ECOLE 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 estimation 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.

avant-propos

Le lecteur saura excuser, si toutefois il se trouve déjà familier avec certaines notions (telle que la dépendance faible), de les voir réintroduites et ré-expliquées parfois de façon très détaillée car leur (ou plutôt ma) compréhension était importante pour le stage.

correctif

ENSAI-stage_fin_etude-datastorm_fda_regularite-rapport/issues

contact

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

Notations

Notation	Explanation
Analyse	
x_0	une valeur spécifique de x
$\mathcal{V}(x_0)$	un voisinage de x_0
$\mathcal{C}(E,F)$	fonction continue de E dans F
$\mathcal{H}_{\mathcal{V}(x_0)}(\alpha_{x_0}, L_{\alpha_{x_0}})$	Classe de Hölder de paramètre $\alpha_{x_0}, L_{\alpha_{x_0}}$
Algèbre	
$\mathrm{sp}_{\mathbb{K}}(\varphi)$	valeurs propres d'un opérateur ou endomorphisme linéaire φ sur le corps $\mathbb{K}$
$\mathrm{sp}(\varphi)$	valeurs propres d'un opérateur ou endomorphisme linéaire φ sous-entendu sur le corps $\mathbb{R}$
$\overrightarrow{sp}_{\substack{\perp \\ \|\cdot\|}}(\varphi)$	vecteurs propres d'un opérateur ou endomorphisme linéaire φ formant une famille orthogonale
$\overrightarrow{sp}^{[1,p]}_{\substack{\perp \\ \|\cdot\|}}(\varphi)$	p premiers vecteurs propres d'un opérateur ou endomorphisme linéaire φ formant une famille orthogonale
Statistique	
X	La « vraie »distribution
$\widetilde{X}$	Quantité intangible/inobservable
$\widehat{X}$	un estimation empirique X
$X_{(k)}$	Statistique d'ordre de X, k^{eme} terme : $X_{(k)} \leq X_{(k+1)}$
$X_n^{(k)}$	Statistique d'ordre de X avec n observations, k^{eme} terme : $X_n^{(k)} \leq X_n^{(k+1)}$
Probabilités	
$C_X(s,t)$	Covariance du processus X entre le temps s et le temps t
$c[f]$	opérateur de covariance évalué en f
$\mathrm{VAR}[E]$	Variable aléatoire à valeur dans E : $\mathbf{m}\big((\Omega,\mathcal{F},\mathbb{P}),(E,\mathcal{A},\mu)\big)$

Table des matières

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 de 2014 à 2017). Cette énergie produite est dénommée la courbe de charge (que l'on abbrégera par **CDC** par la suite). Il est cependant plus utile de s'intéresser au facteur de charge (ou **FDC**) qui est défini comme 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.

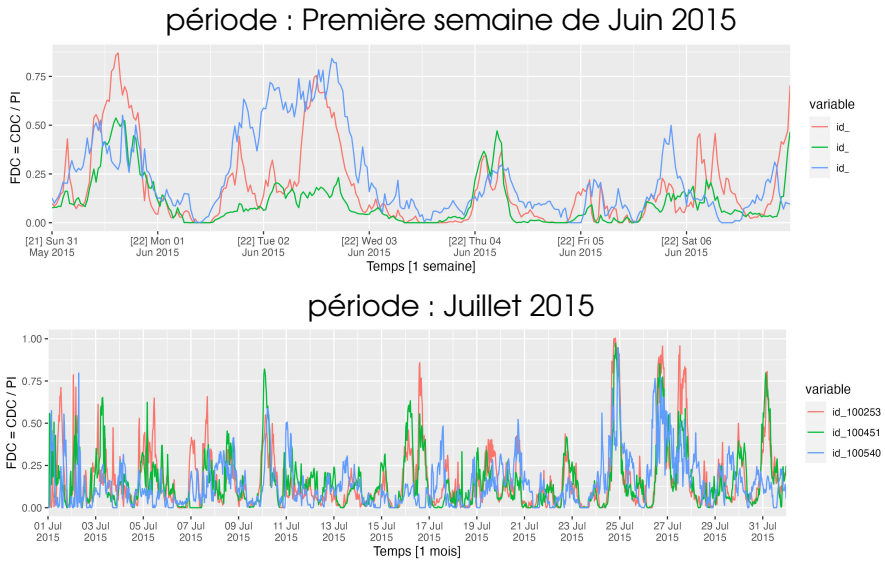


FIGURE 1.1 – Courbes de charges éoliennes sur 3 premiers parcs éoliens

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.

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 (23)

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 standarde 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 vraissemblance dans les années $1980, \ldots$).

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 (25, 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.

Même si 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 considérait 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 $\mathcal{T}$.

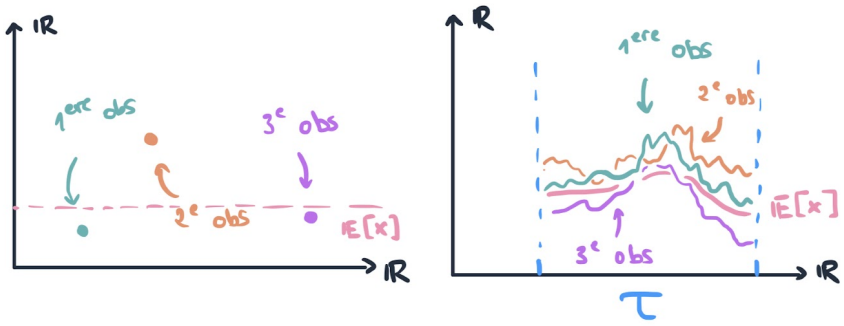


FIGURE 1.2 – Différence entre donnée fonctionnelle et donnée réelle

Si le statisticien est déjà à l'aise avec l'idée qu'une variable aléatoire réelle identiquement distribuéé 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 $t = 12$h. Formellement :

$$\mathcal{T} \underset{\text{déf}}{\equiv} [0, 24[\quad = 1 \text{ jour avec } t \text{ en heure}$$

La consommation du foyer i à midi, notée y_i, suit la loi d'un phénomène général Y, comme une loi normale $\mathcal{N}\left(0.27\ kW\text{h}, 0.1^2\right)$ [1] par exemple. Travailler sur des données fonctionnelles dans ce cadre c'est étudier non plus la consommation y_i à midi, mais regarder

1. ordre de grandeur de la consommation électrique d'un foyer en France calculé à partir des données d'ENGIE disponibles librement (6). **La variance est arbitraire**, tout comme le choix de la loi juste afin de servir d'exemple

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 $x_i : \mathcal{T} \longrightarrow \mathbb{R}$ 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 :

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

On définit de même une donnée fonctionnelle comme une application :

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

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 $(t_k^{(\text{individu } i)}, y_k^{(\text{individu } i)})_{i \in [\![1,m]\!]}$, 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

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

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 (24)

Bien que l'histoire du développement de l'Analyse de Données Fonctionnelles (FDA) puisse être retracée jusqu'aux travaux de Grenander et Karhunen (13) 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 (19) 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 connait 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 (20) ou scalaire (3) en 1999. Les modèles linéaires généralisés ont également été étudiés (12, 18), 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 (21), et leur estimation directe a été étudiée une décennie auparavant par M.Hristache, Juditsky et Spokoiny (11). De même, les modèles additifs ont été étendus aux données fonctionnelles en 1999 par Lin et Zhang (15). 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.

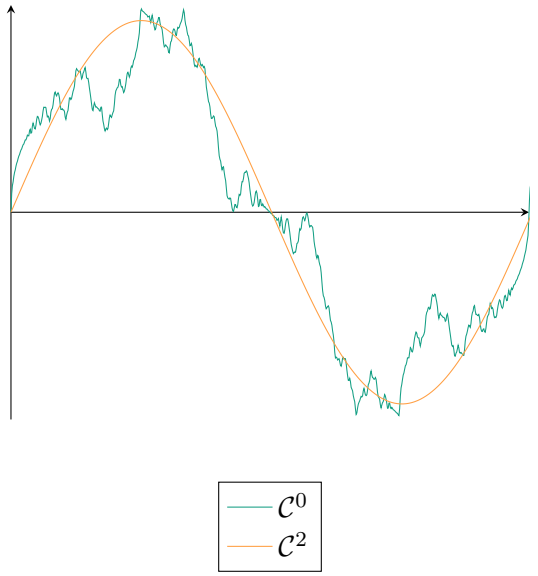
Depuis lors, des ressources telles que l'ouvrage de Kokoszka et Reimherr, *Introduction to Functional Data Analysis (2017)* (14), rendent la théorie et la mise en production des méthodes d'analyse et de prédiction de données fonctionnelles plus accessibles.

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 considérait 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

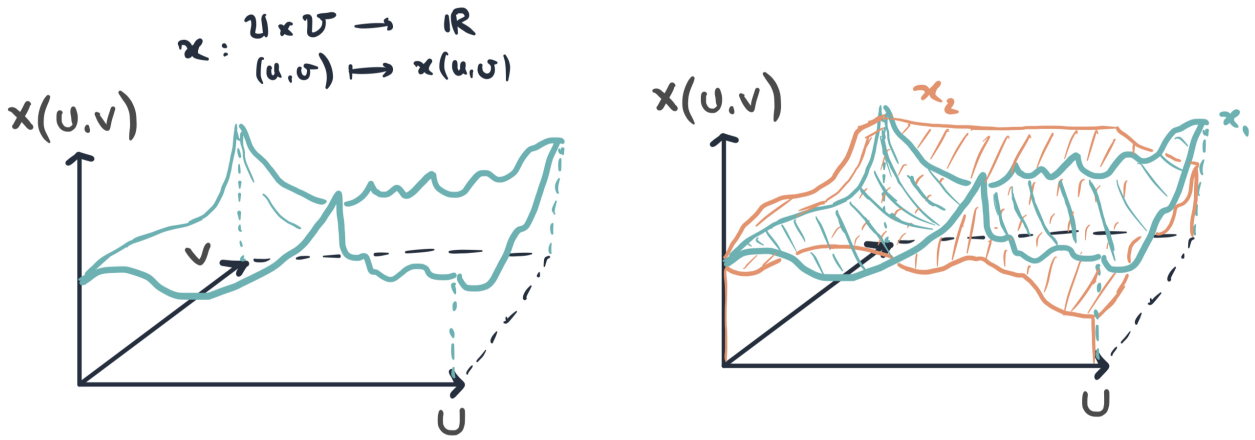
2.1.1 Définitions et propritétés informelles

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.



Gauche : exemple de surface
Droite : échantillon de deux observations de la surface suivant une loi fonctionnelle

FIGURE 2.1 – Donnée fonctionnelle : relation fonctionnelle avec 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 *de fonctions* adaptée aux données (au sens de la covariance) que l'on appelle base ACP fonctionelle (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 connait.

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.

2.1.2 estimation adaptative informelle

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 (Continuité de Kolmogorov) : Un théorème (Continuité de Kolmogorov) permet à partir de l'espérance d'incréments d'un processus aléatoire de déduire sa régularité. C'est pourquoi les estimateurs sont définis à partir de l'espérance des incréments quadratiques du processus. C'est entre autres *la raison pour laquelle les données fonctionnelles permettent de récupérer la régularité locale des trajectoires*.

Les motivations de l'obtention de la régularité étaient en partie de pouvoir mieux estimer les quantités qui nous intéressent dont 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 (7))

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 »*.

Cependant, bien qu'une largeur de bande optimale existe, elle est inconnue. Il est donc important de savoir si le praticien peut l'estimer, et avec quelle précision (c'est à dire à quel point l'estimateur sera biaisé ou non). C'est ce que nous affirme le théorème suivant :

Théorème (expression de la largeur de bande optimale — informel (7))

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 la cas pour les données éoliennes notamment). Ce cas est traîté par le théorème suivant dérivé par MPV :

Théorème (Estimation adaptative de séries temporelles fonctionnelles — informel (16))
On peut estimer la régularité d'une série temporelle de données fonctionnelles à condition que la mémoire temporelle de la série soit courte. (La décroissance de la dépendance temporelle doit être au moins aussi rapide qu'une décroissance géométrique)

2.1.3 Résumé informel de la méthodologie

Les données fonctionnelles permettent de travailler sur un modèle où la *relation* entre plusieurs quantités est sujet à une loi (*cf* données fonctionnelles — informel) . Ce point de vue de réplication de courbes est notamment utile car il permet d'extraire des observations leur régularité (*cf* Continuité de Kolmogorov, Regularité locale — informel) . L'estimation de cette régularité permet, entre autres, de lisser les courbes de façon appropriée en fonction de la quantité que l'on souhaite estimer, telle que la moyenne et la covariance avec une plus grande précision (*cf* Estimateurs de la moyenne et de la covariance — informel (7)) .

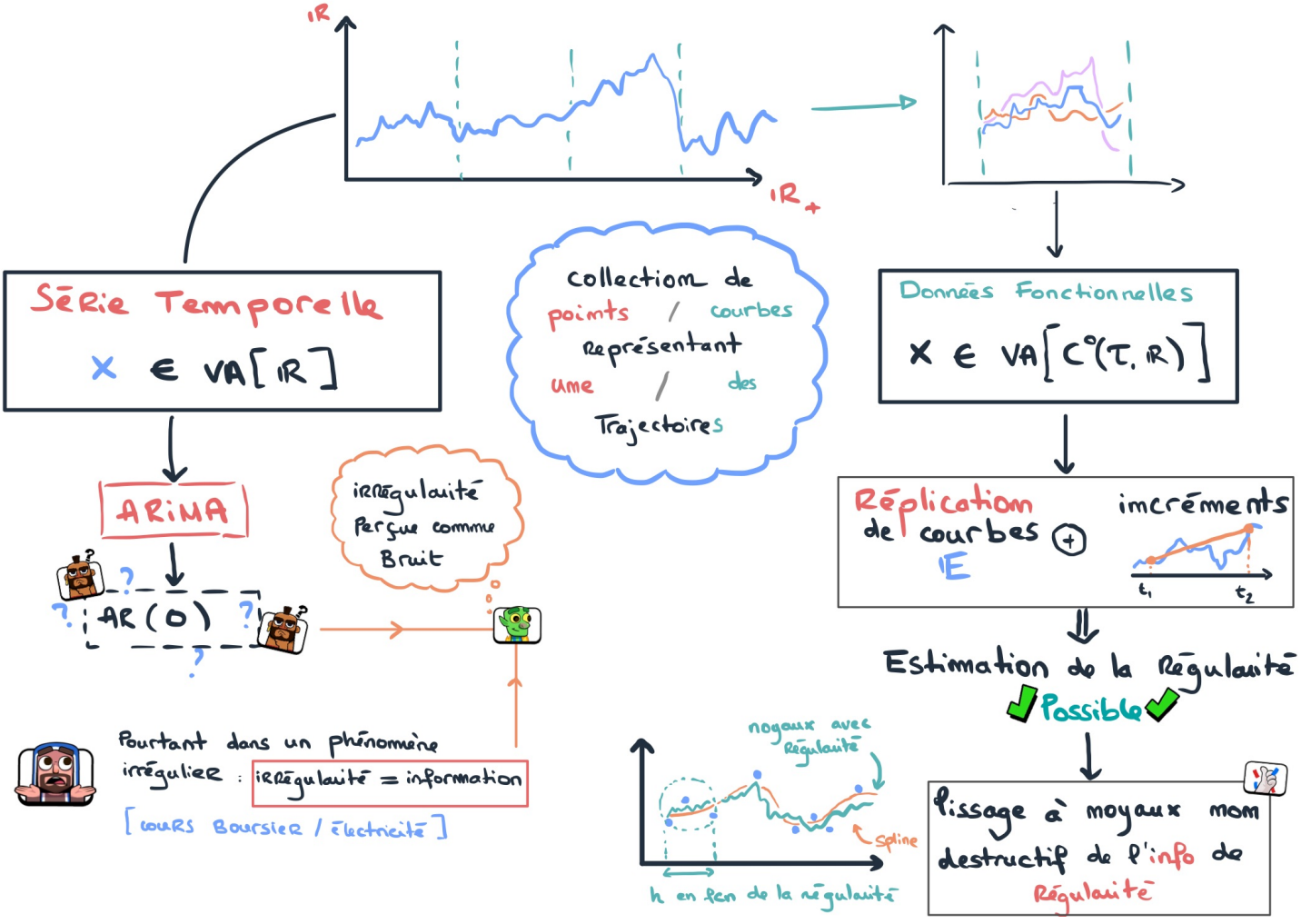


FIGURE 2.2 – Résumé des motivations du de l'estimation de la régularité locale des trajectoires

2.1.4 Données fonctionnelles : formellement

2.1.4 ☐ A⟩ Définition formelle

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 $x_i : I \to \mathbb{R}^d$ issues d'un processus X défini comme ci-dessous :

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

2.1.4 ☐ B⟩ 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 :

$$C_X : (s,t) \mapsto \mathbb{E}\left[\left[X-\mu\right](s) \cdot \left[X-\mu\right](t)\right]$$

On travaille plutôt avec l'**opérateur** de covariance :

$$c : \begin{array}{ccc} \mathbb{L}^2 & \longrightarrow & \mathbb{L}^2 \\ f & \longmapsto & \int\limits_I f(u) C_X(u, \cdot)\, du \end{array}$$

C'est parceque 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.

Etant donné que l'on traîte des données fonctionnelles, on considère la géométrie usuelle de $\mathbb{L}^2(\mathbb{R},\, \lambda)$ et on note ainsi

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

le produit scalaire que l'on considère pour manipuler les données fonctionnelles.

Théorème 1 (Karhunen-Loeve)
référence : (14, 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 : $\forall p \geq 1, \lambda_{p+1} \leq \lambda_p \qquad \lambda_p, \lambda_{p+1} \in \mathrm{sp}(c_X)$
- ✱ on pose $\overrightarrow{sp}^{[1,p]}_{\underset{\|\cdot\|}{\perp}}(c_X) \underset{\text{déf}}{\equiv} \left\{ \phi_k \in \overrightarrow{sp}_{\underset{\|\cdot\|}{\perp}}(\, c_X\,) \text{ associé à } \lambda_k, k \in [\![1\,,\,p]\!] \right\}$

alors :

- $\forall p \geq 1 \quad \underset{u_k \in \mathcal{C}(I,\mathbb{R})}{\operatorname{argmin}} \; \mathbb{E} \left\| X - \sum_{k=1}^{p} \langle\, X - \mu \,|\, u_k \,\rangle_{\mathbb{L}^2}\, u_k \right\|^2 = \overrightarrow{sp}^{[1,p]}_{\underset{\|\cdot\|}{\perp}}(\, c_X\,)$
- $$X = \mu + \sum_{k=1}^{+\infty} \langle\, X - \mu \,|\, \phi_k \,\rangle\, \phi_k$$
  $$\text{avec } \phi_k \in \overrightarrow{sp}_{\underset{\|\cdot\|}{\perp}}(\, c_X\,)$$

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

$$\forall p \geq 1 \quad \underset{u_k \in \mathcal{C}(I,\mathbb{R})}{\operatorname{argmin}} \; \mathbb{E} \left\| X - \sum_{k=1}^{p} \langle\, X - \mu \,|\, u_k \,\rangle\, u_k \right\|^2 = \overrightarrow{sp}^{[1,p]}_{\underset{\|\cdot\|}{\perp}}(\, c_X\,)$$

dans le théorème de Karhunen-Loeve.

2.1.5 Cas non indépendant : séries temporelles de données fonctionnelles

Une large partie de la théorie des données fonctionnelles suppose que l'on observe des courbes $X_i : \Omega \to \mathcal{C}^0(I, \mathbb{R})$ **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 autorégressifs 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 $X_n = \sum_{k=1}^{n-1} \varphi_k X_k + \xi_n$ où $\varphi_k \in \mathbb{R}$ et $\xi_n \begin{cases} \in \mathrm{VA}(\mathbb{R}) \\ \perp\!\!\!\perp \sigma\left(X_i\right)_{1:n-1} \end{cases}$, dans le cadre fonctionnel on capture la même idée en considérant $X_n = \sum_{k=1}^{n-1} \phi_k\left(X_k\right) + \xi_n$ 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 $\phi : \mathbb{R} \to \mathbb{R}$ 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'occurence les données de courbe de charge des parcs éoliens) semblent être naturellement corrélées dans le temps.

Pourquoi se soucier en particulier des séries temporelles fonctionnelles lorsque l'on souhaite incorporer la régularité du processus dont est issu nos données dans l'estimation des quantités qui nous intéressent ?

Rappelons-nous que les données fonctionnelles sont la clé pour déterminer la régularité, et que cela est en réalité permis par le théorème de Continuité de Kolmogorov (que nous n'avons pas énoncé en détails, mais mentionné dans la section 2.1.1). Malheureusement, dans le monde réel où vit le praticien, nous n'avons pas accès à l'espérance de la loi dont sont issues nos données. Il nous faut donc estimer cette espérance, et c'est là que les séries temporelles fonctionnelles entrent en jeu. Puisque l'estimateur usuel de l'espérance est la moyenne empirique, qui nous est fourni par la loi des grands nombres, cela devient très problématiques lorsque l'on dispose de données corrélées. Ce que nous allons voir, c'est que l'on peut tout de même utiliser l'estimateur usuel de l'espérance, et que l'on obtiendra des estimateurs des paramètres de régularité convergents ponctuellement vers ceux du processus dont sont issues nos données. Ce résultat, dû à MPV, nécessitera toutefois de faire d'abord un détour sur ce qu'on entend par dépendance.

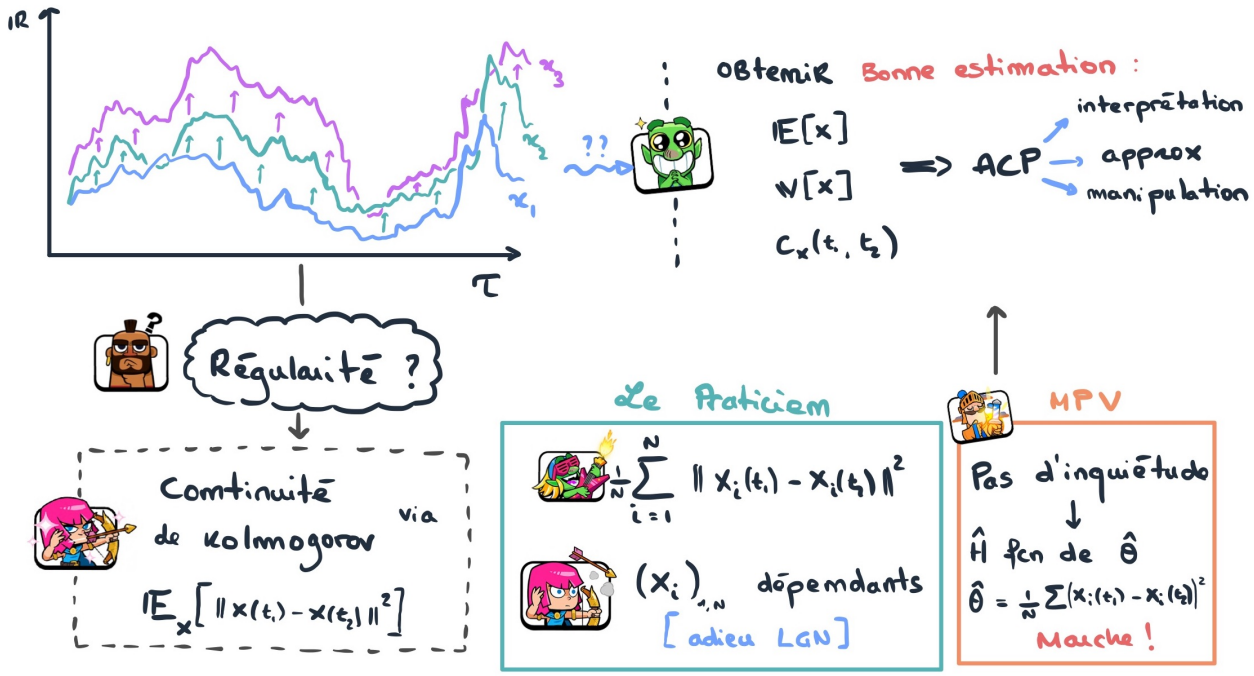


FIGURE 2.3 – Schéma grossièrement récapitulatif : Estimation de la régularité pour une série temporelle fonctionnelle

Il faut faire attention lorsque l'on manipule ou interprète des séries temporelles fonctionnelles. (comme par exemple tout résultat utilisant la loi de $\sum_n X_n, \dots$)

Une série temporelle discrète est le fait que l'observation suivante dépend linéairement de l'observation précédente, dans le cadre fonctionnel *l'observation est une fonction*. La dépendance se fait sur l'indice de la fonction, et non pas sur l'argument de la fonction interprété dans notre caps comme étant le temps.

Dans le cadre éolien c'est d'autant plus trompeur de parler de temps car on observe des courbes de charge sur une année : à la fois l'indice de la fonction et l'argument de la fonction ont des interprétations temporelles.

dans l'expression « $X_n(t)$ », la série temporelle (discrète) concerne bien l'indice n et non pas l'argument t.

La question devient alors :

Lorsque l'on a une dépendance dans les observations fonctionnelle $\{X_1 \dots X_n\}$, possède-t-on une dépendance dans les observations ponctuelles à t fixé $\{X_1(t) \dots X_n(t)\}$? Cette dépendance est-elle la même ?

Et la réponse, c'est qu'**on ne sait pas**. En tout cas, dans le cadre général. Il y a en effet plusieurs façon de définir ce qu'on appelle par « dépendance temporelle ». Toutes les définitions de dépendance ne mènent pas à cette conclusion, mais celle adoptée par

(MPV) l'est, étant plus faible. De manière générale, lorsque l'on traîte des données avec de la dépendance, il convient d'être extrêmement précautionneux avec les théorèmes et « faits »que l'on invoque. Toujours bien vérifier les hypothèses.

Il y a dans un premier temps ce qu'on appelle la dépendance « forte », comme la dépendance dite de « α-mixing »comme définie dans (22) :

Définition ($\alpha-$mixing) une suite $X = (X_i)_{i\geq 0}$ de variables aléatoire est dite α-mixing si pour tout $n \in \mathbb{N}$

$$\alpha(n) \xrightarrow[n\to\infty]{} 0$$

avec : $\alpha(n) = \sup_k \{|\mathbb{P}[A \cap B] - \mathbb{P}[A]\,\mathbb{P}[B]| \quad | \quad A \in \sigma(X_{1:k}),\, B \in \sigma(X_{k+n:\infty})\}$

en d'autres termes, la « dépendance »($|\mathbb{P}[A \cap B] - \mathbb{P}[A]\,\mathbb{P}[B]|$) entre les variables aléatoires X_k et X_{k+n} tend vers 0 lorsque n tend vers l'infini.

Ce point de vue est « fort »dans le sens où l'on manipule directement les tribus, et que l'on regarde leur degré d'indépendance via la mesure de probabilité. Il ne s'agit pas de l'approche considérée par MPV qui est plus faible, en se reposant non pas sur l'indépendance des tribus engendrées par la série temporelle mais en exploitant la qualité d'approximation de la série temporelle que l'on étudie par un autre processus, indépendant de la série temporelle étudiée à partir d'un certain rang. La définition de dépendance temporelle est alors dite « faible ». Le point de vue faible offre un comportement plus sympathique pour l'aspect *local* dans l'estimation de la régularité : qui est le coeur de l'approche de MPV.

Les processus qui nous intéressent et ceux auxquels on va se limiter dans un premier temps sont les processus causaux. Comme dans le cas réel, on peut étudier les séries temporelles en posant l'opérateur :

$$B : x_n \mapsto x_{n-1}$$

et la relation de dépendance encodée par :

$$X_{n-1} = \phi(X_n) + \xi_n \qquad \phi \text{ linéaire}$$

Si le processus est inversible, on peut écrire X_n comme le développement en série entière suivant :

$$\begin{aligned} X_n &= && \phi \circ B(X_n) + \xi_n \\ [I - (\phi \circ B)](X_n) &= && \xi_n \\ X_n &\underset{\|\phi\circ B\|<1}{=} && [\phi \circ B]^{-1}(\xi_n) \\ X_n &\underset{\sum \mathsf{E}}{=} && \sum_{k=0}^{\infty} \underbrace{[\phi \circ B]^k}_{\phi^k \circ B^k}(\xi_n) \end{aligned}$$

En effet, les opérateurs ϕ et B commutent car :

$$x = (x_n)_{n\in\mathbb{Z}} = (\dots, x_0, x_1, x_2, \dots)$$

$$\phi(x) = (\dots, \phi(x_0), \phi(x_1), \phi(x_2), \dots)$$

on a bien $\phi \circ B = B \circ \phi$

$$\begin{aligned}
\phi \circ B(x) &= (\dots, \phi \circ B(x_0), \phi \circ B(x_1), \phi \circ B(x_2), \dots) \\
&= (\dots, \phi(x_{-1}), \phi(x_0), \phi(x_1), \dots) \\
&= (\dots, B(\phi(x_0)), B(\phi(x_1)), \dots) \\
&= B(\phi(x))
\end{aligned}$$

et ainsi

$$\boxed{X_n = \sum_{k=0}^{\infty} \phi^k(\xi_{n-k}) = f(\dots \xi_{n-k} \dots \mid k \geq 0)}$$

Définition 2 (copie indépendante) on appelle V une copie indépendante de U si $V \sim U \sim \mathcal{L}$ ET $V \perp\!\!\!\perp U$.

i.e : U et V sont de même loi et indépendantes. Exemple : même étude réalisée à deux laboratoires différents avec des patients différents.

soit maintenant

$$\Xi_n \underset{\text{déf}}{\equiv} \{\xi_n\}_{-\infty:n} \text{ la suite de bruits blancs dans l'inversion précédente}$$

on va regarder le niveau de dépendance de X_n à l'ordre a. pour cela nous allons commencer par effectuer une copie indépendante du bruit pour chaque ordre a que nous allons regarder. L'idée est que l'on ne va garder que les a derniers termes de notre processus dont on souhaite savoir jusqu'à combien de termes la dépendance avec le passé est significative. Les termes qui les précèdent seront remplacés par une copie indépendante qui n'a donc pas pu avoir d'influence sur les a derniers termes (par copie *indépendante*) : les termes que l'on a conservé ne peuvent pas dépendre de la copie.

$$\begin{aligned}
\Xi^{[1]} &= \underset{\perp\!\!\!\perp}{\text{copy}}\, \Xi \\
&\vdots \quad \vdots \\
\Xi^{[a]} &= \underset{\perp\!\!\!\perp}{\text{copy}}\, \Xi \\
&\vdots \quad \vdots \\
\Xi^{[\infty]} &= \underset{\perp\!\!\!\perp}{\text{copy}}\, \Xi
\end{aligned}$$

$$X_n^{(a)} = f\left(\underbrace{\xi_n, \xi_{n-1}, \dots}_{a \text{ termes}} \;,\; \overbrace{\underbrace{\xi_{n-a}^{[a]}, \dots, \xi_1^{[a]}}_{\text{tronqué } a \text{ derniers termes}}}^{a^{\text{ème}} \underset{\perp\!\!\!\perp}{\text{copy}} \text{ de } (\Xi_n)} \right)$$

Ensuite il nous suffit de regarder si on a perdu beaucoup d'information sur le processus en le comparant au processus initial, dont on souhaite déterminer l'ordre de dépendance. On regarde le pire cas pour $t \in \mathcal{T}$:

$$L_p(X_n|a) = \mathbb{E}\|X_n - X_n^{[a]}\|_{\infty(\mathcal{T})}^p$$

On parle alors de $\mathbb{L}^p - a$ approximation en étudiant la convergence de la série :

$$\sum_{a=1}^{\infty} L_p(X_n|a)^{\frac{1}{p}} = \sum_{a=1}^{\infty} \left(\mathbb{E}\|X_n - X_n^{[a]}\|^p_{\infty(\mathcal{T})} \right)^{\frac{1}{p}}$$

Définition 3 ($\mathbb{L}^p - a$ approximation) une suite de variables aléatoires $(X_i)_{i\geq 0}$ est dite $\mathbb{L}^p - a$ approximable si la série $\sum\limits_{a=1}^{\infty} L_p(X_n|a)^{\frac{1}{p}}$ converge.

Il s'agit de la définition de dépendance faible proposée pour les données fonctionnelles par Hörmann et Kokoszka(10). Une autre définition est aussi populaire : aulieu de remplacer tout le passé par la copie, on ne remplace que ξ_0 par la $a^{\text{ème}}$ copie.

L'idée est qu'après inversion du processus causal on obtient :

$$\begin{aligned}
X_n &= \sum_{k=0}^{a-1} \phi^k(\xi_{n-k}) + \sum_{k=a}^{\infty} \phi^k(\xi_{n-k}) \\
X_{\substack{n \\ [k\leq a]}}^{[a]} &= \sum_{k=0}^{a-1} \phi^k(\xi_{n-k}) + \sum_{k=a}^{\infty} \phi^k(\xi_{n-k}^{[a]}) \\
X_{\substack{n \\ [k=n]}}^{[a]} &= \sum_{k\neq n}^{\infty} \phi^k(\xi_{n-k}) + \phi^n(\xi_0^{[a]})
\end{aligned}$$

Le reste dans l'approximation $\mathbb{L}^p - a$ ($X_n - X_n^{[a]}$) devient alors le suivant :

$$\begin{aligned}
X_n &= \sum_{k=0}^{a-1} \phi^k(\xi_{n-k}) + \sum_{k=a}^{\infty} \phi^k(\xi_{n-k}) \\
R_{\substack{n \\ [k\leq a]}}^{[a]} &\underset{\phi \text{ lin}}{=} \sum_{k=a}^{\infty} \phi^k(\xi_{n-k}^{[a]} - \xi_{n-k}) \\
R_{\substack{n \\ [k=n]}}^{[a]} &\underset{\phi \text{ lin}}{=} \phi^n(\xi_0^{[a]} - \xi_0)
\end{aligned}$$

et on peut alors montrer que pour une certaines métrique ν_2 basée sur la norme $\mathbb{L}^2$,

$$\nu_2 \left(R_{\substack{n \\ [k\leq a]}}^{[A]} \right) \leq C \sum_{a\in A} \nu_2 \left(R_{\substack{n \\ [k=n]}}^{[a]} \right)$$

ce qui fait de la dernière version introduite est une version plus forte. Avec la dernière définition introduite, il avait été démontré différentes inégalités qui se trouvent très utiles pour déterminer les bornes de concentration de différents estimateurs. La question est désormais la suivante :

« est ce que ces inégalités restent vraies pour la définition $X_{\substack{n \\ [k\leq a]}}^{[a]}$? »

La réponse, déterminée par MPV (16) est oui. C'est important de l'avoir aussi pour cette définition car MPV a réussi à étendre la notion de $\mathbb{L}^p - a$ approximation au cas $\mathbb{L}^\infty$ (16) pour avoir un héritage local de la notion de dépendance définie sur les trajectoires.

N'est-il pas bizarre qu'une norme infinie permette de définir une notion de dépendance locale ?

Il semble en effet plus que contre-intuitif qu'une norme infinie, c'est à dire une norme invoquant le supremum sur un intervalle, permette d'obtenir une notion de dépendance locale.

en notant $\nu_p : x \mapsto \mathbb{E}\left[\,|x|^p\,\right]^{\frac{1}{p}}$

$$\sum_n \mathbb{E}\left[\,|X_n(t) - X_n^{[a]}(t)|^p\,\right] \leq \sum_n \mathbb{E}\left[\,\| X_n - X_n^{[a]} \|^p_{\infty(\mathcal{T})}\,\right]$$

La somme des $\nu_P(\,|\cdot(t)|\,)$ étant bornée par la somme des $\nu_P(\|\cdot\|_\infty)$, la dépendance locale (ie à t fixé) est directement héritée. Si la démarche consistait juste à obtenir une notion de dépendance locale, on remarque que ce qui la fait marcher est le fait que l'on a la convergence en considérant les pires cas sur chaque trajectoire.

Démontrer que $\sum_n \mathbb{E}\left[\,|X_n(t) - X_n^{[a]}(t)|^p\,\right] < \infty \quad t$ par t ne suffit pas pour que les résultats sur l'obtention de la régularité découlent :
il est important de définir les hypothèses de données fonctionnelles sur les fonctions et non pas sur les valeurs prises par les fonctions. Puisque c'est la réplication des courbes qui est la clé.

L'idéal serait d'avoir une notion de dépendance faible qui permettrait d'obtenir une inégalité du genre :

$$\sum_n \nu_p(\| X_n - X_n^{[a]} \|_{\text{hypothétique}_{inf}}) \leq \sum_n \nu_p(|X_n(t) - X_n^{[a]}(t)|) \leq \sum_n \nu_p(\| X_n - X_n^{[a]} \|_{\text{hypothétique}_{sup}})$$

Qui donnerait une sorte d'équivalence entre le point de vu fonctionnel et le point de vue local en terme de dépendance, mais à ce jour, et à notre connaissance, il n'existe pas de telle notion de dépendance.

Si l'on souhaite juste regarder l'ordre de dépendance, en remplaçant l'information après le $a^{\text{ème}}$ dernier terme par quelquechose dont le processus qui nous intéresse ne dépend pas, pourquoi s'embêter avec des copies indépendantes aulieu de simplement tronquer (c'est-à-dire remplacer par des 0) ?

Il s'avère que les deux définitions sont en quelques sorte « équivalentes »mais que celles avec les copies est plus générale et donc est évidemment privilégiée pour plus de flexibilité et de puissance dans les résultats dérivés.

$$
\begin{aligned}
X_n &= \sum_{k=0}^{a-1} \phi^k(\xi_{n-k}) + && \sum_{k=a}^{\infty} \phi^k(\xi_{n-k}) \\
X_{n\ [k\leq a]}^{[a]} &= \sum_{k=0}^{a-1} \phi^k(\xi_{n-k}) + && \sum_{k=a}^{\infty} \phi^k(\xi_{n-k}^{[a]}) \\
X_{n\ [k=n]}^{[a]} &= \sum_{k=0}^{a-1} \phi^k(\xi_{n-k}) + && 0
\end{aligned}
$$

et ainsi lorsque l'on va regarder

$$\|X_n - X_n^{[a]}\|_{\mathbb{L}^\infty}^p = \|\sum_{k=a}^{p} \phi^k(\xi_{n-k} - \xi_{n-k}^{[a]})\|_{\mathbb{L}^\infty}^p$$

que ce soit avec une méthode ou l'autre, on remarque que lorsque l'on va développer les sommes, les termes en $\|\xi \cdot \xi^{[a]}\|_{\mathbb{L}^\infty}$ seront nuls.

On en déduit ainsi que la dépendance faible comme définie dans (16) nous donne bien le résultat naturel suivant : $(X_n)_{n\geq 1}$ faiblement dépendant $\implies$ $\{X_n(t)\}_{n\geq 1}$ faiblement dépendant (au sens local introduit précédemment). On peut donc travailler localement sur les trajectoires tout en utilisant des hypothèses fonctionnelles (que ce soit pour la dépendance ou autres) pour obtenir la régularité.

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 irréguliers [1] (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'emphase 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 :

$$\text{Lipschitz} \implies \text{Hölder} \implies \underbrace{\text{Localement Hölder}}_{\text{ce qui nous intéresse}} \implies \text{Uniformément continue} \implies \text{Continue}$$

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.

— 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 :

$$\exists L_I \quad (\forall x, y \in I) \quad |f(x)-f(y)| < L_I |x-y|$$

— Hölder :

$$\exists \alpha \in (0,1] \quad \exists L_{\alpha(I)} \quad (\forall x, y \in I) \quad |f(x)-f(y)| < L_{\alpha(I)} |x-y|^{\alpha}$$

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

— Localement Hölder :

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

1. les fonctions dérivables nulle part sont même denses dans les fonctions continues pour la topologie de la convergence uniforme (9). A epsilon près on rencontre toujours une fonction dérivable nulle part lorsque l'on considère la distance maximale réalisée entre deux fonctions continues sur leur support I...

? | 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. 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]

2.2.2 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 $\mathbb{E}\left[\,|X(u)-X(v)|^2\,\right]$ utilisés pour l'estimation de la régularité locale.

Théorème 2 (Continuité de Kolmogorov)
référence : (2, thm : 2.197 | page : 145)

- $X : \begin{array}{ccc} \mathbb{R}_+ \times \Omega & \longrightarrow & \mathbb{R} \\ (t,\omega) & \longmapsto & X(t,\omega) = x(t) \end{array}$ séparable
- $\exists r, c, \varepsilon, \delta \in \mathbb{R}_+ \quad (\forall h < \delta)(\forall t \in \mathbb{R}_+) \quad \mathbb{E}\left[\,|X(t+h)-X(t)|^r\,\right] \leq c \cdot h^{1+\varepsilon}$

$$\Downarrow$$

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

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

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

2. Afin de ne pas alourdir l'essence du propos, une simplification par rapport à l'article de MPV (16) a été faite, si le lecteur souhaite aller dans le détail, il est possible de se référer à l'Annexe A.1.

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

La méthode de Golovkine et al. (8, pages : 7—9) n'utilise que les points observés, et construit un estimateur des incréments quadratiques à base de statistique d'ordre.

$$\theta(T_{(l)}, T_{(k)}) = \mathbb{E}\left[\left|X(T_{(l)}) - X(T_{(k)})\right|^2\right] \underset{\text{LGN}}{\approx} \boxed{\frac{1}{N}\sum_{n=1}^{N}\left|Y_n^{(2k-1)}Y_n^{(k)}\right|^2 \underset{\text{déf}}{\equiv} \hat{\theta}_k}$$

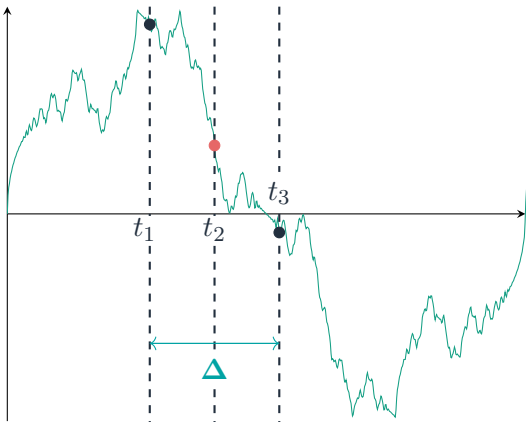
$$\underset{\substack{\text{Hölder}\\ \approx \\ +\mathcal{C}^0 Kol.}}{} L_{t_0}\mathbb{E}\left[\left|T_{(l)} - T_{(k)}\right|^{2H_{t_0}}\right] \quad \rightarrow H_{t_0} = f(\theta)$$

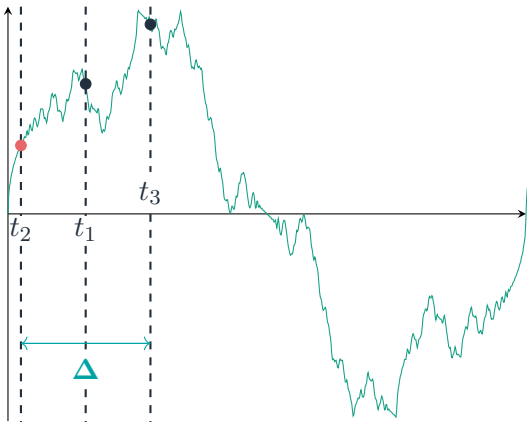
et on obtient ainsi l'estimateur suivant :

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

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 (7, 16), 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 ».





X
Point d'estimation de la régularité locale

X
Point d'estimation de la régularité locale

On se donne un $\Delta \in]\,0,1\,[$, arbitraire pour le moment, comme diamètre de l'intervalle J_Δ que l'on considère pour évaluer la régularité en t_0.
Il est naturel de définir les points d'estimation de la régularité de la façon suivante :

$$\begin{aligned} t_1 &\underset{\text{déf}}{\equiv} t_0 - \frac{\Delta}{2} \\ t_2 &\underset{\text{déf}}{\equiv} t_0 \\ t_3 &\underset{\text{déf}}{\equiv} t_0 + \frac{\Delta}{2} \end{aligned}$$

avec t_0 le point en lequel on souhaite estimer la régularité.

Remarque : Rien n'empêche dans la théorie d'avoir les points $t_1, t_2, t3$ non ordonnés dans le temps, mais dans la pratique, on considère naturellement que $t_1 < t_2 < t_3$.
Seule la condition $t_1, t_2, t_3 \in J_\Delta$ importe pour l'estimation de la régularité locale. ⚠ (vérifier si il faut basolument être équidistant) Ainsi aux bords, si l'on souhaite estimer la régularité au point t_0 tel que la définition précédente nous donne un point t_1 en dehors de $[0,1]$, on peut tout à fait à la place considérer :

$$\begin{aligned} t_2 &\underset{\text{déf}}{\equiv} t_0 \\ t_1 &\underset{\text{déf}}{\equiv} t_0 + \frac{\Delta}{2} \\ t_3 &\underset{\text{déf}}{\equiv} t_0 + \Delta \end{aligned}$$

on pourra se référer à la 2^e image de la figure 2.2.2

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 lisssée à partir des observations $(T_i^{[n]}, Y_i^{[n]})_{n\in 1:N, i\in 1:M_n}$

2.2.3 Prélissage

Comme mentionné précédemment, l'estimation de la régularité locale nécessite l'évaluation de notre processus observé X en 3 points. Il est possible de ne pas observer ces

points, qui sont de plus bruités dû au sampling de X. C'est pourquoi nous décidons de lisser les courbes comme « pré-lissage »pour pouvoir estimer la régularité locale.

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 ?

Comme mentionné précédemment, 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, cette-fois, une fenêtre de lissage appropriée qui dépend de ces paramètres de régularité.

En d'autres termes, le pré lissage utilise un lissage à noyaux tel que la fenêtre de lissage cross-validée nous donne :

$$h^{*[\mathrm{CV}]}_{\mathrm{pre}} \text{ estimateur de } h^*_{\mathcal{R}_{\mathrm{quadr}}}(t) = \underset{\mathbb{P}}{\mathcal{O}}\left(\lambda^{-\frac{1}{2H_t+1}}\right)$$

à partir duquel on peut lisser les courbes observées $(T_i^{[n]}, Y_i^{[n]})_{n\in 1:N, i\in 1:M_n}$ pour estimer la régularité locale H_t. On peut désormais obtenir la fenêtre de lissage adaptée à la quantité que l'on souhaite estimer :

$$h^*_\mu(t) = \underset{h}{\operatorname{argmin}}\, \mathcal{R}_\mu(\underbrace{t}_{\substack{\text{Régularité, sparsity, ...}\\ \to H_t\ ,\ L_t\ ,\ \mathcal{W}_t}}, h)$$

Le coeur de ce stage est la détermination du comportement dé l'hyper-paramètre Δ, diamètre de l'intervalle que l'on considère dans lequel on vient prendre la valeur de notre processus en 3 points régulièrement espacés. MPV affirme déjà que pour un Δ donné, on a bien la convergence ponctuelle des estimateurs. Ces points ne sont pas nécessairement observés, et on va donc effectuer un pré-lissage. (16)

Toutefois, le praticien est en droit de se demander quel Δ explicitement choisir ? Est ce qu'il y a une procédure simple pour déterminer la valeur optimale de Δ qu'il faut choisir pour obtenir un biais le plus petit possible pour l'estimation des paramètres de régularité ?

la méthode de pré-lissage a-t-elle une importance ? SI oui, laquelle faut-il choisir ?

C'est pourquoi nous allons établir une première heuristique avant d'aborder le comportement du Δ :

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.

On rappelle que l'utilisation de Splines comme méthode de lissage nécessite tout de même de faire des choix : elle est sensible aux nombre de noeuds et leur emplacement. Il est donc nécessaire de les déterminer par validation croisée. Une méthode fréquemment utilisée est d'utiliser un nombre de noeuds || égal au nombre d'observations, et de les placer aux points d'observations. Puis on utilise des splines pénalisées sur leur dérivée seconde ($L = L_{quad} + \lambda \int_0^1 f''(u)du$) et on détermine le paramètre de pénalisation par validation croisée afin de s'affranchir du choix du nombre de noeuds et de leur emplacement. La validation croisée sur la pénalisation est supposée compenser ce choix. Il s'agit de la méthode qui a été utilisée dans le cadre de ce stage, car très populaire et simple à mettre en place.

Il est à noter qu'une autre méthode de lissage spline est de déterminer le nombre de noeuds || par validation croisée, et de placer les points de façon uniforme sur les quantiles de la distribution des observations. Ce qui ne sera pas utilisé dans le cadre de ce stage.

En effectuant un pré-lissage de splines cubiques naturelles sur une courbe Höldérienne, on ne s'attend pas à obtenir de bonnes performances sur l'estimation de la régularité locale. En effet les courbes splines sont par construction de classe $\mathcal{C}^2$ (fonctions polynômiales $\mathcal{C}^\infty$ avec des raccordements $\mathcal{C}^2$), et la courbe lissée écrasera complétement l'information de régularité. Même si il s'agit de ce que l'on souhaite obtenir et qu'on ne connait pas encore la régularité, il est raisonnable de penser qu'être précautionneux dans le choix de la technique de lissage de telle façon à être le plus proche de la régularité d'une fonction qui pourrait potentiellement ne même pas être dérivable est une bonne idée.

2.2.3 ☐ B⟩ pré-lissage à noyaux

Considérer un lissage non paramétrique à noyaux est une alternative au lissage spline. L'espoir est la détermination lors du pré-lissage d'utiliser une fenêtre de lissage qui permette de mieux conserver l'information irrégulière que les splines via la détermination du $h_{\text{pre}}^{*[\text{CV}]}$ optimal par validation croisée.

Pour rappel, la fenêtre de lissage retenue est une fenêtre de lissage déterminée par validation croisée, qui est un estimateur de la fenêtre de lissage optimale pour le risque quadratique qui peut s'exprimer en fonction de la régularité locale si l'on suppose les hypothèses retenues sur le processus par MPV (16). Même si le $h^*_{\mathcal{R}_{quadr}}$ est techniquement une fonction de $t \in \mathcal{T}$, l'estimateur que l'on considère lui sera sélectionné pour l'ensemble du support de la courbe $\mathcal{T}$. On peut espérer que si la courbe change de régularité sur son support mais que celui-ci ne varie pas trop, alors la fenêtre de lissage sélectionnée sera adaptée à la régularité locale de la courbe peu importe où l'on se trouve sur le support.

2.2.3 ☐ C⟩ Lisser en utilisant une base de fonction sans écraser l'information irrégulière ?

Le lissage spline donne une fonction de classe $\mathcal{C}^2$, ce qui est un désavantage dans le cadre du prélissage 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 traîtement 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 1946 par Denis Gabor 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.

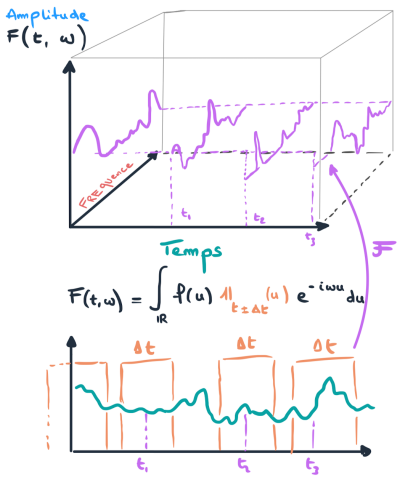


FIGURE 2.4 – Transformée de Fourier à court terme d'une fonction

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.

? 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 observent 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 temporaire (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

2.2.4 ☐ B.a⟩ 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 :

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

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

Proposition 3 (base d'ondelette dichotomique)

Base d'ondelettes

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

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.

2.2.4 ☐ B.b⟩ Propriétés principales des ondelettes

Approximation dans l'espace fréquentiel-temporel La transofrmée en ondelettes

$$\mathcal{W} : f \mapsto \langle f \,|\, \psi_{t,s} \rangle$$

est une isométrie de $\mathbb{L}^2$. Etant donné qu'elle est de plus une application linéaire, nous pouvons donc d'affirmer que

$$\boxed{||f - \hat{f}||_{\mathbb{L}^2} = ||\mathcal{W}f - \mathcal{W}\hat{f}||_{\mathbb{L}^2}}$$

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.

Propriété de Fast Decay : (ref : (17)) Une caractérisation des fonctions Hölderiennes, fournie par Antoniadis et Gijbels en 2002 (citation requise)est :

$$f \in \mathcal{H}_{\mathcal{V}(t_0)}(\alpha, L_\alpha) \cap \mathbb{L}^2 \iff \begin{array}{l} \exists P \in \mathbb{R}[X],\ \deg P \leq \alpha \leq \deg P + 1 \\ \exists f_{loc} \underset{t\to 0}{=} \mathcal{O}(t^\alpha) \end{array} \qquad f(t_0 + h) \underset{t\to 0}{=} P(h) + f_{loc}(h)$$

Définition 4 (vanishing moment) on dit qu'une ondelette ψ possède n vanishing-moments si :

$$\forall k < n \left\langle\, t \mapsto t^k \,|\, \psi \,\right\rangle_{\mathbb{L}^2} = 0 = \int_{\mathbb{R}} t^k \psi(t)\, dt$$

Proposition 4 (vanishing-moment et polynômes)

il suffit donc de choisir une ondelette avec $n > \alpha$ vanishing-moments pour obtenir :

$$\mathcal{W}f_{|\mathcal{V}(t_0)} = \mathcal{W}(P + f_{loc}) = \mathcal{W}P + \mathcal{W}f_{loc} = \mathcal{W}f_{loc}$$

enfin

Théorème 5 (Fast Decay | ref : (17) - thm 6.3)

$$f \in \mathcal{H}_{\mathcal{V}(t_0)}(\alpha, L_\alpha) \cap \mathbb{L}^2 \implies \exists A > 0,\ |[\mathcal{W}f](t,s)| \leq A \cdot s^{\alpha + \frac{1}{2}}$$

et inversement en supposant f bornée (ce qui est le cas pour une fonction continue sur un segment : notre cas) et f Hölder juste après les bords. (C'est à dire que ça ne marche pas pour les points extrémaux $t \in \{0, 1\}$)

Ainsi lorsque $s \in \left\{2^{-k}\right\}_{k \in \mathbb{N}}$:

$$|[\mathcal{W}f](t,s)| \leq A \cdot 2^{-k(\alpha+\frac{1}{2})}$$

La magnitude de la transformée en ondelette décroit exponentiellement vers 0, et beaucoup plus rapidement là où f est plus régulière. Ainsi, la transformée en ondelette agit comme un encodeur efficace d'information d'irrégularité

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

La capacité de capturer de façon efficiente les irrégularités de la fonction lissée est une motivation pour l'utilisation de la base d'ondelettes pour effectuer le pré-lissage de données, dont on espère qu'il n'écrase pas la majorité de l'information irrégulière de nos données. Si une des méthodes possibles, comme mentionnée précédemment, est d'utiliser un lissage non paramétrique à noyaux, les bases de fonctions ont de nombreux avantages. Un des avantage est le fait qu'une fois les projections sur la base déterminées, il n'y a plus besoin de se référer de nouveau aux données originales par la suite. Cela donne une représentation très parcimonieuse des données. Alors pour déterminer la valeur de $\widehat{X}(t)$ en un point t non observé, il suffit d'évaluer l'expression $\sum_k \langle X \mid \psi_k \rangle \psi_k(t)$.

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

? Peut on quantifier le biais introduit par le lissage en utilisant les ondelettes sur l'estimation de la régularité locale ?

(regarder ce que ça donne, en utilisant les différents théorèmes et bornes disponibles sur les ondelettes pour un processus Holder LORSQUE J AI LE TEMPS - certainement en Septembre)

2.2.5 Résumé de la méthodologie d'estimation de la régularité locale

2.3 Estimation adaptative

Dans la section précédente, nous avons déterminé comment obtenir des estimateurs de la régularité locale des trajectoires. Cette régularité locale nous permet désormais de lisser les courbes observées de manière à ne pas détruire l'information irrégulière. L'obtention d'un tel lissage était motivé notamment par l'obtention de quantités capitales pour l'analyse de nos données, l'interprétation et la prise de décision : la moyenne, la covariance, et l'auto-corrélation des séries temporelles fonctionnelles observées.

Un meilleur lissage nous donne ainsi une meilleure estimation de ces quantités. Toutefois, il est possible d'aller plus loin dans l'adaptation de notre lissage. En effet, il faut dans un premier temps constater que les différentes quantités que l'on souhaite estimer représentent des concepts différents, préférant chacun un lissage différent.

2.3.1 Estimation adaptative de la fonction moyenne

L'idée du lissage adaptatif est que chaque quantité évaluée en un point $t \in \mathcal{T}$ tire parti différemment des information du voisinage de t. Il semble intuitif que le processus moyen ($\mu = \mathbb{E}[X]$) considère des informations d'un voisinage assez large du processus et que celui-ci soit « assez lisse ». Pour déterminer une fenêtre adaptée à l'estimation de la moyenne, on définit une grille de fenêtres à évaluer $\mathfrak{H} = (h_i)_{1:r}$ que l'on choisit en minimisant un risque spécifiquement adapté :

$$\widehat{h^*_\mu} = \underset{h \in \mathfrak{H}}{\operatorname{argmin}}\, R_\mu(t, h)$$

Déterminons maintenant ce risque.

2.3.1 ☐ A⟩ Méthode Golovkine et al. : indépendance

Dans le cadre de données indépendantes, on peut invoquer la Loi des Grands Nombres pour approximer l'espérance par la moyenne empirique.

On effectue une suite d'approximations de la façon suivante :

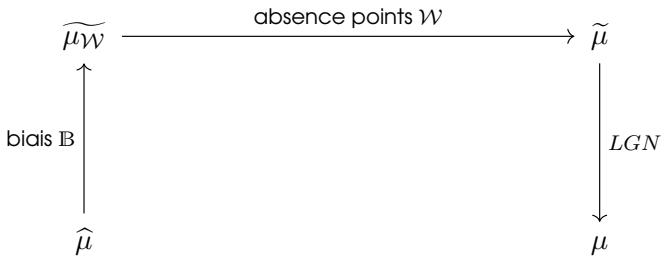


FIGURE 2.5 – Schéma du découpage du contrôle des erreurs

On détermine alors fenêtre de lissage en minimisant le risque suivant (7) :

$$R_\mu^{[Golovk.]}(t,h) = \underbrace{q_1^2 h^{2H_t}}_{\text{contrôle du biais}} + \underbrace{\frac{q_2^2}{\mathcal{N}_\mu(t,h)}}_{\text{contrôle de la variance}} + \underbrace{q_3^2\Big[\frac{1}{\sum_k w_k} - \frac{1}{n}\Big]}_{\text{pénalise absence de points}}$$

Il est tout à fait possible de regarder directement l'erreur d'approximation entre $\widetilde{\mu_{\mathcal{W}}}$ et μ. Toutefois, le choix de Golovkine est avant tout un choix pédagogique, pour signaler et renforcer l'idée qu'il faut faire attention à l'erreur d'approximation entre l'inobservable et le véritable processus ($\mathbb{E}$ vs $\frac{1}{N}\sum X_i \neq \frac{1}{N}\sum \widehat{X}_i$)

Afin de prendre en compte la dépendance, que l'on doit contrôler aussi, on raisonne plutôt de la façon suivante.

2.3.1 ☐ B⟩ Méthode MPV : dépendance

Lorsque l'on traite le cas de la dépendance, il est tout de suite plus délicat d'obtenir la convergence d'estimateurs de moments d'une loi. MPV utilise ce découpage du risque

pour déterminer une fenêtre de lissage adaptée à l'estimation de la fonction moyenne :

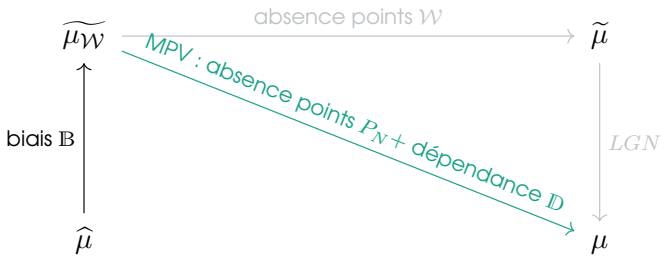


FIGURE 2.6 – Schéma du découpage du contrôle des erreurs

On détermine cette fois-ci la fenêtre de lissage en minimisant le risque suivant (16) :

$$R_\mu(t,h) = \underbrace{L_t^2 h^{2H_t}\mathbb{B}(t,h,2H_t)}_{\text{contrôle du biais}} + \underbrace{\sigma^2\mathbb{V}_\mu(t,h)}_{\text{contrôle de la variance}} + \underbrace{\frac{\mathbb{D}_\mu(t)}{P_N(t,h)}}_{\text{contrôle de la dépendance}}$$

2.3.2 Estimation adaptative de l'opérateur de covariance

On souhaite désormais estimer la quantité la covariance de la loi de notre processus. Si il semblerait naturel d'évaluer :

$$C_X(s,t) = \mathbb{E}\left[\big(X(t)-\mu(t)\big)\cdot\big(X(s)-\mu(s)\big)\right]$$

la quantité qui nous intéresse, in-fine est l'opérateur de covariance :

$$c[f] = \int_I f(u)c(u,\cdot)\,du$$

C'est parceque c'est cet opérateur qui nous donnera, notamment, les vecteurs et valeurs propres de la décomposition dans la base FPCA, aussi connue sous le nom de décomposition de Karhunen-Loève.

On comprend bien que la covariance est une quantité qui mesure la dispersion des données et qu'il est donc naturel de s'intéresser beaucoup plus aux fines variations dans un voisinage proche du couple (t,s) des différents temps qui nous intéressent. Cela vient motiver, une fois de plus l'intérêt de l'utilisation d'un lissage adaptatif qui nous est donné par la minimisation du risque suivant (16) :

$$R_\Gamma(t,h) =$$

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

Chapitre 3

Détermination du diamètre optimal des intervalles à considérer pour l'estimation de la régularité locale

Contents

3.1 Introduction et Objectifs de la simulation

3.2 Simulation de données $\mathrm{FAR}(1)$ Localement Hölderiennes

3.2.1 Fonction de Hurst

On appelle $H : t \mapsto H_t$ la fonction de Hurst, celle qui a été choisie est la suivante :

$$H^{[0.4,0.8,5,0.5]}_{\text{logistic}} : \begin{array}{ccc} [0,1] & \longrightarrow & [0.4, 0.8] \\ t & \longmapsto & 0.4 + \frac{(0.8-0.4)}{1+e^{-5(t-0.5)}} \end{array}$$

On dispose donc d'une régularité locale qui varie sur $\mathcal{T}$, tout en ayant une évolution pas trop brusque. Nous allons étudier le comportement du Δ lors de l'estimation de la régularité locale en les points suivants :

$$\vec{t} = \begin{bmatrix} 0.3 \\ 0.4 \\ 0.5 \\ 0.6 \\ 0.7 \\ 0.8 \end{bmatrix} \qquad H(\vec{t}) = \begin{bmatrix} 0.51 \\ 0.55 \\ 0.6 \\ 0.65 \\ 0.69 \\ 0.73 \end{bmatrix}$$

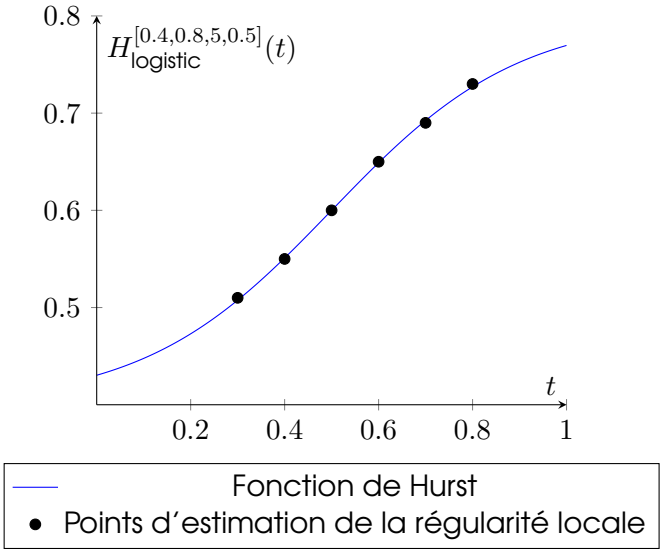


FIGURE 3.1 – Hurst Function : Logistic

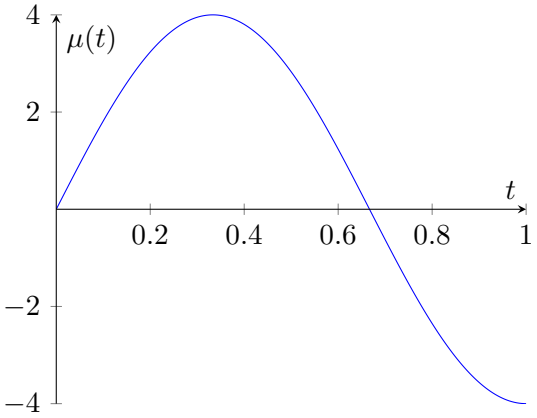
3.2.2 Constante de Hölder

$$\forall t \quad L_t = 1$$

3.2.3 Moyenne

La fonction moyenne du processus que l'on souhaite retrouver après bruitage est la suivante :

$$\mu : \begin{array}{ccc} [0,1] & \longrightarrow & \mathbb{R} \\ t & \longmapsto & 4 \cdot \sin\left(\frac{3}{2}\pi \cdot t\right) \end{array}$$



3.2.4 Noyau de la relation FAR(1)

On décide de simuler un FAR(1) basé sur un opérateur linéaire intégral :

$$X_{n+1} = \phi(X_n) + \varepsilon_{n+1}$$

avec : $\phi : f \mapsto \int_{\mathcal{T}} f(u)\beta(u,\cdot)\,du$

C'est une modélisation fréquente des FAR(1). Le noyaux que l'on considère dans l'opérateur intégral pour les simulations est le suivant :

$$\beta : \begin{array}{ccc} [0,1]^2 & \longrightarrow & \mathbb{R} \\ (t,s) & \longmapsto & \frac{9}{4} t\sqrt{s(1-s)} \end{array}$$

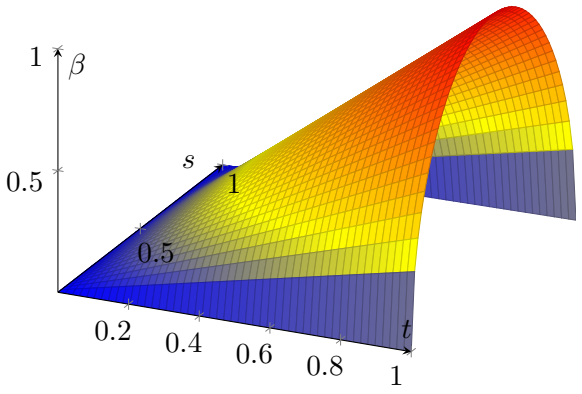


FIGURE 3.2 – Graphique du noyau intégral pour la relation FAR(1)

on notera que le noyaux utilisé pour la relation de FAR(1) est une fonction lisse. ⚠ (il me semble qu'il est important que le noyaux soit plus régulier que le processus, mais c'est à vérifier)

3.2.5 Nombre de courbes

Afin d'étudier le lien potentiel qu'il pourrait y avoir entre le nombre de courbes observées et le Δ optimal pour l'estimation de la régularité locale, on choisit plusieurs valeurs de nombres de courbes observées de telle sorte à avoir un « petit »et un « grand »nombre de courbes observées.

On choisit les valeurs suivantes concernant le nombre de courbes observées :

$$\vec{N} = [100, 200, 300, 400]$$

3.2.6 Ensemble des Δ testés

On souhaite obtenir plusieurs graphiques avec Δ sur l'axe des abscisses afin de pouvoir étudier le comportement de diverses quantitées, telle que le risque quadratique, lorsque l'on fait varier Δ avec certains paramètres fixés (nombre de courbes observées, nombre moyen de points observés par courbe, ...). Toutefois plus on va considérer de Δ, et plus la simulation sera coûteuse.

En effet, pour simuler un mouvement brownien multi fractionnaire, il est nécessaire d'inverser une matrice de covariance, cette opération est d'odre de complexité $\mathcal{O}(T^3)$, avec T le nombre de points où l'on doit évaluer nos $(X_i)_{i,1}\,n$. Dans notre cas, le nombre de points considérés pour la simulation est :

$$\underbrace{\dim \vec{\Delta}}_{30} \times \underbrace{3}_{t_1/t_2/t_3} \times \underbrace{\dim \vec{t}}_{6} + \underbrace{n_{Grid\_\int}}_{100} + \underbrace{\lambda}_{\leq 480} \leq \underbrace{640}_{fixe} + \underbrace{480}_{pts\,aleat} = 1\,120$$

En effet, afin d'avoir une approximation correcte de l'intégrale requise pour la relation FAR(1), on choisit d'effectuer la méthode des rectangles en découpant $[0,1]$ en 100 sous intervalles réguliers. On a besoin aussi d'évaluer la vraie valeur des X_i en t_1, t_2, t_3 et ce,

pour chaque valeur de Δ afin de pouvoir comparer l'estimateur $\hat{\theta}(u,v) = \frac{1}{N}\sum_i(\widehat{X}_i(u) - \widehat{X}_i(v))^2$ obtenu via le pré-lissage avec la valeur intangible $\tilde{\theta}(u,v) = \frac{1}{N}\sum_i(X_i(u) - X_i(v))^2$; et ce sur l'ensemble des différents points $t_2 \in \vec{t}$

Afin d'avoir un nombre raisonnable de points pour travailler tout en ayant des temps de simulation raisonables, on décide de prendre 30 Δ uniformément répartis entre 0.01 et 0.2, au delà duquel de toute façon le diamètre de l'intervalle des points considérés pour la régularité devient tellement grand comparé à la taille du support, qu'il serait inconfortable de parler de « régularité locale ».

$$\vec{\Delta} = [0.01 \cdots (0.01 + k \cdot 0.01) \cdots 0.2]_{k \in 0:30}$$

3.2.7 Bruit blanc

Une fois que l'on a simulé

$$(X_i)_{i,1}\, n \qquad X_{n+1} = \phi(X_n) + \varepsilon_n$$

on doit désormais reproduire l'erreur de mesure, pour cela chaque courbe est ensuite bruitée en rajoutant un bruit blanc :

$$\eta \sim \mathcal{N}(0, 0.04)$$

Il est important d'avoir un bruit blanc d'écart type d'un ordre de grandeur en dessous de celui des valeurs prises par le processus, sinon l'estimation serait mauvaise quoi qu'il arrive.

3.2.8 Résumé des Paramètres

	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

3.2.9 Les courbes obtenues

(ajouter graphe : même courbe avec et sans bruit, et lissée)

3.3 Prélissage des données simulées

3.4 Qualité de l'estimation des incréments quadratiques moyens

Il y a différentes manières de définir les paramètres de régularité $\hat{H}_t$ et $\hat{L}_t$. En effet il est possible de définir $\hat{H}_t$ en utilisant $\hat{\theta}(t_1, t_2)$ mais aussi en utilisant $\hat{\theta}(t_2, t_3)$ ($\theta(t_1, t_3)$ est forcément

utilisé [1]). De même pour $\hat{L}_t$. On peut donc se demander quels sont les meilleurs $\theta(u,v)$ avec $u,v \in \{t_1,t_2,t_3\}$ à utiliser pour obtenir la meilleure estimation de H_t et L_t ainsi que leur Δ optimal associé pour l'estimation de ces paramètres.

Le problème est que le proxy θ est défini comme une espérance, et donc n'est pas observable. On ne peut donc pas directement comparer $\hat{\theta}(u,v) = \sum_i |\widehat{X}_i(u) - \widehat{X}_i(v)|^2$ et $\theta(u,v) = \mathbb{E}_X\left[|X(u)-X(v)|^2\right]$, à moins d'avoir fait le calcul de l'expression explicite en connaissant la loi du processus initial.

On peut cependant comparer $\hat{\theta}(u,v)$ et $\widetilde{\theta}(u,v) = \frac{1}{N}\sum_i |X_i(u)-X_i(v)|^2$ qui est un estimateur de $\theta(u,v)$, et que l'on obtient aisément avec la simulation. On peut ainsi déterminer pour quelle valeur de Δ et quel couple (u,v) on dispose de la meilleur estimation du $\tilde{\theta}$, qui est entre-autre le meilleur estimateur que l'on pourrait espérer de θ. Le meilleur couple (au sens donné dans cette section) est pris comme étant les deux $\hat{\theta}(u,v)$ réalisant les risques minimaux par rapport au $\tilde{\theta}$ sur les 3 couples (u,v) possibles.

	$\lambda < 120$	$\lambda \geq 120$
$H_t < 0.6$	$\Theta_{1\to 2\to 3}$ $\Delta^- \to 0.01$	$\Theta_{1\to{}^{3}_{2}}$ $\Delta^+ \to 0.2$
$H_t \geq 0.6$	$\Theta_{1\to{}^{3}_{2}}$ $\Delta^- \to 0.2$	$\Theta_{1\to 2\to 3}$ $\Delta^+ \to 0.01$

TABLE 3.2 – Tableau récapitulatif des Θ optimaux : Risque individuel sur $\tilde{\theta}(u,v)$

3.5 Qualité de l'estimation de la régularité locale

Les simulations de Monte Carlo permettent d'avoir accès directement à la véritable régularité de la courbe en chaque point. Nous allons dans l'étude du comportement du Δ essayer de tirer profit de cet avantage que ne possède pas le praticien qui utilise des données réelles.

	$\lambda < 120$	$\lambda \geq 120$
$H_t \leq 0.65$	$\overset{\checkmark}{\iff} \mathcal{R}$ $\overset{\checkmark}{\iff} \Delta^*$ $\Delta^- \downarrow 0.01$	$\simeq \overset{\checkmark}{\iff} \mathcal{R}$ $\overset{\times}{\iff} \Delta^*$ $\Delta^+ \to [\leq 0.6]0.1/0.2[\geq 0.6]$
$H_t > 0.65$	$\overset{\checkmark}{\iff} \mathcal{R}$ $\overset{\times}{\iff} \Delta^*$ ⚠$H = 0.7 : \Delta^- = 0.02$ ⚠$H = 0.73 : \Delta^- = 0.2$	$\Theta_{1\to{}^{3}_{2}}$ ⚠$H = 0.7 : \Delta^+ = 0.02$ ⚠$H = 0.73 : \Delta^+ = 0.2$

TABLE 3.3 – Tableau récapitulatif des Δ optimaux : Risque sur H_t

1. $\hat{H}_t$ ne serait même pas bien défini pour le couple $\theta(t_1,t_2)$, $\theta(t_2,t_3)$

3.6 Qualité de l'estimation des couples d'incréments utilisés dans l'estimation de la régularité

on rappelle les notations suivantes :
- vrai : $\theta = \mathbb{E}\big[\, f(X) \,\big]$
- intangible : $\widetilde{\theta} = \frac{1}{N}\sum_i f(X_i)$
- observable : $\hat{\theta} = \frac{1}{N}\sum_i f(\widehat{X}_i)$

L'estimateur du paramètre de régularité H_t est donné par :

$$\hat{H}_t = \frac{\log\hat{\theta}(t_1,t_3) - \log\hat{\theta}(t_1,t_2)}{2\log 2}$$

ou bien encore :

$$\hat{H}_t = \frac{\log\hat{\theta}(t_1,t_3) - \log\hat{\theta}(t_2,t_3)}{2\log 2}$$

autrement dit :

en posant $\Theta_{1\to\substack{3\\2}} = \begin{bmatrix}\theta(t_1,t_3)\\ \theta(t_1,t_2)\end{bmatrix}$ et $\Theta_{\substack{1\\2}\to 3} = \begin{bmatrix}\theta(t_1,t_3)\\ \theta(t_2,t_3)\end{bmatrix}$

$$\hat{H}_t : \Theta \longmapsto \frac{\log\Theta_1 - \log\Theta_2}{2\log 2}$$

C'est pourquoi, étant donné que le meilleur estimateur que l'on puisse espérer soit $\big(\hat{H}_t(\Theta_{1\to\substack{3\\2}})$ ou $\hat{H}_t(\Theta_{\substack{1\\2}\to 3})\big)$, on va s'intéresser désormais à l'estimation conjointe des deux $\theta(u,v)$ utilisés dans l'estimation de H_t comme critère de sélection du diamètre Δ.

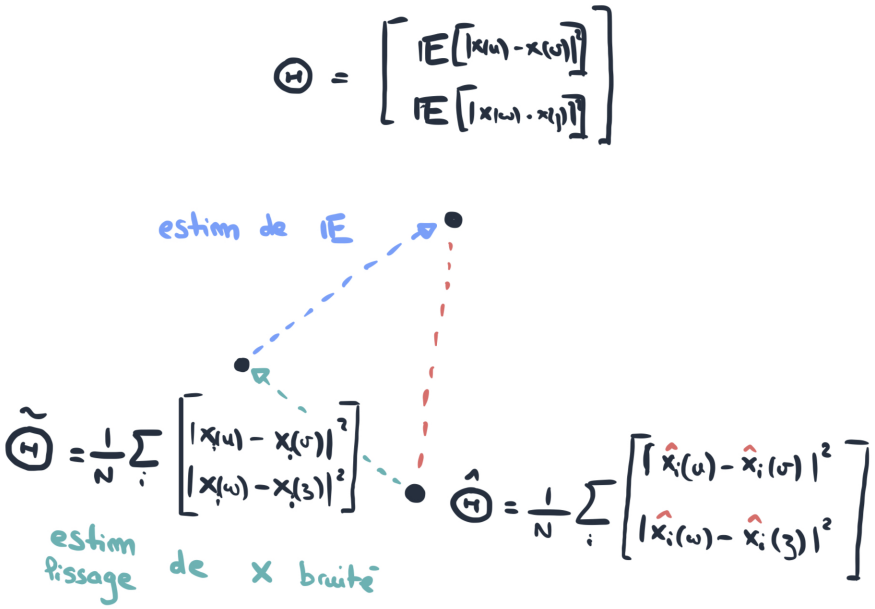


FIGURE 3.3 – title

Pour cela on considère la distance euclidienne usuelle pour des vecteurs de $\mathbb{R}^2$

$$R(\Theta, \Delta) = \| \widehat{\Theta}(\Delta) - \widetilde{\Theta}(\Delta) \|_2$$

et on nomme $R_{1\to\substack{3\\2}} = R(\Theta_{1\to\substack{3\\2}}, \cdot)$ et $R_{\substack{1\\2}\to 3} = R(\Theta_{\substack{1\\2}\to 3}, \cdot)$

	$\lambda < 120$	$\lambda \geq 120$
$H_t < 0.6$	$\overset{\checkmark}{\Longleftrightarrow}\ \mathcal{R}, \Delta^*$ $\Delta^*_- = 0.01$	$\Theta_{\substack{1\\2}\to 3}$ $\Delta^*_+ = 0.2$
$H_t \geq 0.6$	$\Theta_{\substack{1\\2}\to 3}$ $\Delta^*_- = 0.2$ ⚠$H = 0.7 : \Delta^- = 0.01 \oplus \overset{\checkmark}{\Longleftrightarrow}\ \mathcal{R}$ ⚠$H = 0.8 : \Theta_{1\to\substack{3\\2}}$	$\overset{\checkmark}{\Longleftrightarrow}\ \mathcal{R}, \Delta^*$ $\Delta^*_+ = 0.01$

TABLE 3.4 – Tableau récapitulatif des Δ optimaux : Risque euclidien sur $\tilde{\Theta}$

3.7 Détermination d'un critère de choix du diamètre Δ des intervalles à considérer pour l'estimation de la régularité locale

Maintenant que l'on a déterminé que l'on souhaite travailler sur un les couples $\Theta_{\substack{1\\2}\to 3} = \begin{bmatrix}\theta(t_1,t_3)\\ \theta(t_2,t_3)\end{bmatrix}$ et $\Theta_{1\to\substack{3\\2}} = \begin{bmatrix}\theta(t_1,t_3)\\ \theta(t_1,t_2)\end{bmatrix}$, il nous faut déterminer un critère pour déterminer quel couple est plus judicieux pour la méilleure estimation en pratique des paramètres de régularité locale.

L'heuristique est la suivante : dans nos simulations, on a le luxe de pouvoir faire 200 simulations de monte carlo et obtenir le Δ^* le plus proche du Δ optimal pour estimer la régularité. Dans la pratique, obtenir un tel Δ optimal n'est pas réaliste, on se trouvera soit un peu en dessous, soit un peu au dessus. L'idée est donc de favoriser le couple de $\theta(u,v)$ qui possède le plus grand plateau autour du Δ^* pour le risque quadratique *si l'écart de risque quadratique entre les deux couples n'est pas trop important*. Si l'un est beaucoup plus performant que l'autre, on choisira le plus performant. Mais si la performance des deux est à peu près équivalente, autant sélectionner celui qui dans la pratique (sans avoir 200 réplications indépendantes) nous donnera le plus de flexibilité sur l'erreur commise en sélectionnant un Δ autour du Δ^* dû à la fluctuation statistique.

3.7.1 Détermination d'un seuil pour l'équivalence de risque quadratique

Il nous faut maintenant déterminer ce que l'on considère comme étant deux risques "équivalents". Pour cela on va déterminer pour différentes valeurs du véritable H le seuil ε sur le risque tel que $R_{1\to\substack{3\\2}}(\Delta + \delta) + \varepsilon$ induit une erreur d'au maximum 10% sur le H estimé. On viendra ensuite déterminer les δ qui en moyenne correspondent à ce seuil ε pour les différentes valeurs de H.

3.7.2 Détermination du meilleur couple à risque « équivalent »

3.7.2 ☐ A⟩ en utilisant les pentes

Une méthode possible serait de définir la pente à gauche et la pente à droite de la façon suivante :

$$a_g : \Delta, \delta \mapsto \frac{R(\Delta) - R(\Delta - \delta)}{\delta}$$
$$a_d : \Delta, \delta \mapsto \frac{R(\Delta + \delta) - R(\Delta)}{\delta}$$

On peut définir les pénalisations suivantes pour déterminer le meilleur couple à risque équivalent en terme de plateau, en pénalisant les larges différence entre la pente à gauche et à droite :

$$m_q(\Delta, \delta) = \frac{a_g^2(\Delta, \delta) + a_d^2(\Delta, \delta)}{2}$$

3.7.2 ☐ B⟩ en utilisant les valeurs de risque

une autre méthode est de regarder :

$$R_2(\Delta_2^*) \geq R_1(\Delta_1^*)$$
$$dR = \left|R_1(\Delta_1^*) - R_2(\Delta_2^*)\right|$$

on compare désormais les valeurs de :

$$r_g^{[2]} = R_2(\Delta_2^* - \delta) - dR$$
$$r_d^{[2]} = R_2(\Delta_2^* + \delta) - dR$$

aux valeurs

$$r_g^{[1]} = R_1(\Delta_1^* - \delta)$$
$$r_d^{[1]} = R_1(\Delta_1^* + \delta)$$

avec le critère de sélection suivant :

$$\operatorname{argmin}\left(\frac{r_g^{[1]} + r_d^{[1]}}{2}, \frac{r_g^{[2]} + r_d^{[2]}}{2}\right)$$

Pour pénaliser les solutions où la pente à gauche est très différente de la pente à droite en magnitude, on peut considérer d'élever r_g et r_d au carré.

$$\operatorname{argmin}\left(\frac{(r_g^{[1]})^2 + (r_d^{[1]})^2}{2}, \frac{(r_g^{[2]})^2 + (r_d^{[2]})^2}{2}\right)$$

3.7.2 ☐ C⟩ résultat

Chapitre 4

Application sur les données simulées

Contents

4.1 Estimation de la fonction moyenne

4.2 Estimation de la fonction de covariance

4.3 Estimation de l'auto-corrélation du modèle $\mathrm{FAR}(1)$

4.4 base FPCA

4.5 Conclusion

Chapitre 5

Application sur les données réelles de courbes de charge éolienne

Contents

5.1 Présentation du jeu de données

5.2 Pré-traitement des données

5.3 Pré-lissage et estimation de la régularité locale

5.4 Estimation de la fonction moyenne

5.5 Estimation de la fonction de covariance

5.6 Estimation de l'auto-corrélation du modèle $\mathrm{FAR}(1)$

5.7 base FPCA et interprétation

5.8 Conclusion

Annexe A

Détails techniques et théoriques

A.1 Régularité Locale

Nous avons mentionné que les processus auxquels on allait s'intéresser étaient les processus localement Höldériens de paramètres $\big(\alpha(t), L_\alpha(t)\big)$. Ce n'est pas tout à fait vrai. Si le coeur de ce que l'on considère sont bel et bien les processus Höldériens, on élargit encore plus la classe des processus que l'on considère en considérant les processus qui sont **presque** Höldériens.

Qu'est ce qu'on entend exactement par presque Höldérien ?

pour u et v dans un voisinage de t de dimaètre Δ :

Ce que l'on demandait pour un processus X est que pour tout u, v dans un voisinage de t de diamètre Δ, il existe L_t et H_t telle qu'on ait :

$$\theta(u,v) \underset{\text{déf}}{\equiv} \mathbb{E}\left[\,\big|\,X(u)-X(v)\,\big|^2\right] \leq L_t^2\,|u-v|^{2H_t}$$

on peut alors retrouver la régularité du processus comme un processus Höldérien de paramètres $\big(H_t, L_t\big)$ d'après le théorème de Continuité de Kolmogorov.

en réalité il suffit que $\theta(u,v)$ soit suffisamment proche d'un processus localement Höldérien de paramètres $\big(H_t, L_t\big)$ et que l'on puisse contrôler l'écart entre les deux. Cet écart dépend de Δ et de la régularité. C'est ce qu'affirme les deux hypothèses suivantes qui sont en fait les hypothèses de régularité qui sont considérées par MPV(16).

$$\big|\theta(u,v) - L_t^2|u-v|^{2H_0}\big| \leq S_t^2|u-v|^{2H_0}\Delta^{2\beta_0}$$

(16, H6)

$$\left|\nu_2\left(\nabla^\delta X(u) - \nabla^\delta X(v)\right)^2 - L_{\delta,t}^2|u-v|^{2H_\delta}\right| \leq S_{\delta,t}^2|u-v|^{2H_\delta}\Delta^{2\beta_\delta}$$

(16, D1-7)

On remarquera que si le processus est localement Höldérien, alors on a un contrôle optimal de l'écart entre $\theta(u,v)$ et $L_t^2\,|u-v|^{2H_t}$.

L'auteur saura donc reconnaître, que bien que ce qui ait été exposé ne soit pas la forme exacte, cela ne change rien à l'idée générale. De plus, cela alourdirait considérablement la rédaction et rendrait la compréhension bien plus difficile de l'objectif du stage.

Annexe B

Algorithmique

B.1 Optimisation Algorithmique

B.1.1 Génération du bruit blanc

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

$$\underbrace{\dim \vec{\Delta}}_{50} \times \underbrace{3}_{t_1/t_2/t_3} \times \underbrace{n_{points\_estim}}_{6} + \underbrace{n_{Grid\_\int}}_{100} + \underbrace{\lambda}_{\leq 480} \leq \underbrace{1000}_{fixe} + \underbrace{480}_{pts\,aleat}$$

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 :

$$U \underset{\text{déf}}{\equiv} BD^{-1}$$
$$V \underset{\text{déf}}{\equiv} CA^{-1}$$

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

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

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 :
$U = \Sigma_{[\text{alea}/\neg\ \text{alea}]}\Sigma^{-1}_{[t\ \text{alea}]}$
$V = \Sigma^T_{[\text{alea}/\neg\ \text{alea}]}\Sigma^{-1}_{[t\neg\ \text{alea}]}$

B.1.1 □ A⟩ Intégrale

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

il est important lorsque l'on effectue autant de simulations d'avoir des calculs efficients 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.

Annexe C

Code & Implémentations

C.1 packages utilisés

```
    # --- install --- #
    install.packages(c("data.table","wavethresh","fda", "fda.usc"))
    # --- general packages --- #
    library(data.table)
    # --- FDA packages --- #
    library(fda)
    library(fda.usc)
    # --- Wavelet packages --- #
    library(wavethresh)
```

C.2 Simulation des FAR

C.3 Lissage des courbes

C.4 Détermination de la régularité locale

C.5 Détermination des risques

C.6 Lissage adaptatif

Dans la section 2.3, nous avons mentionné qu'il était judicieux de lisser les courbes de façon adaptative à la quantité que l'on souhaite estimer. Si l'on a mentionné le risque à minimiser pour chaque quantité que l'on souhaite estimer, aucun détail n'a été fourni car il alourdit considérablement la trame de l'objectif du stage sans apporter des informations cruciales.

Cependant pour l'implémentation d'un tel lissage adaptatif, il fallait évidemment se référer au détail de l'expression pour pouvoir évaluer ce risque et déterminer la meilleure fenêtre.

$$R_\mu(\,t\,,h\,) = \underbrace{L_t^2 h^{2H_t}\mathbb{B}(\,t,h,2H_t\,)}_{\text{contrôle du biais}} + \underbrace{\sigma^2\mathbb{V}_\mu(\,t,h\,)}_{\text{contrôle de la variance}} + \underbrace{\frac{\mathbb{D}_\mu(\,t\,)}{P_N(t,h)}}_{\text{contrôle de la dépendance}}$$

Développons maintenant les différentes quantités présentes dans l'expression :

$$\mathbb{D}_\mu(\,t\,) \underset{\text{déf}}{\equiv}$$

$$\mathbb{V}_\mu(\,t,h\,) \underset{\text{déf}}{\equiv}$$

$$\mathbb{B}(\,t,h,2H_t\,) \underset{\text{déf}}{\equiv}$$

$$P_N(t,h) \underset{\text{déf}}{\equiv}$$

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 probabiblités et en statistique : Essai d'étude synthétique*. PhD thesis, Éditeur inconnu, 1976.

(6) ENGIE. Quelle est la consommation d'électricité par personne dans un foyer ? (lire en ligne) , 2022.

(7) Steven Golovkine, Nicolas Klutchnikoff, and Valentin Patilea. Adaptive estimation of irregular mean and covariance functions, 2021.

(8) 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.

(9) 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.

(10) Siegfried Hörmann and Piotr Kokoszka. Weakly dependent functional data. *The Annals of Statistics*, 38(3) :1845 – 1884, 2010.

(11) 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.

(12) Gareth M James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(3) :411–432, 2002.

(13) Kari Karhunen. Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae, AI*, 34, 1946.

(14) Piotr Kokoszka and Matthew Reimherr. *Introduction to functional data analysis*. CRC press, 2017.

(15) 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.

(16) Patilea Maissoro, Vimond. Learning smoothness of functional times series under weak dependency assumption. 2023. not available yet.

(17) Stéphane Mallat. Wavelet tour — part vi — wavelet zoom. (lire en ligne) .

(18) Hans-Georg Müller and Ulrich Stadtmüller. Generalized functional linear models. 2005.

(19) James O Ramsay. When the data are functions. *Psychometrika*, 47 :379–396, 1982.

(20) 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.

(21) William F Sharpe. A simplified model for portfolio analysis. *Management science*, 9(2) :277–293, 1963.

(22) Yousri Slaoui. Recursive nonparametric regression estimation for dependent strong mixing functional data. *Statistical Inference for Stochastic Processes*, 23(3) :665–697, 2020.

(23) 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.

(24) 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.

(25) 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.