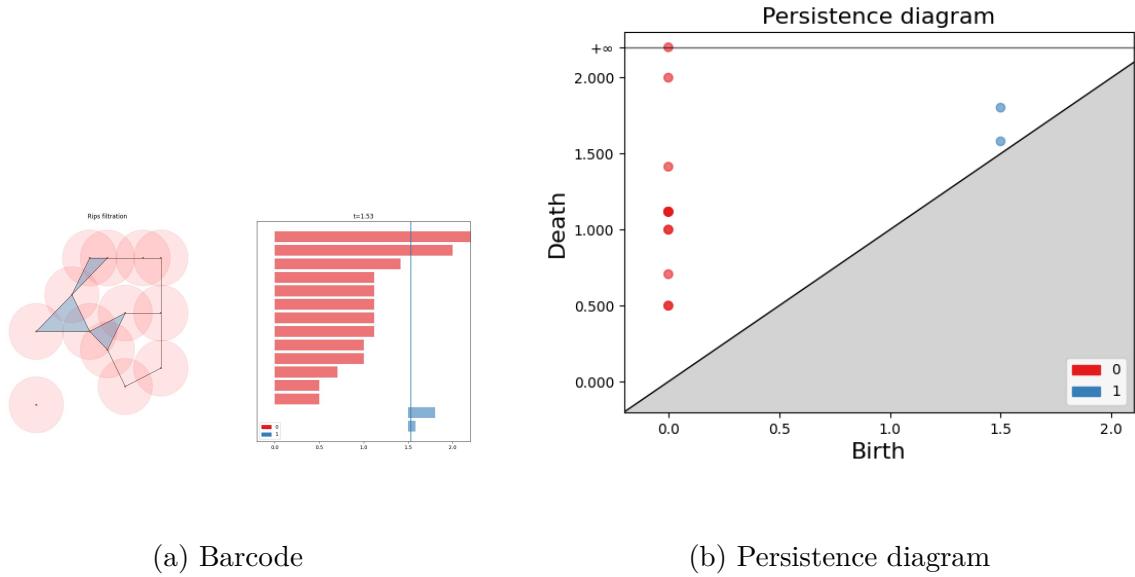
Definition 4.1.11 (Metrics on the space of persistence diagrams). Let dgm_1 and dgm_2 two persistence diagrams. A matching between dgm_1 and dgm_2 is a subset $m \subseteq dgm_1 \times dgm_2$ such that every points in $dgm_1 \cup dgm_2$ appears exactly once in m .

1. The **Bottleneck distance** between dgm_1 and dgm_2 is defined by

$$d_B(dgm_1, dgm_2) = \inf_{m \text{ a matching}} \max_{(p,q) \in m} \|p - q\|_\infty$$

2. The **Wasserstein distance** between dgm_1 and dgm_2 is defined by

$$W_p(dgm_1, dgm_2) = \inf_{m \text{ a matching}} \sum_{(p,q) \in m} \|p - q\|_\infty^p$$



(a) Barcode

(b) Persistence diagram

4.1.5 Our statistical testing procedure

Hypothesis testing via signature coefficients

Under the assumption that no malignant tumor is present, we can make the assumption that the expected level of ctDNA is constant. Moreover, assuming that bloodsamples for a single patient are taken at sufficient distant times, we can assume with a small approximation error that C_t^B has reached its large time distribution and that $C_{t_i}^B$ is independent of $C_{t_{i-1}}^B$. For the sake of simplicity, we will set

$$\Lambda = \frac{B_0 \cdot \lambda_{\text{bn}}}{\varepsilon}.$$

Consider $X_t = (t, C_t^B)$. According to the usual framework of signature computation, we preprocess the signal. Let denote $\{t_0, t_1, \dots, t_n\}$ the sampling times of X , the signals are centered w.r.t. times :

$$\tilde{C}_t^B \leftarrow C_t^B - \mathbb{E}_{t \in \{t_0, \dots, t_n\}}[C_t^B]$$

Another precomputation consists in reducing $[0; T]$ to $[0, 1]$, so that $T = 1$.

Construction of the pivotal variable :

We begin by computing the distribution of the Signature coefficients up to level 2. The first Signature tensor is

$$\left(\int_{t=0}^1 dt = 1, \quad \int_{t=0}^1 dC_t^B = C_1^B - C_0^B \right).$$

We suppose $C_0^B \sim \text{Poisson}_c(0, \Lambda)$ with Poisson_c being the **centered Poisson(Λ) distribution** : if $Y \sim \text{Poisson}(\lambda)$ and Y_c satisfies

$$\mathbb{P}[Y_c = k - \lambda] = \mathbb{P}[Y = k], \quad \forall k \in \mathbb{N},$$

then we say that $Y_c \sim \text{Poisson}_c(0, \lambda)$. Based on this, we deduce that

$$S^2(X) = C_1^B - C_0^B \sim \text{Skellam}(\Lambda, \Lambda)$$

Let us now address the question of computing the 2nd order signature coefficients for the signal X_t . The distribution of the integral of t with respect to the increments of a Poisson process C_t^B , denoted dC_t^B , sampled in $\sigma_1 = \{0 = t_0, \dots, t_n = 1\}$, is given by

$$S_{[0,1]}^{12}(X) = \int_0^1 t dC_t^B = C_1^B - S_{[0;1]}^{21}(X), \quad (4.13)$$

using the shuffle product $S^1 S^2 = S^{12} + S^{21}$. In order to compute $S^{21}(X)$, we use a linear interpolation over σ_1

$$S_{[0;T]}^{21}(X) = \int_0^T C_t^B dt = \sum_{i=1}^n \int_{t_{i-1}}^{t_i} C_t^B dt = \frac{1}{2} \sum_{i=1}^n (C_{t_i}^B - C_{t_{i-1}}^B)(t_i - t_{i-1}). \quad (4.14)$$

From (4.13) and (4.14), it follows that

$$\begin{aligned} 2\mathcal{LA}_{[0;T]}(X) &= S^{12}(X) - S^{21}(X) \\ &= T C_T^B - 2 S^{21}(X) \\ &= T C_T^B - \sum_{i=1}^n (C_{t_i}^B - C_{t_{i-1}}^B)(t_i - t_{i-1}). \end{aligned}$$

Finally, letting $t_i - t_{i-1} = \Delta t$ for all $1 \leq i \leq n$, we conclude that

1.

$$\begin{aligned} S^{12}(X) &= (T - \frac{\Delta t}{2}) C_T^B + \frac{\Delta t}{2} C_0^B \\ &= (1 - \frac{\Delta t}{2}) C_1^B + \frac{\Delta t}{2} C_0^B \sim \text{Poisson}_c(0, \Lambda), \end{aligned}$$

2.

$$\begin{aligned} S^{21}(X) &= \frac{\Delta t}{2} (C_T^B - C_0^B) \\ &= \frac{\Delta t}{2} (C_1^B - C_0^B) \sim 0.5 \text{ Skellam}(\Lambda, \Lambda), \end{aligned}$$

3.

$$\begin{aligned} 2\mathcal{LA}(X) &= ((T - \Delta t) C_T^B + \Delta t C_0^B) \\ &= (1 - \Delta t) C_1^B + \Delta t C_0^B \sim \text{Poisson}_c(0, \Lambda). \end{aligned}$$

Empirical distribution of signature coefficients for benign trajectories.

We now compare the theoretical approximate distributions established in the former paragraph with the empirical distributions obtained using the dataset. The empirical and theoretical distributions are displayed in Figure 4.6 below.

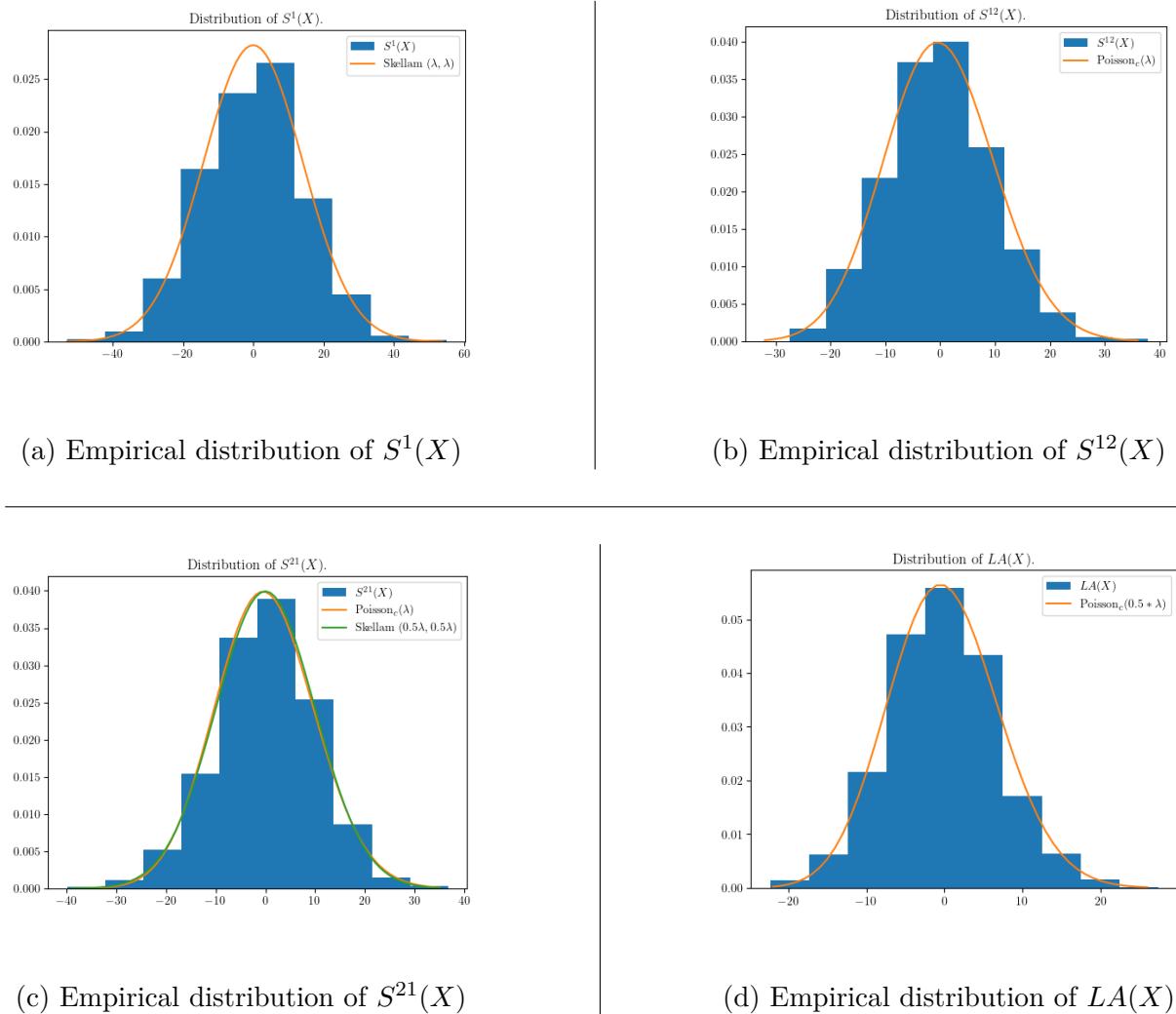


FIGURE 4.6 – Empirical distribution against theoretical distribution.

Our testing scheme :

Based on the explicit form of the distribution of our pivotal quantities under the Null Hypotheses, i.e. Benign state of ctDNA dynamics we obtain the following testing procedure.

Algorithm 4 Testing for S^w with $w \in \{2, 12, 21, 12 - 21\}$

Require: X and $w \in \{2, 12, 21, 12 - 21\}$ (with the notation $12 - 21$ referring to the Levy Area).

Ensure: (H_0) : the tumor is benign.

Compute $\mathcal{C} = \frac{1}{n} \sum_{i=0}^n C_{t_i}$ with n the number of observations of X

Compute $S^w(\tilde{X})$ with $\tilde{X} = (t, C_t - \mathcal{C})$

Compute the p -value corresponding to the observed value of S^w with $\Lambda = \mathcal{C}$

Output : The selected tests for which H_0 is not corroborated by the observed data, using the multiple testing correction proposed by, i.e. Benjamini-Hochberg [5].

4.1.6 Visualisation of the data and analysis of the delineability of the benign/malignant classes

In this section, we present several experiments based on simulated data to illustrate the efficiency of our method.

Data :

We sampled 4000 benign ($\mu \in \text{Params}_\mu = \{0.1, 0.12, 0.14, 0.17\}$, denoted by C_t^B) and 36000 benign then malignant (with a random transition time leading $\lambda \in \text{Params}_\lambda = \{0.04, 0.07, 0.1, 0.15, 0.2, 0.25\}$ and $r = b - d = 0$ to get

$$r \in \mathcal{R} = \{0.001, 0.002, 0.004, 0.007, 0.01, 0.015\},$$

corresponding to an aggressive tumorous dynamic trajectory, denoted by C_t^{BM} . For the sake of completeness, we also sampled pure malignant trajectories, noted C_t^M , see Fig. 4.7.

We used the BirDePy python package to sample 7 observations times from a continuous trajectory. These times are distributed every 300 days, as coefficients for birth and death models are designed for a day by day evolution. We interpret this experiment as exploring the distribution of simulated dynamics with 1 blood sample every year in a 7 years time range.

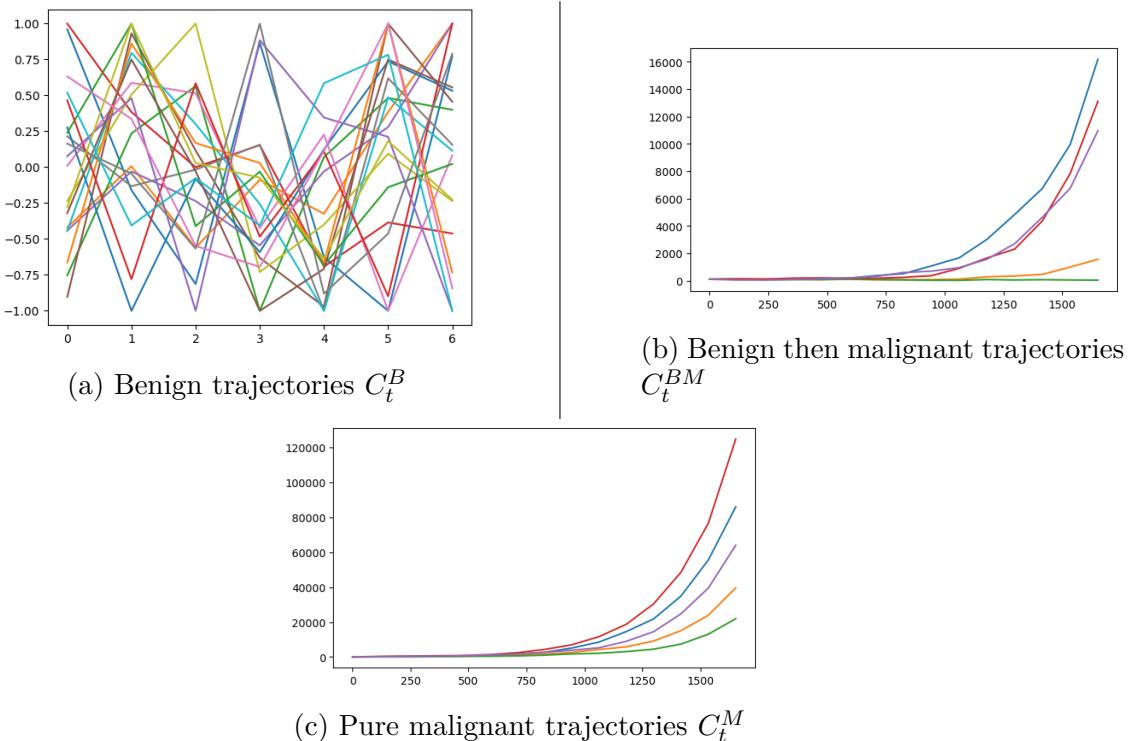


FIGURE 4.7 – Samples of benign trajectories, benign then malignant trajectories and pure malignant trajectories

In this part, we try and find to discover a delineation between benign and malignant trajectories in an appropriate feature space. As the signature transform is an homeo-

morphism, and signature space is a Lie group, it is quite natural to explore topology of signals.

- In the first stage, we compute Rips complex of Time Delay embedding version of signal :

Definition 4.1.12. Let $X : I \rightarrow \mathbb{R}$ be a 1-dimensional signal. The **time delay embedding** of X is the point cloud defined by

$$TD = \{(X_{t_i}, X_{t_{i+1}})\}$$

This embedding allows us to compute a simplicial complex for each signal. Then, we compute a distance matrix between persistence diagrams for both *bottleneck* and *Wasserstein* distance. Finally, we pass this distance matrix through various manifold learning algorithms. Our results are displayed in Figure 4.8¹.

- In a second stage, we tried and compute several Principal Geodesics. First, we display the projections of the benign trajectories ; see Fig.4.9. Next, we display projections of benign and malignant ; see Fig. 4.10
- In the third stage, we use the Mapper algorithm on the Signature set. Concerning the choice of the lens function, we tried Levy Area, $\text{dist}(S(X), \mathbf{1})$ and $\text{dist}(S(X), \hat{\mu}_{\text{Lie}}(X_{\text{ben}}))$; see [33] for a formal definition of μ_{Lie} . Our results are displayed in Figure 4.11. Benign data are labeled by -1 and Malignant data are labeled by indices in the range starting from 0 to 20, depending on the simulations parameters.

Comments on the features visualisations :

We can draw several conclusions from the experiments displayed in Figure 4.8, Figure 4.12 and Figure 4.9. Some of these representations are very encouraging with respect to the possibility of discriminating the two classes.

Nonlinear embeddings : As we can see in Figure 4.8, it is quite clear that MDS, t-SNE and Isomap embedding are able to separate benign from malignant. Moreover, t-SNE seems to separate some of the malignant trajectories with respect to the growing rates.

The Mapper results : Our experiments using Mapper show that this representation of the data is also able to discriminate benign from malignant trajectories through an embedding in signature space. As each lens function produce some kind of filament structure, possibly using more than 1 complex component, one could expect an extension of this filament to multiple rate. Indeed, we confirmed this intuition by modifying some hyper-parameters, and obtained Figure 4.12, in which the colorscale indicates the rate level of tumor growth. Interesting structures can be observed as when varying the rate parameter.

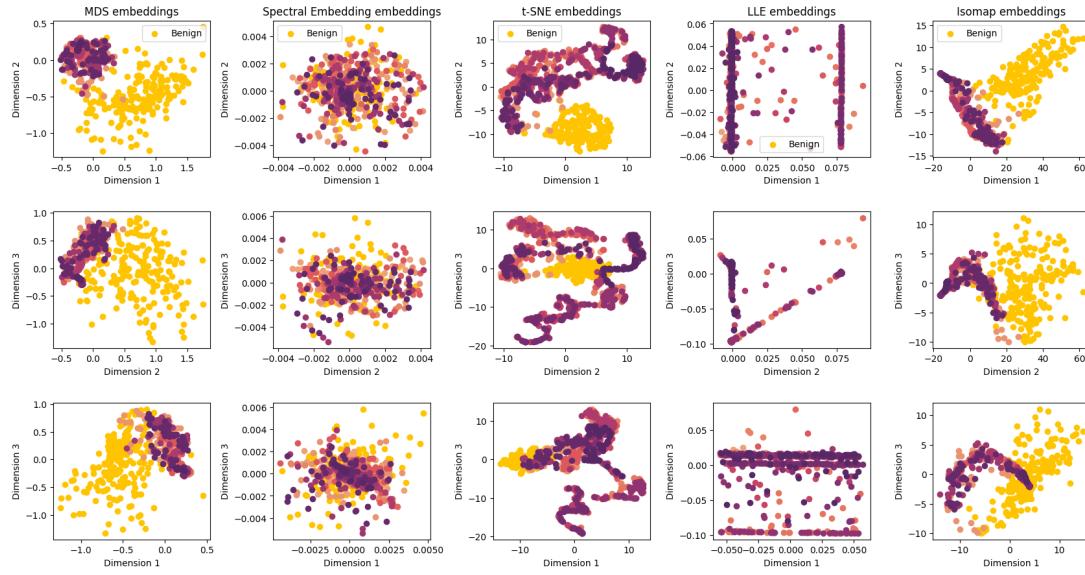
Dimensionality reduction using principal geodesics : Figure 4.9 confirms that a vast majority of malignant signatures can be delineated from the benign ones, which is again a very encouraging observation.

4.1.7 Empirical performance of our testing procedure

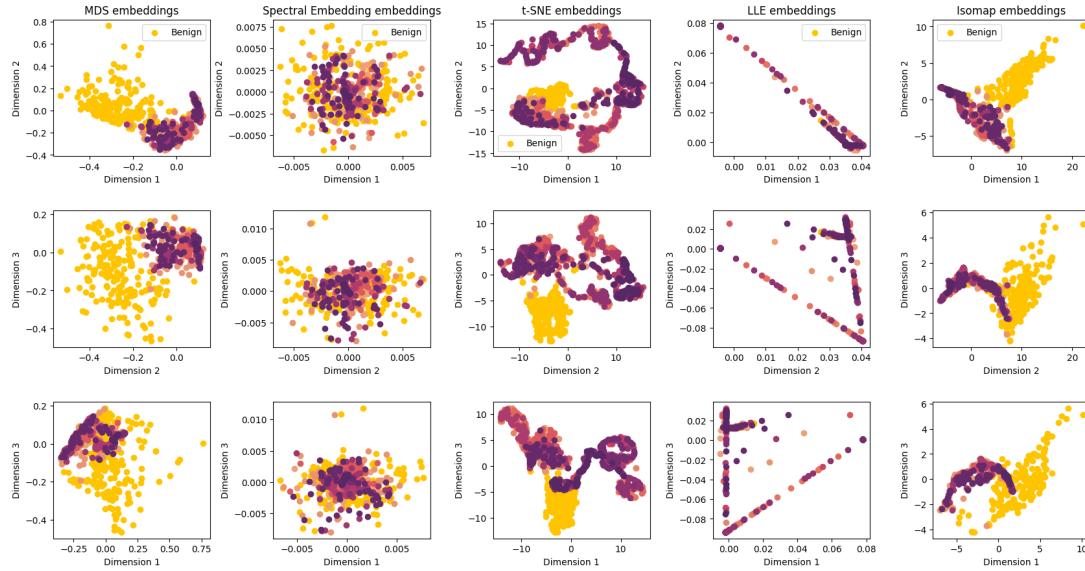
Results

We used different bespoke scores, one for each test. Results are rounded to the third decimal 10^{-3} and are recorded in the following table.

1. More precise figures with the code are available on <https://github.com/RemiVaucher/Thesis>



(a) Embedding with Wasserstein Distance



(b) Embedding with Bottleneck Distance

FIGURE 4.8 – Non linear embeddings of the benign and malignant trajectories.

Table 4.1 shows that all our tests exhibit very high detection capability over simulated datasets.

4.1.8 Future works

We have devised a specific testing methodology for detecting malignant trajectories from irregularly sampled measurements of blood. We have applied our methodology to simulated trajectories with 7 samples over 2000 days.

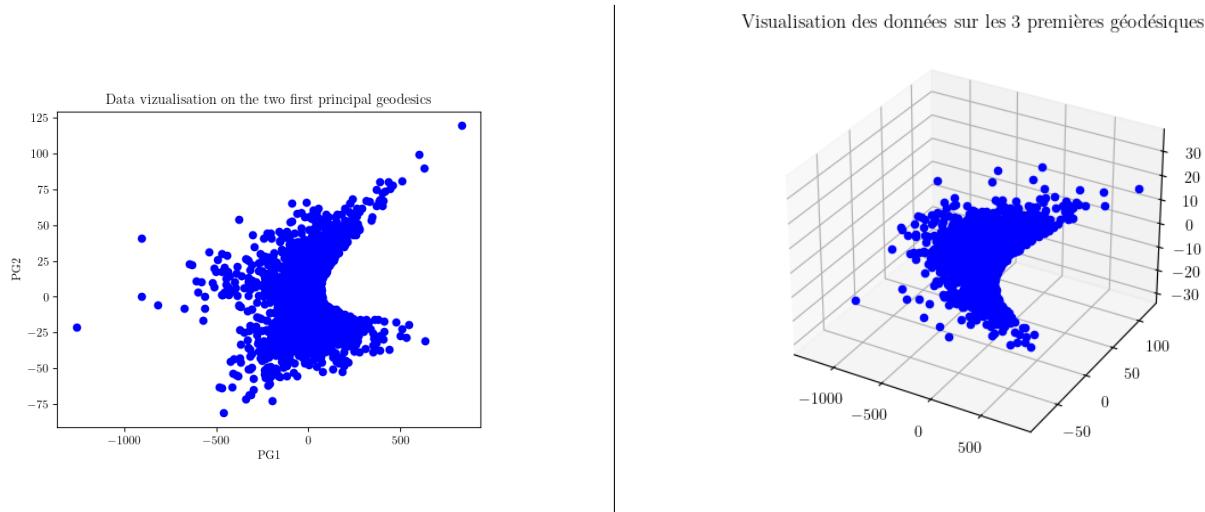


FIGURE 4.9 – Projections of benign trajectories on the two and three first tangent principal geodesic for depth 3.

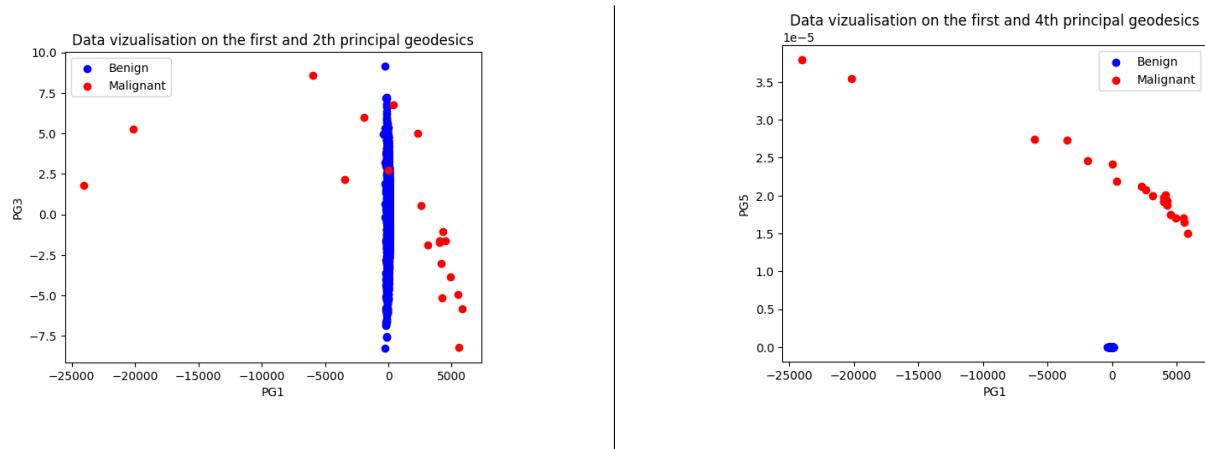


FIGURE 4.10 – Projections of all trajectories on two tangent principal geodesic for depth 3.

TABLE 4.1 – Scores for each coefficient tests.

Test	Accuracy	Precision	Recall	F1-score
T_2	0.946	0.999	0.999	0.999
T_{12}	0.945	0.999	1	0.999
T_{21}	0.943	0.999	0.951	0.974
T_{LA}	0.945	0.999	0.999	0.999
Corrected multiple test	0.958	0.999	0.999	0.999

The main future prospects are :

- *Theoretical power of the tests* : As always with medicine problem, minimizing False Negative is a priority. With this in mind, the first objective would be to find the theoretical power of each test.

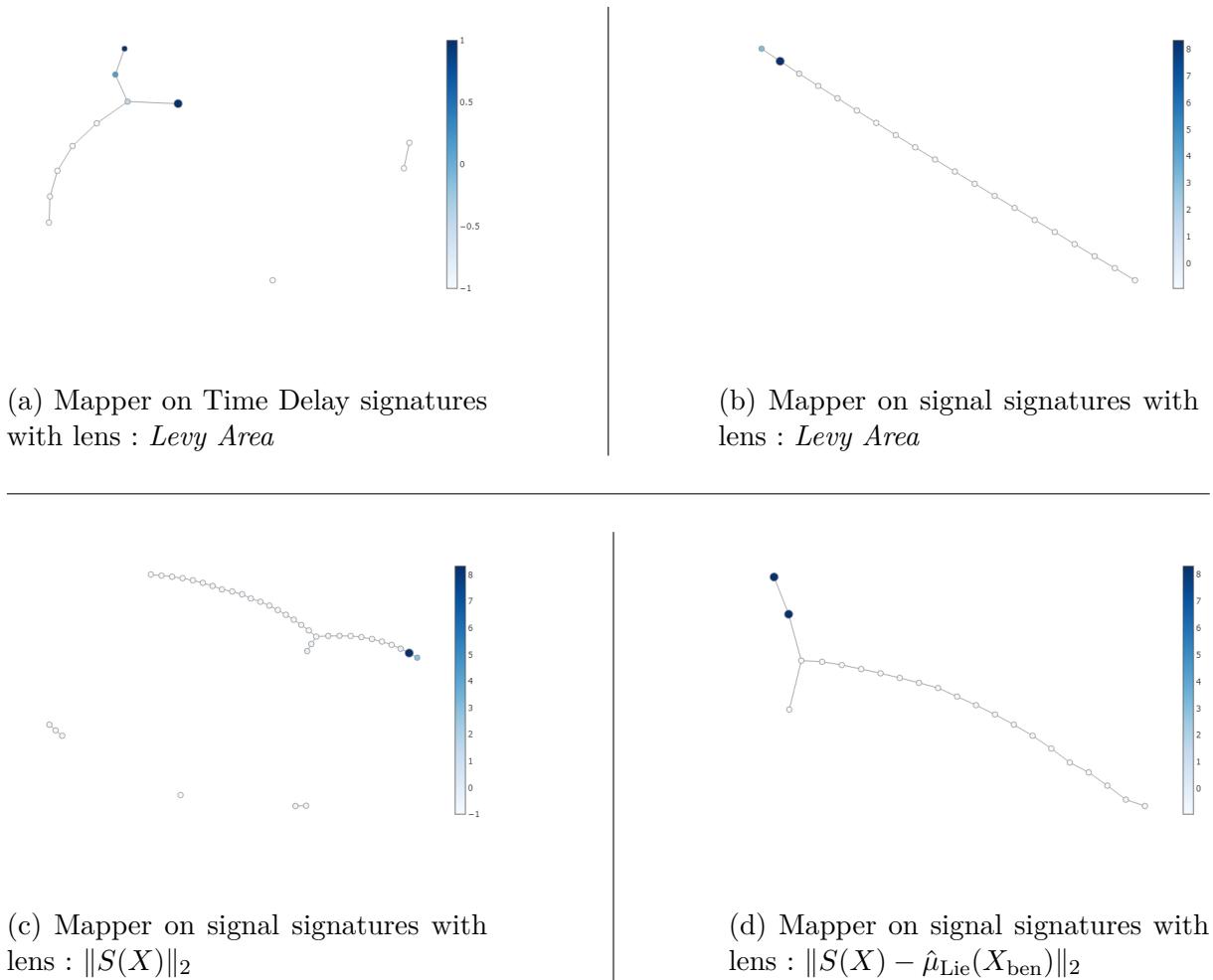


FIGURE 4.11

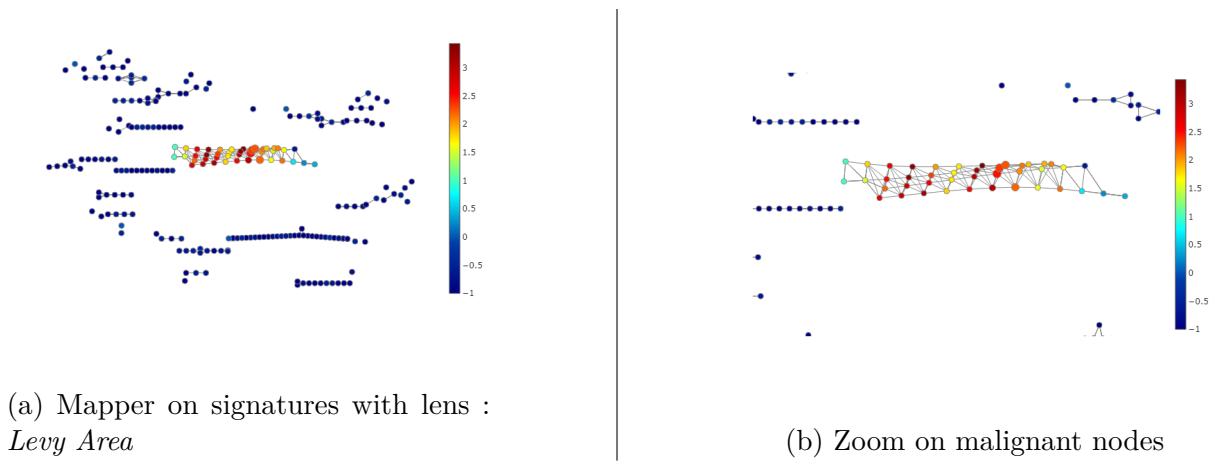


FIGURE 4.12

- *Lowering to less than 5 samples* : 7 samples over 2000 days may fit reality. However, it could be interesting to lower the sample number as much as we can (possibly with keeping the same Δt and lowering T).

- *Computing the distribution with random Δt* : As close as it gets to reality, a patient will never be so regulated on the intervals between two blood samples.
- *Studying the distribution under (H_0) for other coefficient* : We reduced the problem on $S^{(2)}(X)$ in order to simplify theoretical computation. Next step is to generalize these results to every coefficients in $S(X)$.

Acknowledgements. This work was funded by the grant 900R17NEU project "Méthodes neuromorphiques pour la détection précoce de pathologies cancéreuses" from La Région Rhône-Alpes, France (Principal Investigator : Stéphane Chrétien).

4.2 Time topological analysis of EEG using signature theory

4.2.1 Introduction

Topological data analysis [18] is a rapidly growing field that has emerged recently, based on the intriguing observation that data come with shape-like properties. In general, existing topological structures that are built from data, such as Cech or Vietoris Rips complexes, make essential use of a metric that may not be fully suitable because inherited from the space where our data are imbedded instead of the geodesic distance on the manifold where our data truly live. In a recent paper [26], we proposed a novel way of buiding simplicial complexes based on performance in prediction rather than distance. The main tool for performing prediction is the signature transform [22], that easily extract meaningful features from multidimensional signals. In this setting, each node represents a component of the signal and a node belongs to a simplex if the signature of all the nodes from the simplex accurately explains the considered node, statistically speaking. Similarly, a face belongs to a simplex if the signature of the signals represented by the face is accurately predicted by the adjacent faces of various orders in the simplex. In this approach, Signatures faithfully represent faces and simplices together with a natural orientation and make the topological construction better motivated and more statistically meaningful. From a computational viewpoint, selecting faces that explain another one using their respective signatures can be done efficiently using the LASSO, in the spirit of the celebrated method proposed in [79] for estimating graphical models. More recently, [85] solved this problem more precisely by considering a signal correlation matrix. This method allows to create high-dimensionnal structures.

In the present paper, we examine a more refined aspect of our topological construct : the dynamic evolution of the topology as a function of time as complexes undergo potential structural transformations at specific change points in time, reflecting the apperance of certain phenomena. In the area of neuroscience, this approach will be instrumental for detecting change points at which electroencephalograms reflects known "neuroscientific" behaviors.

Signature of rough paths (in a nutshell)

Consider a d -dimensional path $X = (X^1, \dots, X^d) : \mathbb{R} \rightarrow \mathbb{R}^d$. In the following, we will note $S_I^{(k)}(X)$ the k -th degree signature applied to X on an time interval $I = [a, b]$.

The **signature** of X is given by the tensors sequence $S_I^{(k)}(X) \in \underbrace{\mathbb{R}^{d \times d \times \dots \times d}}_{k \text{ times}}$ for all $k \in \mathbb{N}$. We will use the notion of truncated signature of order K to define the sequence of signature tensor $S_I^{(k)}$, $k \leq K$. These tensors are given, for $\{i_1, \dots, i_k\} \subset \{1, \dots, k\}$ by :

$$(S_I^{(k)}(X))_{i_1, \dots, i_k} = S_I^{i_1, \dots, i_k}(X) = \int_{a < t_1 < t_2 < \dots < t_k < b} dX_{t_1}^{i_1} dX_{t_2}^{i_2} \dots dX_{t_k}^{i_k}$$

The signature of a rough path serves as a powerful geometric feature extractor. To compute the signature of a multidimensional signal that is discretely sampled, we must consider linear interpolation between each consecutive time point, thus invoking Chen's theorem :

Theorem 4.2.1 (Chen's identity). *Consider $X : [a, b] \rightarrow \mathbb{R}^d$ and $Y : [b, c] \rightarrow \mathbb{R}^d$. Define :*

$$X * Y = \begin{cases} X_t, & \text{si } t \in [a, b] \\ Y_t - Y_0 + X_b, & \text{si } t \in [b, c] \end{cases}$$

as the **concatenation** of X and Y . Then :

$$S_I^{(k)}(X * Y) = S_I^{(k)}(X) \otimes S_I^{(k)}(Y)$$

To center/rescale the signals, one has to consider the following properties

- For any constant $\gamma \in \mathbb{R}^d$, $S_I^{(k)}(X + \gamma) = S_I^{(k)}(X)$
- For any constant $\lambda \in \mathbb{R}$, $S_i^{(k)}(\lambda X) = \lambda^k S_I^{(k)}(X)$

Furthermore, to ensure a "pseudo" unicity (because it is computationally infeasible to create the whole signature) we will consider the next result :

Theorem 4.2.2. *If $\exists i \in \{1, \dots, k\}$ such that X^i strictly monotonic, then $S_I(X)$ uniquely defines X .*

Finally, defining $\mathcal{C}_{\text{mon}}^{1-var}(I, \mathbb{R}^d)$ the space of d -dimensional signals (linearly interpolated on a subdivision $D(I)$ of I) with $|X|_{1-var} = \sup_{t_i \in D(I)} \sum_i d(X_{t_i}, X_{t_{i+1}}) < \infty$ and

such that $\exists i \in \{1, \dots, d\}$ with X^i strictly monotonic, then the signature transform $S_I : \mathcal{C}_{\text{mon}}^{1-var}(I, \mathbb{R}^d) \rightarrow S_I(\mathcal{C}_{\text{mon}}^{1-var}(I, \mathbb{R}^d))$ is an homeomorphism. [47]

Simplicial complex

Consider a set of d vertices $V = \{X^1, \dots, X^d\}$.

Definition 4.2.1. — For $k < d$, a k -simplex σ_k is a $n + 1$ set and all its subsets [95].

- A geometric realization of σ_k is given by the convex hull E_k of its $n + 1$ points (eventually embedded) in \mathbb{R}^d such that $\dim(E_k) = k$ (so $d \geq k$)

With this tool, one can build a new structure on V [18].

Definition 4.2.2. An abstract simplicial complex \mathcal{C} on a finite vertex set V is a collection $\{\sigma_k, k < d\}$ of simplices such that :

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

The **dimension** of \mathcal{C} is the highest k such there exists a k -simplex in \mathcal{C} .

A simplex $\sigma \in \mathcal{C}$ is called a **face** of \mathcal{C} . The upcoming definition is the most important for our algorithm :

Definition 4.2.3. Consider $k \in N^*$ and σ a k -simplex in a simplicial complex \mathcal{C} . Its **link** $Lk(\sigma, \mathcal{C})$ in \mathcal{C} is the set of all faces $\tau \subset \mathcal{C}$ such that :

- $\sigma \cap \tau = \emptyset$
- $\sigma \cup \tau$ is a face of \mathcal{C}

Remarks :

- The link of a vertex is called **neighborhood** in graph theory.
- In the following algorithm, for a vertex v and a fixed k , we build the subset of all the k -simplex in $Lk(v, \mathcal{C})$ that we call k -dimensional Link (for simplicity).

Some tools for simplicial complex analysis.

The main goal of this subsection is to introduce some useful topological invariants, the first of these being the **Betti numbers** [18]. Informally, each b_k count the number of $k + 1$ -dimensional holes in \mathcal{C} . In our experiments, we focus on 2-dimensional complexes, and thus only consider b_0 (the number of connected components) and b_1 (the number of 2-dimensional holes).

The second invariant that we used is the **Persistence Diagram**. To get a numerical summary of this, we compute the notion of **Persistence Entropy** [23, 92]. A more robust notion of Persistence Summary presented in [32] could be put to work, but the implementation raises much more complex issues that we will not address in the present work.

Definition 4.2.4. Consider a k -dimensional simplicial complex \mathcal{C} . A **filtration** over \mathcal{C} is a sequence $(\mathcal{C}_i)_{0 \leq i \leq n}$ of simplicial complexes such that :

- For all $1 \leq i \leq j \leq n$, $\mathcal{C} \subset \mathcal{C}_j$
- $\mathcal{C}_n = \mathcal{C}$

A natural way of building a filtration is by defining an increasing (non necessary strictly) sequence of time of arrival $(b_i)_{1 \leq i \leq n}$ for each simplex in \mathcal{C} . In Cech complex, the time of arrival correspond to the radius r at which the simplex appear in \mathcal{C} .

Definition 4.2.5. Consider a simplicial complex \mathcal{C} and his persistence diagram DP

$$DP = \{[b_i, d_i], i \in \{1, \dots, \#\mathcal{C}\}\}$$

with b_i and d_i respectively birth and death times for sub-simplex of \mathcal{C} .

Define

$$p_i = \frac{b_i - d_i}{\sum_j b_j - d_j}$$

Persistence entropy (PE) is given by the Shannon entropy of $\{p_i\}$:

$$PE = - \sum_i p_i \log(p_i)$$

Simplistically, the persistence entropy measures the similarity (or dissimilarity) between closure speed of k -dimensional holes : high level PE indicates that all holes are filled at same speed, in opposition to a PE close to 0. Since a quickly filled hole is likely just noise, PE quantifies how many significatives sub-structures lies in \mathcal{C} .

Beginning with subsection 2, we will consider these structures and invariants as evolving as a function of time : for any timestamp $t \in I$, we will consider the simplicial complex $\mathcal{C}(t)$, its associated betti numbers $b_0(t), \dots, b_n(t)$ and its related persistence entropy denoted by $PE(t)$.

Remark : many other features may be extracted from the filtration we introduce for \mathcal{C} . More will be studied in a long version of this contribution.

Our simplicial complex construction algorithm

The algorithm Our algorithm build the simplicial complex on $X = \{X^1, \dots, X^d\}$ by constructing the k -dimensional link (iteratively on k) of a channel X^i . Here is a simplified version of the algorithm [26].

The induced filtration : In order to retrieve the persistence diagram, we need to construct a filtration and so birth time.

- Take all the weights $(\beta_\sigma)_{\sigma \in \mathcal{C}}$ attached to each simplex.
- Create $b_\sigma = 1 - \frac{\beta_\sigma}{\sum_{\sigma \in \mathcal{C}} \beta_\sigma}$.
- For each $\tau, \sigma \in \mathcal{C}$ such that $\tau \subset \sigma$ and $b_\tau > b_\sigma$, then fix $b_\tau = b_\sigma$.

This (non necessarily strictly) increasing sequence ensures that a highly significative simplex appears early in the filtration. In the following subsection, we show how to exploit this filtration.

4.2.2 Empirical study

We performed a set of computational experiments using EEG signals from the CHB-MIT Scalp Database [53]. The main benefit of using real-world signals is to test the stability of our algorithm against a uncontrolled noise.

Algorithm 5 Lasso explicability Simplicial Complex algorithm

Require: Set $\ell = 1$ and set $\mathcal{C}^{(1)} = \{1, \dots, d\}$.

Ensure: The time series interaction simplicial complex

```

while No more simplex is selected do
    Select an (augmented)  $k$ -subset of nodes  $C^J = \{X^{j_1}, X^{j_2}, \dots, X^{j_k}\}$ ,
    with  $J = \{j_1, \dots, j_k\}$  in  $\mathcal{C}^{(\ell)}$ , and compute  $S_{[t,t+L]}(C^J)$ .
    for (augmented)  $k'$ -combination  $C^{J'} = \{X^{j'_1}, X^{j'_2}, \dots, X^{j'_{k'}}\}$  with  $J' = \{j'_1, \dots, j'_{k'}\}$ 
    in  $\overline{\mathcal{C}^{(\ell)}}$  do
        Compute  $S_{[t,t+L]}(C^{J'})$ 
    end for
    Predict  $S_{[t,t+L]}(C^J)$  from  $(S_{[t,t+L]}(C^{J'}))_{J' \cap J = \emptyset}$  with LASSO
    Compute  $S^w(\tilde{X})$ 
    if  $R^2 > 0,67$  then
        Select all non-zero  $\beta_j$  LASSO coefficients
    else
        Set  $\beta_j = 0$  for all  $j$ .
    end if
end while

```

Importance of hyperparameter values

We will evaluate the impact of the choice of the main parameters based on a one-hour trajectory, averaged every second. The signals reflect the occurrence of an epilepsy seizure which t lasts less than one minute (on the interval [49', 51']). We focus our exploration on the LASSO parameters and the sliding Window size for the computation of the Signature. For the sake of simplicity, we will specify the maximal dimension of $\mathcal{C}(t)$ as equal to 2.

LASSO parameters For every channel X^i , we solve the following minimization problem :

$$\min_{\beta \in \mathbb{R}} \|S_I^{deg}(\tilde{X}^i) - \sum_{j \neq i} \beta_j S_I^{deg}(\tilde{X}^j)\|_2^2 + \lambda_1 \sum_{j \neq i} |\beta_j| \quad (4.15)$$

LASSO allows us to select which signatures (and then by homeomorphism which X^j) lies in the neighbourhood of X^i . This selection depends entirely on λ_1 . However, this minimization only creates the 1-dimensional neighbourhood of X^i . to creates the 2-dimensional NH, we solve :

$$\min_{\beta \in \mathbb{R}} \|S_I^{deg}(\tilde{X}^i) - \sum_{j_1, j_2 \neq i, j_1 \neq j_2} \beta_{j_1 j_2} S_I^{deg}(\tilde{X}^{j_1 j_2})\|_2^2 + \lambda_2 \sum_{j \neq i} |\beta_{j_1 j_2}| \quad (4.16)$$

This naturally brings a second parameter.

Experimental results included in the version posted on ArXiv (more precisely, see Figure 3 and 5 from the appendix of the ArXiv version), show that fixing λ_1 and varying λ_2 change b_1 and b_2 either. With λ_2 to close to 0, epileptic seizure is no longer explicit. On another set of data, figure 4.18 show the evolution of betti numbers by moving λ_2 .

Then, fixing λ_2 and varying λ_1 does not give the same type of results : b_1 's trajectory becomes noisier when one enforces sparsity in the LASSO.

Sliding window size for computing the Signature : At each time t we build a simplicial complex \mathcal{C}_t , and for a path P defined on I , we compute $S_{[t-L,t]}^k(P)$ (under the condition $[t-L,t] \subset I$). Length L of I is then a hyperparameter that can impact the topological structure of $\mathcal{C}(t)$. First results show that choosing L is key to safe detection of the sought for patterns.

Experiments with multiple EEG trajectories.

In the proposed experiments, we focus on a multivariate EEG signal sampled during 15 consecutive hours, and that includes 3 seizures. The method's hyperparameters are specified as follow :

- $\lambda_1 = 1$ and $\lambda_2 = 10^{-5}$
- $L = 50$.

The main goal is to retrieve pre-critical and/or critical behaviour based on the Betti numbers and the Persistence Entropy, our main topological invariants of interests. According to [84],

Pre-critical behaviour : are characterized by "area partitioning" (loss of connectivity between neuronal areas, loss of synchrony). Our main hope is that this behaviour is reflected in the trajectory of b_1 : a loss of connectivity should result in poorly interconnected cortical areas, resulting in several stabilized 2-dimensional holes. An expected dynamics is then to observe the steady growth of b_1 in time. It would be relevant to extend the dimension to $k = 3$ in order to see if this fact is confirmed with $b_2(t)$. Another idea would be to characterise every hole in \mathcal{C} and track the stable ones among them.

Critical behaviour : Neuronal populations will abruptly synchronise throughout the duration of the crisis, leading to hypersynchrony before returning to a more 'local' level of synchrony.

In Figure 4.13 below, one observes the global behavior of b_1 and PE on 1 hour of time (second trajectory). We computed an average measure on a clean trajectory (the green line) for both b_1 and PE . The grey area represent the interval $[x(t) - \hat{\sigma}_x, x(t) + \hat{\sigma}_x]$ with $x = b_1$ or PE and $\hat{\sigma}_x$ computed on $[t - h, t]$.

b1 time variation through different configuration

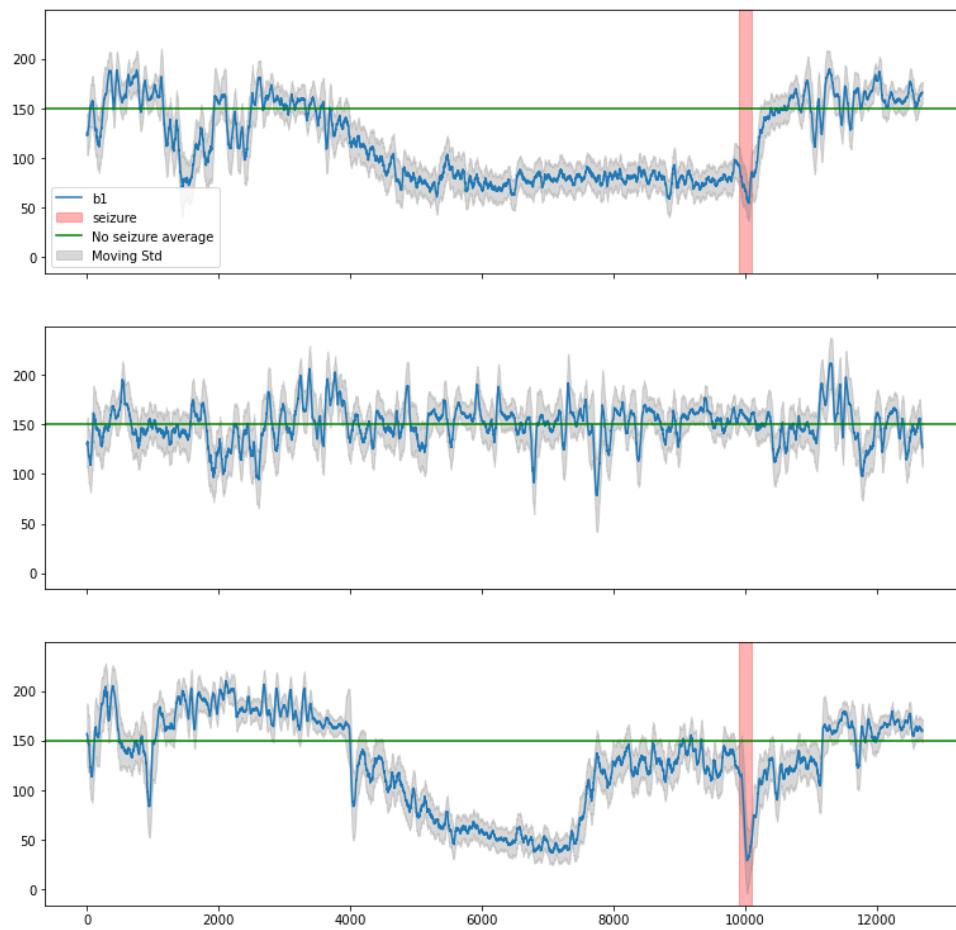


FIGURE 4.13 – $b_1(t)$ on 3 trajectories

Persistent entropy time variation through different configuration

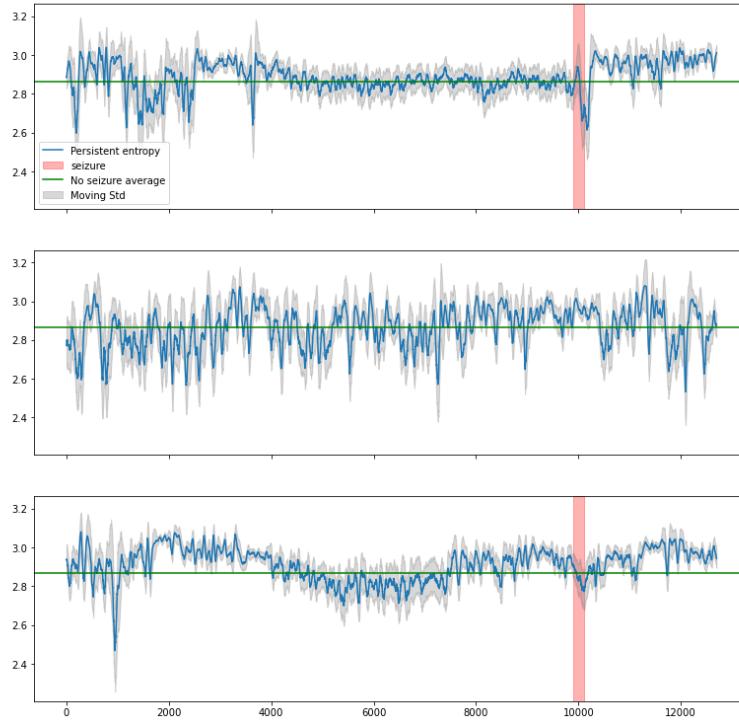


FIGURE 4.14 – $PE(t)$ on 3 trajectories

General behavior (second trajectory) : We see a volatility around the average value along time, with small perturbations that we cannot explain. The volatility goes from 100 to 200 for b_1 and from 2.6 to 3 for PE .

Critical behavior (red area in first and third trajectory) : The critical behavior is characterized by an abrupt diminution to 50 in b_1 's trajectory (more connected structure in the simplicial complex before getting back to near the average value. Unfortunately, it cannot really be distinguished from PE .

Pre-critical behavior (6000 times before critical) : The precritical behavior is characterised by a steep decrease of b_1 (between 50 and 100) on the same time interval for the two seizure trajectories. A sudden increase in noticeable just before the next seizure, whilst remaining under the average value.

Heuristically, we can observe a decrease of the PE 's standard deviation during the precritical stage. We plan to extend this study in a future work.

Discussion

Many improvements can still be made to the proposed methodology, in particular (i) using Knockoff filters to guarantee that the faces are selected with confidence, (ii) using SLOPE instead of the LASSO, (iii) using higher-order sub-complexes, etc. One may also use the path $(b_1(t), PE(t))$ and its signature as well for extracting more sensitive detection pipelines.

4.3 Etude de données acoustiques.

4.3.1 Séparation des anomalies

Les données étudiées proviennent de [55, 88]. On dispose de 100 enregistrements .wav dont 50% sont anormaux (pour chaque dataset).

L'objectif ici est de voir expérimentalement si la filtration associée aux coefficients LASSO est consistante avec la détection d'anomalie. Pour constater cela, nous appliquons la même procédure qu'à la section 4.1 :

- Préparation des données : transformation des données soit par spectrogramme Mel soit par Short Fourier Transform.
- Création d'un complexe simplicial sur chacune des données, puis calcul des diagramme de persistance.
- Création d'une matrice de distance entre les diagrammes de persistance ou les complexes.
- Création de plongement des données grâce à la matrice de distance suivant plusieurs méthode de Manifold Learning.

Les distances utilisées sont : Bottleneck Distance ou Wasserstein distance pour les diagrammes de persistance, et distance entre les tenseurs d'adjacences pour les complexes.

Résultats : Plusieurs faits émergent de cette analyse.

- Les hyperparamètres jouent un rôle prépondérant dans l'efficacité de l'algorithme (voir 4.15 et 4.16).
- De même, une configuration donnée ne parvient pas à effectuer la séparation sur tous les jeux de données (voir fig 4.17).
- La précision de la séparation dépend majoritairement du nombres de canaux de la décomposition et des coefficients de parcimonie du LASSO.

Ces résultats, quoique préliminaires, restent encourageant. De plus amples investigations seront menées ultérieurement.

4.3.2 Conclusion

Les premiers résultats tendent à montrer que la filtration induite par les coefficients du LASSO est assez fine pour séparer des données normales de données anormales. Pour autant, la recherche des paramètres adaptés à cette tâche peut s'avérer fastidieux.

Il serait intéressant de voir si une filtration construite plus intelligemment permettrait une économie de temps. Les pistes envisagées actuellement sont :

- S'affranchir du paramètre λ en utilisant un LASSO avec validation croisée.
- Dans le cadre supervisée : rajouter un algorithme de prediction, par exemple une régression logistique, puis déterminer le meilleur coefficient de parcimonie par validation croisée.

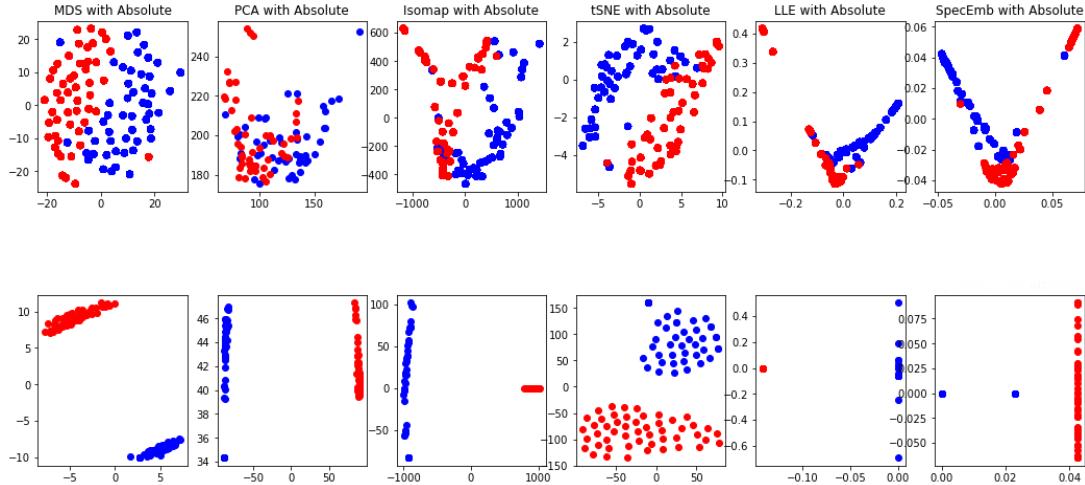


FIGURE 4.15 – Projection des données dans l'espace des complexes simpliciaux avec la distance absolue pour 2 jeux d'hyperparamètres différents. (Bleu :Normal, Rouge :Anormal, données MMII)

— Créer une nouvelle filtration dont le paramètre temporel serait exactement le paramètre λ . La figure 4.18 permet de se rendre compte de ce potentiel en faisant varier le coefficient λ_{2D} .

Cette dernière idée est plutôt intuitive : la variation de λ influence la taille du support de $\hat{\beta}$ et donc la taille du Link de chaque sommet. Seul un problème persiste : la suite de complexes ainsi obtenue n'est pas garantie d'être croissante.

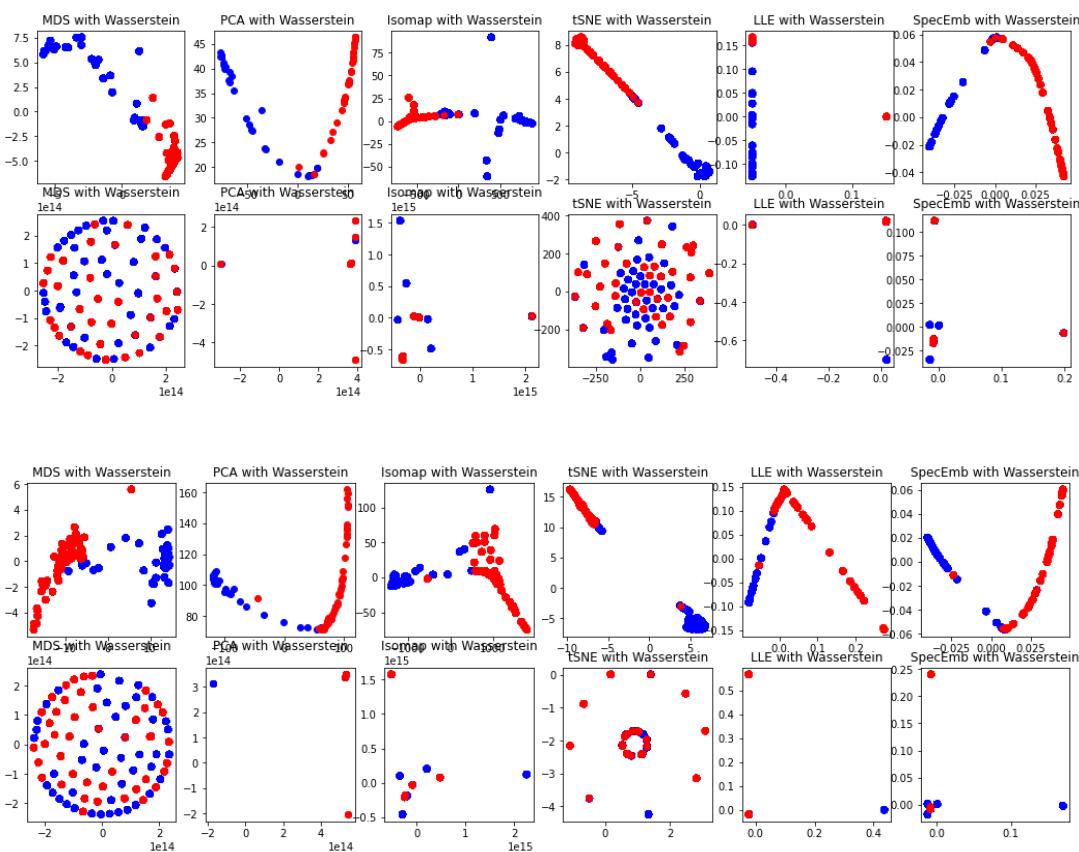


FIGURE 4.16 – Projection des données dans l'espace des diagrammes de persistance pour la distance de Wasserstein avec 2 jeux d'hyperparamètres différents. (Bleu :Normal, Rouge :Anormal, données MMII)

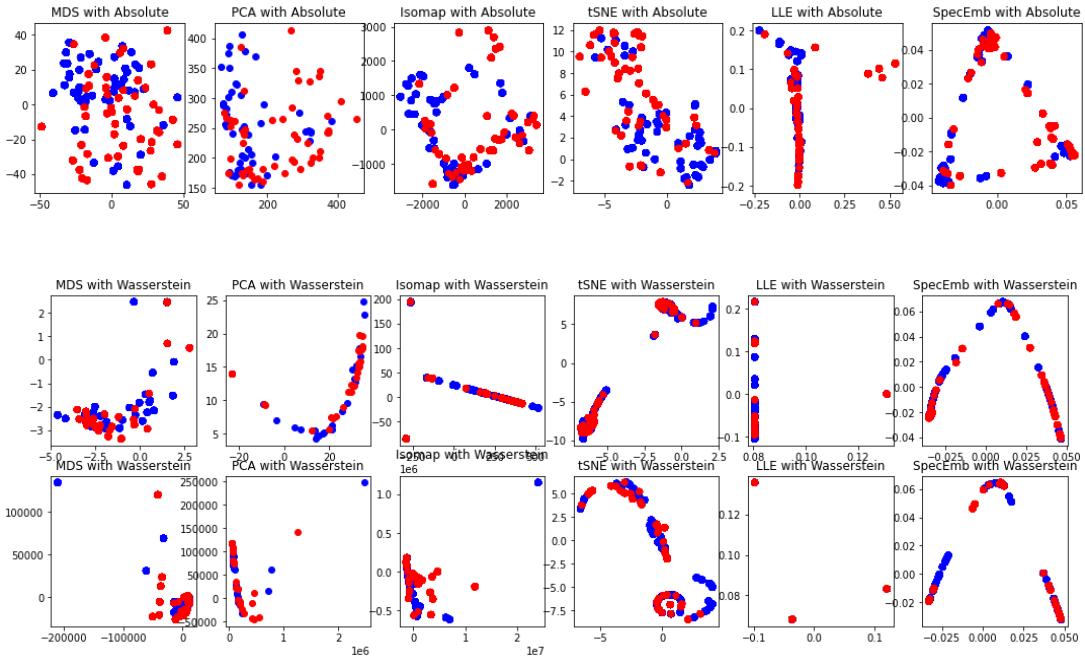


FIGURE 4.17 – Projection des données dans l'espace des diagrammes de persistance avec 1 jeu de paramètres ne permettant pas la séparation. (Bleu :Normal, Rouge :Anormal, données MMII)

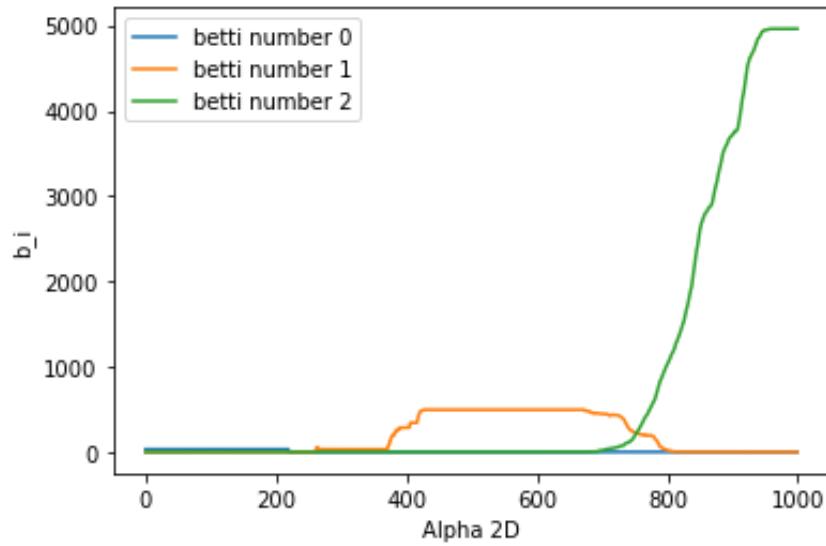


FIGURE 4.18 – Evolution des nombres de Betti b_0 , b_1 et b_2 en fonction du coefficient λ_{2D}

Chapitre 5

Générer des signaux multivariés sous une loi donnée à partir des signatures : le problème de l'inversion.

*Des siècles et des siècles que tous les
enchanteurs courrent après, hé ben
c'est bibi qui a trouvé !*

Merlin,
Kaamelott Livre III, Episode 32

5.1 Introduction

Dans le cadre de la détection d'anomalie, il peut être intéressant de disposer d'un outil de génération de données sans anomalies :

- soit pour enrichir le jeu de données actuel,
- soit pour reconstruire les anomalies contextuelles et ainsi fournir un pipeline immédiat de détection d'anomalie (voir [103] pour les anomalies sur images).

Pour cela, une idée actuellement étudiée de multiple manière est d'utiliser l'espace des signatures pour générer des signaux temporels multivariés [3, 72]. Quelle que soit l'approche, un problème difficile doit être abordé, celui d'inverser l'application calculant la Signature afin de reconstruire le chemin auquel il correspond. L'inversibilité elle même est un problème délicat et a été étudié avec profondeur dans [54], mais aussi plus récemment dans [13] avec une approche différente utilisant le transport optimal. Ce problème d'inversion a été traité dans plusieurs publication jusqu'à présent [3, 69, 86].

Dans [86], la technique utilisée permet de reconstruire une approximation du chemin dont uniquement le degré 3 de Signature est connu. Dans ce chapitre, nous nous basons sur ce point de vue en étudiant les limites statistiques, puis en proposant une amélioration.

Plan du chapitre : Dans cette première section, nous exposons et étendons les résultats publiés dans [28]. Dans ce travail, nous étudions le phénomène d'anti-concentration

des coefficients de signature, permettant de quantifier les fluctuations incontournables de l'estimateur de la Signature lorsque le chemin est perturbé par une certaine forme de bruit. Pour y remédier, nous modifions la fonction objectif de l'algorithme présenté dans [86] en intégrant un estimateur robuste aux outliers suivant l'approche appelée Median of Means [68].

Dans la section suivante, nous proposons une modification du problème d'optimisation utilisé en y incluant un certain de nombres de contraintes sensées faciliter l'obtention du minimum visé.

5.2 Signature estimation and signal recovery using Median of Means

5.2.1 Signature of paths from their coefficients in a dictionary

The Signature Transform [22] of general multivariate signals has recently become a tool of prominent importance for the analysis of multidimensional dynamical phenomena which are pervasive in many applications of machine learning. Signatures are formal series of tensors of increasing degree which capture meaningful features about the various time dependencies of groups of components of increasing cardinal, and it can apply to settings where the signal is known through irregular sampling with possibly unsynchronised components.

In practice however, the signals under study are often corrupted by additive observation noise, which can be modelled by

$$X = X^* + \epsilon. \quad (5.1)$$

When X^* (resp. X) are only observed at a finite number of sampling times, denoted by $\{t_1, \dots, t_\nu\}$, we will assume that X^* (resp. X) can also be considered as continuous paths using linear interpolation between the successive observed positions. For the sake of simplicity, we will assume that ϵ is a Gaussian white noise process.

Our study will address in particular

- the problem of estimating the Signature tensor of X^* and
- the problem of recovering the original signal X^* in an appropriate basis, based on the Signature tensor of a subsampled noisy version of X^* .

When the signal is corrupted by an observation noise, the computed signature may not be as accurate as expected for estimating the Signature of the true signal. In the first part of this section, we study how much of anticoncentration the noisy Signature inherits from the observation noise of the signal. Our results intend to raise awareness about the impact of noise when using Signatures for downstream Machine Learning tasks. In the second part, we show how to use a new technique from the field of Robust Statistics, named Median of Means, in order to build a robust version of the Signature tensor.

5.2.2 Anti-concentration for the 3-Signature coefficients

One standard approach to estimating $S^{(3)}(\mathbb{E}[X])$, i.e. the signature of the original signal is to solve the following least-square regression problem :

$$\min_{A \in \mathbb{R}^{n \times m}} \|S^{(3)}(X) - \llbracket C; A, A, A \rrbracket\|_F^2. \quad (5.2)$$

where $A \in \mathbb{R}^{d \times m}$, $C \in \mathbb{R}^{m \times m \times m}$ and

$$\llbracket C; A, A, A \rrbracket_{\alpha, \beta, \gamma} = \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m c_{ijk} a_{\alpha, i} a_{\beta, j} a_{\gamma, k} \quad (5.3)$$

Using that, after [59], for an orthogonal basis ψ ,

$$\epsilon = E\psi \quad (5.4)$$

where E is a i.i.d. Gaussian matrix, and using the expansion

$$X^* = \mathbb{E}[X] = A^*\psi \quad (5.5)$$

of X^* in the basis ψ , we obtain

$$\begin{aligned} S^{(3)}(X^* + \epsilon) &= S^{(3)}(A^*\psi + E\psi) \\ &= S^{(3)}((A^* + E)\psi) \\ &= \llbracket C; A^* + E, A^* + E, A^* + E \rrbracket. \end{aligned}$$

For the sake of making the problem finite dimensional, we will further approximate $S^{(3)}(X^* + \epsilon)$ with $S^{(3)}(XT_\nu^* + \epsilon T_\nu)$.

Expanding the expression of the Signature tensors

To simplify, we will note

- $\llbracket C \rrbracket_{A+E} = \llbracket C; A+E, A+E, A+E \rrbracket$
- $\llbracket C \rrbracket_{A+E, \alpha, \beta, \gamma} = \llbracket C; A+E, A+E, A+E \rrbracket_{\alpha, \beta, \gamma}$

First, it is easy to see that

$$\begin{aligned} \llbracket C \rrbracket_{A+E} &= \llbracket C; A, A, A \rrbracket + \llbracket C; E, E, E \rrbracket \\ &\quad + \sum_{i=0}^2 (\llbracket C; \sigma^i(A, A, E) \rrbracket + \llbracket C; \sigma^i(A, E, E) \rrbracket) \end{aligned} \quad (5.6)$$

with $\sigma = (1, 2, 3) \in S_3$. So each coefficients from $\llbracket C \rrbracket_{A+E}$ takes the form of a polynomial of coefficients $e_{\alpha, i}$ with $\alpha \in \llbracket 1; n \rrbracket$, $i \in \llbracket 1, m \rrbracket$.

We deduce from (5.3) and (5.6) :

$$\llbracket C \rrbracket_{A+E, \alpha, \beta, \gamma} = P_1(E) + P_2(E) + P_3(E) + R = P(E) \quad (5.7)$$

where :

- $P_1(E) = \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m c_{ijk} e_{\alpha, i} e_{\beta, j} e_{\gamma, k}$
- $P_2(E) = \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m c_{ijk} (a_{\alpha, i} a_{\beta, j} e_{\gamma, k} + a_{\alpha, i} e_{\beta, j} a_{\gamma, k} + e_{\alpha, i} a_{\beta, j} a_{\gamma, k})$
- $P_3(E) = \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m c_{ijk} (a_{\alpha, i} e_{\beta, j} e_{\gamma, k} + e_{\alpha, i} e_{\beta, j} a_{\gamma, k} + e_{\alpha, i} a_{\beta, j} e_{\gamma, k})$
- $R = \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m c_{ijk} a_{\alpha, i} a_{\beta, j} a_{\gamma, k}$

Anti-concentration of coefficients

We use here Theorem 1.8 from [80] or Theorem 1.2 from [73] depending on multilinearity (or not) of P (i.e. depending on $\alpha \neq \beta \neq \gamma$ or not) :

- The case $\alpha \neq \beta \neq \gamma$ (P is multilinear) :

Theorem 5.2.1. *There is an absolute constant B such that the following holds. Let ξ_1, \dots, ξ_n be independent (but not necessarily iid) random variables. Let P be a polynomial of degree d with the form*

$$P(\xi_1, \dots, \xi_n) = \sum_{S \subset \{1, \dots, n\}, |S| \leq d} a_S \prod_{i \in S} \xi_i$$

whose rank $r \geq 2$. Assume that there are positive numbers p and ε such that for each $1 \leq i \leq n$, there is a number y_i such that $\min \{\mathbf{P}(\xi_i \leq y_i), \mathbf{P}(\xi_i > y_i)\} = p$ and $\mathbf{P}(|\xi_i - y_i| \geq 1) \geq \varepsilon$. Assume furthermore that $\tilde{r} := (p\varepsilon)^d r \geq 3$. Then for any interval I of length 1

$$\mathbf{P}(P(\xi_1, \dots, \xi_n) \in I) \leq \min \left(\frac{Bd^{4/3}(\log \tilde{r})^{1/2}}{(\tilde{r})^{1/(4d+1)}}, \frac{\exp(Bd^2(\log \log(\tilde{r})^2))}{\sqrt{\tilde{r}}} \right) \quad (5.8)$$

The application here is simple. P is here a multilinear polynomial of degree $d = 3$, and all $e_{\alpha,i}$ are independants. The existence of p and ε such that, for all $e_{\alpha,i}$ (with $(\alpha, i) \in [\![1, n]\!] \times [\![1, m]\!]$) there exist $y_{\alpha,i}$ verifying :

$$\min \{\mathbb{P}(e_{\alpha,i} \leq y_{\alpha,i}), \mathbb{P}(e_{\alpha,i} \geq y_{\alpha,i})\} = p \quad (5.9)$$

and

$$\mathbb{P}(|e_{\alpha,i} - y_{\alpha,i}| \geq 1) = \varepsilon \quad (5.10)$$

depends on the $e_{\alpha,i}$ distributions.

Finally, the rank r depends on the dictionary $\psi : r = \#\{c_{ijk}, |c_{ijk}| > 1\}$. For the last hypothesis ($\tilde{r} \geq 3$), we need $r \geq \frac{3}{(p\varepsilon)^3}$. Under these assumptions on ψ and $a = \{a_{\alpha,i}\}$, the equation (5.8) applies.

- The two other cases :

For the two other cases, as the polynomials are no more multilinear, we need an anti-concentration inequality for multivariate polynomials. With this in mind, we use Theorem 5.2.2. First, we need to define PSD anti-concentration :

A distribution \mathcal{D} has *PSD* anti-concentration if there exist $C, c > 0$ such that the following holds. Let A be an $n \times n$ positive semi-definite matrix with $\text{Tr}(A) = 1$. Then for any $\varepsilon > 0$,

$$\mathbb{P}_{\mathbf{x} \in \mathcal{D}^n} [\mathbf{x}^t A \mathbf{x} \leq \varepsilon] \leq C \cdot \varepsilon^c.$$

We now fix a notation for sum and subtraction of independent variables for the same distribution :

Define $d\mathcal{D} := \mathcal{D} + \dots + \mathcal{D}$ to be the distribution of the sum of d independent elements sampled from D , and $\mathcal{D} - \mathcal{D}$ to be the distribution of the difference of two independent elements sampled from D .

And now the theorem :

Theorem 5.2.2. *Let \mathcal{D} be a distribution over \mathbb{R} such that $\mathcal{D} - \mathcal{D}$ has PSD anti-concentration. Then there exist $C_d, c_d > 0$ such that the following holds. Let $f(\mathbf{x}) = f(x_1, \dots, x_n)$ be a degree d polynomial, normalized to have $\text{Var}_{(d\mathcal{D})^n}[f] = 1$. Then for any $t \in \mathbb{R}$ and $\varepsilon > 0$,*

$$\mathbb{P}_{\mathbf{x} \sim (d\mathcal{D})^n}[|f(\mathbf{x}) - t| \leq \varepsilon] \leq C_d \cdot \varepsilon^{1/c_d}, \quad (5.11)$$

where $c_d = O(d \cdot 2^{O(d)})$.

Under this assumption, which is clearly satisfied for i.i.d. Gaussian matrices, on

$$E = \{e_{\alpha,i}, e_{\beta,j}, e_{\gamma,k}, \quad \alpha, \beta, \gamma \in \llbracket 1, n \rrbracket, \quad i, j, k \in \llbracket 1, m \rrbracket\}, \quad (5.12)$$

we can apply (5.11).

5.2.3 Estimation of the signal decomposition using Median of Means (MoM)

We now turn to the problem of estimating the Signature coefficients.

For this purpose, we will first consider the general problem of estimating the expectation $\mu_P = P[X]$ of a distribution P from the observation of an i.i.d. sample $\mathcal{D}_N = (X_1, \dots, X_N)$ of real valued random variables with common distribution P .

In this part, we will note ϵ for the noise of an observation (so $X = X^* + \epsilon$) to avoid confusion with Rademacher variables.

The MoM principle

Let K and b such that $N = Kb$ and let B_1, \dots, B_K denote a partition of $\{1, \dots, N\}$ into subsets of cardinality b . For any $k \in \{1, \dots, K\}$, let $P_{B_k}X = b^{-1} \sum_{i \in B_k} X_i$. The MoM estimators of μ_P are defined by

$$\text{MOM}_K[X] \in \text{median}\{P_{B_k}X, k \in \{1, \dots, K\}\}.$$

Recall that the Rademacher complexity of a class \mathcal{F} of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ is defined by

$$D(\mathcal{F}) = \left(\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i f(X_i) \right\} \right] \right)^2. \quad (5.13)$$

where $\xi_i, i = 1, \dots, N$ are independant ± 1 Rademacher random variables.

Theorem 5.1

(Concentration for suprema of MoM processes). Let \mathcal{F} denote a separable set of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that $\sup_{f \in \mathcal{F}} \sigma^2(f) = \sigma^2 < \infty$, where $\sigma^2(f) = \text{Var}(f(X))$. Then, for any $K \in \{1, \dots, N/2\}$,

$$\mathbb{P} \left(\sup_{f \in F} |\text{MoM}_K[f] - Pf| \geq 128 \sqrt{\frac{D(\mathcal{F})}{N}} \vee 4\sigma \sqrt{\frac{2K}{N}} \right) \leq e^{-K/32}. \quad (5.14)$$

Application to signal decomposition

Let $X = X^* + \epsilon$ be a continuous noisy observation of X^* , i.e. corrupted by a continuous noise ϵ on the interval $[0, T]$. and C_Ψ be the third degree signature of an orthogonal basis ψ on $[0, T]$. We will assume that C_Ψ is computable without subsampling ; otherwise, we can approximate C_Ψ by subsampling the basis functions in Ψ as well. Given a positive integer ν , and a random set of timestamps T_ν , our goal is to estimate the quantity

$$\mathbb{E}_{T_\nu, \epsilon} \left[\|S^{(3)}(X_{T_\nu}^* + \epsilon_{T_\nu}) - [C_\psi; A, A, A]\|_F^2 \right] \quad (5.15)$$

In order to put the MoM principle to work, we need n samples of the variable

$$Y = \|S^{(3)}(X_{T_\nu}^* + \epsilon_{T_\nu}) - [C_\psi; A, A, A]\|_F^2. \quad (5.16)$$

Let $\{T_\nu^{(1)}, T_\nu^{(2)}, \dots, T_\nu^{(N)}\}$ be a set of N sets of timestamps with same distribution as T_ν . For all $i = 1, \dots, N$, define

$$Y^{(i)} = \|S^{(3)}(X_{T_\nu^{(i)}}^* + \epsilon_{T_\nu^{(i)}}) - [C_\psi; A, A, A]\|_F^2. \quad (5.17)$$

Notice that the vectors $\epsilon^{(i)}$, $i = 1, \dots, N$ with

$$\epsilon^{(i)} = \begin{bmatrix} \epsilon_{t_1^{(i)}}^{(i)} \\ \vdots \\ \epsilon_{t_\nu^{(i)}}^{(i)} \end{bmatrix} \quad (5.18)$$

are i.i.d. and therefore, the signals

$$X_{T_\nu^{(i)}}^* + \epsilon_{T_\nu^{(i)}}, \quad (5.19)$$

$i = 1, \dots, N$ are i.i.d. as well. Based on these assumptions, we can use the MoM approach to estimate the expectation (5.16) Instead of estimating (5.16) using the mean of $Y^{(i)}$, $i = 1, \dots, N$, we will turn to the Median of Means technique. Let $N = Kb$, Writing $P_{B_k} Y = \frac{1}{b} \sum_{i \in B_k} Y^{(i)}$:

$$\begin{aligned} MOM(Y) &= \text{median} \{P_{B_k} Y, k \in \{1, \dots, K\}\} \\ &= \text{median} \left\{ \frac{1}{b} \sum_{i \in B_k} \|S^{(3)}(X_{T_\nu^{(i)}}^* + \epsilon_{T_\nu^{(i)}}) - [C_\Psi; A, A, A]\|_F^2, k \in \{1, \dots, K\} \right\} \end{aligned} \quad (5.20)$$

In order to apply Theorem 5.1, we need to compute $D(\mathcal{F})$, where

$$\begin{aligned} \mathcal{F} = \left\{ (\epsilon, T) \mapsto \|S^{(3)}(X_T^* + \epsilon)\|_F^2 - 2 \langle S^{(3)}(X_T^* + \epsilon), [C_\Psi; A, A, A] \rangle \right. \\ \left. + \| [C_\Psi; A, A, A] \|_F^2 \right\}. \end{aligned} \quad (5.21)$$

Thus, we have

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i f(X_i) \right\} \right] &\leq \mathbb{E} \left[\sup_{A \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^n} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i \left(\|S^{(3)}(b + \epsilon)\|_F^2 \right. \right. \right. \\ &\quad \left. \left. \left. - 2 \langle S^{(3)}(b + \epsilon), [C_\Psi; A, A, A] \rangle + \| [C_\Psi; A, A, A] \|_F^2 \right) \right\} \right] \end{aligned} \quad (5.22)$$

conditioning on ϵ gives

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i f(X_i) \right\} \right] &\leq \mathbb{E} \left[\mathbb{E} \left[\sup_{A \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^n} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i \left(\|S^{(3)}(b + \epsilon)\|_F^2 \right. \right. \right. \right. \\ &\quad \left. \left. \left. \left. - 2 \langle S^{(3)}(b + \epsilon), [C_\Psi; A, A, A] \rangle + \| [C_\Psi; A, A, A] \|_F^2 \right) \right\} \right] \mid \epsilon \right] \\ &= \mathbb{E}[\mathcal{R}_\epsilon(\mathcal{F}) \mid \epsilon] \end{aligned} \quad (5.23)$$

In (5.23), $\mathcal{R}_\epsilon(\mathcal{F})$ is the empirical Rademacher complexity. In order to bound this quantity, it is important to note that (5.16) is the squared Frobenius norm of a matrix where every coefficient is a degree 3 polynomial of ϵ . Hence, the final quantity is a degree 6 polynomial in the variable ϵ . Rademacher complexity of polynomials has been addressed in earlier sources such as, e.g. [83]. At this point, we are referring to [30] to “convert” a polynomial function to a polynomial network and [105] for the Rademacher complexity.

Consider $\mathcal{E} = (\epsilon_1, \dots, \epsilon_n)$ a set of n i.i.d. samples from the same distribution as $\epsilon \in \mathbb{R}^d$. Let \mathcal{C} denote the event

$$\mathcal{C} = \{\|\epsilon_k\|_\infty \leq 1, k \in \{1, \dots, n\}\} \quad (5.24)$$

Then, it is proved in [105] that there exist two constants μ and λ such that

$$\mathcal{R}_{\mathcal{E}}(\mathcal{F}) \leq 2\mu\lambda\sqrt{\frac{12\log(d)}{n}} \quad (5.25)$$

on \mathcal{C} . It follows from (5.23) that

$$\begin{aligned} D(\mathcal{F}) &\leq \mathbb{E} [\mathbb{E}[\mathcal{R}_{\mathcal{E}}(\mathcal{F}) \mid \mathcal{E}]] \\ &\leq \mathbb{E} [\mathbb{E}[\mathcal{R}_{\mathcal{E}}(\mathcal{F}) \mid \mathcal{E}] \mid \mathcal{C}] \mathbb{P}(\mathcal{C}) + \mathbb{E} [\mathbb{E}[\mathcal{R}_{\mathcal{E}}(\mathcal{F}) \mid \mathcal{E}] \mid \overline{\mathcal{C}}] \mathbb{P}(\overline{\mathcal{C}}) \\ &\leq 2\mu\lambda\sqrt{\frac{12\log(d)}{n}}\mathbb{P}(\mathcal{C}) + \frac{c}{\sqrt{n}} \end{aligned} \quad (5.26)$$

where we used that $\sqrt{n}\mathbb{E} [\mathbb{E}[\mathcal{R}_{\mathcal{E}}(\mathcal{F}) \mid \mathcal{E}] \mid \overline{\mathcal{C}}] \mathbb{P}(\overline{\mathcal{C}})$ can be shown to be bounded by a constant using a peeling argument. Combining (5.1), (5.26), we obtain the following result



Theorem 5.2

Let Y be defined by (5.16), $Y^{(I)}$, $i = 1, \dots, N$ be defined (5.17), and the MoM estimator be defined by (5.20). Then, we have

$$\mathbb{P} \left(\sup_{Y \in \mathcal{F}} |\text{MOM}_K[Y] - PY| \geq 128 \sqrt{\frac{2\mu\lambda\sqrt{\frac{12\log(d)}{n}}\mathbb{P}(\mathcal{C}) + \frac{c}{\sqrt{n}}}{N}} \vee 4\sigma\sqrt{\frac{2K}{N}} \right) \leq e^{-K/32}. \quad (5.27)$$

5.3 Inversion de la signature : pistes vers une amélioration de l'algorithme Pfeffer-Seigal-Sturmfelds (PSS).

Dans cette section, n'ayant pour l'instant pas fait l'object d'une publication, nous nous intéressons maintenant au problème inverse d'estimer un signal connaissant sa Signature d'ordre fini. Notre objectif ici est de trouver une manière d'améliorer l'algorithme Pfeffer-Seigal-Sturmfelds (PSS) de manière à :

- *Fonctionner sur des chemins bruités* : comme nous l'expliquons ci-dessus, le problème (5.2) est sujet à une propriété d'anti-concentration. Pour répondre à ce problème, nous apportons une réponse en terme d'optimisation en modifiant la fonction objectif pour y inclure Median of Means.
- *Ajuster la fonction objectif à la structure de groupe de $G^N(\mathbb{R}^d)$* : La fonction objectif (5.2) s'affranchit de la structure de $G^N(\mathbb{R}^d)$ pour viser une fonction d'erreur classique (en effet, le shuffle product empêche la stabilité par somme). Nous proposons une fonction objectif plus adaptée à la structure de groupe, incluant les paradigmes de l'algorithme PSS.
- *S'affranchir des restrictions sur la dimension pour le dictionnaire de fonctions Ψ* : En effet, les résultats présenter dans [86] ne peuvent éviter l'impossibilité d'augmenter le nombre de fonctions m constituant la base, au delà du nombre de canaux d du chemin visé. Nous étudions ce problème et regardons si nos modifications permettent de s'affranchir de cette contrainte embarrassante.

- *S'affranchir de la restriction au plus court chemin :* Comme cet algorithme n'utilise "que" le degré 3 de signature, l'application Signature perd sa bijectivité, et donc l'uniquité du minimum global. Pour contourner ce problème, [86] et d'autres algorithmes d'inversion de la Signature se focalise sur le chemin de plus courte longueur dont la Signature tronquée correspond à celle visée. Nous réfléchissons à une nouvelle manière de viser une longueur désirée.

5.3.1 Ajuster la fonction objectif à la structure de groupe de $G^N(\mathbb{R}^d)$

Pour tenter de contourner cette difficulté, nous nous basons sur [47, Proposition 7.18] que nous restreignons à $G^N(\mathbb{R}^d)$:

Proposition 5.3.1

On considère g_n et g deux éléments de $G^N(\mathbb{R}^d)$. Alors :

$$\lim_{n \rightarrow +\infty} \|g_n - g\| = 0 \iff \lim_{n \rightarrow 0} \|g_n^{-1} \otimes g - \mathbf{1}\| \rightarrow 0$$

Revenons à notre problème de départ. Si on définit $\pi_3(g) : G^N(\mathbb{R}^d) \rightarrow (\mathbb{R}^d)^{\otimes 3}$ la projection canonique sur le tenseur d'ordre 3, le problème d'optimisation devient

$$\begin{aligned} (5.2) &= \min_{A \in \mathbb{R}^{d \times m}} \left\| \pi_3 \left(S^N(X) \otimes \left(\bigoplus_{k=0}^N [\![S^{(k)}(\Psi); A, A, \dots, A]\!] \right)^{-1} \right) - (\pi_3 \mathbf{1}) \right\|_F^2 \\ &= \min_{A \in \mathbb{R}^{d \times m}} \left\| \pi_3 \left(S^N(X) \otimes \left(\bigoplus_{k=0}^N [\![S^{(k)}(\Psi); A, A, \dots, A]\!] \right)^{-1} \right) \right\|_F^2 \end{aligned} \quad (5.28)$$

De plus, en utilisant le fait que $g^{-1} = (1+a)^{-1} = \sum_{k=0}^N (-1)^k a^{\otimes k}$ (avec $a = \bigoplus_{i=1}^N g^{(i)}$), nous obtenons deux réécritures de (5.28) :

$$\begin{aligned} \left(\bigoplus_{k=0}^N [\![S^{(k)}(\Psi); A, A, \dots, A]\!] \right)^{-1} &= S^N(X \star A \overleftarrow{\psi}) \\ &= \sum_{i=0}^N (-1)^k \left(\bigoplus_{k=0}^N [\![S^{(k)}(\Psi), A, \dots, A]\!] \right)^{\otimes i} \end{aligned}$$

On en déduit alors que :

$$5.28 = \min_{A \in \mathbb{R}^{d \times m}} \left\| \pi_3 \left(S^N(X \star A \overleftarrow{\psi}) \right) \right\|_F^2 \quad (5.29)$$

$$= \min_{A \in \mathbb{R}^{d \times m}} \left\| \pi_3 \left(S^N(X) \otimes \sum_{i=0}^N (-1)^k \left(\bigoplus_{k=0}^N [\![S^{(k)}(\Psi), A, \dots, A]\!] \right)^{\otimes i} \right) \right\|_F^2 \quad (5.30)$$

Nous cherchons donc à ce que tous les coefficients d'un tenseur de Signature donné atteignent 0. Néanmoins, par le Shuffle Product, si tous les coefficients de degré k sont nuls, alors tous les tenseurs de degrés inférieurs sont identiquement nuls, et c'est en particulier vrai pour $k = 3$. Par ailleurs, les fonctions objectif (5.29) et (5.30) peuvent se généraliser au cas où on considère π_k , pour des valeurs de k supérieures à 3. Considérons la base suivante :

$$\psi_i(t) = \begin{cases} 0 & \text{si } t \leq \frac{i-1}{m} \\ mt - (i-1) & \text{si } \frac{i-1}{m} < t < \frac{i}{m} \\ 1 & \text{si } t \geq \frac{i}{m} \end{cases}$$

qui est la base des chemins linéaires par morceaux. Le tenseur $S^{(k)}(\Psi)$ est très facilement calculable en amont, et permet de choisir la valeur de k à sa guise.

5.3.2 S'affranchir de la restriction au plus court chemin

Comme notre fonction objectif (5.2) n'implique qu'un faible degré de Signature, la bijectivité n'est donc pas assurée. Il existe donc de multiples minima globaux : plusieurs chemins peuvent avoir la même signature de degré 3. En utilisant un degré de signature plus élevé, on peut s'attendre à obtenir une reconstruction plus proche de l'original si les problèmes numériques associés sont bien conditionnés. Pour autant, nous devrons continuer à faire face à la non-unicité du minimum global. Une parade est de d'utiliser la propriété que, pour un degré fixé, il existe un **unique** chemin de **plus courte longueur**. C'est pour cette raison qu'une large majorité des algorithmes d'inversion utilise une pénalisation assurant la reconstruction du chemin le plus court.

Dans notre cas, nous aimerais suivre une autre direction, celle d'assurer plutôt que le chemin reconstruit soit de longueur probable plutôt que de plus courte longueur. Une des étapes consiste à créer une méthode permettant de générer des chemins sous une loi donnée **par inversion de la Signature**. Dans cette optique, en plus de jouer sur l'algorithme d'optimisation, nous allons créer une augmentation du chemin de manière à ce que la Signature intègre naturellement l'information sur la longueur :

$$\tilde{X} = (\text{Length}(X|_{[0,t]}), X_t) \in \mathbb{R}^{d+1}$$

où $\text{Length}(X|_{[0,t]})$ représente la longueur du chemin X entre les points X_0 et X_t . Cette augmentation a le bon goût d'amener une nouvelle dimension croissante monotone **intrinsèquement liée** à X . Evidemment, cette augmentation peut être accompagnée de l'augmentation par le temps pour s'assurer une dimension strictement monotone. Maintenant, rappelons que, dans notre contexte où il existe $A \in \mathbb{R}^{d \times m}$ telle que $X = A\Psi$ avec $\Psi \in (BV([0, T], \mathbb{R}^d))^m$ une base de linéarisation par morceaux, alors :

$$\text{Length}(X) = \text{Length}(A\Psi) = \sum_{i=1}^m \sqrt{\sum_{j=1}^d a_{ij}^2}$$

Mais

$$S^{111}(\tilde{X}) = \frac{1}{6} (\text{Length}(X|_{[0,T]}) - \text{Length}(X|_{[0,0]}))^3 = \frac{1}{6} \text{Length}(X)^3.$$

On peut donc tenter de rajouter une contrainte sur $S^{111}(X)$. Notre problème d'optimisation (5.29) peut donc se formuler comme :

$$\begin{cases} \min_{A \in \mathbb{R}^{d \times m}} \|\pi_3(S^N(\tilde{X} \star A \overleftarrow{\Psi}))\|_F^2 \\ S^{111}(X) = \left(\sum_{i=1}^m \sqrt{\sum_{j=2}^{d+1} a_{ij}^2} \right)^3. \end{cases} \quad (5.31)$$

Maintenant que nous avons établi une contrainte de longueur, nous allons définir des contraintes de forme.

- **Contraintes de bord :** Les coefficients diagonaux des différents tenseurs de signatures correspondent à des puissances de la variations totales :

$$S^{\overbrace{i i \dots i}^{k \text{ fois}}}(X) = \frac{1}{k!} (X_T^i - X_0^i)^k$$

On peut donc exprimer très facilement ces coefficients en fonction de A :

$$(S^{ii\dots i}(X))_{1 \leq i \leq d} = f_k(A(\Psi(T) - \Psi(0)))$$

où $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ envoie x_i sur $\frac{1}{k!}x_i^k$

- **Contrainte d'aire de Levy :** Le shuffle product nous autorise à récupérer les coefficients de l'aire de Levy à partir de la signature de degré 3 : comme $S^i(X) = (S^{iii}(X))^{1/3}$, on en déduit, sous la condition $S^i = S^i(X) \neq 0$ que :

$$LA^{jk} = S^{jk} - S^{kj} = (S^{ijk} + S^{jik} + S^{ikj} - S^{ikj} - S^{kij} - S^{kji})/S^i, \quad \forall i \text{ t.q. } S^i \neq 0.$$

Quitte à considérer le chemin $\tilde{X} = (t, X_t)$, $t \in [0, 1]$, on peut considérer $i = 1$ et $S^1(X) = 1$. Une nouvelle contrainte s'impose donc :

$$LA = [\![S^2(\Psi); A, A]\!] - [\![S^2(\Psi); A, A]\!]^T.$$

Au final, le problème devient :

$$\begin{cases} \min_{A \in \mathbb{R}^{d \times m}} \|\pi_3(S^N(\tilde{X} \star A \overleftarrow{\Psi}))\|_F^2 \\ S^{111}(X) = \left(\sum_{i=1}^m \sqrt{\sum_{j=2}^{d+1} a_{ij}^2} \right)^3 \\ \text{Diag}(S^{(3)}(X)) = f_k(A(\Psi(T) - \Psi(0))) \\ LA = [\![S^2(\Psi); A, A]\!] - [\![S^2(\Psi); A, A]\!]^T \end{cases}$$

5.4 Expériences

Nous avons tenté de reconstruire plusieurs types de chemins :

- Deux chemins bidimensionnels sur $[0, 1]$:

$$X_t = (\text{Length}(f|_{[0,t]}), f(t))$$

avec $f(t) = \cos(5t) + \varepsilon_t$ et $f(t) = \exp\left(-(\frac{5}{t} - 0.5)^2\right) + \varepsilon_t$, avec $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ pour tout $t \in [0, 1]$

- Un cercle

$$(\cos(t) + \varepsilon_t, \sin(t) + \nu_t)$$

avec $\varepsilon_t, \nu_t \sim \mathcal{N}(0, \sigma^2)$

- Un mouvement brownien uni- et bi-dimensionnel :

$$dX_t \sim \mathcal{N}(0, \Sigma^2), \quad \forall t \in [0, 1].$$

Dans l'optique où l'on voudrait générer des séries temporelles à partir d'une base de données connues, on peut connaître à l'avance le nombre de temps d'échantillonnage et leur répartition. On souhaite donc en pratique reconstruire X sur l'ensemble $\{t_0, \dots, t_m\}$. Comme on reconstruit une interpolation linéaire par morceaux du chemin, la base considérée est :

$$\Psi = (\psi_1, \dots, \psi_m)$$

avec

$$\psi_i = \begin{cases} 0 & \text{si } t \leq \frac{i-1}{m} \\ mt - (i-1) & \text{si } \frac{i-1}{m} \leq t \leq \frac{i}{m} \\ 1 & \text{si } t \geq \frac{i}{m} \end{cases}.$$

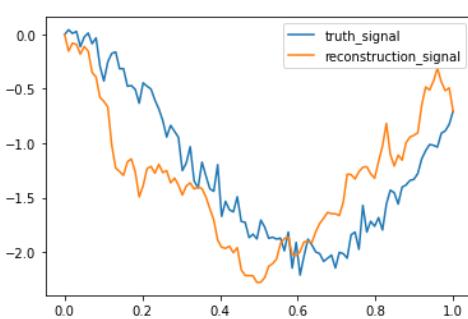
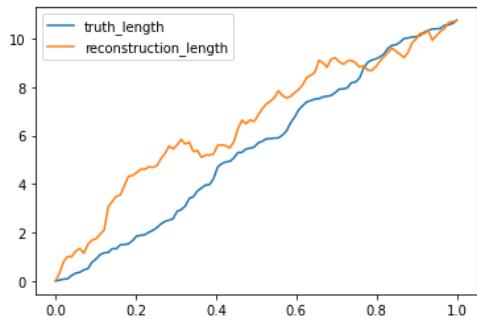
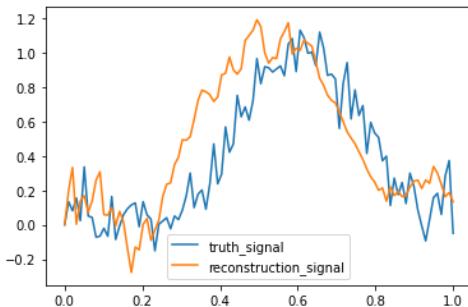
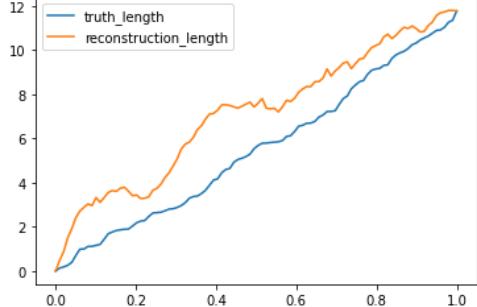
Nous allons optimiser la fonction :

$$\begin{aligned} \mathcal{L}(A) = & \left\| \pi_3 \left(S^{(3)}(\tilde{X}) \otimes \bigoplus_{i=0}^3 [\![S^{(3)}(\Psi^{-1}); A, A, A]\!] \right) \right\|_F^2 \\ & + \lambda_1 \left| S^{111}(\tilde{X}) - \sum_{j=1}^m \sqrt{\sum_{i=2}^{d+1} a_{ij}^2} \right| \\ & + \lambda_2 \left\| \text{Diag}\left(S^{(3)}(\tilde{X})\right) - \left[\left(A(\Psi(1) - \Psi(0)) \right)^3 \right]_{1 \leq i \leq d+1} \right\|_2 \\ & + \lambda_{\text{Ridge}} \|A\|_2. \end{aligned}$$

Nous laissons pour l'instant de côté la contrainte de forme qui a tendance à faire diverger l'algorithme d'optimisation. Nous utilisons l'algorithme Adam et le laissons évoluer pendant 2×10^5 étapes. Les Figures 5.1 à 5.4 montrent nos résultats préliminaires. Un notebook permettant de tenter l'aventure est disponible sur <https://remivaucher.github.io/manuscrit>. L'imperfection de la reconstruction est notable mais nous pouvons déjà observer que notre approche avance dans la bonne direction, ce que nous considérons comme un succès étant donnée la difficulté du problème¹ :

1. dont le nombre d'heures passé à trouver une fonction de coût pertinente est pour nous un témoin

- La reconstruction reproduit bien (pour des valeurs pertinentes des hyperparamètres) la forme globale des chemins sans la contrainte de forme, et donc uniquement grâce à la Signature de degré 3.
- La reconstruction paraît être une version un peu plus lisse que la version originale, ce qui est rassurant, les Signatures de plus haut degré permettant de capturer les oscillations les plus subtiles.
- L'écart entre le départ et l'arrivée est bien reconstruit.
- La courbe de longueur semble bien reconstruite.
- La moyenne des écarts temps par temps est centrée en 0. (voir la Figure 5.5)

(a) $\cos(t)$ (b) $\sin(t)$ FIGURE 5.1 – Reconstruction d'une fonction $\cos(t)$ bruitée.(a) $\cos(t)$ (b) $\sin(t)$ FIGURE 5.2 – Reconstruction d'une fonction $\exp(t)$ bruitée.

5.5 Perspectives

Plusieurs pistes sont encore à explorer :

- Améliorer la reconstruction grâce à la TDA : que ce soit pour stabiliser la reconstruction ou bien être capable de reproduire le bruit spécifique au type de signal

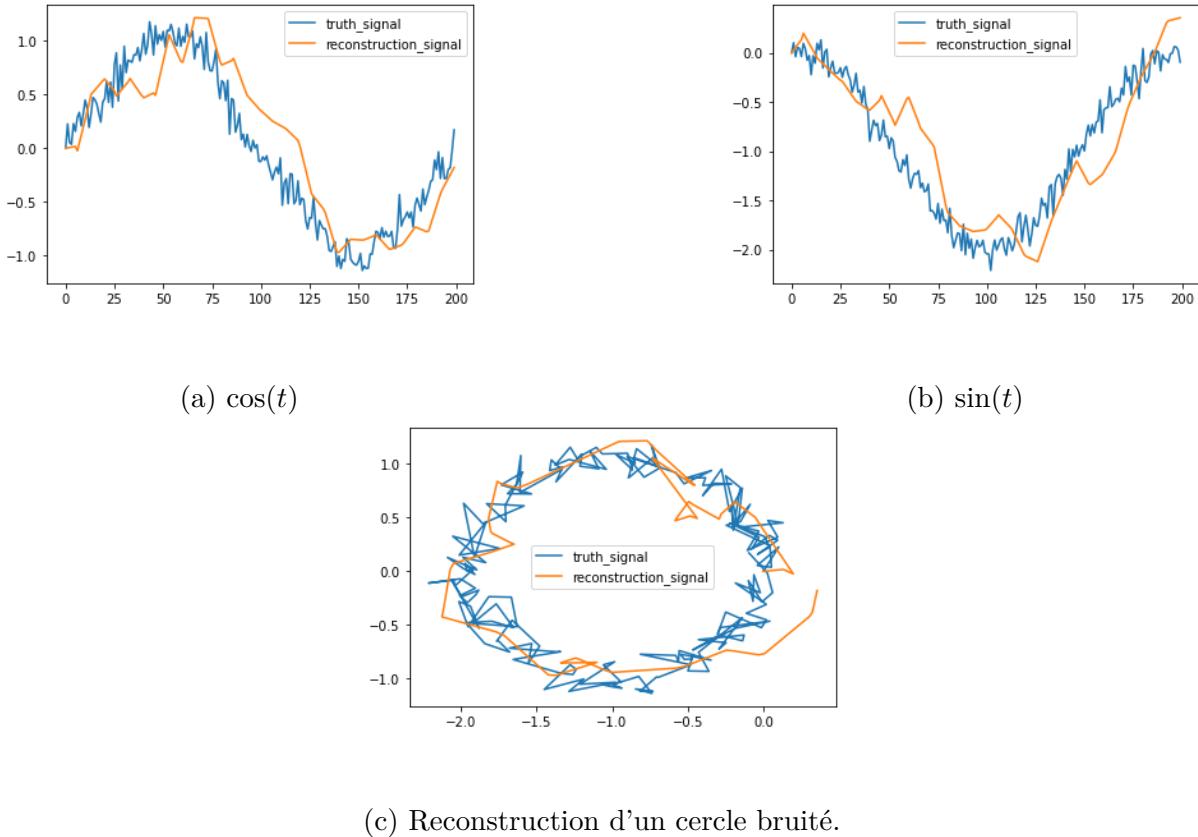


FIGURE 5.3

considéré, il serait intéressant d'intégrer des structures invariantes (notamment topologiques) dans la fonction objectif.

- Génération aléatoire de signaux : la possibilité d'inverser la Signature offre l'opportunité de générer des Signatures sous la loi des signaux, puis les inverser pour obtenir une génération sous la loi des signaux.
- Amélioration du processus de débruitage pour les **diffusions stables** : nous avons constaté que cette méthode d'inversion permet de reconstruire des chemins browniens multidimensionnels. Le processus de génération aléatoire par diffusion repose sur un réseau de neurones pour reconstruire un chemin brownien. Par notre méthode, il est possible de reconstruire le chemin inverse en prédisant la Signature du chemin de bruitage. Ce problème très excitant nécessite une meilleure compréhension de la stabilité numérique du schéma de reconstruction, et sera étudié en profondeur dans un futur proche.

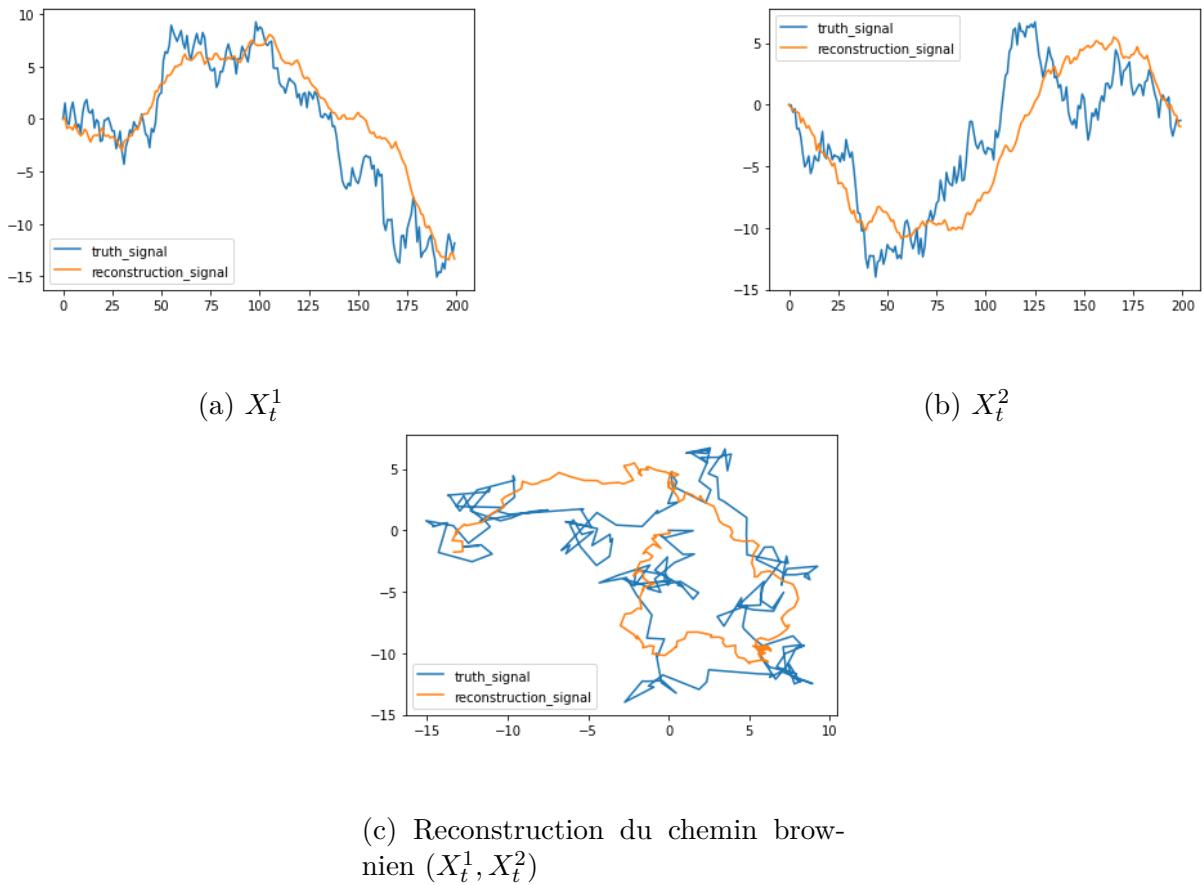


FIGURE 5.4

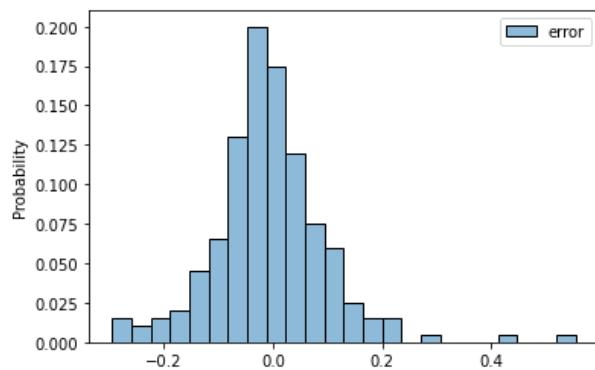


FIGURE 5.5 – Distribution des erreurs pour 200 reconstructions d'un chemin brownien 5-dimensionnel.

Conclusion

Au cours de ces travaux, nous avons essayé d'exploiter la structure topologique des signatures dans le cadre de la topologie des données. En premier lieu, nous avons proposé un nouvelle méthodologie pour la création de complexe simplicial sur un ensemble de signaux uni ou multivariés en utilisant la topologie induite par la sélection statistique de variable d'intérêts. Dans le Chapitre 3, nous étudions un algorithme de sélection de variable à coefficient entier fondé sur une variante du célèbre algorithme de Lenstra, Lenstra et Lovasz (LLL), dont l'intérêt est justifié par la filtration expliquée lors du Chapitre 4. Cette filtration échelonne l'inclusion des différents sous-complexes grâce au coefficients attribués par l'algorithme de régression associé à la sélection de variables. Dans ce même chapitre, nous évaluons l'efficacité de la combinaison TDA/signature sur des processus de vie et de mort dans un premier temps, puis sur des données de neurosciences et d'acoustiques dans un second temps. La première application n'implique pas notre algorithme, mais permet d'évaluer la pertinence du raisonnement. Les deux autres applications, elles, sont une directe utilisation de notre algorithme et permettent une première évaluation de la filtration associée. Leur utilisation sur les signaux acoustiques démontrent un réel impact de cette méthode dans la séparation des données normales et anormales. Pour finir, nous proposons une contribution au problème difficile de l'inversion algorithmique de la Signature. Nous décrivons les limites de l'algorithme Pfeiffer-Seigal-Sturmfelds. Nous étudions les fluctuations de l'estimateur de la Signature à l'aide de résultats d'anti-concentration et proposons une modification robuste à la présence d'outliers en utilisant l'estimateur Median of Means. Nous finissons en introduisant une nouvelle formulation du problème d'inversion de la Signature basé sur les propriétés de conservation de la longueur. Les résultats sont encourageant et nous proposons divers enrichissements théoriques et algorithmiques, utilisant notamment la TDA.

Chacun de ces travaux ouvre un grand nombre de piste de recherche, et nous allons en proposer une liste non-exhaustive que nous n'avons pas mentionné dans ce qui précède !

- L'algorithme de création de complexe simplicial mérite encore quelques améliorations, toutefois, nous nous proposons de regarder au delà : nous avons élaboré une méthode pour construire un complexe simplicial puis une filtration sur des signaux multivariés. Cette méthode utilise un espace de features qui, dans la dernière année, s'est vu étendre à des concepts plus généraux comme des surfaces aléatoires [65, 66], ou même des images. Dès lors, en adaptant le principe du Chapitre 2 à ces nouveaux objets, la création d'un complexe simplicial sur un ensemble d'image devient envisageable.

- Pour rester dans le domaine des interactions "graphiques" découvertes par le biais des Signatures, nous pouvons citer les relations de causalités. Sujet déjà étudié en surface [50, 52], l'injection d'équation différentielle forcée dans la boucle pourrait permettre d'éclaircir le rôle des signatures dans la causalité effective entre deux signaux temporels. Pour l'instant, nous avons essayé d'établir un graphe de causalité en utilisant les coefficient de l'aire de Levy. Cette méthode suit le principe de [50] pour la découverte de relation de causalité, mais rajoute un algorithme de Metropolis Hastings pour réaliser des échantillons de graphes de causalités suivant la loi invariante amenée par les coefficients de signatures. La Figure 5.6 représente les résultats obtenus qui ont été présentés sous forme de poster à l'école d'automne Bayes@CIRM [29].

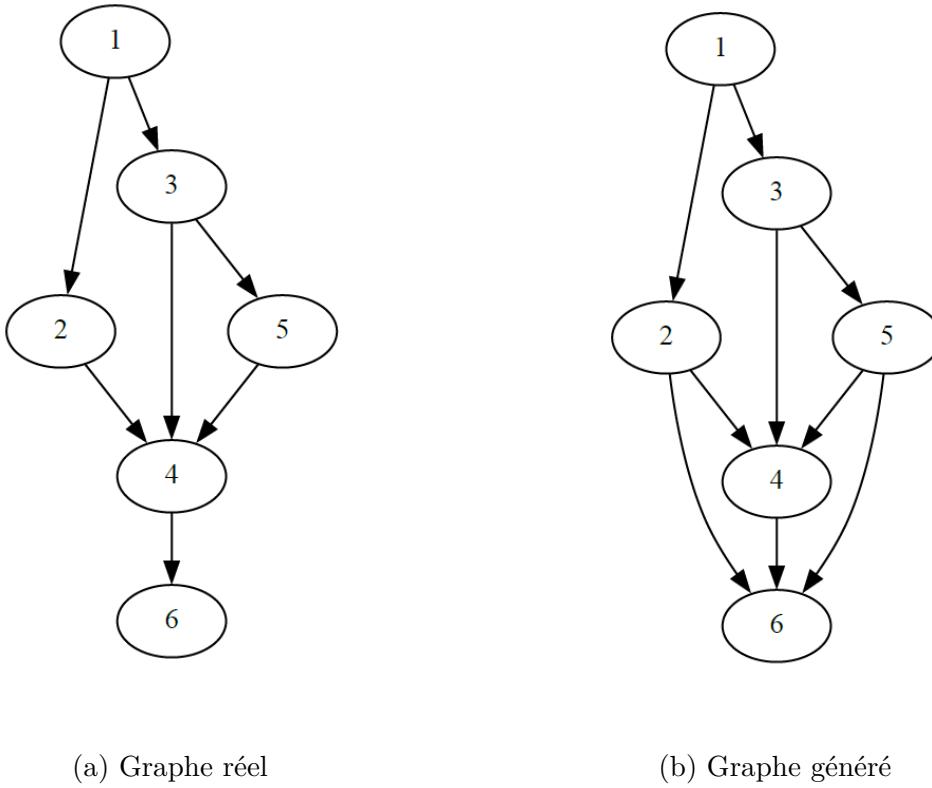


FIGURE 5.6 – Comparaison entre le graphe réel et le graphe obtenu après algorithme de Metropolis Hastings pour des données générées.

- La topologie des données est un domaine très récent dans le large horizon des statistiques. Beaucoup de choses restent encore à explorer, les outils à disposition provenant des mathématiques théoriques continuant à évoluer en parallèle. A titre d'exemple, le géomètre que je suis ne peux s'empêcher de faire un parallèle avec les réalisations algorithmiques de plongements isométriques réussis par les équipe d'HEVEA. Ces objets considérés comme un équilibre entre la géométrie riemannienne et fractale paraissent un terreau fertile pour l'étude des données, les statistiques allouant un cadre assez flexible pour introduire de nouveaux outils.

Bibliographie

- [1] E. Aleskerov, B. Freisleben, and B. Rao. Cardwatch : A neural network based database mining system for credit card fraud detection. In *Proceedings of the IEEE/IAFE 1997 computational intelligence for financial engineering (CIFEr)*, pages 220–226. IEEE, 1997.
- [2] S. Avanzini, D. M. Kurtz, J. J. Chabon, E. J. Moding, S. S. Hori, S. S. Gambhir, A. A. Alizadeh, M. Diehn, and J. G. Reiter. A mathematical model of ctDNA shedding predicts tumor detection size. *Science advances*, 6(50) :eabc4308, 2020.
- [3] B. Barancikova, Z. Huang, and C. Salvi. Sigdiffusions : Score-based diffusion models for long time series via log-signature embeddings. *arXiv preprint arXiv:2406.10354*, 2024.
- [4] G. B. Bastian Julien, Chrétien Stéphane and V. Rémi. Détection non supervisée d'anomalies dans les images satellites pour le monitoring des surfaces océaniques à l'aide de l'ACP robuste et du test de goodness of fit basé sur la distance de wasserstein entre processus ponctuels. In *Journées de la Statistique*, 2024, <https://jds2024.sciencesconf.org/531034>.
- [5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal statistical society : series B (Methodological)*, 57(1) :289–300, 1995.
- [6] R. J. Bolton, D. J. Hand, et al. Unsupervised profiling methods for fraud detection. *Credit scoring and credit control VII*, pages 235–255, 2001.
- [7] K. Borkar, A. Chaturvedi, P. K. Vinod, and R. Bapi. Ayu-characterization of healthy aging from neuroimaging data with deep learning and rsfmri. *Frontiers in Computational Neuroscience*, 16, 09 2022.
- [8] Boulet. *Notes, Tome 1*. Delcourt, 2008.
- [9] S. Broyd, C. Demanuele, S. Debener, S. Helps, C. James, and E. Sonuga-Barke. Default-mode brain dysfunction in mental disorders : A systematic review. *Neuroscience and biobehavioral reviews*, 33 :279–96, 10 2008.
- [10] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data : methods, theory and applications*. Springer Science & Business Media, 2011.
- [11] P. Caldero and J. Germoni. *Histoires hédonistes de groupes et de géométries*. Calvage et Mounet, 2016.
- [12] E. J. Candès and Y. Plan. Near-ideal model selection by l1 minimization. 2009.

- [13] T. Cass, R. Messadene, and W. F. Turner. Signature asymptotics, empirical processes, and optimal transport. *Electronic Journal of Probability*, 28 :1–19, 2023.
- [14] T. Cass and C. Salvi. Lecture notes on rough paths and applications to machine learning. *arXiv preprint arXiv :2404.06583*, 2024.
- [15] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection : A survey. *ACM computing surveys (CSUR)*, 41(3) :1–58, 2009.
- [16] J. Chang and T. Lyons. Insertion algorithm for inverting the signature of a path. *arXiv preprint arXiv :1907.08423*, 2019.
- [17] F. Chazal. High-dimensional topological data analysis. In *Handbook of Discrete and Computational Geometry*, pages 663–683. Chapman and Hall/CRC, 2017.
- [18] F. Chazal and B. Michel. An introduction to topological data analysis : fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4 :108, 2021.
- [19] K.-T. Chen. Integration of paths, geometric invariants and a generalized baker-hausdorff formula. *Annals of Mathematics*, 65(1) :163–178, 1957.
- [20] K.-T. Chen. Integration of paths—a faithful representation of paths by noncommutative formal power series. *Transactions of the American Mathematical Society*, 89(2) :395–407, 1958.
- [21] K.-T. Chen. Iterated path integrals. *Bulletin of the American Mathematical Society*, 83(5) :831–879, 1977.
- [22] I. Chevyrev and A. Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv :1603.03788*, 2016.
- [23] H. Chintakunta, T. Gentimis, R. Gonzalez-Diaz, M.-J. Jimenez, and H. Krim. An entropy-based persistence barcode. *Pattern Recognition*, 48(2) :391–401, 2015.
- [24] W.-L. Chow. Über systeme von linearen partiellen differential-gleichungen erster ordnung. In *The Collected Papers Of Wei-Liang Chow*, pages 47–54. World Scientific, 2002.
- [25] S. Chrétien and S. Darses. Sparse recovery with unknown variance : a lasso-type approach. *IEEE Transactions on Information Theory*, 60(7) :3970–3988, 2014.
- [26] S. Chrétien, B. Gao, A. T. Guiochon, and R. Vaucher. Leveraging the power of signatures for the construction of topological complexes for the analysis of multivariate complex dynamics.
- [27] S. Chrétien, B. Gao, A. Thebault-Guiochon, and R. Vaucher. Time topological analysis of eeg using signature theory. *arXiv preprint arXiv :2404.15328*, 2024.
- [28] S. Chrétien and R. Vaucher. Signature estimation and signal recovery using median of means. In *International Conference on Geometric Science of Information*, pages 323–331. Springer, 2023.
- [29] G. B. Chrétien Stéphane and V. Rémi. Sampling signature-induced causality chain in time series. In *Autumn school in Bayesian Statistics*, 30 October-3 November, 2023.
- [30] G. G. Chrysos, S. Moschoglou, G. Bouritsas, Y. Panagakis, J. Deng, and S. Zafeiriou. P-nets : Deep polynomial neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7325–7335, 2020.

- [31] S. Chrétien. An alternating l_1 approach to the compressed sensing problem. *IEEE Signal Processing Letters*, 17(2) :181–184, 2009.
- [32] Y.-M. Chung and A. Lawson. Persistence curves : A canonical framework for summarizing persistence diagrams. *Advances in Computational Mathematics*, 48(1) :6, 2022.
- [33] M. Clausel, J. Diehl, R. Mignot, L. Schmitz, N. Sugiura, and K. Usevich. The barycenter in free nilpotent lie groups and its application to iterated-integrals signatures. *SIAM Journal on Applied Algebra and Geometry*, 8(3) :519–552, 2024.
- [34] D. R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society : Series B (Methodological)*, 17(2) :129–157, 1955.
- [35] M. P. Deisenroth, A. A. Faisal, and C. S. Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.
- [36] R. Durrett and R. Durrett. *Branching process models of cancer*. Springer, 2015.
- [37] H. Dutta, C. Giannella, K. Borne, and H. Kargupta. Distributed top-k outlier detection from astronomy catalogs using the demac system. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 473–478. SIAM, 2007.
- [38] J.-P. Eckmann and U. Genève. Martin hairer got the fields medal for his study of the kpz equation.
- [39] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & computational geometry*, 28 :511–533, 2002.
- [40] A. Fermanian. Embedding and learning with signatures. *Computational Statistics & Data Analysis*, 157 :107148, 2021.
- [41] A. Fermanian. *Learning time-dependent data with the signature transform*. PhD thesis, Sorbonne Université, 2021.
- [42] A. Fermanian, J. Chang, T. Lyons, and G. Biau. The insertion method to invert the signature of a path. *arXiv preprint arXiv :2304.01862*, 2023.
- [43] A. Fermanian, T. Lyons, J. Morrill, and C. Salvi. New directions in the applications of rough path theory. *IEEE BITS the Information Theory Magazine*, 2023.
- [44] A. Fermanian, P. Marion, J.-P. Vert, and G. Biau. Framing rnn as a kernel method : A neural ode approach. *Advances in Neural Information Processing Systems*, 34 :3121–3134, 2021.
- [45] A. M. Frieze. On the lagarias-odlyzko algorithm for the subset sum problem. *SIAM Journal on Computing*, 15(2) :536–539, 1986.
- [46] P. K. Friz and M. Hairer. *A course on rough paths*. Springer, 2020.
- [47] P. K. Friz and N. B. Victoir. *Multidimensional stochastic processes as rough paths : theory and applications*, volume 120. Cambridge University Press, 2010.
- [48] D. Gamarnik, E. C. Kızıldağ, and I. Zadik. Inference in high-dimensional linear regression via lattice basis reduction and integer relation detection. *arXiv preprint arXiv :1910.10890*, 2019.
- [49] D. Gamarnik and I. Zadik. High dimensional linear regression using lattice basis reduction. *arXiv preprint arXiv :1803.06716*, 2018.

- [50] C. Giusti and D. Lee. Iterated integrals and population time series analysis. In *Topological Data Analysis : The Abel Symposium 2018*, pages 219–246. Springer, 2020.
- [51] C. Giusti, D. Lee, V. Nanda, and H. Oberhauser. A topological approach to mapping space signatures. *arXiv preprint arXiv :2202.00491*, 2022.
- [52] W. Glad and T. Woolf. Path signature area-based causal discovery in coupled time series. In *Causal Analysis Workshop Series*, pages 21–38. PMLR, 2021.
- [53] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet : components of a new research resource for complex physiologic signals. *circulation*, 101(23) :e215–e220, 2000.
- [54] B. Hambly and T. Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, pages 109–167, 2010.
- [55] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito. ToyAD-MOS2 : Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 1–5, Barcelona, Spain, November 2021.
- [56] S. Hariri, M. C. Kind, and R. J. Brunner. Extended isolation forest. *IEEE transactions on knowledge and data engineering*, 33(4) :1479–1489, 2019.
- [57] T. Hastie. *The elements of statistical learning : data mining, inference, and prediction*, volume 2. Springer.
- [58] S. Hautphenne and B. Patch. Birth-and-death processes in python : The birdepy package. *arXiv preprint arXiv :2110.05067*, 2021.
- [59] I. M. Johnstone. Function estimation and gaussian sequence models. *Unpublished manuscript*, 2(5.3) :2, 2002.
- [60] P. Kidger and T. Lyons. Signatory : differentiable computations of the signature and logsignature transforms, on both CPU and GPU. In *International Conference on Learning Representations*, 2021. <https://github.com/patrick-kidger/signatory>.
- [61] H. Kim, J. Hahm, H. Lee, E. Kang, H. Kang, and D. S. Lee. Brain networks engaged in audiovisual integration during speech perception revealed by persistent homology-based network filtration. *Brain connectivity*, 5(4) :245–258, 2015.
- [62] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3) :455–500, 2009.
- [63] A. Kormilitzin, N. Vaci, Q. Liu, H. Ni, G. Nenadic, and A. Nevado-Holgado. An efficient representation of chronological events in medical texts. *arXiv preprint arXiv :2010.08433*, 2020.
- [64] J. C. Lagarias and A. M. Odlyzko. Solving low-density subset sum problems. *Journal of the ACM (JACM)*, 32(1) :229–246, 1985.
- [65] D. Lee. The surface signature and rough surfaces. *arXiv preprint arXiv :2406.16857*, 2024.

- [66] D. Lee and H. Oberhauser. Random surfaces and higher algebra. *arXiv preprint arXiv :2311.08366*, 2023.
- [67] W. Lee and D. Xiang. Information-theoretic measures for anomaly detection. In *Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001*, pages 130–143. IEEE, 2000.
- [68] M. Lerasle. Lecture notes : Selected topics on robust statistical learning theory. *arXiv preprint arXiv :1908.10761*, 2019.
- [69] S. Li, Z. Lyu, H. Ni, and J. Tao. On the determination of path signature from its unitary development. *arXiv preprint arXiv :2404.18661*, 2024.
- [70] L.-H. Lim. Tensors in computations. *Acta Numerica*, 30 :555–764, 2021.
- [71] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.
- [72] H. Lou, S. Li, and H. Ni. Pcf-gan : generating sequential data via the characteristic function of measures on the path space. *Advances in Neural Information Processing Systems*, 36, 2024.
- [73] S. Lovett. An elementary proof of anti-concentration of polynomials in gaussian variables. In *Electron. Colloquium Comput. Complex.*, volume 17, page 182, 2010.
- [74] T. Lyons and A. D. McLeod. Signature methods in machine learning. *arXiv preprint arXiv :2206.14674*, 2022.
- [75] T. Lyons and Z. Qian. *System control and rough paths*. Oxford University Press, 2002.
- [76] T. J. Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2) :215–310, 1998.
- [77] T. J. Lyons, M. Caruana, and T. Lévy. *Differential equations driven by rough paths*. Springer, 2007.
- [78] C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec. The gudhi library : Simplicial complexes and persistent homology. In *Mathematical Software–ICMS 2014 : 4th International Congress, Seoul, South Korea, August 5-9, 2014. Proceedings 4*, pages 167–174. Springer, 2014.
- [79] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. 2006.
- [80] R. Meka, O. Nguyen, and V. Vu. Anti-concentration for polynomials of independent random variables. *arXiv preprint arXiv :1507.00829*, 2015.
- [81] M. Michelen and J. Sahasrabudhe. Anti-concentration of random variables from zero-free regions. *arXiv preprint arXiv :2102.07699*, 2021.
- [82] R. Mignot, M. Clausel, and K. Usevich. Principal geodesic analysis for time series encoded with signature features. 2024.
- [83] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- [84] V. Navarro, M. Le Van Quyen, S. Clemenceau, C. Adam, C. Petitmengin, F. Du-beau, J. Gotman, J. Martinerie, and M. Baulac. Seizure prediction : from myth to reality. *Revue Neurologique*, 167(3) :205–215, 2010.

- [85] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and F. Vaccarino. Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface*, 11(101) :20140873, 2014.
- [86] M. Pfeffer, A. Seigal, and B. Sturmels. Learning paths from signature tensors. *SIAM Journal on Matrix Analysis and Applications*, 40(2) :394–416, 2019.
- [87] M. Posner and S. Petersen. The attention system of the human brain. *Annual review of neuroscience*, 13 :25–42, 02 1990.
- [88] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi. Mimii dataset : Sound dataset for malfunctioning industrial machine investigation and inspection. *arXiv preprint arXiv :1909.09347*, 2019.
- [89] J. Reizenstein and B. Graham. The iisignature library : efficient calculation of iterated-integral signatures and log signatures. *arXiv preprint arXiv :1802.08252*, 2018.
- [90] P. Rigollet and J.-C. Hütter. High-dimensional statistics. *arXiv preprint arXiv :2310.19244*, 2023.
- [91] V. Roth. Outlier detection with one-class kernel fisher discriminants. *Advances in Neural Information Processing Systems*, 17, 2004.
- [92] M. Rucco, F. Castiglione, E. Merelli, and M. Pettini. Characterisation of the idiosyncratic immune network through persistent entropy. In *Proceedings of ECCS 2014 : European Conference on Complex Systems*, pages 117–128. Springer, 2016.
- [93] A. Santoro, F. Battiston, G. Petri, and E. Amico. Unveiling the higher-order organization of multivariate time series. *Nature Physics*, 19(2) :221–229, 2023.
- [94] A. Schaefer, R. Kong, E. Gordon, T. Laumann, X.-N. Zuo, A. Holmes, S. Eickhoff, and B. Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. 07 2017.
- [95] H. Schenck. *Algebraic foundations for applied topology and data analysis*. Springer, 2022.
- [96] G. Singh, F. Mémoli, G. E. Carlsson, et al. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, 2 :91–100, 2007.
- [97] A. E. Sizemore, C. Giusti, A. Kahn, J. M. Vettel, R. F. Betzel, and D. S. Bassett. Cliques and cavities in the human connectome. *Journal of computational neuroscience*, 44 :115–145, 2018.
- [98] B. J. Stolz, H. A. Harrington, and M. A. Porter. Persistent homology of time-dependent functional networks constructed from coupled time series. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 27(4), 2017.
- [99] G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. M. Medina-Mardones, A. Dassatti, and K. Hess. giotto-tda : : A topological data analysis toolkit for machine learning and data exploration. *Journal of Machine Learning Research*, 22(39) :1–6, 2021.
- [100] F. Testard and R. Mneimné. Introduction à la théorie des groupes de lie classiques. (*No Title*), 1997.

- [101] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 58(1) :267–288, 1996.
- [102] M. J. Wainwright. *High-dimensional statistics : A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.
- [103] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks. Anoddpm : Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 650–656, 2022.
- [104] B. T. Yeo, F. Krienen, J. Sepulcre, M. Sabuncu, D. Lashkari, M. Hollinshead, J. Roffman, J. Smoller, L. Zollei, J. Polimeni, B. Fischl, H. Liu, and R. Buckner. The organization of the human cerebral cortex estimated by functional correlation. *Journal of neurophysiology*, 106 :1125–65, 06 2011.
- [105] Z. Zhu, F. Latorre, G. G. Chrysos, and V. Cevher. Controlling the complexity and lipschitz constant improves polynomial nets. *arXiv preprint arXiv :2202.05068*, 2022.
- [106] A. Zomorodian and G. Carlsson. Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 347–356, 2004.

Annexe A

"Rudiments" sur les groupes de Lie

Sinon on serait dans de beaux draps !

OSS 117

On se propose ici de redéfinir les groupes de Lie et quelques outils s'y attenant. Nous ponctuerons ces rappels par des d'applications à la théorie des signatures. Pour construire cet appendice, nous nous référerons largement à [11, 100].

A.1 Rappels rapides pour non algébristes.

Nous avons besoin de deux prérequis :



Definition A.1.1

Soit G un ensemble. Soit $+$ une loi de composition interne sur G (pour tous g_1, g_2 dans G , $g_1 + g_2 \in G$). La donnée de $(G, +)$ est **un groupe** si :

- $+$ est associative.
- Il existe $e \in G$ tel que, pour tout $g \in G$, $g + e = e + g = g$. e est appelé l'élément neutre de G .
- Pour tout $a \in G$ il existe $b \in G$ tel que $a + b = b + a = e$. b est appelé le symétrique de a dans G (pour $+$).



Definition A.1.2

Une **variété topologique de dimension n** est un espace topologique \mathbb{M} tel que :

- \mathbb{M} est séparé.
- \mathbb{M} admet un recouvrement dénombrable de compacts.
- Tout $x \in \mathbb{M}$ admet un voisinage (dans \mathbb{M}) homéomorphe à \mathbb{R}^n .

Definition A.1.3

Un **atlas** C^k de la variété topologique \mathbb{M} est la donnée d'un recouvrement de \mathbb{M} par des ouverts U_i et d'application ϕ_i pour chaque U_i telles que $\phi_i : U_i \rightarrow O_i \subset \mathbb{R}^n$ est un homéomorphisme. On appelle (U_i, ϕ_i) les *cartes*. De plus, les **changements de cartes**

$$\forall i, j, \quad \phi_j \circ \phi_i^{-1} : \phi_i(U_i \cap U_j) \rightarrow \phi_j(U_i \cap U_j)$$

sont C^k .

Tous les atlas C^k dont la réunion forment toujours un atlas C^k forment une classe d'équivalence (sous entendu, les applications $\phi_i \circ \psi_j^{-1}$ sont C^k).

Definition A.1.4

La donnée d'une variété topologique \mathbb{M} et d'une classe d'équivalence atlas C^k est une **variété différentielle**.

A.2 Groupes de Lie

Definition A.2.5

Un **groupe de Lie** est un groupe G muni d'une structure de variété différentielle telle que les opérations Mult (composition interne) et inverse définies par

$$\begin{aligned} \text{Mult} : G \times G &\rightarrow G \\ (g_1, g_2) &\mapsto g_1 g_2 \end{aligned}$$

et

$$\begin{aligned} \text{Inv} : G &\rightarrow \\ g &\mapsto g^{-1} \end{aligned}$$

sont au moins C^1 .

Remarque(s)

| L'inversion $g \mapsto g^{-1}$ vérifiant $\text{Inv}^2 = Id$, c'est un homéomorphisme.

Exemple(s)

- $GL_n(\mathbb{R}) = \{A \in M_n(\mathbb{R}), \det(A) \neq 0\}$
- $O_n(\mathbb{R}) = \{A \in M_n(\mathbb{R}), A^T A = Id\}$

 **Exemple(s)** *Appliqué au signature*

Considérons l'ensemble suivant

$$1 + \mathfrak{t}^N(\mathbb{R}^d) = \{g \in (\mathbb{R}^d), \quad \pi_0(g) = 1\}$$

Il est facile de montrer que cet ensemble est un groupe pour \otimes . De plus, il est trivialement difféomorphe avec $\bigoplus_{k=1}^{\infty} (\mathbb{R}^d)^{\otimes k}$ et est donc une variété différentielle. Pour finir, les opérations \otimes et $g \rightarrow g^{-1}$ sont continues (elles sont en fait polynomiales). C'est donc un groupe de Lie.

 **Theorem A.1** *Cartan*

 Tout sous groupe fermé d'un groupe de Lie est un sous groupe de Lie (et donc en particulier un groupe de Lie).

 **Exemple(s)**

- Tout espace vectoriel de dimension finie sur \mathbb{R} est un groupe de Lie muni du produit donné par l'addition.
- \mathbb{S}^1 (le cercle unitaire) est un groupe de Lie comme sous-groupe fermé de \mathbb{R}^2 .
- Soit \mathfrak{A} une algèbre associative de dimension finies sur \mathbb{R} , d'unité $1_{\mathfrak{A}}$. Alors le groupe \mathfrak{A}^\times des éléments inversibles de \mathfrak{A} est un groupe de Lie. En effet, si l'on considère l'application $\phi : \mathfrak{A}^\times \rightarrow \mathfrak{A}$ définie comme

$$a \mapsto (a, a^{-1}) \mapsto a \cdot a^{-1}$$

alors \mathfrak{A}^\times est l'image réciproque d'un fermé par une application continue, donc un fermé de \mathfrak{A} . C'est donc un sous groupe de Lie.

 **Exemple(s)** *Appliqué au signature*

$\mathfrak{G}^N(\mathbb{R}^d)$ l'image par la signature tronquée à l'ordre N de $BV(\mathbb{R}^d)$ est un groupe de Lie comme sous groupe fermé de $1 + \mathfrak{t}^N(\mathbb{R}^d)$. Nous le montrerons plus bas.

A.3 Algèbre de Lie

La notion d'algèbre de Lie est l'une des notions les plus fondamentales pour comprendre comment fonctionne "en gros" un groupe de Lie. En effet, même si G est localement homéomorphe à un espace vectoriel euclidien, cette propriété n'est absolument

global. L'exemple de \mathbb{S}^1 résume bien toute cette ambiguïté.

Or, le groupe de Lie étant en particulier une variété différentielle, nous pouvons faire appel à certaines méthodes toutes droit issues de la géométrie différentielle. Notamment, il est possible de se placer dans l'espace tangent pour linéariser un problème. Avant de parler d'espace tangent, donnons une définition plus générale.



Definition A.3.6

Une **algèbre de Lie** sur un corps commutatif \mathbb{K} est un \mathbb{K} -espace vectoriel \mathfrak{g} muni d'une application bilinéaire

$$[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$$

antisymétrique et vérifiant *l'identité de Jacobi* :

$$[\xi, [\nu, \zeta]] + [\nu, [\zeta, \xi]] + [\zeta, [\xi, \nu]] = 0, \quad \forall \xi, \zeta, \nu \in \mathfrak{g}$$

cette application est appelée **crochet de Lie**.



Exemple(s)

- $\mathfrak{gl}_n(\mathbb{R}) = M_n(\mathbb{R})$ est un algèbre de Lie si on la munit du commutateur :

$$\forall A, B \in M_n(\mathbb{R}), [A, B] = AB - BA$$

-

$$\mathfrak{sl}_n(\mathbb{R}) = \{A \in M_n(\mathbb{R}), \text{tr}(A) = 0\}$$

et

$$\mathfrak{o}_n(\mathbb{R}) = \left\{ A \in M_n(\mathbb{R}), A^T = -A \right\}$$

sont deux algèbres de Lie matricielles, i.e. deux sous algèbres de $\mathfrak{gl}_n(\mathbb{R})$.



Exemple(s) Application aux signatures

On définit $\mathfrak{t}^N = \{g \in T^N(\mathbb{R}^d), \pi_0(g) = 0\}$. On définit un crochet de Lie pour $\mathfrak{t}^N(\mathbb{R}^d)$ par le commutateur :

$$[g, h] = g \otimes h - h \otimes g$$

Alors $\mathfrak{t}^N(\mathbb{R}^d)$ est une algèbre de Lie.

A.3.1 Action d'un groupe (de Lie) sur une variété : algèbre de Lie associée à un groupe de Lie.

Considérons maintenant un groupe de Lie G



Definition A.3.7

Soit G un groupe de Lie. Son action à gauche sur une variété M est l'application

$$\begin{aligned} L : G \times M &\rightarrow M \\ (g, m) &\mapsto g.m \end{aligned}$$

Nous allons ici nous intéresser principalement à l'application $R_m(\cdot) = L(\cdot, m)$. Comme L est une action différentiable, à fortiori, R_m l'est aussi. Considérons donc sa différentielle dR_m . C'est une application linéaire

$$dR_m : T_g G \rightarrow T_{g.m} M$$

Regardons donc ce qui se passe si l'on prends $g = e_G$ (l'élément neutre de G). On obtient :

$$(dR_m)_{e_G} : T_{e_G} G \rightarrow T_m M$$

Pour tout $X \in T_{e_G} G$, on obtient un vecteur dans $T_m M$, noté $X^L(m)$. On en tire un champ de vecteur naturel, défini pour tout $X \in T_{e_G} G$:

$$\begin{aligned} X^L : M &\rightarrow T_m M \\ m &\mapsto X^L(m) \end{aligned}$$



Definition A.3.8 Champ fondamental.

∇ X^L est le champ de vecteur fondamental gauche associé à X de $T_{e_G} G$.



Remarque(s)

Par la suite, nous noterons souvent $X^L(m) = \left. \frac{d}{dt} \right|_0 e^{tX} \cdot m$, pour se rapprocher de l'esprit "géométrie différentielle" : un espace tangent en x sur G est l'espace engendré par les vecteurs tangent de toutes les courbes (sous-entendu C^1) de G passant par x .

On va définir l'algèbre de Lie du groupe de Lie G comme l'espace vectoriel

$$\text{Lie}(G) = T_{e_G} G = \langle X^L(e_G) \rangle.$$

Pour définir un crochet sur cette algèbre, on regarde l'action de conjugaison de G sur G et surtout sa différentielle en l'identité

$$\begin{aligned} G \times T_e G &\rightarrow T_e G \\ (x, Y) &\mapsto \text{Ad}_x(Y) \end{aligned}$$

Redifférentions ensuite en l'identité pour obtenir

$$\begin{aligned} T_e G \times T_e G &\rightarrow T_e G \\ (X, Y) &\mapsto \text{ad}_X(Y) \end{aligned} \tag{A.1}$$

On définit $[X, Y]_{\text{Lie}(G)} = \text{ad}_X(Y)$. On peut donc dire qu'à tout $X \in \text{Lie}(G)$ on peut associer un champ fondamental.

Récapitulons. On a :

$$\left(\begin{array}{ccc} G & \rightarrow & M \\ g & \mapsto & g.m \end{array} \right) \xrightarrow{\text{On différentie}} (T_g G \rightarrow T_{g.m} M) \xrightarrow{\text{On particularise}} \left(\begin{array}{ccc} \text{Lie}(G) & \rightarrow & T_m M \\ X & \mapsto & X^L(m) \end{array} \right)$$

Regardons maintenant quelques propriétés sur les actions qui nous seront utiles lors de la réduction.

Exemple(s) Algèbre de Lie de $1 + \mathfrak{t}^N(\mathbb{R}^d)$

Soit $X \in 1 + \mathfrak{t}^N(\mathbb{R}^d)$. On calcule $\frac{d}{dt}\Big|_0 e^{tX}$:

$$\begin{aligned} \frac{d}{dt}\Big|_0 e^{tX} &= \frac{d}{dt}\Big|_0 \sum_{i=0}^N \frac{(tX)^{\otimes k}}{k!} \\ &= \frac{d}{dt}\Big|_0 \left(1 + \sum_{i=1}^N t^k \frac{X^{\otimes k}}{k!} \right) \\ &= \left(0 + X + \sum_{i=1}^N t^{k-1} \frac{X^{\otimes k}}{(k-1)!} \right)\Big|_0 \\ &= X \end{aligned}$$

On a donc que $\text{Lie}(1 + \mathfrak{t}^N(\mathbb{R}^d)) = \{X \in T^N(\mathbb{R}^d), \quad \pi_0(X) = 0\}$. De même, nous avions vu précédemment que, sur $\mathfrak{t}^N(\mathbb{R}^d)$, le commutateur définissait un crochet de Lie :

$$[x, y] = x \otimes y - y \otimes x$$

A.3.2 Application \exp et \log

On esquive habilement la définition générale de l'exponentielle pour un groupe de Lie, et on se concentre sur le cas des signatures. On va considérer $a \in \mathfrak{t}^N(\mathbb{R}^d)$ un élément de l'algèbre de Lie de $1 + \mathfrak{T}^N(\mathbb{R}^d)$. On a donc $\pi_0(a) = 0$.

On définit l'exponentielle comme l'application :

$$\begin{aligned}\exp : \mathfrak{t}^N(\mathbb{R}^d) &\rightarrow 1 + \mathfrak{t}^N(\mathbb{R}^d) \\ a &\mapsto 1 + \sum_{k=1}^N \frac{a^{\otimes k}}{k!}\end{aligned}$$

De la même manière, on va définir le logarithme :

$$\begin{aligned}\log : 1 + \mathfrak{t}^N(\mathbb{R}^d) &\rightarrow \mathfrak{t}^N(\mathbb{R}^d) \\ (1 + a) &\mapsto \sum_{k=1}^N (-1)^{k+1} \frac{a^{\otimes k}}{k}\end{aligned}$$

On a assez facilement que, pour $a \in \mathfrak{t}^N(\mathbb{R}^d)$:

$$\exp(\log(1 + a)) = \log(\exp(a)) = a$$

Et donc $\log = \exp^{-1}$ (on admettras que \log est bien défini globalement). On a alors le résultat suivant :



Theorem A.2



L'application $\exp : \mathfrak{T}^N(\mathbb{R}^d) \rightarrow 1 + \mathfrak{t}^N(\mathbb{R}^d)$ est **un difféomorphisme local** d'un voisinage ouvert de $\mathbf{0} \in \mathfrak{t}^N(\mathbb{R}^d)$ sur un voisinage ouvert de $\mathbf{1} \in 1 + \mathfrak{t}(\mathbb{R}^d)$.

La preuve est rapide en passant par le théorème d'inversion locale.



Remarque(s)

On aimerais que l'exponentielle ait la même propriété de morphisme que dans les réels, i.e. $\exp(t + s) = \exp(t) \exp(s)$. Ce n'est pas le cas, il suffit de se placer à $N = 2$ pour s'en convaincre :

$$\begin{aligned}\exp(a) \otimes \exp(b) &= (1 + a + a^{\otimes 2}) \otimes (1 + b + b^{\otimes 2}) \\ &= 1 + a + b + \frac{1}{2}(a \otimes b - b \otimes a) + \frac{(a + b)^{\otimes 2}}{2} \\ &= \exp(a + b + \frac{1}{2}[a, b])\end{aligned}$$

De même, en étendant ce calcul à $N > 2$, on trouve une expression dépendante du crochet de Lie, si bien que l'on obtient la proposition suivante :

Proposition A.3.1

Soit a et b deux éléments de $\mathfrak{t}^N(\mathbb{R}^d)$. Si $[a, b] = a \otimes b - b \otimes a = 0$, alors

$$\exp(a + b) = \exp(a) \otimes \exp(b)$$

Il existe toutefois une manière de gagner un équivalent de cette formule en passant par l'application ad.

La définition (A.1) entraîne la création de :

$$\begin{aligned}\text{ad} : T_e G &\rightarrow \mathcal{L}(T_e G) \\ X &\mapsto \text{ad}_X\end{aligned}$$

Où $\text{ad}_X(Y) = [X, Y]$, et dans le cas de $\mathfrak{t}^N(\mathbb{R}^d)$, $\text{ad}_a(b) = a \otimes b - b \otimes a$. Un simple calcul permet de voir $(\text{ad}_X)^n = 0$ si $n > N$. De même :

$$\begin{aligned}\text{ad}_X \circ \text{ad}_Y &= [X, [Y, .]] \\ &= -[Y, [., X]] - [., [X, Y]]\end{aligned}$$

par l'identité de Jacobi. Dans le cadre de $\mathfrak{t}^N(\mathbb{R}^d)$, nous avons aussi que l'action de $1 + \mathfrak{t}^N(\mathbb{R}^d)$ sur $\mathfrak{t}^N(\mathbb{R}^d)$ par conjugaison donne :

$$\forall d \in \mathfrak{t}^N(\mathbb{R}^d), \quad \exp(a) \otimes d \otimes \exp(b) = e^{\text{ad}_a}(d)$$

où

$$e : \mathcal{L}(T_{e_G} G) \rightarrow \mathcal{L}(T_{e_G} G) \tag{A.2}$$

$$\text{ad}_a \mapsto \sum_{i=0}^N \frac{(\text{ad}_a)^n}{n!} \tag{A.3}$$

alors, en notant $d = \log(\exp(a) \otimes \exp(b))$, on obtient :

$$e^{\text{ad}_c} = e^{\text{ad}_a} \circ e^{\text{ad}_b}$$

Le théorème de Campbell-Baker-Hausdorff¹ porte justement sur la quantité c :

Theorem A.3

On considère $a, b \in \mathfrak{t}^N(\mathbb{R}^d)$. Alors

$$\log(\exp(a) \otimes \exp(b)) = b + \int_0^1 a H(e^{t \text{ad}_a} \circ e^{t \text{ad}_b}) dt$$

Où $H(z) = \sum_{n \geq 0} \frac{(1)^n}{n+1} (z-1)^n$.

En particulier, $\log(\exp(a) \otimes \exp(b))$ est une somme de crochet de Lie itérés en a et b

1. Une preuve de ce résultat et des suivants est trouvable dans [47] p.139-140

De ce théorème, on déduit le corollaire suivant :

Corollary A.3.1

On définit \mathfrak{g}^N comme la plus petite algèbre de Lie contenant $\pi_1(\mathfrak{t}^N(\mathbb{R}^d)) \simeq \mathbb{R}^d$, c'est à dire :

$$\mathfrak{g}^N(\mathbb{R}^d) = \mathbb{R}^d \oplus [\mathbb{R}^d, \mathbb{R}^d] \oplus \cdots \oplus \underbrace{\left[\mathbb{R}^d, [\dots, [\mathbb{R}^d, \mathbb{R}^d]] \right]}_{(N-1) \text{ crochets}} \quad (\text{A.4})$$

$\mathfrak{g}^N(\mathbb{R}^d)$ est appelé **l'algèbre de Lie nilpotente libre de degré N** .

Alors pour tous $a, b \in \mathfrak{g}^N(\mathbb{R}^d)$,

$$\log(\exp(a) \otimes \exp(b)) \in \mathfrak{g}^N(\mathbb{R}^d), \quad (\text{A.5})$$

et donc $\exp(\mathfrak{g}^N(\mathbb{R}^d))$ est un sous-groupe de $1 + \mathfrak{t}(\mathbb{R}^d)$ muni du produit tensoriel.

De plus, l'application $\log : 1 + \mathfrak{t}^N(\mathbb{R}^d) \rightarrow \mathfrak{t}^N(\mathbb{R}^d)$ étant continue, $\exp = \log^{-1}$ conserve le caractère fermé des sous-ensembles. En particulier, une sous-algèbre est fermée en tant que sous-espace vectoriel, et donc $\exp(\mathfrak{g}^N(\mathbb{R}^d))$ est fermée. On en déduit

Corollary A.3.2

$\exp(\mathfrak{g}^N(\mathbb{R}^d))$ est un sous-groupe de Lie de $1 + \mathfrak{t}^N(\mathbb{R}^d)$

Nous terminerons ici par le principal résultat :

Theorem A.4

On définit $G^N(\mathbb{R}^d) = \{S^N(X), \quad X \in BV(\mathbb{R}^d)\}$. Alors

$$G^N(\mathbb{R}^d) = \exp(\mathfrak{g}^N(\mathbb{R}^d)) \quad (\text{A.6})$$

et donc l'espace des signatures **est un groupe de Lie** dont l'algèbre de Lie est $\mathfrak{g}^N(\mathbb{R}^d)$

Proof. Voir [47]. L'esprit est assez simple : On montre que chacun des deux ensembles est égal à un ensemble tiers que l'on construit grâce à l'expression explicite de la signature et de la log-signature d'un chemin linéaire par morceau. \square

Abstract :

In recent years, anomaly detection has become a major issue in the industrial sector. At the same time, companies have implemented numerous online monitoring solutions. By combining these two elements, anomaly detection algorithms must be able to apply on time-evolving signals.

With this goal in mind, we introduce the signature of a rough path as the main tool. This tool, developed in 1954 by K.T. Chen within the framework of theoretical geometry, was brought back to the forefront by T. Lyons in 1998 in the context of stochastic differential equations involving rough paths. Over the past twenty years, this tool has led to a drastic improvement in the performance of some algorithms applied to time series data. Furthermore, its introduction enabled the success of an algorithm for Chinese handwriting recognition. An interesting aspect is that the signature space $G(\mathbb{R}^d) \subset T(\mathbb{R}^d)$ is equipped with a Lie group structure : this provides us with a group homomorphism for key operations, along with a homeomorphism for topology transport.

Throughout this manuscript, the main thread is the study of temporal signals through their topological organization. For this reason, we map these signals to the signature space. We then introduce a series of tools from the field of Topological Data Analysis (TDA). This field from statistics emerged from Edelsbrunner's work in 2002. TDA is based on a highly intuitive principle : a point cloud has a shape. A variation of this principle in our context would be that functional data have an underlying, difficult-to-visualize shape, whose alteration can indicate an anomaly.

In this thesis, we begin by demonstrating the statistical (and topological) power of signatures on stochastic processes modeling a cancer biomarker rate. Next, we propose an algorithm to create a topological structure in the signature space, ensuring that this structure reflects that of the signals being considered. We build on this algorithm by introducing a variable selection method and presenting some obtained results. Finally, a concluding chapter focuses on anomaly detection by randomly generating multivariate time signals using the two previously mentioned tools.

Résumé :

Ces dernières années, la détection d'anomalies s'est imposée comme une problématique majeure dans le domaine industriel. Parallèlement, les entreprises ont mis en place de nombreuses solutions de surveillance en ligne. En combinant ces deux aspects, les algorithmes de détection d'anomalies doivent pouvoir s'appliquer aux signaux évolutifs dans le temps.

Dans cette optique, nous introduisons comme outil principal la signature d'un chemin rugueux (rough path). Cet outil, développé en 1954 par K.T. Chen dans le cadre de la géométrie théorique, a été remis en avant par T. Lyons en 1998 dans le cadre des équations différentielles stochastiques impliquant des chemins rugueux. Depuis une vingtaine d'années, cet outil a entraîné une amélioration considérable de la performance de certains algorithmes appliqués aux séries temporelles. Par ailleurs, son introduction a permis la réussite d'un algorithme de reconnaissance d'écriture chinoise. Ce qui est particulièrement intéressant, c'est que l'espace des signatures $G(\mathbb{R}^d) \subset T(\mathbb{R}^d)$ est doté d'une structure de groupe de Lie : nous obtenons donc un homomorphisme de groupe pour les opérations importantes, accompagné d'un homéomorphisme pour le transport de la topologie.

Tout le long de ce manuscrit, le fil conducteur principal est l'étude des signaux temporels par le biais de leur organisation topologique. C'est pourquoi nous transposons ces

signaux dans l'espace des signatures. Nous introduisons alors une série d'outils issus du champ de l'analyse topologique des données (TDA). Ce domaine de l'analyse statistique a émergé avec le travail d'Edelsbrunner en 2002. La TDA repose sur un principe très intuitif : un nuage de points possède une forme. Une application de ce principe dans notre contexte serait que les données fonctionnelles ont une forme sous-jacente, difficilement visualisable, dont une modification peut traduire une anomalie.

Dans cette thèse, nous débutons par une démonstration de la puissance statistique (et topologique) des signatures sur des processus stochastiques modélisant un taux de biomarqueurs cancéreux. Ensuite, nous proposons un algorithme de création de structure topologique sur l'espace des signatures, en veillant à ce que cette structure reflète celle des signaux considérés. Nous approfondissons cet algorithme en introduisant un algorithme de sélection de variables et en présentant certains résultats obtenus. Enfin, un dernier chapitre s'intéresse à la détection d'anomalies par génération aléatoire de signaux temporels multivariés, en utilisant les deux outils précédemment mentionnés.