

ÉCOLE NATIONALE DE LA STATISTIQUE
ET DE L'ANALYSE DE L'INFORMATION



STAGE DE FIN D'ÉTUDE
pour l'entreprise DataStorm

Estimation adaptative en analyse des données fonctionnelles

rédigé par
Hugo Brunet
Tuteur
Hassan Maissoro

Avril—Septembre 2023

Résumé

Les séries temporelles sont des données omniprésentes dans l'analyse et la prédiction de données. Elles concernent de nombreux secteurs critiques allant du secteur de l'énergie à la finance. Leur étude systématique depuis 1927 (Yule) est ainsi motivée par leur importance et utilité pour la mise en production.

Les données fonctionnelles quant à elles sont particulièrement présentes dans les données de capteurs ou à composante temporelle. Elles permettent grâce au point de vue qu'elles offrent, d'obtenir notamment de meilleures estimations sur le long terme que le point de vue réel multivarié classique. Cependant, la littérature jusqu'alors ne prenait pas en compte les différences de régularité des données traitées, ce qui pose problème pour des données peu régulières pourtant fréquemment observées.

Ce stage porte sur l'estimation de la régularité locale des trajectoires des séries temporelles de données fonctionnelles afin d'obtenir une meilleure estimation de leur fonction moyenne et de l'opérateur d'auto-covariance. Plus spécifiquement, le stage consiste à étudier le comportement d'un hyper-paramètre utilisé lors de l'estimation de la régularité locale, et à proposer une méthode de sélection de ce dernier. Enfin cette méthode sera appliquée sur des données réelles du secteur énergétique.

contribution

si jamais vous apercevez des fautes méthodologiques ou orthographiques dans le rapport, merci de rédiger une *issue* sur Github à l'adresse :

correctif



ENSAI-stage_fin_etude-datastorm_fda_regularite-rapport/issues

contact



mail étudiant : hugo.brunet@eleve.ensai.fr

Table des matières

1 Motivations	2
2 Méthodologie	8
2.1 Données Fonctionnelles : l'essentiel	8
2.1.1 Généralités sur les données fonctionnelles	8
2.1.2 Cas non indépendant : séries temporelles de données fonctionnelles	12
2.2 Estimation de la régularité locale des trajectoires	13
2.2.1 Ce qu'on entend par régularité locale	13
2.2.2 Estimation des paramètres régularité locale des trajectoire	14
2.2.3 Prélissage	16
2.2.4 Ondelettes	17
2.3 Estimation adaptative	19
2.3.1 Estimation adaptative de la fonction moyenne	19
2.3.2 Estimation adaptative de l'opérateur de covariance	19
2.3.3 Estimation adaptative de l'auto-covariance des séries temporelles fonctionnelles	19
3 Applications et comparaison des différentes méthodologies	20
3.1 Données simulées	21
3.1.1 Objectifs de la simulation	21
3.1.2 Simulation d'un processus Brownien (multi)-Fractionnaire	21
3.1.3 Optimisation Algorithmique	21
3.2 Données Réelles	23
3.2.1 Courbes de charge éolienne	23
3.2.2 Données Hydrauliques	23
3.3 Conclusion sur l'efficacité des différentes méthodologies	23
A Note et réflexions personnelles sur le stage	24
A.1 Apprentissage	24
A.1.1 Implémentation d'un package R	24
A.1.2 Théorie et Pratique : il n'y a pas d'ordre total	24
A.2 Communication	25
A.2.1 Communication scientifique : Breizh Data Day	25

Chapitre 1

Motivations

Dans le cadre de ce stage, les données que l'on traite sont des données du secteur de l'énergie, et plus particulièrement des données de production électrique. On dispose ainsi de plusieurs éoliennes identifiées par le tag "id_(identifiant de l'éolienne)" dont l'énergie produite est mesurée toutes les demies heures, et ce pendant 4 ans (de 2014 à 2017). Cette énergie produite est dénommée la courbe de charge (que l'on abrégera par **CDC** par la suite). Il est cependant plus utile de s'intéresser au facteur de charge (ou **FDC**) qui est défini comme $\text{Facteur de Charge} = \frac{\text{Courbe de Charge}}{\text{Puissance Installée}}$. On en déduit que **FDC** doit nécessairement être compris entre 0 et 1. C'est entre autre aussi une manière de détecter des anomalies et données atypiques comme la surproduction d'énergie par rapport à ce qui était attendu de la part d'un parc éolien ou encore un défaut de capteur (tension / intensité, ...) qui mesure la courbe de charge.

Ainsi, les données qui sont traitées dans le cadre de ce stage sont, entre autres, des courbes de charge éoliennes observées chaque demie-heure. Le schéma d'observation est donc le 'common-design'. C'est-à-dire que les temps d'observations sont ici déterministes à intervalle de temps fixe.

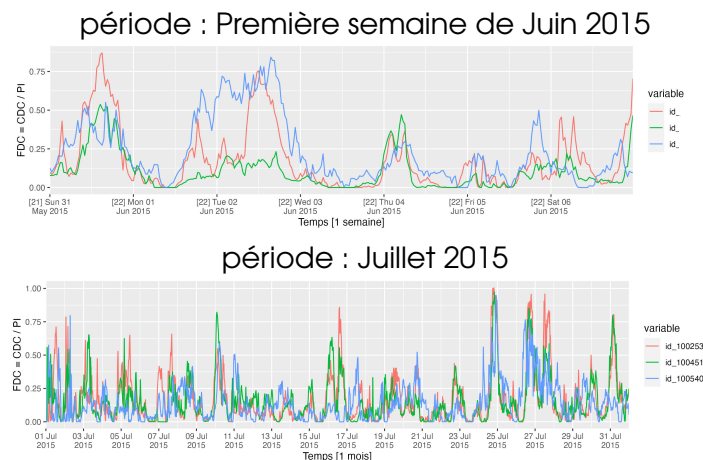


FIGURE 1.1 – Courbes de charges éoliennes sur 3 premiers parcs éoliens

Bien que la différenciation en analyse de séries temporelles soit une méthode efficace pour éliminer la tendance, qu'elle soit saisonnière ou non, permettant ainsi une bonne analyse des données; ces modèles présentent des limites en termes de prédiction à long

terme, les rendant moins utiles lorsque l'objectif est de prédire à moyen ou long terme. De plus, ces modèles, ainsi que différents modèles de machine learning populaires, estiment les données courbe par courbe ce qui ne tire pas profit du fait que les observations aient une forme similaire entre les courbes.

Une première idée serait d'utiliser un modèle de série temporelle ARIMA afin de modéliser la dynamique des courbes de charge.



Un peu d'histoire sur les séries temporelles ...



une grande partie des informations présentées dans cette section histoire provient de la référence (19)

Parmi les étapes importantes du développement des séries temporelles, on peut noter l'article *Time Series Analysis : Forecasting and Control* de Box et Jenkins (1970) qui introduit le modèle ARIMA et une approche aujourd'hui standard d'évaluation du modèle à utiliser ainsi que son estimation. Ce développement est dû en grande partie à l'utilisation de telles données dans les secteurs économiques et des affaires afin de suivre l'évolution et la dynamique de différentes métriques

L'étude des séries temporelle a été divisée en l'étude du domaine fréquentiel, qui étudie le spectre des processus pour le décomposer en signaux principaux, et du domaine temporel, qui étudie les dépendances des indices temporels. L'utilisation de chacune des approches était sujet à débats mouvementés jusqu'aux alentours de l'an 2000.

Le développement des capacités de calcul a été une révolution notamment pour l'identification des modèles (le critère AIC, l'estimation par vraisemblance dans les années 1980, ...).

À partir des années 1980, les modèles non linéaires émergent (ARCH par Engle, modèles à seuil ...) et trouvent application en économie notamment. Enfin l'étude multivariée (modèle VAR) fait surface dans les années 1980 par Christopher Sims (21, lien de l'article)

Une large partie de la théorie s'appuie notamment sur l'étude des racines de l'unité, en considérant un polynôme d'opérateur $P(B) = (I + \sum_k a_k B^k)$ à partir duquel les relations d'autocorrélations peuvent se ré-écrire.

Bien que naturelle, l'utilisation d'un modèle ARIMA ne permet de modéliser la dynamique du phénomène étudié. En effet, la sélection d'un modèle ARIMA sur le critère du BIC sélectionnait, peu importe le parc éolien, un modèle auto-régressif d'ordre 0. Ainsi le modèle sélectionné considèrerait les irrégularités de la courbe de charge, dont on attend que le processus duquel elle est issue soit très irrégulier (de par sa complexité), comme étant du bruit. On en conclut que ces modèles peuvent ne pas capturer efficacement la structure complexe des données.

Afin de prédire sur le long terme, nous allons donc adopter une approche basée sur les données fonctionnelles pour capturer la structure de la consommation. Cette approche permettra de d'exploiter une information clé : la similarité entre les courbes observées.



Qu'est-ce qu'une donnée fonctionnelle ?

Une donnée est dite fonctionnelle lorsque la variable aléatoire qui nous intéresse n'est plus une variable aléatoire à valeur dans \mathbb{R}^d , comme le statisticien a l'habitude de manipuler, mais une variable aléatoire à valeur dans un espace de fonction. Concrètement, chaque réalisation n'est plus un nombre mais bien une courbe toute entière indexée (le plus souvent) sur un intervalle \mathcal{T} .

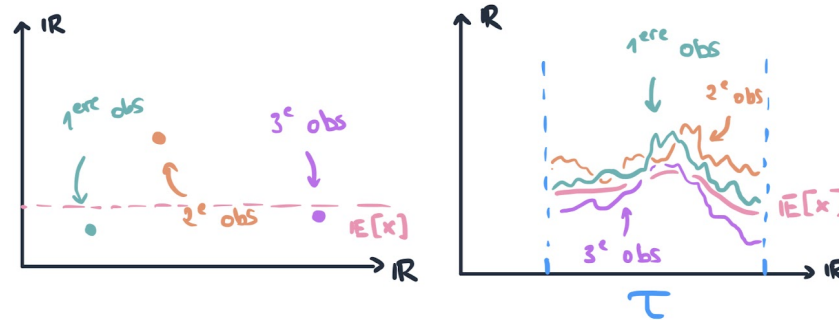


FIGURE 1.2 – Différence entre donnée fonctionnelle et donnée réelle

Si le statisticien est déjà à l'aise avec l'idée qu'une variable aléatoire réelle identiquement distribuée puisse modéliser une expérience répétable provenant d'un même phénomène, il pourra se convaincre que les données fonctionnelles permettent elles aussi de modéliser des expériences en lien (fonctionnel) avec un certain paramètre. Et c'est le lien entre les deux valeurs, cette fois-ci, qui provient d'un même phénomène.

Donnons en un exemple : observons la consommation électrique d'un foyer dans une journée. Lorsque l'on travaille sur \mathbb{R} , on s'intéresse à sa consommation électrique disons en l'instant $t = 12h$ ($\in \mathcal{T} \stackrel{\text{déf}}{=} [0, 24[= 1 \text{ jour avec } t \text{ en heure}$). La consommation du foyer i à midi, notée y_i , suit la loi d'un phénomène général Y , comme un normale $\mathcal{N}(50 \text{ kWh}, 9)$ par exemple. Travailler sur des données fonctionnelles dans ce cadre c'est étudier non plus la consommation y_i à midi, mais regarder l'ensemble de sa consommation en même temps sur toute la journée $y_i(t) = x_i(t)$ avec $t \in \mathcal{T}$.

On remarque ainsi que toutes les consommations électriques le long de la journée d'un foyer à l'autre suivent la même tendance : on consomme plus le matin avant le travail et le soir alors que pendant la journée on consomme moins car on est au travail. Ainsi c'est la fonction $x_i : \mathcal{T} \rightarrow \mathbb{R}$ qui suit la loi d'un phénomène X général. Ce que l'on vient de dire c'est que la **relation** entre le temps $t \in \mathcal{T}$ et la consommation électrique $y_i(t)$ est elle même sujet à une loi plus générale. Grossièrement, les courbes auront la même allure, mais chaque individu a sa consommation propre.

Plus formellement : comme on a défini une variable aléatoire réelle comme une application :

$$X : \begin{array}{ccc} \Omega & \longrightarrow & \mathbb{R} \\ \omega & \longmapsto & x = X(\omega) \end{array}$$

On définit de même une donnée fonctionnelle comme une application :

$$X : \begin{array}{ccc} \Omega & \longrightarrow & \mathcal{C}^0(\mathcal{T}, \mathbb{R}) \\ \omega & \longmapsto & x = X(\omega) \end{array}$$

Ce que l'on observe sont donc les valeurs des paramètres $t \in \mathcal{T}$ ainsi que l'image de t par $x : y = x(t)$. Les points que le statisticien observe sont donc les couples de la forme $(t_k^{(\text{individu } i)}, y_k^{(\text{individu } i)})_{i \in \llbracket 1, m \rrbracket}$, générés par le processus aléatoire X dont la réalisation est la véritable courbe x_i de l'individu i que l'on souhaite estimer pour travailler avec.



Certaines ressources sur l'analyse de données fonctionnelles définissent les données fonctionnelles de la manière suivante

$$X : \begin{array}{ccc} \Omega \times \mathcal{T} & \longrightarrow & \mathbb{R} \\ (\omega, t) & \longmapsto & X(\omega, t) = y \end{array}$$

qui selon mon humble avis, ne permet pas une interprétation clé en main du concept mais certainement plus commode à manipuler pour les mathématiciens.



...et un peu d'histoire sur les données fonctionnelles



Pour une description plus complète de l'histoire du développement de l'analyse fonctionnelle, on pourra se référer à [cet article de Wang, Chiou et Müller \(20\)](#)

Bien que l'histoire du développement de l'Analyse de Données Fonctionnelles (FDA) puisse être retracée jusqu'aux travaux de Grenander et Karhunen (11) dans les années 1940 et 1950, où l'outil a été utilisé pour étudier les courbes de croissance en biométrie, ce sous-domaine de la statistique a été étudié de manière systématique à partir des années 1980.

En effet, c'est J.O. Ramsay qui a introduit l'appellation de "données fonctionnelles" en 1982 (16) et qui contribuera en partie à sa popularisation. La thèse de Dauxois et Pousse en 1976 sur l'analyse factorielle dans le cadre des données fonctionnelles(5) a ouvert la voie à l'analyse par composante principale fonctionnelle (FPCA), un outil clé pour l'étude des données fonctionnelles. La FPCA permet d'étudier des objets fonctionnels qui sont de dimension infinie, difficiles à manipuler et impossibles à observer empiriquement, en dimension finie et surtout sur \mathbb{R}^d que l'on connaît bien.

Au cours des années 2000, de nombreux outils statistiques déjà développés pour des données à valeurs dans \mathbb{R}^d depuis un siècle, tels que la régression linéaire (éventuellement généralisée), les séries temporelles ou encore les modèles additifs, ont été adaptés aux données fonctionnelles. Par exemple, les modèles de régression linéaire fonctionnelle ont été développés avec une réponse fonctionnelle (17) ou scalaire (3) en 1999. Les modèles linéaires généralisés ont également été étudiés (10, 15), avec l'estimation de la fonction de lien par méthode non paramétrique à direction révélatrice (*Single Index Model*) récemment étudiée en 2011 (4). Cette méthode avait déjà été utilisée en économétrie pour des données de \mathbb{R}^d depuis 1963 (18), et leur estimation directe a été étudiée une décennie auparavant par M.Hristache, Juditsky et Spokoiny (9). De même, les modèles additifs ont été étendus aux données fonctionnelles en 1999 par Lin et Zhang (13). Enfin, le livre de Bosq, [Linear Processes in Function Spaces : Theory and Applications](#) (1), publié en 2000, a contribué au développement des séries temporelles pour les données fonctionnelles.

Depuis lors, des ressources telles que l'ouvrage de Kokoszka et Reimherr, [Introduction to Functional Data Analysis \(2017\)](#) (12), rendent la théorie et la mise en production des méthodes d'analyse et de prédiction de données fonctionnelles plus accessibles.

Maintenant que l'on possède une meilleure intuition de ce que sont les données fonctionnelles, il est naturel de se demander pourquoi le choix de modéliser notre phénomène par des données fonctionnelles serait particulièrement judicieux. Pour cela, rappelons nous les difficultés que l'on avait rencontrées dans le cadre de nos données de production électrique en utilisant un modèle de série temporelle classique :



Rappel :

"Ainsi le modèle ([arima](#)) sélectionné considèrerait les irrégularités comme étant du bruit [...]. Afin de prédire sur le long terme, nous allons donc adopter une approche basée sur les données fonctionnelles pour capturer la structure de la consommation [...]."



Pourquoi est-ce que l'on s'intéresse autant à la régularité des données que l'on étudie ici ? Et surtout, en quoi est ce que les données fonctionnelles vont nous permettre de mieux capturer la régularité ?

Comme mentionné auparavant, la production électrique est un phénomène très irrégulier (1.1) étant influencé par la consommation, la météo, etc. Par conséquent, la prévision de ces courbes de charge doit prendre en compte la nature fondamentalement irrégulière du phénomène afin de proprement le modéliser et, en définitive, mieux le prédire. Ce qui est notamment contraire à la plupart des méthodes qui utilisent des fonctions de classe \mathcal{C}^2 pour lisser les points observés en données fonctionnelles, ce qui limite la prédiction à des courbes de nature \mathcal{C}^2 . Cela est d'autant plus critique lorsque l'on cherche à estimer le processus moyen ou l'opérateur de covariance du processus, car ces derniers sont estimés à partir des courbes lissées, qui détruisent toute l'information irrégulière si elle n'est pas prise en compte, impactant significativement l'estimation des objets qui nous intéressent en tant que statisticien.

Il est ainsi important pour des phénomènes de nature irrégulière de ne pas négliger des précautions lors du lissage afin de ne pas perdre l'information irrégulière. L'idée est donc d'estimer dans un premier temps la régularité de notre processus afin de lisser nos données de manière adaptée pour débruiter et prédire des valeurs non observées tout en préservant les informations irrégulières critiques pour la bonne estimation du processus moyen et de l'opérateur de covariance. L'approche fonctionnelle est clé dans l'estimation de cette régularité, car c'est la **réplication de courbes** de même nature qui permet in-fine d'**estimer la régularité** du phénomène.

Chapitre 2

Méthodologie

Contents

2.1	Données Fonctionnelles : l'essentiel	8
2.1.1	Généralités sur les données fonctionnelles	8
2.1.2	Cas non indépendant : séries temporelles de données fonctionnelles	12
2.2	Estimation de la régularité locale des trajectoires	13
2.2.1	Ce qu'on entend par régularité locale	13
2.2.2	Estimation des paramètres régularité locale des trajectoire	14
2.2.3	Prélassage	16
2.2.4	Ondelettes	17
2.3	Estimation adaptative	19
2.3.1	Estimation adaptative de la fonction moyenne	19
2.3.2	Estimation adaptative de l'opérateur de covariance	19
2.3.3	Estimation adaptative de l'auto-covariance des séries temporelles fonctionnelles	19

2.1 Données Fonctionnelles : l'essentiel

2.1.1 Généralités sur les données fonctionnelles

2.1.1 □ A) définitions et propriétés informelles

Commençons par introduire les données fonctionnelles de manière informelle afin de mieux intégrer la définition formelle, plus utile pour la manipulation.

Cette section regroupe l'ensemble des messages essentiels à retenir des données fonctionnelles pour la pratique, sans alourdir les notions avec des notations mathématiques. Le cadre formel sera traité juste après.

Définition (données fonctionnelles — informel) Les données fonctionnelles sont des données dont les observations sont des fonctions, c'est-à-dire des courbes, des surfaces, des images, ...

i.e : toute donnée ayant une dépendance de type "relation fonctionnelle" avec un ou plusieurs paramètres.

Maintenant introduites, les théorèmes suivant permettent de manipuler ces données à la fois pour la théorie et la pratique :

Théorème (Karhunen-Loeve — informel)

Il est possible pour une large classe de données fonctionnelles de les décomposer dans une base adaptée aux données (au sens de la covariance) que l'on appelle base ACP fonctionnelle (FPCA).

⚙️ *preuve informelle.* La covariance est un opérateur bilinéaire symétrique défini positif, on peut donc appliquer le théorème de Mercer (équivalent du théorème spectral) qui nous donne une base orthonormale de \mathbb{L}^2 sur laquelle on va décomposer notre processus **centré**. □

Remarque : La classe de fonctions pouvant être décomposées est large, puisqu'elle regroupe l'ensemble des processus qui nous intéressent la plus part du temps en tant que statisticien : celles qui sont à support sur un intervalle, admettant une covariance continue et finie sur le support.

On en déduit que pour travailler avec des données fonctionnelles, il suffit de les décomposer dans la base ACP fonctionnelle puis de travailler sur les composantes de chaque élément de la base. On travaille désormais avec des réels et non plus des fonctions, ce qu'on aime manipuler. On peut alors faire de la statistique traditionnelle avec les outils que l'on connaît.

Propriété (intérêt de la base FPCA — informel)

la base ACP fonctionnelle est la plus économe, c'est à dire qu'elle explique au mieux la covariance des données pour un nombre de composantes fixées, ce qui est utile car on ne sait manipuler numériquement que des objets de dimension finie.

On a mentionné qu'il serait judicieux de lisser les observations en tenant compte de la régularité du processus dont est issu nos données. La question est désormais la suivante :



Est-il possible de récupérer la régularité locale des trajectoires à partir des données ? Si oui, comment ?

C'est ce qu'affirme le théorème suivant à partir des travaux de Golovkine et al. ainsi que Maissoro-Patilea-Vimond (MPV) :

Théorème (Regularité locale — informel)

Les données fonctionnelles permettent de récupérer la régularité locale des trajectoires. Les estimateurs définis **ponctuellement** convergent.

Remarque : Les estimateurs sont définis à partir de l'espérance des incréments quadratiques du processus.

Les motivations de l'obtention de la régularité étaient en partie de pouvoir mieux estimer la fonction moyenne du processus, ainsi que son opérateur de covariance. Ce qui est à la fois important pour l'analyse (via l'interprétation de la base ACP déterminée par

la covariance) et pour la prédiction. On peut alors se demander si il existe des estimateurs de la moyenne et de la covariance prenant en compte la régularité locale. C'est ce qu'affirme les théorèmes suivants :



demander à Hassan la dernière version de son papier car la partie d estimation adaptative a beaucoup changé

Théorème (Estimateurs de la moyenne et de la covariance — informel (6))

Il est possible en lissant les observations par méthode à noyaux avec une largeur de bande *spécifique à l'objet que l'on souhaite estimer*, de dériver des estimateurs de la moyenne et de la covariance qui convergent. La largeur de bande optimale *pour l'objet que l'on souhaite estimer* est celle qui minimise un risque qui effectue un compromis biais-variance, qui dépend de la régularité locale du processus, en pénalisant les largeurs de bande menant à des "trous" dans les fonctions lissées. On parle d'*estimation adaptative*.

Théorème (expression de la largeur de bande optimale — informel (6))

Sous certaines hypothèses de régularité du processus, et d'indépendance des temps observés, la largeur de bande optimale peut être approchée (avec forte probabilité de bonne approximation) par une expression ne dépendant que du nombre de courbes observées, du nombre moyen de temps observés par courbe, et de la régularité locale du processus. Ce biais de l'estimateur de la fonction moyenne est alors contrôlé en fonction de ces mêmes quantités.

Sous des hypothèses un peu plus fortes sur le nombre d'observations par courbe, et le nombre de courbe on dispose de résultats similaires pour l'estimateur de la covariance.

Enfin, on peut se demander ce qu'il en est des estimateurs dans le cadre où l'on dispose de la dépendance dans les données (ce qui est le cas pour les données éoliennes notamment). Ce cas est traité par le théorème suivant dérivé par MPV :

Théorème

2.1.1 □ B) définition formelle et premières propriétés

Pour éviter d'alourdir les notations, on se place dans le cas où les fonctions sont à valeurs dans \mathbb{R} et à support sur un intervalle fermé I de \mathbb{R} . Toutefois, on peut très bien considérer des fonctions à valeurs dans \mathbb{R}^d et à support sur un compact K de \mathbb{R}^p sans perte de généralités.

Définition 1 (données fonctionnelles) On appelle données fonctionnelles, un échantillon $(x_i)_{1,n}$ de fonctions continues $x_i : I \rightarrow \mathbb{R}^d$ issues d'un processus X défini comme ci-dessous :

$$X : \begin{array}{ll} \Omega & \longrightarrow \mathcal{C}(I, \mathbb{R}) \\ \omega & \longmapsto X(\omega) = x \end{array}$$

2.1.1 □ C) résultats importants

On énonce désormais le théorème central de l'analyse de données fonctionnelles qui n'est autre que la décomposition dans la base FPCA de notre processus.

Remarque : on notera que dans le cadre des données fonctionnelles, on ne travaille pas de façon générale avec la covariance $C_X : (s, t) \mapsto \mathbb{E}[(X - \mu)(s) \cdot (X - \mu)(t)]$ mais avec un **opérateur** covariance $f \mapsto \int_I f(u) C_X(u, \cdot) du$. C'est parce que cet opérateur est linéaire continu (car Hilbert-Schmidt donc borné pour la norme d'opérateur) symétrique semi-défini positif (pour le produit scalaire de \mathbb{L}^2) et que l'on peut donc en faire une décomposition spectrale sur une base orthonormale de vecteurs propres associés à des valeurs propres positives. Cette décomposition est à la base des approximations que le praticien effectuera ainsi qu'à la base de la dérivation de nombreux théorèmes et propriétés.

$$\text{on note } \langle \cdot | \cdot \rangle_{\mathbb{L}^2} : \begin{array}{ccc} \mathbb{L}^2 \times \mathbb{L}^2 & \longrightarrow & \mathbb{R} \\ (f, g) & \longmapsto & \int f(u)g(u) du \end{array}$$

Théorème 1 (Karhunen-Loeve)

référence : [12, pages : 238-239-241]

Hypothèses :

▣ $X \in \mathbb{L}^2(\Omega, \mathcal{C}(I, \mathbb{R}))$

▣ covariance : $C : \begin{array}{ccc} \mathbb{L}^2(\Omega, \mathcal{C}(I, \mathbb{R})) & \longrightarrow & \mathcal{C}(I^2, \mathbb{R}) \\ X & \longmapsto & C_X \end{array}$

ie : $C_X : (s, t) \mapsto C_X(s, t)$ est continue

* opérateur covariance $c_X[\cdot] : \begin{array}{ccc} \mathcal{C}(I, \mathbb{R}) & \longrightarrow & \mathcal{C}(I, \mathbb{R}) \\ f & \longmapsto & \int_I f(s) C_X(s, \cdot) ds \end{array}$

▣ valeurs propres ordonnées : $\forall p \geq 1, \lambda_{p+1} \leq \lambda_p \quad \lambda_p, \lambda_{p+1} \in \text{sp}(c_X)$

* on pose $\vec{s\hat{p}}_{\|\cdot\|}^{[1,p]}(c_X) \stackrel{\text{déf}}{=} \left\{ \phi_k \in \vec{s\hat{p}}_{\|\cdot\|}(c_X) \text{ associé à } \lambda_k, k \in \llbracket 1, p \rrbracket \right\}$

alors :

▣ $\forall p \geq 1 \quad \underset{u_k \in \mathcal{C}(I, \mathbb{R})}{\text{argmin}} \mathbb{E} \left\| X - \sum_{k=1}^p \langle X - \mu | u_k \rangle_{\mathbb{L}^2} u_k \right\|^2 = \vec{s\hat{p}}_{\|\cdot\|}^{[1,p]}(c_X)$

▣
$$X = \mu + \sum_{k=1}^{+\infty} \langle X - \mu | \phi_k \rangle \phi_k$$

avec $\phi_k \in \vec{s\hat{p}}_{\|\cdot\|}(c_X)$

Remarque : pour pouvoir ordonner les valeurs propres dans l'ordre décroissant, et sélectionner les composantes principales les plus informatives, il faut pouvoir réarranger l'ordre

de la somme. Pour cela il faut que les valeurs propres forment une famille sommable, une condition suffisante et souvent utilisée est que $\mathbb{E}\|X\|^2 < \infty$

Remarque : la propriété de la section précédente sur l'aspect économe de la base FPCA découle directement de l'assertion $\forall p \geq 1 \quad \operatorname{argmin}_{u_k \in \mathcal{C}(I, \mathbb{R})} \mathbb{E} \left\| X - \sum_{k=1}^p \langle X - \mu | u_k \rangle u_k \right\|^2 = \overrightarrow{sp}_{\|\cdot\|}^{[1,p]}(c_X)$ dans le théorème de Karhunen-Loeve.

2.1.2 Cas non indépendant : séries temporelles de données fonctionnelles

Une large partie de la théorie des données fonctionnelles suppose que l'on observe des courbes $X_i : \Omega \rightarrow \mathcal{C}^0(I, \mathbb{R})$ **indépendantes** et identiquement distribuées. Cependant une partie non négligeable des données que l'on observe ont des dépendances avec les valeurs passées. Par exemple, il est raisonnable de penser que la consommation électrique d'un foyer au cours d'une année croît avec l'ajout successif de nouveau appareils électroniques. L'hypothèse d'indépendance entre les données n'est donc plus pertinente pour les données que l'on traite et il devient important de considérer des processus auto-régressifs adaptés aux données fonctionnelles. Si dans le cadre des données de \mathbb{R} cette relation de *dépendance linéaire* avec le passé pouvait s'écrire sous la forme suivante

$$X_n = \sum_{k=1}^{n-1} \varphi_k X_k + \varepsilon_n \quad \text{où } \varphi_k \in \mathbb{R} \text{ et } \varepsilon_n \begin{cases} \in \text{VA}(\mathbb{R}) \\ \perp\!\!\!\perp \sigma(X_i)_{1:n-1} \end{cases}, \text{ dans le cadre fonctionnel on cap-}$$

ture la même idée en considérant $X_n = \sum_{k=1}^{n-1} \phi_k(X_k) + \varepsilon_n$ où ϕ_k est un *opérateur linéaire* de $\mathbb{L}^2(I, \mathbb{R})$, le plus souvent intégral.



Il s'agit d'une généralisation naturelle de la relation dans le cadre réel, puisqu'on peut démontrer que sur l'espace des nombres réels l'ensemble des fonctions linéaires $\phi : \mathbb{R} \rightarrow \mathbb{R}$ sont de la forme $x \mapsto ax$ avec $a \in \mathbb{R}$. La relation sur \mathbb{R} que l'on a vue juste avant peut alors se ré-écrire de façon similaire à la version fonctionnelle.

On considère lors de ce stage des séries temporelles de données fonctionnelles car les données que l'on manipule (en l'occurrence les données de courbe de charge des parcs éoliens) semble être naturellement corrélées dans le temps.



Il faut cependant faire attention avec l'interprétation de séries temporelles fonctionnelles. En effet l'interprétation que l'on fait d'une série temporelle fonctionnelle ne peut être véritable calquée sur l'interprétation dans la cadre réel.

Si sur \mathbb{R} la corrélation temporelle s'interprète comme le fait que l'observation suivante (c'est à dire le facteur de charge du parc éolien) dépend de l'observation précédente (c'est à dire dépend du jour précédent par exemple), dans le cadre fonctionnel l'observation est une fonction. On rappelle que dans le cadre des données éoliennes, chaque observation est la courbe de charge observée sur un an. Ce que cela signifie, c'est que la dépendance à l'observation passée que l'on va constater sur l'observation suivante n'affecte pas un unique point. Elle affecte tous les points de la courbe de charge simultanément. La dépendance temporelle concerne l'année entière, d'une année à l'autre. Fixer le t dans $X_n(t)$ ne garantit pas l'observation de la même structure de dépendance comparé à la série temporelle fonctionnelle.

Par exemple, en utilisant un opérateur intégral pour ϕ_k , comme $\phi_k : X \mapsto \int_I \beta_k(u, t) X(u) du$, la vision de regarder les t individuellement et de regarder leur évolution selon les années n'est pas évidente à interpréter car $\phi_k(X)$ intègre des informations de X sur l'ensemble des $t \in I$, dû à l'intégrale.

2.2 Estimation de la régularité locale des trajectoires

2.2.1 Ce qu'on entend par régularité locale

Longtemps, il était cru que les fonctions continues étaient dérivables presque partout. C'est notamment Weierstrass qui a démontré qu'il existe des fonctions continues partout mais dérivable nulle part. Poincaré notamment disait de tels objets qu'ils n'existaient que pour contredire le travail des pères. Cependant, des objets manipulés tous les jours comme le monde de la finance notamment traitent des processus qui sont fondamentalement irréguliers¹ (au point de vue de l'analyse, où l'on traite souvent des fonctions au moins dérivables). Il est donc important de pouvoir quantifier la régularité d'une fonction de façon plus fine que le nombre de dérivées qu'elle possède.

Nous allons repasser rapidement en revue les différents concepts de régularité pour mettre l'accent dans ce que l'on considère comme régularité locale.

Afin de savoir à quel niveau de régularité nous souhaitons estimer, il est important de garder en tête un ordre de différents niveaux de régularité résumé par les relations suivantes :

Lipschitz \implies Hölder \implies Localement Hölder \implies Uniformément continue \implies Continue
ce qui nous intéresse

Afin de mieux discerner ce que chaque propriété signifie, et quelles sont les différences entre chaque niveau de régularité, nous allons rappeler rapidement les définitions de ces propriétés.

1. ces fonctions sont même nombreuses, car dense dans les fonctions continues pour la topologie de la convergence uniforme (8)

— Continuité :

$$(\forall \varepsilon > 0) (\forall x) (\exists \delta_x > 0) (\forall y) |x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

— Uniforme Continuité :

$$(\forall \varepsilon > 0) (\exists \delta > 0) (\forall x, y) |x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

— Lipschitz :

$$(\forall x, y) |f(x) - f(y)| < L|x - y|$$

— Hölder :

$$\begin{cases} (\forall x, y) |f(x) - f(y)| < L_\alpha |x - y|^\alpha \\ 0 < \alpha \leq 1 \end{cases}$$



une fonction lipschitz est une fonction Holderienne avec $\alpha = 1$

— Localement Hölder :

$$\forall x_0 \begin{cases} (\forall x) |f(x) - f(x_0)| < L_{\alpha(x_0)} |x - x_0|^{\alpha(x_0)} \\ 0 < \alpha(x_0) \leq 1 \end{cases}$$



Pourquoi se concentrer sur des processus localement Hölder ?

La nature des phénomènes rencontrés dans la vie réelle est souvent complexe. Influencés par de nombreux phénomènes, certains d'entre eux sont, comme mentionnés précédemment, irréguliers. C'est notamment le cas des courbes de charge électriques, qui dépendent de multitudes de phénomènes physiques ou comportementaux, dont on peut attendre une certaine régularité, mais qui ne sont pas nécessairement uniformes tant sur leur niveau régularité que l'intervalle de temps sur lequel ils ont une influence. On pourrait par exemple attendre une différence de régularité de la production électrique en plein été (soleil et température stables ...) comparé au mois de mars (plus grande instabilité des conditions climatiques).

De plus, les fonctions Hölderiennes représentent une classe suffisamment large de fonctions. ⚠ (citer le papier sur le fractional brownian motion (how large)) L'espace de fonctions sur lequel on travail est donc devrait être en pratique suffisamment grand pour inclure l'ensemble des processus qui nous intéressent. Enfin les fonctions que le praticien sera amené à manipuler seront des fonctions d'un intervalle dans \mathbb{R} , qui lorsque continues sont automatiquement uniformément continues en vertu du théorème de Heine. Il est donc naturel de se concentrer sur des fonctions localement Hölderiennes.

2.2.2 Estimation des paramètres régularité locale des trajectoire

2.2.2 □ A) Deux méthodes d'obtention de la régularité locale des trajectoires

Il existe deux méthodes différentes pour estimer la régularité des trajectoires. Si la clé des deux méthodes pour extraire la régularité locale est le théorème de continuité de

Kolmogorov énoncé ci-dessous, les deux méthodes diffèrent par les points $t \in \mathcal{T}$ considérés dans l'estimation des accroissements quadratiques $\mathbb{E} [|X(u) - X(v)|^2]$ utilisés pour l'estimation de la régularité locale.

Théorème 2 (Continuité de Kolmogorov)

référence : (2, thm : 2.197 | page : 145)

$$\blacksquare \quad X : \begin{array}{ccc} \mathbb{R}_+ \times \Omega & \longrightarrow & \mathbb{R} \\ (t, \omega) & \longmapsto & X(t, \omega) = x(t) \end{array} \quad \text{séparable}$$

$$\blacksquare \quad \exists r, c, \varepsilon, \delta \in \mathbb{R}_+ \quad (\forall h < \delta)(\forall t \in \mathbb{R}_+) \quad \mathbb{E} [|X(t+h) - X(t)|^r] \leq c \cdot h^{1+\varepsilon}$$

\Downarrow

$$\ast \quad X \text{ est continu en } t \in \mathbb{R}_+ \text{ pour presque tout } \omega \in \Omega$$

ie : il existe une version \tilde{X} de X continue en t telle que $\mathbb{P} [\tilde{X}(t) = X(t)] = 1$

$$\ast \quad \tilde{X} \text{ est } \gamma\text{-Hölderienne en } t \text{ pour tout } 0 < \gamma < \frac{\varepsilon}{r}$$

Etant donné que notre estimateur utilise les incréments quadratiques, on se place dans le cas où $r = 2$.

⚠ (Mettre la version avec Hölder, qui permet de dériver des fda la régularité locale)

La méthode de Golovkine et al. (7, pages : 7—9) n'utilise que les points observés, et construit un estimateur des incréments quadratiques à base de statistique d'ordre.

$$\theta(T_{(l)}, T_{(k)}) = \mathbb{E} [|X(T_{(l)}) - X(T_{(k)})|^2] \underset{\substack{\text{LGN} \\ \text{Hölder} \\ \text{+C}^0 \text{Kol.}}}{\approx} \boxed{\frac{1}{N} \sum_{n=1}^N |Y_n^{(2k-1)} Y_n^{(k)}|^2 \stackrel{\text{déf}}{=} \hat{\theta}_k} \underset{L_{t_0} \mathbb{E} [|T_{(l)} - T_{(k)}|^{2H_{t_0}}]}{\rightarrow} H_{t_0} = f(\theta)$$

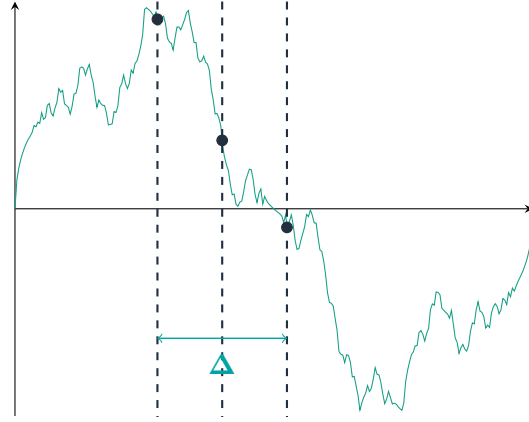
$$\hat{H}_{t_0}(k) = \begin{cases} \frac{\log(\hat{\theta}_{4k-3} - \hat{\theta}_{2k-1}) - \log(\hat{\theta}_{2k-1} - \hat{\theta}_k)}{2 \log 2} & \hat{\theta}_{4k-3} > \hat{\theta}_{2k-1} > \hat{\theta}_k \\ 1 & \text{sinon} \end{cases}$$



Cette méthode peut s'avérer spécifiquement utile lorsque l'on traite un flux de données, car l'arrivée de nouvelles données ne nécessite pas spécifiquement de recalculer les incréments quadratiques sur l'ensemble des points observés.

L'autre méthode proposée par (6, 14), elle se base sur l'utilisation de points non observés, inférés par lissage des courbes, à une distance $\Delta/2$ les uns des autres pour estimer les incréments quadratiques. Cette dernière méthode implique le choix d'un hyperparamètre lors de l'estimation Δ et pourrait être sensible à la qualité du lissage de la

courbe. Etant donné que l'objectif de la détermination de la régularité locale est de pouvoir faire un lissage à noyaux adaptatif en fonction de l'objet que l'on souhaite estimer, on appelle le lissage effectué pour estimer la régularité « pré-lissage ».



soit $\Delta \in \mathbb{R}_+^*$, on pose $t_1 \stackrel{\text{déf}}{=} t_0 - \frac{\Delta}{2}$ et $t_3 \stackrel{\text{déf}}{=} t_0 + \frac{\Delta}{2}$.

alors on approche $\theta(t_1, t_3) = \mathbb{E} \left[|X(t_3) - X(t_1)|^2 \right] = \theta_{13}$ par :

$$\tilde{\theta}_{13} = \frac{1}{N} \sum_{n=1}^N |X(t_3) - X(t_1)|^2$$

qui n'est pas observable, étant donné qu'il n'est pas garanti d'observer $X(t_1)$ et $X(t_3)$, et qu'il faut donc lisser dans un premier temps les courbes pour pouvoir évaluer X en t_1 et t_3 . L'estimateur que l'on considère est donc une approximation de $\tilde{\theta}_{13}$, et est défini par :

$$\hat{\theta}_{13} = \frac{1}{N} \sum_{n=1}^N \left| \hat{X}(t_3) - \hat{X}(t_1) \right|^2$$

où \hat{X} est la courbe lissée à partir des observations $(T_i^{[n]}, Y_i^{[n]})_{n \in 1:N, i \in 1:M_n}$

2.2.3 Prélissage

Afin de pouvoir estimer la régularité locale des trajectoires, nous allons lisser les trajectoires. En effet, celles-ci sont souvent bruitées, et il est nécessaire de lisser les trajectoires afin de pouvoir estimer la régularité locale de façon pertinente. De plus si on veut estimer la régularité en un point non observé, il devient alors nécessaire de lisser les trajectoires afin de pouvoir estimer la régularité en ce point. Cette étape de lissage est appelée pré-lissage de la courbe.



Pourquoi parle-t-on de **pré**-lissage ? Le but de considérer la régularité n'était-il pas justement de l'utiliser dans le lissage des trajectoires ? Lisser avant même d'estimer la régularité n'est-il pas contre-productif ?

L'objectif de l'obtention des paramètres de régularité des trajectoires est de pouvoir effectuer un lissage de ces trajectoires qui préserve les irrégularités fondamentales du processus dont elles sont issues, tout en éliminant le bruit. Les paramètres de régularité sont donc dans un premier temps estimés en utilisant des trajectoires lissées puis utilisés pour effectuer un nouveau lissage à noyaux en utilisant une fenêtre de lissage appropriée qui dépend de ces paramètres. ⚠ (inclure l'expression de $h_\alpha(t)$)

2.2.3 □ A) pré-lissage Spline

Le lissage spline est certainement une des méthodes de lissage les plus répandues de par sa simplicité d'implémentation. De plus la détermination des hyper-paramètres de lissage via la méthode de GCV permet de déterminer une approximation de base optimale à un coût computationnel relativement faible. Un des plus grands avantages du lissage B-Spline est l'obtention d'une base de fonctions, qui permet à coût de stockage faible de pouvoir prédire des points non observés. Une fois la base déterminée, il ne reste plus qu'à prédire les points non observés en utilisant la base de fonctions et les coefficients de la décomposition de la courbe sur cette base.

2.2.3 □ B) pré-lissage à noyaux

2.2.3 □ C) écrasement des irrégularités? vers le pré-lissage à ondelettes

Le lissage spline donne une fonction de classe \mathcal{C}^2 , ce qui est un désavantage dans le cadre du pré-lissage qui sert à déterminer les paramètres de régularité de courbes issues d'un processus que l'on ne suppose pas plus régulier que continu. Toutefois, le fait d'utiliser une base de fonctions pour effectuer le lissage a de nombreux avantages par rapport au lissage à noyaux qui peuvent éventuellement s'avérer utiles dans certaines situations spécifiques pour la mise en production de modèles.

En effet, une fois que l'on a déterminé les composantes de la décomposition de notre signal sur la base de fonctions, on n'a plus besoin de se référer aux données pour prédire une valeur. Il s'agit d'une méthode très économe en mémoire, ce qui peut être très avantageux dans le cadre de la mise en production de modèles lorsqu'il y a de nombreuses courbes observées.



2.2.4 Ondelettes

2.2.4 □ A) Une brève introduction aux ondelettes

Les ondelettes proviennent du monde du traitement du signal. Elles répondent à un problème de représentation des données à la fois dans le domaine temporel et dans le domaine fréquentiel. En effet, la transformée de Fourier nous donne accès aux fréquences présentes dans un signal mais ne nous permet pas de localiser à quel moment sont intervenues les fréquences spécifiques. Le théorème d'indétermination de Heisenberg stipule que l'on ne peut avoir une résolution parfaite à la fois dans le domaine fréquentiel et le domaine temporel, il y a un compromis qui doit être fait. La question devient alors :



Comment représenter une fonction dans le domaine temporel et dans le domaine fréquentiel de façon optimale ? En d'autres termes, quelle résolution temporelle et quelle résolution fréquentielle choisir ?

Une première approche proposée en  (compléter date) par  (compléter auteur) est la transformée de Fourier à court terme (STFT). Celle-ci consiste à regarder la transformée de Fourier d'une fonction sur une fenêtre de taille fixe et à faire glisser cette fenêtre sur la fonction. On obtient ainsi la représentation fréquentielle de la fonction sur un intervalle de temps centré en un point que l'on peut faire varier.

Cependant contrairement à ce que peut suggérer le dessin présenté ici, la résolution fréquentielle n'est pas parfaite. Elle est d'ailleurs dans le cadre de la Transformée de Fourier à court terme constante, que ce soit sur le domaine temporel ou le domaine fréquentiel. La résolution fréquentielle est donc constante quelque soit la fréquence considérée.

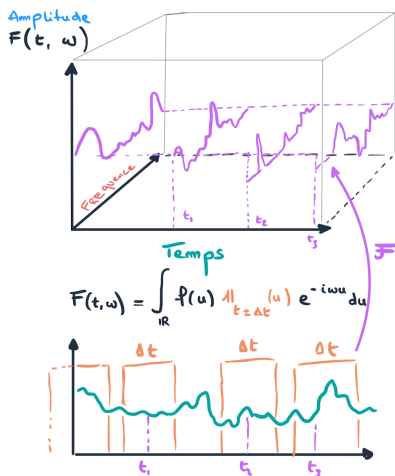


FIGURE 2.1 – Transformée de Fourier à court terme d'une fonction



Quel est le problème avec cette approche ?

le problème ne vient pas du monde mathématique mais plutôt du monde réel : les signaux que l'on observe présentent la caractéristique suivante : Les signaux de basse fréquence ont tendance à s'étendre sur la durée, et les signaux de hautes fréquences ont tendance à être très localisées, sous forme d'impulsion. Il devient alors clair que pour correctement identifier et localiser les fréquences présentes dans un signal, il est judicieux (voire parfois nécessaire) de varier la résolution fréquentielle et temporelle (limitées par le théorème d'indétermination de Heisenberg) en fonction de ce qui est le plus difficile à distinguer. C'est ce que proposent les ondelettes.

2.2.4 □ B) Théorie de la base ondelettes

Transformée en ondelettes

Introduisons maintenant de façon plus formelle les ondelettes et regardons leurs propriétés intéressantes dans le cadre du lissage de trajectoires.

on définit la transformée en ondelettes vis à vis de l'ondelette mère ψ d'une fonction f par :

$$F : \begin{array}{ccc} \mathbb{R} \times \mathbb{R}_+ & \longrightarrow & \mathbb{R} \\ (t, s) & \longmapsto & \frac{1}{\sqrt{|s|}} \int_{\mathbb{R}} f(u) \psi\left(\frac{u-t}{s}\right) du \end{array}$$



on peut remarquer que la formule de la transformée en ondelettes ressemble à une projection : $\frac{\langle f, \psi_{t,s} \rangle_{\mathbb{L}^2}}{\|\psi_{t,s}\|}$. Cela vient en quelque sorte motiver la section suivante

Base d'ondelettes

$$\left(\psi_{k,n} : t \mapsto \frac{1}{\sqrt{2^k}} \psi\left(\frac{t - 2^k n}{2^k}\right) \right)_{(k,n) \in \mathbb{Z}^2} \text{ est une base } \frac{1}{\|\cdot\|} \text{ de } \mathbb{L}^2$$



notons que les résolutions sont des puissances de 2, ceci est un détail qui demandera une implémentation particulière dans le cadre des données réelles : il faudra faire attention à ce que le nombre de points que l'on donne dans l'algorithme de transformée rapide en ondelettes soit aussi une puissance de 2.

Propriétés principales des ondelettes

- **Approximation dans l'espace fréquentiel-temporel** : La transformée en ondelettes ($\mathcal{W} : f \mapsto \langle f | \psi_{t,s} \rangle$) est une isométrie de \mathbb{L}^2 . Cela nous permet donc d'affirmer que $\|f - \hat{f}\|_{\mathbb{L}^2} = \|\mathcal{W}f - \mathcal{W}\hat{f}\|_{\mathbb{L}^2}$. Ainsi on peut travailler dans l'espace des ondelettes pour approximer (dans notre cas lisser les trajectoires) des fonctions et contrôler l'approximation directement dans le domaine fréquence-temporel tout en le conservant dans le domaine temporel. (citation requise)
- **Propriété de Fast Decay**

2.2.4 ☐ C) Motivation dans le cadre de l'analyse de données fonctionnelles

2.2.4 ☐ D) Effets du lissage à ondelettes sur la régularité locale

2.3 Estimation adaptative

2.3.1 Estimation adaptative de la fonction moyenne

2.3.2 Estimation adaptative de l'opérateur de covariance

2.3.3 Estimation adaptative de l'auto-covariance des séries temporelles fonctionnelles

Chapitre 3

Applications et comparaison des différentes méthodologies

Contents

3.1	Données simulées	21
3.1.1	Objectifs de la simulation	21
3.1.2	Simulation d'un processus Brownien (multi)-Fractionnaire	21
3.1.3	Optimisation Algorithmique	21
3.2	Données Réelles	23
3.2.1	Courbes de charge éolienne	23
3.2.2	Données Hydrauliques	23
3.3	Conclusion sur l'efficacité des différentes méthodologies	23



toute la rédaction de ce chapitre est une ébauche grossière, destinée à former le squelette du rapport. Le processus de rédaction est itératif sur toute la durée du stage. avancement du stage : 3 / 6 mois

— Noyau de lissage : Epanechnikov :

$$K : x \mapsto \frac{3}{4}(1 - u^2) \cdot \mathbb{1}_{|x| \leq 1}$$

— Kernel d'auto-régression :

$$\beta : (s, t) \mapsto \frac{9}{4} \sqrt{s(1-s)} \times t$$

— méthode d'approximation d'intégrale : méthode des points médians (Newton-Cotes)

— Processus de référence : Brownien multi-Fractionnaire  (citation requise 

— Méthodes de lissage :

— Nadaraya-Watson

— base B-Spline

— base Ondelettes

— méthode de sélection des hyper-paramètres de lissage :

— NW : sélection de la largeur de bande h par validation croisée

— B-Spline : sélection du nombre de noeuds par GCV

— Ondelettes : détermination de la base optimale par validation croisée

	nombre de valeurs testées	de	jusqu'à	valeur
Δ	30	0.01	0.2	
λ	30	30	480	
N	4	100	400	
fonction de Hurst (H_t)	2	logistique	escalier	
nb simulations MC				200

TABLE 3.1 – Hyper-paramètres de la simulation Monte-Carlo

3.1 Données simulées

3.1.1 Objectifs de la simulation

Si la théorie développée par MPV assure la convergence ponctuelle des estimateurs de régularité \hat{L}_t, \hat{H}_t en fonction de $\hat{\theta}(u, v)$ ($u, v \in [t \pm \frac{\Delta}{2}]$), celle-ci se fait à Δ donné. Le praticien devra donc choisir un Δ , que l'on espère judicieux, c'est à dire du bon choix de la taille du voisinage de t pour effectuer les calculs des estimateurs de régularité.

En effet (14, Thm 1) donne une borne de concentration pour la convergence de l'estimateur de H_t proposé :



en posant $\mathfrak{a}_{N,\Delta}^{H_0}(\varphi) \stackrel{\text{déf}}{=} N\varphi^2\Delta^{4H_0}$

$$\mathbb{P}\left(|\hat{H}_0 - H_0| > \varphi\right) \leq \frac{\mathfrak{f}_0}{\mathfrak{a}_{N,\Delta}^{H_0}(\varphi)} + 4\mathfrak{b} \exp\left(-\mathfrak{g}_0 \mathfrak{a}_{N,\Delta}^{H_0}(\varphi)\right)$$

$$\mathbb{P}\left(|\hat{L}_t^2 - L_t^2| > \psi\right) \leq \frac{\mathfrak{c}_0}{\mathfrak{a}_{N,\Delta}^{H_0+\varphi}(\psi)} + \frac{\mathfrak{f}_0}{\mathfrak{a}_{N,\Delta}^{H_0}(\varphi)} + \mathfrak{b} \left[4\exp\left(-\mathfrak{g}_0 \mathfrak{a}_{N,\Delta}^{H_0}(\varphi)\right) + \exp\left(-\mathfrak{l}_0 \mathfrak{a}_{N,\Delta}^{H_0+\varphi}(\psi)\right) \right]$$

Au moment où ce rapport est rédigé l'expression de Δ comme une fonction de paramètres estimables n'est pas encore connue. C'est pourquoi l'on souhaite effectuer des simulations en faisant varier différents paramètres afin d'essayer d'intuiter la forme de l'expression de Δ comme une fonction de paramètres estimables par le praticien.

citation requise

3.1.2 Simulation d'un processus Brownien (multi)-Fractionnaire

3.1.3 Optimisation Algorithmique

3.1.3 □ A) Génération du bruit blanc

pour générer un processus sous gaussien il nous faut inverser la matrice de covariance, qui dans notre cas a une dimension de :

$$\underbrace{\dim \vec{\Delta}}_{50} \times \underbrace{3}_{t_1/t_2/t_3} \times \underbrace{n_{points_estim}}_6 + \underbrace{n_{Grid_f}}_{100} + \underbrace{\lambda}_{\leq 480} \leq \underbrace{1000}_{fixe} + \underbrace{480}_{pts_aleat}$$

on peut donc gagner du temps de calcul en inversant une unique fois la covariance restreinte aux points qui ne sont pas aléatoires et présents sur chaque courbe, ce qui peut faire la différence quand on a 400 courbes.

en posant :

$$U \stackrel{\text{d\'ef}}{=} BD^{-1}$$

$$V \stackrel{\text{d\'ef}}{=} CA^{-1}$$

on obtient l'inversion de la matrice par blocs avec l'algorithme suivant :

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - UC)^{-1} & 0 \\ 0 & (D - VB)^{-1} \end{bmatrix} \begin{bmatrix} I & -U \\ -V & I \end{bmatrix}$$

dans notre cas

$$\Sigma = \begin{bmatrix} \Sigma_{[t \neg \text{alea}]} & \Sigma_{[\text{alea} / \neg \text{alea}]} \\ \Sigma_{[\text{alea} / \neg \text{alea}]^T} & \Sigma_{[t \text{ alea}]} \end{bmatrix}$$

ce qui donnerait la formule d'inversion par bloc suivante :

$$\Sigma^{-1} = \begin{bmatrix} (\Sigma_{[t \neg \text{alea}]} - UC)^{-1} & 0 \\ 0 & (\Sigma_{[t \text{ alea}]} - VB)^{-1} \end{bmatrix} \begin{bmatrix} I & -U \\ -V & I \end{bmatrix}$$

avec :

$$U = \Sigma_{[\text{alea} / \neg \text{alea}]} \Sigma_{[t \text{ alea}]}^{-1}$$

$$V = \Sigma_{[\text{alea} / \neg \text{alea}]}^T \Sigma_{[t \neg \text{alea}]}^{-1}$$

3.1.3 □ B) Intégrale

$$X_{n+1}(t) = \int_{[0,1]} \beta(u, t) \cdot [X_n(u) - \mu(u)] du + \varepsilon_n$$

il est important lorsque l'on effectue autant de simulations d'avoir des calculs efficaces pour limiter le temps de calcul.

Parmi les méthodes d'approximation d'intégrale classiques se trouvent les méthodes des rectangles, trapèze et de Newton-Cotes. On se basera sur la méthode de Newton Cotes d'ordre 0 aussi appelée des points médians pour l'avantage suivant : elle permet d'avoir à évaluer le Brownien fractionnaire en un seul point, ce qu'il signifie qu'on a besoin de générer qu'un seul point par sous-intervalle pour calculer l'intégrale, avec une approximation d'ordre 1 (ie, exacte pour un polynôme de degré ≤ 1), plus précise que la méthode des rectangles à gauche et même des trapèzes.

$$\tilde{E}[g_k, g_{k+1}] = \frac{(g_{k+1} - g_k)^3}{12} f''(\eta_{k,k+1}) = \frac{(\frac{k+1}{G} - \frac{k}{G})^3}{12} f''(\eta_{k,k+1}) = \frac{f''(\eta_{k,k+1})}{12G^3}$$

$$\tilde{E} = \frac{1}{12G^3} \left[\sum_{k=0}^{G-1} f''(\eta_{k,k+1}) \right] \leq \frac{\sup_{[0,1]} f''}{12G^2} = \mathcal{O}\left(\frac{1}{G^2}\right)$$

Bien que nous ne manipulons pas des fonctions 2 fois dérivables, la borne d'approximation nous donne une idée de l'erreur qui sera commise en utilisant cette méthode.

3.2 Données Réelles

3.2.1 Courbes de charge éolienne

3.2.2 Données Hydrauliques

3.3 Conclusion sur l'efficacité des différentes méthodologies

Annexe A

Note et réflexions personnelles sur le stage

A.1 Apprentissage

A.1.1 Implémentation d'un package R

Le stage a été l'occasion pour moi d'apprendre à implémenter un package sous R. C'est donc aussi l'occasion de coder en R d'une façon plus avancée qu'à mon habitude, ce qui est positif lorsque l'on considère qu'une large partie des nouvelles méthodes développées par les chercheurs en statistique sont implémentées en R. Même si la communauté du Machine Learning est plus tournée vers Python, R reste un langage très utilisé dans le monde de la statistique. Ce stage orienté vers la programmation R me permet de manipuler R, étant donné que j'ai une préférence naturelle envers Python, qui je trouve est un langage plus adapté à mon style de programmation. J'ai commencé à apprendre des bribes de programmation quand j'étais plus jeune en C puis en Python aux alentours de mon passage en classe préparatoire, ce qui explique certaines difficultés que je peux avoir à penser dans la logique de R. Néanmoins, il est indispensable aujourd'hui de pouvoir manipuler les deux langages et de choisir celui qui est le plus adapté à la situation. Dans le cadre des données fonctionnelles, il est clair que R est le langage le plus adapté, car des packages ont déjà été développés pour R alors que Python peine à avoir un package de référence pour les données fonctionnelles.

A.1.2 Théorie et Pratique : il n'y a pas d'ordre total

J'ai aussi appris à bien plus apprécier la partie application lors de ce stage. J'ai été tout au long de ma scolarité un élève qui ne se souciait que de la théorie, trouvant les travaux pratiques plus fastidieux et moins intéressants. Je trouve que c'est une façon assez réductrice de voir la statistique et que ce point de vue retire justement la magie de la théorie statistique. La théorie motivée par le réel est bien plus intéressante que la théorie pour la théorie. La pratique et la théorie vont de pair et se nourrissent l'une de l'autre, il n'y en a pas une plus importante que l'autre car chacune fait vivre l'autre. On se rend compte aussi que de la théorie à la pratique il n'y a **pas** qu'un pas, et l'implémentation de la théorie peut parfois être plus complexe que prévu. Je pense qu'il est important de garder en tête lors du développement de concepts l'implémentation pratique de ce

qu'on développe, car à la fin si l'on fait de la recherche en statistique, c'est pour que ce que l'on développe soit utilisé par les praticiens.

A.2 Communication

A.2.1 Communication scientifique : Breizh Data Day

Dès mon arrivée, j'ai eu la chance de me voir proposé par l'équipe DataStorm un passage au Breizh Data Day où des intervenants exposent à des praticiens et chercheurs des concepts et applications de certaines branches de la statistique. On a pu assister à des exposés très intéressants, notamment celui axé autour de l'anonymisation des données pour avoir un processus d'apprentissage dans le machine learning sain et respectueux de la vie privée de chacun. C'est un problème qui semble complexe, tant il doit se battre avec la quête de performance des modèles statistiques, d'autant plus avec les énormes bases de données qui entraînent les modèles de langages qui sont aujourd'hui très populaires. Enfin j'ai pu voir la communication scientifique à l'oeuvre avec Hassan et Sunny, deux étudiants en thèse sur les données fonctionnelles, qui parlaient de leur sujet de recherche et de l'importance pour les praticiens d'un tel concept. C'est un exercice qui n'est pas facile, car il faut savoir vulgariser des concepts qui peuvent être très complexes, tout en gardant un certain niveau de rigueur.

Bibliographie

- (1) Denis Bosq. *Linear processes in function spaces : theory and applications*, volume 149. Springer Science & Business Media, 2000.
- (2) V. Capasso and D. Bakstein. *An Introduction to Continuous-Time Stochastic Processes : Theory, Models, and Applications to Finance, Biology, and Medicine*. Modeling and Simulation in Science, Engineering and Technology. Springer New York, 2015.
- (3) Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Functional linear model. *Statistics & Probability Letters*, 45(1) :11–22, 1999.
- (4) Dong Chen, Peter Hall, and Hans-Georg Müller. Single and multiple index functional regression models with nonparametric link. 2011.
- (5) Jacques Dauxois and Alain Pousse. *Les analyses factorielles en calcul des probabilités et en statistique : Essai d'étude synthétique*. PhD thesis, Éditeur inconnu, 1976.
- (6) Steven Golovkine, Nicolas Klutchnikoff, and Valentin Patilea. Adaptive estimation of irregular mean and covariance functions, 2021.
- (7) Steven Golovkine, Nicolas Klutchnikoff, and Valentin Patilea. Learning the smoothness of noisy curves with application to online curve estimation. *Electronic Journal of Statistics*, 16(1), Jan 2022.
- (8) X. Gourdon. *Les maths en tête. Analyse - 3e édition*. Editions Ellipses, 2020. Théorème et Applications : densité des fonctions dérivables nulle part - pages : 398—399 , ex4 : 401.
- (9) Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.
- (10) Gareth M James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(3) :411–432, 2002.
- (11) Kari Karhunen. Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae, AI*, 34, 1946.
- (12) Piotr Kokoszka and Matthew Reimherr. *Introduction to functional data analysis*. CRC press, 2017.
- (13) Xihong Lin and Daowen Zhang. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 61(2) :381–400, 1999.
- (14) Patilea Maissoro, Vimond. Learning smoothness of functional times series under weak dependency assumption. 2023. not available yet.
- (15) Hans-Georg Müller and Ulrich Stadtmüller. Generalized functional linear models. 2005.

- (16) James O Ramsay. When the data are functions. *Psychometrika*, 47 :379–396, 1982.
- (17) James O Ramsay and CJ1125714 Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society : Series B (Methodological)*, 53(3) :539–561, 1991.
- (18) William F Sharpe. A simplified model for portfolio analysis. *Management science*, 9(2) :277–293, 1963.
- (19) Ruey S. Tsay. Time series and forecasting : Brief history and future research. *Journal of the American Statistical Association*, 95(450), 2000. DOI : <https://doi.org/10.2307/2669408>.
- (20) Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and its application*, 3 :257–295, 2016. télécharger.
- (21) James H. Stock & Mark W. Watson. Vector autoregressions. *Journal of the American Statistical Association / Journal of Economic Perspective*, 15(4), 2001. page 101 - DOI : <https://doi.org/10.1257/jep.15.4.101> - télécharger.