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

STAGE 2A

CREST

Modèles additifs avec covariables contaminées et généralisation du modèle additif aux données fonctionnelles

version corrigée

rédigé par
BRUNET Hugo

Novembre 2023

Résumé

Pendant le cursus 2A réalisé à l'ENSAI, les étudiants découvrent la régression non paramétrique qui permet d'estimer la loi conditionnelle d'une réponse vis à vis de covariables sans hypothèse sur la forme de la relation entre la réponse et les covariables. On peut alors complexifier les modèles de données que l'on considère en tirant avantage des bénéfices de l'approche paramétrique, aux vitesses de convergence rapide des estimateurs, et de l'approche non paramétrique, flexible et plus robuste à l'erreur de choix du modèle. On appelle cela une méthode « semi-paramétrique ». Un modèle semi-paramétrique courant est le modèle partiellement linéaire avec la composante non paramétrique supposée additive. L'ensemble des concepts et des motivations sont introduites dans ce stage.

Il existe un modèle de données appelées données fonctionnelles qui sont de plus en plus présentes dans différents champs d'application de la statistique : santé, sport, industrie .... De la théorie a déjà été produite sur la régression fonctionnelle. Peut-on faire de la régression semi-paramétrique fonctionnelle ? Le sujet de ce stage est l'étude à partir de diverses ressources bibliographiques sur la régression semi-paramétrique, de comprendre dans un premier lieu à partir du savoir d'un étudiant de 2^e année du cursus ingénieur de l'ENSAI la méthodologie derrière la régression semi-paramétrique du modèle partiellement-linéaire dans le cadre réel. Enfin on pourra rendre compte des difficultés rencontrées ainsi que les différences et similarités dans les approches lors de l'extension des méthodes du modèle additif au cadre fonctionnel.

contribution

si jamais vous apercevez des fautes dans le polycopié, merci de rédiger une issue sur Github à l'adresse :

correctif

ENSAI-2A-stage-FGAM/issues

contact

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

Notations

Notation	Signification
Probabilités	
$\mathcal{L}\int$ ou $\int$	Intégrale de Lebesgue
$\mathbb{B}\int$	Intégrale de Böchner
$p \cdot \mu$	mesure p à densité par rapport à la mesure μ
Statistiques	
$\mathbb{V}A\big[\, E \,\big]$	ensemble des Variables Aléatoires à valeurs dans un ensemble E : applications mesurables $(\Omega, \mathcal{F}, \mathbb{P}) \to E$
X	Variable Aléatoire
$\widehat{X}$	Quantité empirique
$\widetilde{X}$	Quantité intangible (« unfeasible » dans la littérature)
$\overline{X}$	Moyenne empirique
$X_d^{[i]}$	d^{eme} composante de l'observation i de la variable aléatoire multivariée X
X^*	Covariable contaminée par une erreur additive indépendante $X^* = X + U$
x	Réalisation de la variable aléatoire X : $x = X(\omega)$ pour un certain $\omega \in \Omega$
Algèbre Linéaire	
E_k	Opérateur d'espérance conditionnelle selon la composante k : $E_k : \begin{array}{ccc} \mathbb{V}A\big[\, E^d \,\big] & \longrightarrow & \mathbb{V}A\big[\, E \,\big] \\ g(X) & \longmapsto & \mathbb{E}\,[\, g(X) \,|\, X_k \,] \end{array}$
$\mathbf{E}\left(\neq \mathbb{E}\right)$ ⚠	Matrice du système d'équation conditionné du modèle additif : cf 1.1.1 ☐ B⟩
$\|\cdot\|_{\mathbb{L}^2}$	Norme de $\mathbb{L}^2$ issue du produit scalaire $\langle f \vert g \rangle_{\mathbb{L}^2} = \int fg \, d\lambda$
Ensembles	
$\mathcal{C}^k(E, F)$	Fonctions de classe k de E dans F
$\mathcal{A}$	Fonctions « additives » : $f : \mathbb{R}^d \to \mathbb{R}$ telles que $\exists (f_k : \mathbb{R} \to \mathbb{R})_{1,d}$ et $f(x) = \sum_{k=1}^d f_k(x_k) \quad \forall x = (x_k)_{1,d} \in \mathbb{R}^d$
$\mathbb{L}^2(\mu)$	Ensemble quotient des classes de fonctions de carré μ-Lebesgue-intégrable pour la relation d'équivalence d'égalité μ presque partout : le μ est omis lorsque sans ambiguïté
$\mathbb{B}(\mu)$	Fonctions Böchner-μ-intégrables contexte : Données Fonctionnelles
Analyse	
D^α	Dérivée multivariée : $D : \begin{array}{ccc} \mathbb{N}^d \times \mathcal{C}^{\max_i \alpha_i}(\mathbb{R}^d, \mathbb{R}) & \longrightarrow & \mathcal{C}^{[\alpha]}(\mathbb{R}^d, \mathbb{R}) \\ (\alpha_i)_{1,d} \ , \ f & \longmapsto & \dfrac{\partial^{\sum_i \alpha_i}}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d} f \end{array}$
$[\alpha]$	Ordre de dérivation : $\sum_i \alpha_i$
$[\alpha]^{max}$	ordre maximal de dérivation d'un opérateur étant combinaison linéaire d'opérateurs de différentiation multivariée : $\max\{\alpha_k$: où $T = \sum_k c_k D^{\alpha_k}\}$ lorsque l'on souhaite estimer $T(m)$

Notation	Signification
Transformée de Fourier	
$\mathcal{F}$	Opérateur de transformée de Fourier : $\mathcal{F} : \begin{array}{ccc} \mathbb{L}^1 & \longrightarrow & \mathbb{L}^1 \\ f & \longmapsto & \omega \mapsto \int e^{-i\omega x} f(x) dx \end{array}$
$\mathcal{F}^{-1}$	Opérateur de transformée de Fourier inverse : $\mathcal{F}^{-1} : \begin{array}{ccc} \mathbb{L}^1 & \longrightarrow & \mathbb{L}^1 \\ f & \longmapsto & \omega \mapsto \frac{1}{2\pi} \int e^{+i\omega x} f(x) dx \end{array}$
$\mathcal{F}_{\text{stat}}$	Opérateur de transformée de Fourier « statistique » : $\mathcal{F}_{\text{stat}} : \begin{array}{ccc} \mathbb{L}^1 & \longrightarrow & \mathbb{L}^1 \\ f & \longmapsto & \omega \mapsto \int e^{+i\omega x} f(x) dx \end{array} \overset{\text{not.}}{=} \phi_f$
ϕ_X	Fonction caractéristique de la variable aléatoire X : $\phi_X : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{C} \\ \omega & \longmapsto & \mathbb{E}\left[e^{i\omega X} \right] \end{array}$
ϕ_{erreur}	Fonction caractéristique de l'erreur de mesure U aussi appelée ϕ_u quand il n'y a pas d'ambiguïté
Noyaux	
K	Noyau de lissage : $K : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R}_+ \\ x & \longmapsto & \omega(x) \mathbb{1}_{[-1,1]} \end{array}$
K_h	Noyau de lissage fenêtré : $K_h : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R}_+ \\ x & \longmapsto & \frac{1}{h} K\left(\frac{x}{h}\right) \end{array}$
$K_h^{[x_0]}$	Noyau de lissage fenêtré centré en x_0 : $K_h^{[x_0]}$: $\begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R}_+ \\ x & \longmapsto & \frac{1}{h} K\left(\frac{x-x_0}{h}\right) = K_h(x - x_0) \end{array}$
$K_{h,\|\cdot\|}^{[x_0]}$	Noyau de lissage normalisé fenêtré centré en x_0 : $K_{h,\|\cdot\|}^{[x_0]}$: $\begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R}_+ \\ x & \longmapsto & \frac{K_h^{[x_0]}(x)}{\int K_h^{[x_0]}(u) du} \end{array}$
$\tilde{K}_h$	Noyau de déconvolution basé sur le noyau de lissage K : $\tilde{K}_h : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R}_+ \\ t & \longmapsto & \frac{1}{2\pi} \int e^{-iut} \frac{\mathcal{F}_{\text{stat}}[K](hu)}{\phi_{\text{erreur}}(u)} du \end{array}$
$\tilde{K}_h^\star$	Noyau de déconvolution normalisé basé sur le noyau de lissage K : $\tilde{K}_h^\star : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R}_+ \\ t & \longmapsto & \frac{1}{2\pi h} \int e^{-i\omega \frac{x - X^*}{h}} \cdot \frac{\phi_{K,\|\cdot\|}^{[x]}(\omega\,;\, h) \phi_K(\omega)}{\phi_u\left(\frac{\omega}{h}\right)} \end{array}$

Notation	Signification
Estimation	
X	Covariable assignée à la partie paramétrique du modèle
Z	Covariable assignée à la partie non paramétrique du modèle
O_i	Observation de l'individu i : $O_i = (X^{*[i]}, Z^{*[i]}, Y^{[i]})$
O	Ensemble des observations : $O = \{O_i\}_{1,n}$
$O_i^{[\text{ideal}]}$	Observation idéale de l'individu i : $O_i^{[\text{ideal}]} = (X^{[i]}, Z^{[i]}, Y^{[i]})$
ε	Erreur de mesure indépendante des données
η	Conditionnement de X par Z projeté sur l'espace des fonctions additives $\mathcal{A}$: $\eta(z) = P_{\mathbb{L}^2 \cap \mathcal{A}} \circ \mathbb{E}[X \mid Z = z]$
$\widehat{\eta}^\star$	Estimation de η par la méthode de déconvolution utilisant le noyau $\tilde{K}_h^\star$
ξ	Conditionnement de Y par Z projeté sur l'espace des fonctions additives $\mathcal{A}$: $\xi(z) = P_{\mathbb{L}^2 \cap \mathcal{A}} \circ \mathbb{E}[Y \mid Z = z]$
$\widehat{\xi}^\star$	Estimation de ξ par la méthode de déconvolution utilisant le noyau $\tilde{K}_h^\star$
$\mathbb{B}$	Biais — contexte : estimation dans $\mathbb{R}$
$\mathbb{B}(\widehat{p})$	Biais de l'estimateur $\widehat{p}$ — contexte : estimation dans $\mathbb{R}$
Autres	
$\mathbb{P}_X$	Loi image réciproque d'une variable aléatoire $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{X}, \mathcal{X}, \mu)$ définie comme $\mathbb{P}_X : \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R}_+ \\ x & \longmapsto & \mathbb{P}[X^{-1}(x)] \end{array}$

Table des matières

Chapitre 1

Modèles Additifs & données imparfaites : le cadre réel

Contents

Le statisticien est amené la plupart de son temps à modéliser les données qu'il traîte afin de pouvoir fournir une analyse et pouvoir faire de la prédiction pour ses clients. Parmis les premiers modèles qu'il apprend à manipuler, se trouve la régression linéaire avec des erreurs gaussiennes en tant que bruit blanc :

$$y = \beta_0 + \sum_{k=1}^{d} \beta_k x_k + \varepsilon$$

On dit alors que l'on fait une modélisation paramétrique. En effet on suppose en premier lieu (et donc on impose) que nos données suivent un lien ici linéaire avec les covariables. On caractérise ainsi la relation entre la réponse Y et les covariables $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}$ par un nombre fini de paramètres : les $(\beta_i)_{1,n}$. De manière générale un modèle paramétrique est un modèle où $y = m_\theta(x) + \varepsilon$ et m_θ est une fonction de x qui est caractérisée par $\theta \in \Theta$ où $\dim \Theta < \infty$.

De manière générale, on voit un problème de régression comme la résolution du problème suivant :

$$\hat{m}(x) = \underset{m \in \mathcal{F}}{\operatorname{argmin}} \left[\, d\big(y, m(x)\big) \, \right] \tag{1.1}$$

où $\mathcal{F}$ est un espace de fonctions et d une certaine métrique ou ce qu'on appelle en machine learning « une fonction de coût ». Une régression à modèle paramétrique est donc le cas où $\mathcal{F}$ est un sous espace de fonctions de dimension finie : le problème devient :

$$\begin{aligned} \hat{m}(x) &= \underset{m \in \mathcal{F}_\Theta}{\operatorname{argmin}} \left[\, d\big(y, m(x)\big) \, \right] \\ &= \underset{\theta \in \Theta}{\operatorname{argmin}} \left[\, d\big(y, m_\theta(x)\big) \, \right] \end{aligned} \tag{1.2}$$

Cette approche permet de modéliser de façon simple les données en obtenant des vitesses de convergence rapides ($n^{-1/2}$ pour la régression linéaire par les moindres carrés). Mais cela est évi-

demment à condition que la modélisation paramétrique choisie ne soit pas trop éloignée du comportement du phénomène étudié, sans quoi le modèle ferait des prédictions, certes, mais qui n'ont pas de sens vis à vis du phénomène étudié. C'est pourquoi il existe une branche de la méthodologie statistique appelée « statistique non paramétrique » qui vise à imposer le moins de restrictions à $\mathcal{F}$.

Cependant la statistique non paramétrique est particulièrement sujette au « fléau de la dimension ». C'est à dire qu'au fur et à mesure que l'on considère de nouvelles covariables, le coût calculatoire ou le nombre de points requis pour obtenir une bonne estimation croît exponentiellement. Prenons le cas d'une régression non paramétrique à noyaux.

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) y_i \begin{cases} w_i(x) = \dfrac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} \\ K : \text{ noyau d'ordre } r \\ h : \text{ fenêtre de lissage} \end{cases}$$

Lorsque $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}$ on considère K^* : $\begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ x = (\, x_k \,)_{1,d} & \longmapsto & \frac{1}{h^d} \prod_{k=1}^{d} K\left(\frac{X_k - x_k}{h}\right) \end{array}$.

On peut maintenant dériver la vitesse de convergence de l'estimateur ponctuellement en regardant la MISE de notre estimateur au point $x \in \mathbb{R}^d$ pour estimer une fonction de classe $\mathcal{C}^k$ avec un noyau d'ordre r :

$$\mathrm{MISE}(\hat{m}) = \left[C(\, r \,|\, K, \partial^r m \,) h^{2r} + \frac{\left(\| K \|^2_{\mathbb{L}^2} \right)^d}{nh^d} \right] \big(1 + o(1)\big) \tag{1.3}$$

$$\mathrm{MISE}(\hat{m}) \underset{h \to 0}{\sim} \frac{1}{n} \left[\frac{\left(\| K \|^2_{\mathbb{L}^2} \right)}{h} \right]^d \xrightarrow[d \to \infty]{} \infty \tag{1.4}$$

On constate alors que l'erreur d'estimation d'estimation explose avec la dimension qui grandit. On peut en fait montrer que la vitesse de convergence (uniformément sur le support) optimale d'une régression non paramétrique est $n^{-\frac{k-[\alpha]^{max}}{2k+d}}$ (20), où $[\alpha]^{max}$ est l'ordre maximal de différentiation de la fonction m pour l'estimation considérée. (par exemple, l'estimation de $\frac{\partial}{\partial x_1 \partial x_2^2} m(x_1, x_2)$ donne un $[\alpha]^{max} = 3 = \underbrace{1}_{\partial x_1} + \underbrace{2}_{\partial x_2^2}$)

ainsi :

$$\widehat{m}(x) = m(x) + O_p\left(n^{-\frac{k}{2k+d}} \right) \tag{1.5}$$

Ici le fléau de la dimension se manifeste dans la vitesse de convergence $n^{-\frac{k}{2k+\mathrm{d}}} \xrightarrow[d \to \infty]{} 0$. En observant l'expression de la vitesse de convergence, on se dit que l'on pourrait obtenir une meilleure vitesse de convergence en réduisant la dimensionalité du problème. C'est ce qui motive l'utilisation de la modélisation suivante dite « additive » :

Définition 1 (modèle additif) On appelle un modèle additif, un modèle où

□ $y = m(x) + \varepsilon$

$$\square\ m : \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ (x_k)_{1,d} & \longmapsto & \sum_{k=1}^{d} m_k(x_k) \end{array}$$

$$\square\ \text{avec } m_k : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x_k & \longmapsto & m_k(x_k) \end{array}$$

L'idée est que l'on va pouvoir désormais faire d régressions non paramétriques en dimension 1 (en lissant les m_k) plutôt qu'une régression en dimension d (en lissant m) pour avoir une vitesse de convergence de la dimension 1. Cette conjecture importante était déjà suggérée par Stone dans son article sur la vitesse de convergence des estimateurs non paramétriques en 1982 :

> Il existe plusieurs questions intéressantes encore ouvertes qui sont liées au Théorème 1 (ce qu'on vient de mentionner, voir (20) pour plus de détails) (. . .) La question suivante est suggérée par le succès pratique de la régression par « projection pursuit » (Friedman & Stuetzle, 1981) :
>
> **Question 2**
>
> soit $d \geq 2$, $\mathcal{F} = \left\{ f : \mathbb{R}^d \to \mathbb{R}^p, f \text{ est } \mathcal{C}^k \right\}$ et $\mathcal{F}_{sub}$ une collection de fonctions m de $\mathbb{R}^d$ additives :
>
> $$m(x_1, \ldots, x_d) = \sum_{p=1}^{d} m_p(x_p)$$
>
> ou une collection de fonctions de la forme :
>
> $$m(x_1, \ldots, x_d) = \psi(\sum_{p=1}^{d} \beta_p x_p)$$
>
> Estimons désormais sur $\mathcal{F} \cap \mathcal{F}_{sub}$ au lieu de $\mathcal{F}$, en posant $r_1 = \frac{k-[\alpha]^{max}}{2k+1}$
>
> > La vitesse de convergence $\boxed{n^{-r_1}}$ est elle une vitesse de convergence atteignable ?
>
> — Stone, 1982 : référence (20)

Une des méthodes que Stone mentionne est notamment l'algorithme dit de « Backfitting » qui permet d'effectuer l'estimation des fonctions m_k du modèle additif. Il se trouve que la réponse à cette question est « oui », en utilisant l'approche de Mammen (1999) que nous allons détailler un peu plus tard (10).

> The asymptotically optimal minimax rates and constants of additive models are the same as they are in nonparametric regression models with one component.
>
> —Horowitz, Klemelä, Mammen : Optimal estimation in additive regression models (2006) (10)

1.1 Modèles additifs

L'idée du modèle additif est de combattre le fléau de la dimension en ajoutant une hypothèse que l'on espère peu coûteuse sur la fonction à estimer : la fonction m est supposée être la somme de d fonctions m_j à une dimension. On peut alors espérer avoir la vitesse de convergence de la régression non paramétrique de la dimension 1 en effectuant d régressions non paramétriques à une dimension.

Comment estime-t-on les d fonctions du modèles ?

1.1.1 Estimation : Algorithme du Backfitting

1.1.1 ☐ A⟩ Les équations d'estimation conditionnelles

Notre modèle d'observation est tel que l'on observe

$$\left(x_1^{[i]}, \dots, x_k^{[i]}, \dots, x_d^{[i]}, y^{[i]}\right) \underset{not.}{=} (x_k, y)_{k\in[\![1\,,\,d]\!]}^{[i]} \underset{not.}{=} O_i$$

avec

$$y = m(x) + \varepsilon$$

en considérant l'hypothèse classique : $\mathbb{E}\left[\,\varepsilon \,|\, X\,\right] = 0$

$$\mathbb{E}\left[\,Y \,|\, X\,\right] = m(X) = \sum_k m_k(X_k) \tag{1.6}$$

on alors pour tout $k \in [\![1\,,\,d]\!]$:

$$\begin{aligned}
\mathbb{E}\left[\,Y \,|\, X_k\,\right] &= \mathbb{E}\left[\,m(X) \,|\, X_k\,\right] \\
&= \mathbb{E}\left[\sum_p m_p(X_p) \,|\, X_k\right] \\
\mathbb{E}\left[\,Y \,|\, X_k\,\right] &= m_k(X_k) + \sum_{p\neq k} \mathbb{E}\left[\,m(X_p) \,|\, X_k\,\right]
\end{aligned} \tag{1.7}$$

1.1.1 ☐ B⟩ Intuition : Formulation Gauss-Seidel

On peut ré-écrire le système d'équations conditionnelles comme un système d'équations linéaires(9) :

$$\begin{cases}
m_1(X_1) + E_1[m_2(X_2)] + \cdots + E_1[m_d(X_d)] &= E_1[Y] \\
\qquad\qquad\qquad\vdots &\quad\vdots \\
E_d[m_1(X_1)] + \cdots + E_d[m_{d-1}(X_{d-1})] + m_d(X_d) &= E_d[Y]
\end{cases} \iff \boxed{\mathbf{E}\circ\mathbb{M}(X) = E\circ Y} \tag{1.8}$$

avec la matrice $\mathbf{E}$, l'application E et le vecteur $\mathbb{M}(X)$ définis par :

$$\mathbf{E} = \begin{bmatrix}
I & E_1 & \dots & \dots & E_1 \\
E_2 & I & E_2 & \dots & E_2 \\
E_3 & E_3 & \ddots & & E_3 \\
\vdots & \vdots & & \ddots & \vdots \\
E_d & E_d & \dots & \dots & I
\end{bmatrix} \qquad E : \begin{array}{ccc}
\mathbb{V}A\left[\,\mathbb{R}\,\right] & \longrightarrow & \mathbb{V}A\left[\,\mathbb{R}^d\,\right] \\
Y & \longmapsto & \begin{bmatrix} E_1[Y] \\ \vdots \\ E_d[Y] \end{bmatrix}
\end{array} \qquad \mathbb{M}(X) = \begin{bmatrix} m_1(X_1) \\ m_2(X_2) \\ \vdots \\ m_d(X_d) \end{bmatrix}$$

On peut désormais décomposer le problème en deux parties et considérer une solution approximative que l'on vient mettre à jour en itérant plusieurs fois : il s'agit de la procédure de Gauss-Seidel :

$$\begin{bmatrix} I & E_1 & \dots & \dots & E_1 \\ E_2 & I & E_2 & \dots & E_2 \\ E_3 & E_3 & \ddots & & E_3 \\ \vdots & \vdots & & \ddots & \vdots \\ E_d & E_d & \dots & \dots & I \end{bmatrix} \qquad \begin{bmatrix} m_1(X_1) \\ m_2(X_2) \\ \vdots \\ \vdots \\ m_d(X_d) \end{bmatrix} = \begin{bmatrix} E_1[Y] \\ \vdots \\ \\ E_d[Y] \end{bmatrix}$$

$$\left(\begin{bmatrix} \mathbf{I} & 0 & \dots & \dots & 0 \\ \mathbf{E_2} & \mathbf{I} & 0 & \dots & 0 \\ \mathbf{E_3} & \mathbf{E_3} & \ddots & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ \mathbf{E_d} & \mathbf{E_d} & \dots & \dots & \mathbf{I} \end{bmatrix} + \begin{bmatrix} 0 & \mathbf{E_1} & \dots & \dots & \mathbf{E_1} \\ 0 & 0 & \mathbf{E_2} & \dots & \mathbf{E_2} \\ 0 & 0 & \ddots & & \mathbf{E_3} \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 0 \end{bmatrix} \right) \begin{bmatrix} m_1(X_1) \\ m_2(X_2) \\ \vdots \\ \vdots \\ m_d(X_d) \end{bmatrix} = \begin{bmatrix} E_1[Y] \\ \vdots \\ \\ E_d[Y] \end{bmatrix}$$

$$\begin{aligned} (L+U) \circ \mathbb{M}(X) &= E \circ Y \\ L \circ \mathbb{M}(X) &= E \circ Y - U \circ \mathbb{M}(X) \end{aligned} \tag{1.9}$$

l'idée « Gauss-Seidel » est alors de pouvoir approcher la solution $\mathbb{M}(X)$ de la façon suivante :

$$\mathbb{M}(X)^{(n+1)} = L^{-1}\left(E \circ Y - U \circ \mathbb{M}(X)^{(n)}\right) \tag{1.10}$$

Attention à la rigueur dans la formule présentée ici, avant de pouvoir tout formuler comme on en a l'habitude dans le cadre de matrices réelles, rappelons que ce qu'on a défini juste avant sont des matrices dont les éléments sont des opérateurs (car les lois de l'algèbre des matrices s'appliquent à la reformulation de notre problème). On se sert ici de la formule uniquement en guise d'outil de compréhension intuitive de l'algorithme de Backfitting, on ne prétend se servir autrement de cette formule.

Si l'on dispose d'un lissage linéaire : comme le lissage à noyaux de Nadaraya-Watson alors :

$$\begin{array}{ccccccc} \mathbb{M}(X)^{(n+1)} & = & L^{-1}[& E \circ Y - & U \circ & \mathbb{M}(X)^{(n)} \,] \\ \downarrow & & \downarrow & \downarrow & \downarrow & \downarrow O_i \text{ observé} \\ \widehat{\mathbb{M}}(x_{obs})^{(n+1)} & = & \widehat{L}^{-1}_{[NW]}[& \widehat{E}_{[NW]} \circ Y - & \widehat{U}_{[NW]} \circ & \widehat{\mathbb{M}}(x_{obs})^{(n)} \,] \end{array} \tag{1.11}$$

où les espérances conditionnelles E_k ont été remplacées par les lisseurs linéaires de Nadaraya-Watson et la variable aléatoire X par ses réalisations observées : $x_{obs} = (x_k^{[i]} : i \in [\![1\,,\,n]\!], k \in [\![1\,,\,d]\!])$.

1.1.1 ☐ C⟩ Algorithme du Backfitting

L'algorithme du Backfitting est une méthode d'approximation de solution de système linéaire utilisé dans l'estimation des modèles additifs. Des équations conditionnelles d'estimation (1.7), on corrige itérativement une première grossière estimation de la solution sur chaque composante que l'on utilise immédiatement pour estimer la composante suivante, motivé par la section précédente :

Algorithm 1: backfitting du modèle additif : estimation des fonctions m_k du modèle $\mathbb{E}[Y \mid X = x] = m(x) = \sum_{k=1}^{d} m_k(x_k)$

Data: observations : $O_i \underset{\text{déf}}{\equiv} (x_k, y)_{k\in[\![1\,,\,d]\!]}^{[i]}$

Input: Initial Guess : $(\forall k \in [\![1\,,\,d]\!])\widehat{m}_k^{[r]} \underset{\text{default}}{=} 0$

tolérance pour la convergence : ε

métrique d'évaluation de la différence entre deux fonctions pour les itérations : d à valeurs dans $\mathbb{R}_+$

Result: termes composant l'estimation de la fonction $m : x \mapsto \sum_k m_k(x_k)$:

$$\forall k \in [\![1\,,\,d]\!] \quad \widehat{m}_k : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x_k & \longmapsto & \widehat{m}_k(x_k) \end{array}$$

$r \leftarrow 0$

`# Jusqu'à la convergence : la différence entre deux étapes est faible`

while $d\left(\widehat{m}_k^{[r]}, \widehat{m}_k^{[r-1]}\right) > \varepsilon$ **do**

  for $k \in [\![1\,,\,d]\!]$ **do**

    `# 1. Backfitting`

    $\widehat{m}_k^{[r]} \leftarrow \mathrm{NW}\left(y_i - \left[\sum_{p<k} \widehat{m}_k^{[r]}(x_k^{[i]}) + \sum_{p>k} \widehat{m}_k^{[r-1]}(x_k^{[i]})\right] \,:\, i \in [\![1\,,\,n]\!]\right)$

    `#` $m_k^{[r]}$ `: vient juste d'être estimée`

    `#` $m_k^{[r-1]}$ `: estimation pas encore executée`

    `# 2. Centrage pour l'identifiabilité`

    $\widehat{m}_k \leftarrow x \mapsto \widehat{m}_k(x) - \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_k\left(x_k^{[i]}\right)$

  end

end

On recentre à chaque fois car pour avoir l'identifiabilité on a besoin que $\mathbb{E}_{p_k\cdot\lambda}[m_k(X_k)] = 0$, cependant en appliquant empiriquement l'algorithme itérativement, rien ne guarantit de garder les estimateurs centrés, c'est pourquoi on retire la moyenne empirique (qui est un estimateur de l'espérance) à chaque étape.

On sait désormais quel algorithme on souhaite utiliser pour estimer un modèle additif, comment peut-on être sûr que cet algorithme converge bien vers la solution que l'on espère approcher ? Étant donné que l'on traîte des observations aléatoires, l'algorithme converge-t-il toujours vers la solution que l'on souhaite ?

1.1.1 ☐ D⟩ Convergence de l'algorithme du Backfitting

L'ensemble de cette section se base sur l'article « The existence and asymptotic properties of a backfitting projection algorithm under weak conditions » de Mammen, Linton et Nielson (1999), pour plus de détails on pourra se référer à (17)

L'algorithme de backfitting a été utilisé en pratique (« avec succès » (17)) avant même la preuve de garantie de convergence vers la quantité qui nous intéresse : la fonction de régression m. L'algorithme de Backfitting s'avère en partie compliqué pour la démonstration de convergence vers la solution souhaitée car il est par nature un algorithme itératif sur l'approximation précédente :

$$\widehat{m}^{[r]} = T\left(\widehat{m}^{[r-1]}\right) = \cdots = T^r\left(\widehat{m}^{[0]}\right) \tag{1.12}$$

où T est une itération de l'algorithme de Backfitting.

Il existe différentes stratégies pour prouver la convergence de tels algorithmes. Une des stratégies classiques pour ce genre d'algorithmes itératifs est de reformuler le problème en un problème du point fixe, c'est à dire trouver un opérateur $\tilde{T} = g(T)$ impliquant l'algorithme du Backfitting qui soit continu et contractant (i.e. $|||\tilde{T}||| < 1$) qui convergerait alors vers la solution que l'on recherche. [1]

Enfin en tant que statisticien, on est intéressés par les propriétés asymptotiques d'un tel estimateur. Le point de vue proposé par Mammen, Linton et Nielson permet une interprétation géométrique de l'algorithme de Backfitting qui débloque à la fois la démonstration de la convergence et l'obtention des propriétés asymptotiques de la solution de l'algorithme du Backfitting. Nous allons exposer ici les grandes idées de ce point de vue et des résultats importants et du chemin menant à leur démonstration sans trop rentrer dans les détails pour rester concis. Le lecteur pourra toujours se référer à (17) pour les démonstrations complètes.

1. Estimation comme une projection dans un espace de fonctions

Le point de vue proposé est de voir le modèle additif comme une projection de la fonction de régression m, élément d'un espace de fonction très général, sur le sous-espace des fonctions additives. Puis l'estimation d'une telle fonction additive est elle aussi obtenue comme projection sur un sous-espace que l'on qualifierait grossièrement de « sous-espace empirique ».

modèle	espace associé
$Y_i = m_i(X) + \varepsilon$	$\mathcal{F} = \left\{ \left(f^{[i]} : \mathbb{R}^d \to \mathbb{R},\, i \in [\![1\,,\, n]\!]\right) \right\}$
$Y = m(X) + \varepsilon$	$\mathcal{F}_{\underset{\forall i}{loi}} = \mathcal{F}_{\perp\!\!\!\perp i} = \left\{(f^{[i]})_{1,n} : \mathbb{R}^d \to \mathbb{R} : f^{[i]} \perp\!\!\!\perp i\right\}$
$Y_i = \sum_k m_{i,k}(X_k) + \varepsilon$	$\mathcal{F}_{add} = \left\{ \left(f^{[i]} : \mathbb{R}^d \to \mathbb{R}\right)_{1,n} : \forall i \in [\![1\,,\, n]\!] \right.$ $\left. \exists (g_{i,p})_{[\![1\,,\, n]\!]\times[\![1\,,\, d]\!]} \text{ tq } f^{[i]}(x) = \sum_{p=1}^{d} g_{i,p}(x_p) \right\}$
$Y = \sum_k m_k(X_k) + \varepsilon$	$\mathcal{F}_{\underset{\perp\!\!\!\perp i}{add}} = \mathcal{F}_{\perp\!\!\!\perp i} \cap \mathcal{F}_{add}$

Si l'idée de voir les différentes hypothèses du modèle et l'estimation comme des projections successives sur des sous-espaces, il ne s'agit pas de la projection de la géométrie usuelle de $\mathbb{L}^2$ que l'on a l'habitude manipuler. Afin de pouvoir interpréter le modèle et l'estimation comme des projections, il y a besoin de considérer une géométrie particulière donnée par le produit scalaire suivant :

$$\text{en posant : } f_{add}^{[\text{local}]}(\,x \mid i\,) \underset{\text{déf}}{\equiv} f_0^{[i]}(x) + \sum_{k=1}^{d} \underbrace{f_k^{[i]}(x)}_{\text{additif}} \cdot \underbrace{\frac{x_k - X_k^{[i]}}{h}}_{\text{local}} \tag{1.13}$$

1. on reste ici volontairement flou vis-à-vis de la norme considérée dont dépend évidemment la convergence de notre estimateur, nous y reviendrons plus loin.

$$\begin{array}{rccl} & \left[\mathbb{L}^2 \cap \mathcal{F}_{add}\right]^2 & \longrightarrow & \mathbb{R} \\ \langle \cdot \mid \cdot \rangle_* : & (f,g) & \longmapsto & \displaystyle\int \frac{1}{n}\sum_{i=1}^{n} \left[f_{add}^{[\text{local}]}(x \mid i) \cdot g_{add}^{[\text{local}]}(x \mid i) \cdot \underbrace{\prod_{j=1}^{d} \frac{1}{h} K\left(\frac{X_j^{[i]} - x_j}{h}\right)}_{\text{local : dimension } d} \right] d\lambda_d(x) \end{array} \tag{1.14}$$

avec la norme $\|\cdot\|_*$ qui lui est associée ($\|f\|_*^2 = \langle f \mid f \rangle_*$)

La considération d'un tel produit scalaire provient de l'interprétation des observations comme des fonctions :
en observant $(X_k^{[i]}, Y^{[i]} : k \in [\![1\,,\,d]\!], i \in [\![1\,,\,n]\!])$, on peut voir $Y^{[i]}$ comme une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$, même si elle doit être constante.

$$Y^{[i]} = \begin{bmatrix} \vdots \\ Y_k^{[i]} \\ \vdots \end{bmatrix} = f^{[i]} = \begin{bmatrix} \vdots \\ Y_k^{[i]} : \mathbb{R} \to \mathbb{R} \\ \vdots \end{bmatrix} \tag{1.15}$$

Ce que montre (17), c'est que l'algorithme du backfitting peut se ramener à un opérateur impliquant la projection sur les sous-espaces empiriques qui est **contractant** du point de vue de la norme opérateur induite par $\|\cdot\|_*$. [2]

Alors, **à condition que** $\int \mathbf{K} d\lambda = \mathbf{1}$ et sous hypothèses de dépendance faible [3] et de régularité du noyau et de la fonction cible, l'estimation du modèle par l'algorithme du Backfitting est une projection au sens de la géométrie induite par $\langle \cdot \mid \cdot \rangle_*$, les projecteurs existent et sont bien définis. Les vitesses de convergence de dimension 1 sont bien atteignables.

1.1.2 Points clés

- ☐ Les modèles additifs combattent le fléau de la dimension qui touche en particulier les modèles non paramétriques en se ramenant à des vitesses de convergence de la dimension 1
- ☐ pour estimer les modèles additifs on peut utiliser l'algorithme du Backfitting qui est un algorithme d'approximation de solution itératif
- ☐ On peut voir l'estimation comme une procédure de projection sur un sous-espace de fonctions **selon une géométrie très particulière**

---

2. ⚠ **simplification :** toujours par soucis de concision, seule l'idée générale est ici restituée. C'est en réalité un peu plus compliqué : on montre dans un premier temps que l'opérateur travaillant sur les quantités que l'on souhaite estimer T est contractant pour la norme $\|\cdot\|_*$ sur $\mathcal{F}_{add}$, puis on définit un opérateur « empirique » travaillant sur les données que l'on manipule (en tant que praticien) $\widehat{T}$, enfin on montre que l'opérateur « empirique » converge vers l'opérateur « idéal » : $|||\widehat{T} - T|||_{\mathbb{L}(\hat{p}\cdot\lambda)} = o_P(1)$ pour en déduire que l'on contrôle la norme opérateur « empirique » si l'on contrôle la norme opérateur « idéale ». La probabilité que la solution du Backfitting s'écarte de la solution de notre problème statistique est alors pleinement contrôlée par le paramètre de contrôle de la norme opérateur de T et $\widehat{T}$.

3. La notion de dépendance faible utilisée est la notion dite de « strong alpha-mixing », on pourra se référer à (2) pour une introduction au sujet.

- ☐ il convient de voir alors nos données, elles aussi, comme des fonctions même si constantes
- ☐ $\longrightarrow$ on adapte la géométrie de l'espace au problème
- ☐ la convergence d'algorithmes itératifs peut se démontrer au moyen d'un théorème du point fixe

L'utilisation d'estimateurs à noyaux indique bien que parmis les hypothèses importantes de la méthodologie utilisée par Mammen, Linton et Nielson figure l'admission d'une loi à densité pour les covariables : on travaille avec les géométries de $\mathbb{L}^2(p_X \cdot \lambda), \mathbb{L}^2(\widehat{p}_X \cdot \lambda) \dots$
Si cela ne semble pas être une hypothèse coûteuse dans le cadre des données réelles que l'on observe au quotidien, cela risque de poser problème dans le cadre de lois pour des objets en dimension infinie (comme les données fonctionnelles) comme il sera traîté au chapitre 2 « Modèles additifs : vers le cadre fonctionnel ».

Il est à noter que le point de vue de « projections successives » de l'algorithme du Backfitting n'est pas simplement une interprétation esthétique de l'algorithme, elle est à ce jour **la** méthode qui permet d'obtenir les résultats de convergence souhaitables de cet algorithme. C'est pourquoi il est important lorsque l'on applique l'algorithme du Backfitting de vérifier les hypothèses qui permettent d'avoir le point de vue « projection ». Cela constitue d'ailleurs une difficulté qui devra être surmontée pour l'estimation de données imparfaites, que nous allons détailler désormais.

1.2 Données Imparfaites

On dispose du modèle semi-paramétrique des données suivant :

$$y = \sum_{k=1}^{d} \beta_k \cdot x_k + m(z) \ + \varepsilon \tag{1.16}$$

Il s'agit du modèle « partiellement linéaire », où les résidus de ce qui peut être modélisé par une partie paramétrique linéaire sont déterminés non paramétriquement afin de bénéficier au maximum des vitesses de convergence du paramétrique tout en gardant une partie de la flexibilité sur le modèle offerte par le non paramétrique.

L'idée c'est que si l'on dispose de covariables $X = (\underbrace{X_1, \dots, X_p}_{X_{lin}}, \underbrace{X_{p+1}, \dots, X_d}_{X_{\neg lin}})$ alors au lieu d'estimer :

$$\widehat{m}(X) = m(X) + O_p\left(n^{-\frac{k}{2k+d}} \right)$$

On estime le modèle :

$$m(X) = X_{lin}^T \beta + g(X_{\neg lin})$$

Ce qui nous permet ainsi d'obtenir les vitesses de convergence suivantes :

$$\begin{aligned} \widehat{m}(X) &= X_{lin}^T \widehat{\beta} + O_p\left(n^{-1/2} \right) + \widehat{g}(X_{\neg lin}) + O_p\left(n^{-\frac{k}{2k+(d-p)}} \right) \\ &= m(X) + O_p\left(\underbrace{n^{-\frac{k}{2k+(d-p)}}}_{\text{pire vitesse}} \right) \end{aligned}$$

Limitant ainsi l'exposition au fléau de la dimension.

Comme mentionné précédemment, pour limiter notre exposition au fléau de la dimension dû à la partie non paramétrique du modèle, on rajoute l'hypothèse d'additivité de la partie non paramétrique :

$$y = \sum_{k=1}^{d} \beta_k \cdot x_k + \sum_{k=1}^{d} m_k(z_k) \; + \varepsilon \tag{1.17}$$

Idéalement, un statisticien travaille sur les observations suivantes :

$$\begin{aligned} O^{[ideal]} &= \left(O^{[ideal]}{}_i \right)_{1,n} \\ \text{avec } O_i^{[ideal]} &= \left(\text{—}x_k^{[i]}\text{—}, \text{—}z_k^{[i]}\text{—}, y^{[i]} \right) \end{aligned} \tag{1.18}$$

Malheureusement il est fréquent que les covariables utilisées pour la régression soient elles-mêmes bruitées en tant que données provenant de capteur, ou recensées, sondées par un humain, ... Ce que le praticien observe en réalité est plutôt de la forme :

$$\begin{aligned} O &= \left(O_i \right)_{1,n} \\ \text{avec } O_i &= \left(\text{—}(x_k^{[i]} + u_k^{[i]})\text{—}, \text{—}(z_k^{[i]} + v_k^{[i]})\text{—}, y^{[i]} \right) \end{aligned} \tag{1.19}$$

où $u_k \perp\!\!\!\perp x_k$(et $v_k, z_k, \dots$) $\qquad v_k \perp\!\!\!\perp z_k$ (et $u_k, x_k, \dots$)
sont des erreurs liées à l'observation (gaussiennes par exemple)

Cela complique considérablement l'estimation : que ce soit pour le paramètre β ou les fonctions m_k.

1.2.1 Erreur des covariables et déconvolution

Avant de vouloir résoudre notre problème du modèle partiellement linéaire, nous allons d'abord nous intéresser à l'estimation non paramétrique lorsque l'on dispose de données avec erreur de mesure dans les covariables. Considérons donc le modèle de données suivant :

$$\begin{cases} Y = m(Z) + \varepsilon \\ \varepsilon \perp\!\!\!\perp Z \\ \mathbb{E}\left[\varepsilon \,|\, Z \right] = 0 \end{cases} \tag{1.20}$$

avec le modèle d'observation suivant :

$$\begin{cases} \left(Z^{*[i]}, Y^{[i]} \right)_{i \in [\![1\,,\,n]\!]} \\ Z^{*[i]} = Z^{[i]} + V^{[i]} \\ V \perp\!\!\!\perp Z \end{cases} \tag{1.21}$$

La méthode de lissage non paramétrique à noyaux classique ne peut plus marcher car le bruit dans la covariable perturbe l'information et un moyennage local introduirait un biais non négligeable :

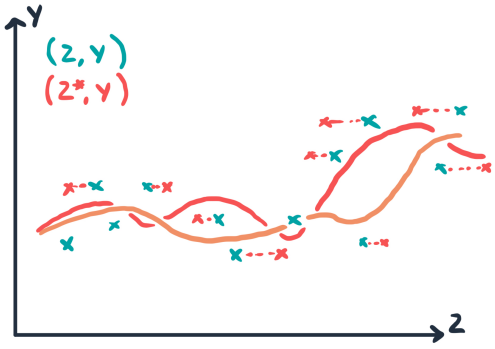

FIGURE 1.1 – Lissage non paramétrique à noyaux avec erreur de mesure dans les covariables sans prise en compte de l'erreur.

L'estimateur de Nadaraya-Watson de la fonction m est un estimateur que l'on pourrait interpréter comme un estimateur plug-in utilisant l'estimation de la densité conditionnelle de Y sachant Z à partir de l'estimateur de la densité de Z et de la densité jointe de (Z, Y). Sauf que la densité que nous observons est celle de $(Z + V, Y)$ et non celle de (Z, Y). L'erreur V étant indépendante de Z, la densité de $Z + V$ est la convolution des deux densités :

$$f_{Z+V} = \left[f_Z * f_V \right] \tag{1.22}$$

La convolution de fonctions intégrables est une opération qui a été étudiée massivement dû à ses nombreuses applications. Elle se comporte algébriquement bien dans le monde fréquentiel : en effet la transformée de Fourier d'une convolution et le produit des transformées de Fourier

$$\mathcal{F}\left[f * g \right] = \mathcal{F}[f] \times \mathcal{F}[g] \tag{1.23}$$

On aimerait à partir de nos données observées avec erreur de mesure dans les covariables, utiliser un lissage à noyau qui ne prendrait en compte que l'information dans les covariables. En d'autre termes juste prendre l'information de f_Z dans f_{Z+V}. En se ramenant dans le monde fréquentiel, on remarque que :

$$\begin{aligned} \mathcal{F}\left[f_{Z+V}\right] &= \mathcal{F}\left[f_Z\right] \times \mathcal{F}\left[f_V\right] \\ \mathcal{F}\left[f_Z\right] &= \frac{\mathcal{F}\left[f_{Z+V}\right]}{\mathcal{F}\left[f_V\right]} \qquad (1.24) \\ \mathcal{F}^{-1}\left(\mathcal{F}\left[f_Z\right]\right) &= \mathcal{F}^{-1}\left(\frac{\mathcal{F}\left[f_{Z+V}\right]}{\mathcal{F}\left[f_V\right]}\right) \\ f_Z &= \mathcal{F}^{-1}\left(\frac{\mathcal{F}\left[f_{Z+V}\right]}{\mathcal{F}\left[f_V\right]}\right) \qquad (1.25) \end{aligned}$$

C'est ce qui va venir motiver l'ensemble de la méthodlogie qui sera exposée dans la suite de cette section : on appelle ainsi naturellement cette opération « la déconvolution ». Si la transformée de Fourier semble être un outil obscur pour le statisticien, il n'en est rien. Le statisticien a probablement utilisé à nombreuses reprises la transformée de Fourier sans le savoir.

Définition 2 (Transformée de Fourier) Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction intégrable. On appelle transformée de Fourier de f la fonction φ_f définie par :

$$\varphi_f : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{C} \\ \omega & \longmapsto & \int_{\mathbb{R}} e^{-i\omega x} f(x) dx \end{array}$$

Cette définition devrait sembler familière :

Définition 3 (Fonction caractéristique) Soit X une variable aléatoire réelle. On appelle fonction caractéristique de X la fonction ϕ_X définie par :

$$\phi_X : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{C} \\ \omega & \longmapsto & \mathbb{E}\left[e^{i\omega X}\right] \end{array}$$

Proposition (Fonction caractéristique d'une variable aléatoire réelle à densité)
Soit X une variable aléatoire réelle à densité. Alors la fonction caractéristique de X est la « transformée de Fourier » de la densité de X.

$$\phi_X(\omega) = \int_{\mathbb{R}} e^{i\omega x} f_X(x) dx$$

Comment ça « transformée de Fourier » ?

Il est fréquent de voir la notion de fonction caractéristique appelée transformée de Fourier en statistique, sauf que cela est **formellement faux**. La fonction caractéristique est en réalité la transformée de Fourier **inverse** :

$$\mathcal{F}^{-1}\left[\varphi_f\right] : x \mapsto \frac{1}{2\pi} \int \varphi_f(\omega) e^{+i\omega x} d\omega = f(x)$$

Ce qui peut mener à de multiples confusions notamment dans les articles de statistiques où le terme transformée de Fourier est utilisé pour mentionner l'application :

$$\mathcal{F}_{\text{article stat}} : \begin{array}{ccc} \mathbb{L}^1 & \longrightarrow & \mathbb{L}^1 \\ f & \longmapsto & \omega \mapsto \int_{\mathbb{R}} e^{+i\omega x} f(x) dx \end{array}$$

Qui est une « transformée de Fourier » renversée (transformée de Fourier en parcourant selon $-t$ au lieu de $+t$). On peut désormais revenir à notre problème d'estimation.

Par la suite, de nombreuses notations liées aux noyau seront utilisées, afin de ne pas surcharger le texte, elles sont toutes répertoriées dans le répertoire de Notations au début du rapport.

1.2.1 ☐ A⟩ Un noyau candidat de déconvolution

On souhaite trouver un noyau $\tilde{K}$ qui nous permette à partir de nos données contaminées Z^* d'estimer la densité de Z avec le même biais que l'estimateur inatteignable en réalité de la densité avec la connaissance parfaite des données sans erreur. Il s'agit du meilleur estimateur que l'on puisse espérer, et on l'appelle donc « l'estimateur oracle ».

$$\begin{array}{rcll} \hat{p}_z^{[\text{oracle}]}(z) = \frac{1}{n}\sum_i K_h(z - Z^{[i]}) & \rightarrow & \hat{p}_z = p_z + & \mathbb{B}\left(\hat{p}_z^{[\text{oracle}]}\right) \\ & & & \| ? \\ \tilde{p}_z(z) = \frac{1}{n}\sum_i \tilde{K}_h(z - Z^{*[i]}) & \rightarrow & \hat{p}_z = p_z + & \mathbb{B}\left(\tilde{p}_z\right) \end{array}$$

Pour qu'une telle estimation soit possible, il suffit que le noyau $\tilde{K}$ ait la propriété dite de « scoring non biaisé » :

$$\mathbb{E}\left[\tilde{K}_h(z - Z^*) \,|\, Z\right] = K_h(z - Z) \tag{1.26}$$

Il se trouve qu'il existe un noyau qui possède cette propriété, il s'agit du noyau de déconvolution :

$$\tilde{K}_h(z) = \frac{1}{2\pi} \int e^{-i\omega z} \frac{\phi_K(h\omega)}{\phi_V(\omega)} d\omega \tag{1.27}$$
$$= \mathcal{F}_{\text{stat}}^{-1}\left(\frac{\mathcal{F}_{\text{stat}}[K]}{\mathcal{F}_{\text{stat}}[f_V]} \right)$$

On dispose désormais d'un noyau pour déconvoluer l'erreur des covariables. Nous souhaitons alors l'utiliser pour les estimations de fonctions de régression. Et là, c'est le drame. Parceque le noyau de déconvolution que nous venons de considérer ne possède pas la propriété de normalisation. Comme nous l'avons mentionné dans la section 1.1.2, l'algorithme du Backfitting nécessite que les noyaux utilisés soient normalisés : $\int K \, d\lambda = 1$ pour pouvoir utiliser le backfitting comme une projection. Nous allons donc devoir trouver un noyau qui possède les deux propriétés : scoring non biaisé et normalisation.

1.2.1 ☐ B⟩ Un noyau candidat de Backfitting

Pour obtenir la propriété de normalisation à partir d'un noyau de lissage il suffit simplement de considérer le noyau suivant :

$$K_{h,\|\cdot\|}^{[x_0]} : x \mapsto \frac{K_h^{[x_0]}(x)}{\int K_h^{[x_0]}(u) du} \tag{1.28}$$

On voudrait alors définir un noyau de déconvolution normalisé (fenêtré et centré en x_0) de la façon naturelle suivante :

$$\tilde{K}_{h,\|\cdot\|}^{[x_0]} : x \mapsto \frac{\tilde{K}_h^{[x_0]}(x)}{\int \tilde{K}_h^{[x_0]}(u) du} \tag{1.29}$$

💣Et là, une fois de plus, le ciel nous tombe sur la tête. Puisque considérer un tel noyau nous fait désormais perdre la propriété de scoring non biaisé qui est essentielle pour obtenir

$$\tilde{p}_Z - p_Z \simeq \mathbb{B} \simeq \hat{p}_Z^{[\text{oracle}]} - p_Z.$$

Pour avoir un noyau de déconvolution qui nous permette d'utiliser l'algorithme du Backfitting, il faut donc le trouver un noyau de manière plus astucieuse.

1.2.1 ☐ C⟩ Un noyau normalisé de déconvolution

Existe-t-il un noyau qui soit à la fois normalisé et qui possède la propriété de scoring non biaisé ?

La réponse est oui, le noyau proposé par Han, Park & Byeong (2018) (8) possède les deux propriétés que nous recherchons. De façon informelle, il s'agit de :

$$\mathcal{F}_{\text{stat}}^{-1}\left[\frac{\mathcal{F}_{\text{stat}}\left(K_{h,\|\cdot\|}^{[\bullet]} * K_h \right)}{\mathcal{F}_{\text{stat}}(p_V)} \right] = \mathcal{F}_{\text{stat}}^{-1}\left[\frac{\mathcal{F}_{\text{stat}}\left(K_{h,\|\cdot\|}^{[\bullet]} \right) \mathcal{F}_{\text{stat}}(K_h)}{\mathcal{F}_{\text{stat}}(p_V)} \right]$$

On peut essayer de se convaincre intuitivment de cette proposition :

- ☐ $K_{h,\|\cdot\|}^{[\bullet]}$ « parcourt l'ensemble des observations bruitées $Z^{*[i]}$ » en laissant fixe la position z où l'on est sur la densité p_Z (on fait varier ici le point de centrage : c'est à dire le X dans $K(\frac{x-X}{h})$)
- ☐ K_h représente le poids affecté à un point spécifique.

La convolution représente une somme glissante, $\left[K_{h,\|\cdot\|}^{[\bullet]} * K_h\right]$ représente le fait d'ajouter avec un poids correspondant à la valeur de $K_{h,\|\cdot\|}^{[\bullet]}$ (qui est normalisé), les poids du lissage à noyaux autour de l'observation Z^*, et on fait cela pour toutes les observations.

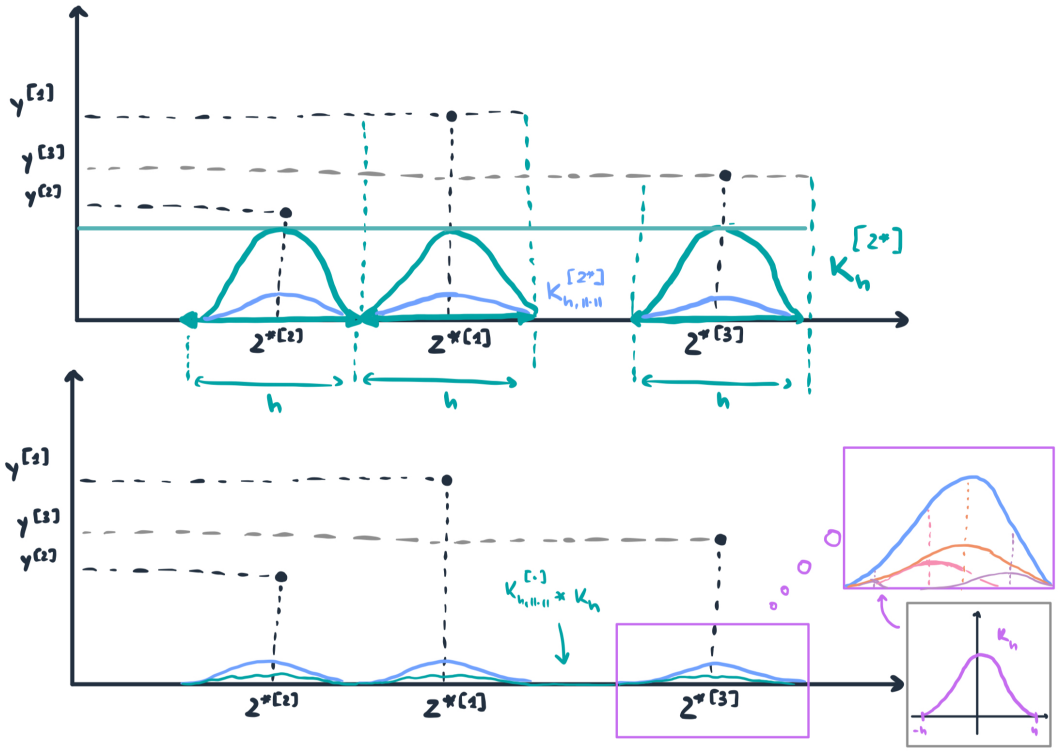

FIGURE 1.2 – Illustration de la convolution de $K_{h,\|\cdot\|}^{[\bullet]}$ et K_h

On obtient ainsi des poids normalisés sur toutes les observations qui se basent sur le noyau de lissage de base K_h que l'on a choisi.

De façon plus formelle :

$$\tilde{K}_h^\star(z, Z^*) = \frac{1}{2\pi h} \int_{\mathbb{R}} e^{-i\omega\left(\frac{z-Z^*}{h}\right)} \cdot \frac{\phi_{K_{h,\|\cdot\|}}^{[z]}(\omega\,;\,h)\phi_K(\omega)}{\phi_V(\omega)}\, d\omega \tag{1.30}$$

où $\phi_{K_{h,\|\cdot\|}}^{[z]}(\omega\,;\,h) = \int e^{i\omega\left(\frac{z-u}{h}\right)} K_{h,\|\cdot\|}^{[u]}(z)du$

1.2.1 ☐ D⟩ Propriétés asymptotiques

Le rapport de stage commençant à être long, nous ne détaillerons pas spécifiquement la méthodologie de cette section et nous renvoyons le lecteur à l'article de Han, Park & Byeong (2018) (8) pour plus de détails. Nous allons juste mentionner les résultats principaux. En supposant que l'on dispose d'un contrôle du comportement de la fonction caractéristique de l'erreur de mesure dans les covariables en s'éloignant de 0 :

$$
\begin{aligned}
|\phi_V| &= \underset{|t|\to\infty}{O}\left(|t|^{-\alpha}\right) \\
&\text{et} \\
|\phi'_V(t)| &= \underset{|t|\to\infty}{O}\left(|t|^{-(\alpha+1)}\right) \text{ pour un } \alpha \geq 0
\end{aligned}
\tag{1.31}
$$

Le noyau proposé permet d'obtenir une estimation des fonctions m_k dans le modèle $y = \sum_{k=1}^{d} m_k(z_k) + \varepsilon$ avec une vitesse de convergence optimale pour le problème de déconvolution en dimension 1 sur l'intérieur du support de la densité p_Z, $I_h = \prod_{k=1}^{d}[2h_k, 1-2h_k]$, en choisissant le bon h (par validation croisée). Par exemple si $\alpha < 1/2$, la sélection du h optimal permet d'obtenir avec l'estimateur du backfitting :

$$
\sup_{z\in I_h} |\hat{m}_k^\star(z) - m_k(z)| \underset{\alpha<1/2}{=} O_{\mathbb{P}}\left(n^{-\frac{2}{5+2\alpha}}\sqrt{\log n}\right) \tag{1.32}
$$

$$
\sup_{z\in [0,1]} |\hat{m}_k^\star(z) - m_k(z)| \underset{\alpha<1/2}{=} O_{\mathbb{P}}\left(n^{-\frac{2}{5+2\alpha}}\right) \tag{1.33}
$$

On peut alors utiliser le noyau de déconvolution normalisé pour l'estimation des fonctions de régression par un estimateur de Nadaraya-Watson dans le cadre d'un modèle additif en utilisant l'algorithme du Backfitting.

1.2.2 Estimation déconvolutive du modèle partiellement linéaire

Nous possédons désormais tous les outils nécessaires pour résoudre le problème de déconvolution dans le cadre du modèle partiellement linéaire.

$$
Y = \beta^T X + \sum_{k=1}^{d} m_k(Z_k) + \varepsilon \tag{1.34}
$$

On souhaite dans un premier temps estimer le paramètre β et dans un second temps estimer les fonctions m_k. Pour cela on peut éliminer les fonctions m_k en soustrayant à (1.34) l'expression conditionnée selon Z en se restrayant aux fonctions additives (on compose par (1.34) par $P_{\mathbb{L}^2\cap\mathcal{A}} \circ \mathbb{E}\left[\bullet \mid Z\right]$) :

$$
\left(Y - \xi(Z)\right) = \beta^T \left(X - \eta(Z)\right) + \varepsilon \tag{1.35}
$$

Où $\eta(Z) = P_{\mathbb{L}^2\cap\mathcal{A}} \circ \mathbb{E}\left[X \mid Z\right] = \sum_k \eta_k(Z_k)$ et $\xi(Z) = P_{\mathbb{L}^2\cap\mathcal{A}} \circ \mathbb{E}\left[Y \mid Z\right] = \sum_k \xi_k(Z_k)$ (ce qui permet de les estimer avec un algorithme de backfitting avec les outils de déconvolution précédents).
Nous obtenons alors :

$$
\beta = \underbrace{\mathbb{E}\left[\left(X - \eta(Z)\right)\left(X - \eta(Z)\right)^T\right]}_{D}^{-1} \underbrace{\mathbb{E}\left[\left(X - \eta(Z)\right)\left(Y - \xi(Z)\right)\right]}_{c} \tag{1.36}
$$

Sauf que nous n'observons ni X ni Z mais $X^* = X + U$ et $Z^* = Z + V$. Regardons naïvement ce que donne D et c dans ce cas :

$$
\begin{aligned}
D^*_{|Z} &= \mathbb{E}\left[\left(X^* - \eta(Z)\right)\left(X^* - \eta(Z)\right)^T\right] && (1.37)\\
&= \mathbb{E}\left[\left(X + U - \eta(Z)\right)\left(X + U - \eta(Z)\right)^T\right] && (1.38)\\
&= \mathbb{E}\left[\left(X - \eta(Z)\right)\left(X - \eta(Z)\right)^T\right] + \mathbb{E}\left[UU^T\right] + \underbrace{\mathbb{E}\left[\left(X - \eta(Z)\right)U^T + U\left(X - \eta(Z)\right)^T\right]}_{=0 \text{ car U } \perp\!\!\!\perp X,Z} && (1.39)\\
&= D + \Sigma_U && (1.40)
\end{aligned}
$$

On peut alors en déduire un estimateur de β en utilisant la formule des probabilités totales $\mathbb{E}[\,\_\_\,] = \mathbb{E}[\,\mathbb{E}[\,\_\_\,|\,Z\,]\,]$ que l'on peut estimer grâce aux outils de déconvolution développés précedemment et en utilisant la propriété de scoring non biaisé du noyau normalisé de déconvolution $\tilde{K}^{\star}$ [4] :

$$\begin{aligned} D &= \mathbb{E}\left[\mathbb{E}\left[\,(X^* - \eta(Z))\,(X^* - \eta(Z))^T \mid Z\right] - \Sigma_U\right] \\ &\underset{Z\sim p_Z\cdot\lambda_d}{=} \int \mathbb{E}\left[\,(X^* - \eta(z))\,(X^* - \eta(z))^T \mid Z\right] - \Sigma_U \, p_Z(z)\, d\lambda_d(z) \\ \widehat{D} &= \int \frac{1}{n}\sum_{i=1}^{n}\left(X^{*[i]} - \hat{\eta}^{\star}(z)\right)\left(X^{*[i]} - \hat{\eta}^{\star}(z)\right)^T \prod_{k=1}^{d} \tilde{K}_h^{\star}(z_k, Z_k^{*[i]}) d\lambda_d(z) - \Sigma_U \end{aligned} \tag{1.41}$$

En procédant de la même manière pour c on obtient :

$$\widehat{c} = \int \frac{1}{n}\sum_{i=1}^{n}\left(X^{*[i]} - \hat{\eta}^{\star}(z)\right)\left(Y^{[i]} - \hat{\xi}^{\star}(z)\right) \prod_{k=1}^{d} \tilde{K}_h^{\star}(z_k, Z_k^{*[i]}) d\lambda_d(z) \tag{1.42}$$

On peut alors estimer β par $\widehat{\beta} = \widehat{D}^{-1}\widehat{c}$. On peut alors estimer simplement les fonctions m_k en utilisant un algorithme de backfitting avec les outils de déconvolution selon $Y - \widehat{\beta}^T X = \sum_k m_k(Z_k) + \varepsilon'$.

L'estimation du modèle partiellement linéaire permet d'obtenir de meilleures propriétés asymptotiques que le modèle additif grâce à la vitesse de convergence de la partie paramétrique. On pourra se référer à (8) pour les détails. L'estimation du modèle partiellement linéaire lorsque les covariables sont contaminées a été permise par le noyau normalisé de déconvolution sous les conditions de contrôler le « spectre »[5]de la distribution de la contamination et la de connaître la variance de la contamination de la variable paramétrique.

1.2.3 Points clés

- ☐ La contamination additive indépendante des covariables résulte en une distribution sous forme de convolution
- ☐ La convolution est une opération à la structure algébrique simple dans le domaine fréquentiel
- ☑ Pour se défaire de la contamination, on travaille donc dans le domaine fréquentiel par transformée de Fourier
- ☐ Définir un noyau qui possède les propriétés pour déconvoluer la distribution contaminée et être compatible avec l'algorithmique de Backfitting n'est pas trivial
- ☐ Une fois le noyau bien défini, on peut estimer le modèle partiellement linéaire contaminé en se ramenant systématiquement à $\mathbb{E}[\,\_\_\,|\,Z\,]$ projeté sur $\mathbb{L}^2 \cap \mathcal{A}$, que l'on sait désormais estimer
- ☑ les vitesses de convergences pour la fenêtre de lissage optimale sont optimales pour les problèmes semi-paramétriques impliquant de la déconvolution
- ☒ On dispose tout de même d'hypothèses de régularité $\mathcal{C}^2$ sur les fonctions m_k et demande de pouvoir contrôler le spectre fréquentiel de la contamination (hypothèse dont je ne connais pas la force en pratique sur des données réelles)

---

4. on pourra se rappeler l'expression (1.26) de la section 1.2.1 ☐ A⟩ : Un noyau candidat de déconvolution $\mathbb{E}\left[\tilde{K}_h(z - Z^*)\,|\,Z\right] = K_h(z - Z)$ nous assure que l'estimation de $\mathbb{E}[\,\hat{\eta}^{\star}(z^*)\,|\,Z\,]$ nous donnera bien une estimation pas plus biaisée de $\eta(z)$ que $\hat{\eta}^{[oracle]}(z)$

5. dans le sens de la transformée de Fourier

Chapitre 2

Modèles additifs : vers le cadre fonctionnel

Contents

2.1 Motivations

Les données fonctionnelles constituent un modèle intéressant par leur capacité à modéliser le fait que la relation entre une réponse et des paramètres (par exemple la pression d'un gaz parfait en fonction de la température ou la débit d'un fleuve en fonction de la position (x, y, z) dans l'espace) est elle même sujette à une loi. Un exemple serait la puissance électrique instantanée consommée par un ménage à un instant t de la journée, si chacun possède sa consommation propre, les ménages ont tendance à consommer plus ou moins de la même manière (moins dans la journée, plus le soir). Elles sont de plus en plus utilisées pour les données de capteur et dans l'industrie notamment. Si il existe déjà des outils pour faire de la régression fonctionnelle paramétrique, notamment avec le modèle linéaire, on voudrait pouvoir bénéficier de la flexibilité des modèles non paramétriques. Nous allons étudier dans cette section la généralisation des modèles additifs aux données fonctionnelles et les difficultés qui découlent d'une telle tentative.

2.2 Complications par rapport aux données à valeurs réelles

2.2.1 Une intégrale fonctionnelle ?

Quand notre objet est une fonction, qu'est ce que l'on entend par loi ? Comment définir l'espérance ? Peut-on intégrer selon une mesure dont la variable est une fonction ?

En voilà bien des questions qui semblent délicates aux premiers abords. Et-ce parceque l'on doit en effet se montrer précautionneux avec ce que l'on manipule. Les données fonctionnelles sont définies comme des variables aléatoires à valeurs dans un espace de fonction qui est en fait structuré comme un espace de Banach. Cet espace est d'autant plus effrayant qu'il est de dimension infinie.

Définition 4 (Données Fonctionnelles) On appelle donnée fonctionnelle une application $(\mathcal{F}, C)$ mesurable :

$$X : \begin{array}{ccc} (\Omega, \mathcal{F}, \mathbb{P}) & \longrightarrow & (\mathcal{C}^0(I, \mathbb{R}), \|\cdot\|_\infty, C, \mathbb{P}_X) \\ \omega & \longmapsto & x = X(\omega) : t \mapsto x(t) \end{array}$$

Dans $(\mathcal{C}^0(I, \mathbb{R}), \|\cdot\|_\infty, C, \mathbb{P}_X)$, quelle est précisément la tribu C ? Quelle mesure sur $\mathbb{P}_X$ pour nous permettre d'intégrer ? L'intégrale est elle une intégrale classique (Lebesgue) ?

On ne peut en réalité pas utiliser l'intégrale de Lebesgue. Pourquoi ? L'intégrale de Lebesgue est définie comme un supremum pris sur des fonctions étagées à valeurs réelles positives. Hors lorsque les applications sont à valeurs dans un espace de fonctions dont la dimension est infinie, la construction proposée précédemment d'intégrale n'est pas bien définie. Il faut donc redéfinir l'intégrale dans le cadre de fonctions : et donc dans un espace de Banach. Il s'agit de l'intégrale de Böchner :

L'idée de l'intégrale de Böchner est de reproduire les étapes de la création de l'intégrale de Lebesgue, en définissant des fonctions simples « équivalentes » [1] aux fonctions étagées dans le cadre réel. Si on peut construire une suite de fonctions simples qui tendent vers la fonction que l'on souhaite intégrer au sens de la topologie induite par la norme de l'espace de Banach considéré [2]. L'intégrale de Böchner de notre fonction est alors la limite des suites d'intégrales des fonctions simples qui tendent vers notre fonction.

Jeon (2020) (11), Park et Byeong formalisent et dérivent les propriétés de l'intégrale de Böchner de façon plus adaptée à la statistique, notamment en considérant des densités. Ils étendent désormais les résultats que l'on a exposés précédemment dans le cadre réel au cadre hilbertien (qui est un cas particulier des espaces de Banach en considérant la norme issue du produit scalaire).(4, 12, 16)

Les auteurs (Park, Byeong, Jeon, . . .) utilisent dans l'ensemble de leurs articles traitant des données hilbertiennes ou dans un espace de Böchner les notations :

- $\oplus$: addition interne
- $\odot$: multiplication externe
- $\ominus = \bullet \oplus [(-1) \odot \bullet]$: soustraction interne

pour mettre sur l'emphase d'avoir des opérations adaptées aux données, qui ne sont pas nécessairement les lois de composition usuelles. Toutefois, dans le reste de ce document, nous n'utiliserons que la notation $\odot$ en dehors des notations usuelles pour ne pas alourdir la lecture.

Afin de mieux comprendre l'opération d'intégration avec laquelle on travaille, on peut comparer la construction de l'intégrale de Lebesgue et celle de l'intégrale de Böchner via le schéma suivant :

---

1. pas au sens mathématique : dans le sens « jouant le rôle de »

2. En réalité c'est plutôt $\int \| f - s_n \|_{\mathbb{B}} \xrightarrow[n\to\infty]{} 0$. Mais l'idée est bien de regarder la convergence au sens de la norme de l'espace considéré

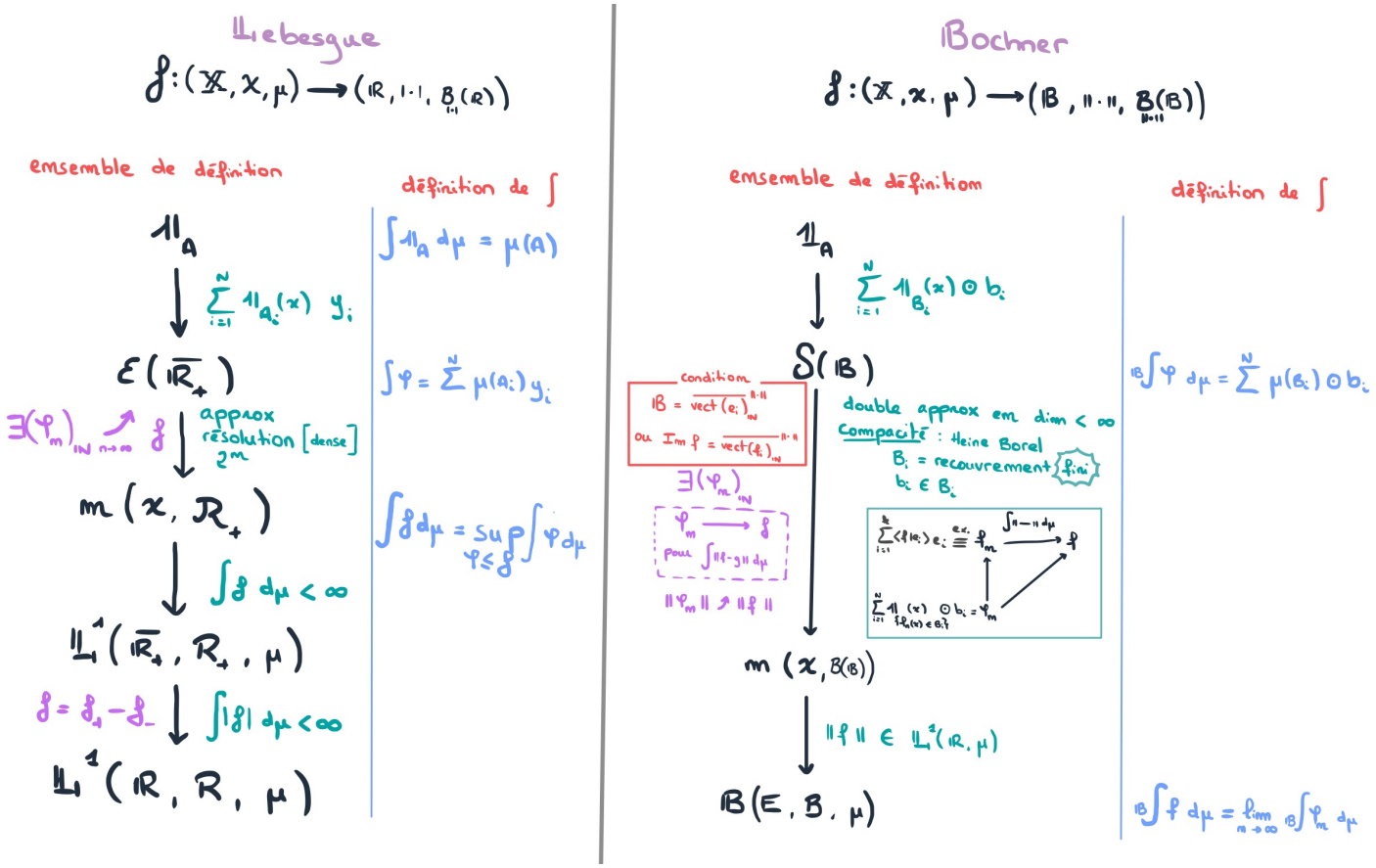

FIGURE 2.1 – Différences entre la construction de l'intégrale de Lebesgue et la construction de l'intégrale de Böchner

2.2.2 Existence de loi à densité

Le problème précédent a été résolu en construisant une intégrale sur des espaces vectoriels normés et en exploitant le fait que la norme est une application d'un espace vectoriel normé dans l'ensemble des réels positifs sur lequel on dispose de l'intégrale de Lebesgue. Pour étendre les travaux sur l'estimation du modèle additif lorsqu'il y a des erreurs de mesure sur les covariables on a désormais besoin de disposer de lois à densité sur nos espaces de fonction.

L'intégrale de Böchner a peu été utilisée en statistiques, c'est pourquoi (11) dérive et expose brièvement les résultats importants sur l'intégrale de Böchner pour les lois à densité, qui sont clés dans l'estimation du modèle additif par l'algorithme du Backfitting avec l'approche de Mammen (1999) (17), dont les auteurs suivent les pas pour l'étendre au cadre hilbertien. L'idée est de montrer qu'en supposant que $\mathbb{P}_X \ll \mu$ et $\mathbb{E}\left[\,\|f(X)\|\,\right] < \infty$, la quantité :

$$\int_{\mathbb{B}} f(x) \odot p_X(x)\, d\mu(x) = \mathbb{E}\left[\, f(X) \,\right] \tag{2.1}$$

et de même si $\mathbb{P}_{X,Z} \ll \mu \otimes \nu$ et $\mathbb{E}\left[\,\|f(X)\|\,\right] < \infty$ alors la quantité :

$$\int_{\mathbb{B}} f(x) \odot \frac{p_{X,Z}(x,z)}{p_Z(z)}\, d\mu(z) \overset{\text{version}}{=} \mathbb{E}\left[\, f(X) \,|\, Z \,\right] \tag{2.2}$$

C'est alors qu'un problème majeur survient :

Sur un espace de Banach quelconque, la propriété de Radon-Nikodym n'est pas vérifiée. Pire, elle ne peut être vérifiée par un ensemble contenant $\mathcal{C}^0([0,1])$

Expliquons un peu plus en détails l'insertion précédente.

2.2.2 ☐ A⟩ Théorème de Radon-Nikodym

Une des méthodes les plus naturelles pour créer de nouvelles mesures est de prendre une mesure que l'on connait déjà et de pondérer la répartition de la masse des objets qu'elle mesure. On appelle cela une mesure à « densité », très utilisée notamment en probabilité pour dériver de multiples « lois continues » à partir de la mesure de Lebesgue particulièrement adaptée à la topologie de $\mathbb{R}$.
On dit que la mesure ν est à densité par rapport à μ si il existe une fonction μ-intégrable $\frac{d\nu}{d\mu}$ telle que :

$$\nu : \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R}_+ \\ A & \longmapsto & \mathcal{L} \int_{\mathbb{X}} \mathbb{1}_A(x) \frac{d\nu}{d\mu}(x)\, d\mu(x) \end{array}$$

Peut-on à partir de n'importe quelle mesure déterminer une mesure à densité ?

Toute mesure[3] qui ne voit pas plus de détails que la mesure d'origine[4] admet une densité par rapport à la mesure d'origine. On appelle cette densité la dérivée de Radon-Nikodym.
Il se trouve que cette propriété qui s'avère pratique pour construire des lois de probabilités sur $\mathbb{R}^d$ n'est plus vraie en général pour les espaces de Banach munis de l'intégrale de Böchner. Le lecteur pourra se référer à (18) pour plus de précisions sur l'existence d'espaces de Banach qui ne disposent pas de la propriété de Radon-Nikodym.

Un point clé pour notre intérêt d'étendre la méthode vue précédemment au cas fonctionnel est le fait que tout espace contenant l'ensemble des fonctions continues sur $[0,1]$ ne possède pas cette propriété, ce qui nous dérange fortement étant donné que l'on travaille sur ces objets.(18)

Ce que l'on vient d'énoncer sur la non validité de la propriété de Radon-Nikodym des espaces de Banach munis de l'intégrale de Böchner ne stipule pas qu'il n'existe pas de loi à densité. Il nous dit juste que la proposition « ν est absolument continue par rapport à $\mu \implies \exists \frac{d\nu}{d\mu},\ \nu = \frac{d\nu}{d\mu} \cdot \mu$ » n'est pas vraie. Ce qui veut juste dire que trouver des densités devient plus compliqué.

Ainsi dans les propositions (2.2) et (2.1) énoncées précedemment, je ne suis pas personnellement convaincu de l'insertion directe $\mathbb{E}[\|f(X)\|] < \infty$ et $\mathbb{P}_X \ll \mu \implies$ (2.1) puisque la non validité de la propriété de Radon-Nikodym dans les espaces de Banach en général nous dit que la continuité absolue n'est pas un critère suffisant pour admettre une densité. Cependant, il se pourrait que je n'ai pas bien compris les arguments des auteurs, et que l'on puisse toujours trouver une densité à partir de l'absolue continuité en se restraignant aux espaces de fonctions qu'ils considèrent. Dans le bénéfice du doute, je remets cette question à mon incompréhension pour le moment.

---

3. toute mesure *σ-finie*

4. On dit que la mesure est absolument continue par rapport à la mesure d'origine, one note $\nu \ll \mu$

2.2.2 ☐ B⟩ Une construction de la théorie autour de l'estimation à noyaux compliquée

Une partie non négligeable de la théorie non paramétrique pour les données fonctionnelles a été développée en se basant sur un estimateur de Nadaraya-Watson impliquant l'étude de la probabilité dite de petite boule qui étudie à quel point deux observations sont proches dans l'espace (ce qui relève évidemment d'une importance capitale pour un estimateur de localisation) :

$$\mathbb{P}\left[\, d(X_a, X_b) < \varepsilon \,\right] \text{ lorsque } \varepsilon \to 0 \tag{2.3}$$

dont l'utilité est parfaitement expliquée dans (19) :

> La (semi-)distance joue également un rôle important pour les propriétés asymptotiques des estimateurs fonctionnels non paramétriques. (...) La probabilité de petite boule définie comme $\mathbb{P}\left[\, d(u, v) < \varepsilon \,\right]$ apparaît dans le taux de convergence de nombreux estimateurs non paramétriques tels que l'estimateur fonctionnel de Nadaraya-Watson. Si la probabilité de petite boule décroît très rapidement lorsque ε tend vers zéro (en d'autres termes, si les données fonctionnelles sont très dispersées), la vitesse de convergence sera faible, alors qu'une probabilité de petite boule décroissant suffisamment lentement conduira à une vitesse de convergence similaire à celles trouvées dans les environnements de dimension finie.
>
> —Selk, Leonie and Gertheiss, Jan (2023) (19)

Il faut faire d'autant plus attention lorsque l'on travaille avec les probabilités de petites boules dans des espaces fonctionnels comme $\mathcal{C}^0\big([0, 1]\big)$ qui est de dimension infinie :

> il a été démontré qu'une hypothèse qui était fréquemment utilisée par certains auteurs en estimation non paramétrique fonctionnelle concernant les probabilités de petites boules implique que l'espace de fonction considéré est de dimension finie. (1) Cela restreint fortement la flexibilité qu'offrent les données fonctionnelles.

Cela illustre bien un point fondamental lorsque l'on travaille avec les objets fonctionnels :

- ☐ Les espaces de fonctions généraux sont de dimension infinie
- ☒ Les espaces de dimension infinie sont **hautement contre-intuitifs** et difficiles à manipuler
- ☐ En statistique, on travaille avec des hypothèses sur le comportement des observations : on doit faire attention aux hypothèses que l'on pose sur les objets fonctionnels qui pourraient restreindre leur flexibilité
- ☑ Travailler avec des sous-espaces de dimension finie n'est pas pour autant inutile, les fonctions polynômiales de degré d sont de dimension finie et s'avèrent être très utiles dans un nombre important de problèmes
- ☒ Il semble toutefois judicieux de développer une théorie générale sur les données fonctionnelles qui s'appliquerait à la plus gande variété de fonctions possible (et donc de dimension infinie), et on doit bien vérifier sur quoi travaille chaque résultat statistique

2.2.2 ☐ C⟩ Définir une densité sur les données fonctionnelles

Pouvoir définir une mesure à densité sur les données fonctionnelles s'avère de plus en plus critique au fur et à mesure que de la théorie sur l'estimation non paramétrique basé sur de telles données voit le jour. Une extension naturelle de l'estimateur de Nadaraya-Watson aux données fonctionnelles est :

$$\widehat{m}: \begin{array}{ccc} \mathcal{C}^0(I,\mathbb{R}) & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \dfrac{\sum_i Y_i \cdot K\left(\frac{d(x,X_i)}{h}\right)}{\sum_i K\left(\frac{d(x,X_i)}{h}\right)} \end{array} \quad \text{où } d \text{ est une semi-distance sur } \mathcal{C}^0(I,\mathbb{R}) \tag{2.4}$$

P. Hall et A. Delaigle (2010) commencent par reconnaître que, comme nous l'avons vu précédemment, dans le cadre des données fonctionnelles la densité n'est en général pas bien définie. Ils proposent alors de définir une « densité » sur un espace fonctionnel à une résolution spécifique qui correspond à un nombre composantes principales à sélectionner dans l'approximation de notre donnée fonctionnelle : cela nous permet de nous ramener à de la dimension finie. Enfin le concept de densité serait lié à la densité des scores [5] dans la base ACP tronquée des données fonctionnelles étudiées, ce qui nous avantage puisque l'on sait bien mieux manipuler les densités dans le cadre des variables aléatoires réelles.(6) La particularité de pouvoir utiliser cette approche même lorsqu'une densité "hilbertienne" n'existe pas rend cette approche attractive. Les auteurs considèrent une log-densité des scores qui représente le comportement à l'ordre 1 de l'effet de la valeurs des scores sur les probabilités que deux fonctions soient proches (probabilité de petite boule).

Soient $(\xi_k)_{\mathbb{N}}$ les scores de la fonction aléatoire X dans la base ACP $(\psi_k)_{k\geq 0}$ et $r(h)$ le nombre de composantes principales à considérer pour l'approximation de X à la résolution h. On note $\xi = (\xi_k)_{1,r(h)}$ et $e = (e_k)_{1,r(h)}$ une réalisation de ξ. On définit alors la log-densité des scores comme :

$$\ell(e|h) = \frac{1}{r(h)} \sum_{k=1}^{r(h)} \log p_{\xi_k}(e_k) \tag{2.5}$$

Alors à condition que les scores ξ_k soient i.i.d et que les $|p''_{\xi_k}| < \infty$ sans s'annuler [6], la probabilité de petite boule à la résolution h est :

$$p(e|h) = \mathbb{P}\left[\sum_{k=1}^{\infty} \lambda_k (\xi_k - e_k)^2 \leq h^2\right] = \widetilde{C} \cdot e^{r(h)\ell(e|h) + o\left(r(h)\right)} \tag{2.6}$$

La « densité » ainsi définie permet de travailler sur les scores des données fonctionnelles, et de donner des informations sur les probabilités de petite boule, ce qui est un point clé pour l'estimation non paramétrique. De plus cette définition de densité se base sur les scores des données fonctionnelles, objets importants et spécifique de ces dernières, ce qui motive d'avantage cette approche.

2.3 Rapide revue de la littérature sur l'estimation du modèle additif dans le cadre hilbertien

Les difficultés concernant la généralisation des concepts statistiques aux données fonctionnelles ayant été mentionnées dans la section précédente, nous nous focaliserons dans cette section à une revue rapide de la méthodologie proposée en partie par J.M. Jeon et B.U. Park concernant le problème de la régression non paramétrique du modèle additifs dans un espace de Banach général.

---

5. les scores sont les composantes de la fonction aléatoire centrée sur la base ACP $(\psi_k)_{k\geq 0}$: $\langle X - \mathbb{E}[X] \,|\, \psi_k \rangle_{\mathbb{L}^2}$. Ce sont des variables aléatoires réelles.

6. c'est un cas particulier des conditions plus générales mentionnées dans (6), ce qui allège la lecture

La méthodologie utilisée pour l'estimation non paramétrique du modèle additif est sensiblement identique à celle proposée par Mammen (1999) (17). Un des apports des auteurs est l'estimation en pratique de l'algorithme du Backfitting basé sur l'intégrale de Böchner dont on rapelle que les lois de composition peuvent être définies de façon non canoniques. L'idée est de ramener l'estimation au monde des réels (que l'on remarque être un thème récurrent dans la méthodologie liée aux données fonctionnelles) en se déplaçant le problème des fonctions aux poids du lissage à noyau. Cela est permis par le fait que :

$$\forall b \in \mathbb{B} \quad {}_{\mathbb{B}}\int f(u) \odot b \, d\mu(u) = {}_{\mathcal{L}}\int f(u) \, d\mu(u) \odot b \tag{2.7}$$

Mettre à jour les fonction $\widehat{m}_k^{[NW|BF]}$ revient à mettre à jours les poids du lissage à noyau $\widehat{w}_{k,i}^{[NW|BF]}$ $\forall i \in [\![1\,,\,n]\!]$ qui sont des réels.

En ce qui concerne la convergence de l'algorithme du Backfitting, les auteurs suivent la méthodologie de Mammen (1999) (17) en définissant les mêmes types d'espaces, et considérant des opérateurs de projection. Notons ici les différences notables avec les travaux de Mammen (1999) (17) :

La méthodologie des projecteurs sur les sous espaces additifs comme définis dans Mammen (1999) fonctionne pour les (sous-)espaces de Hilbert de dimension finie. En revanche dans le cadre de la dimension infinie, utiliser la projection orthogonale sur un sous-espace vectoriel fermé devient plus compliqué. La différence est qu'en dimension infinie les projecteurs comme comme définis dans Mammen (1999) ne sont plus des opérateurs compacts[7]. C'est un problème majeur puisque dans l'algorithme du Backfitting, pour estimer $m = \sum_k m_k$ on doit estimer les fonctions m_k 1 par 1.
À chaque itération r de notre algorithme du Backfitting T nous nous retrouvons face à la situation :

$$\widehat{m}^{[r]} = \sum_{k=1}^{p} \widehat{m}_k^{[r]} + \sum_{k=p+1}^{d} \widehat{m}_k^{[r-1]}$$

En notant $\mathcal{M}_k = \left\{m \in \mathcal{A} : m = m_k \text{ avec } m_k \in \mathbb{L}^2(p_k \cdot \lambda)\right\}$, on doit donc vérifier que pour tout $p \leq d$ l'espace $\sum_{k\leq p} \mathcal{M}_k$ est un fermé pour pouvoir appliquer le théorème de projection orthogonale sur un sous-espace vectoriel fermé. On peut alors estimer une composante après l'autre avec le point de vue de projection (qui pour rappel est essentiel à l'étude de la convergence). La compacité des projecteurs permet d'affirmer que la somme des sous espaces est elle aussi un sous espace vectoriel fermé. On ne peut donc plus invoquer cet argument en dimension infinie il faut donc trouver une autre méthode pour montrer que la somme des sous espaces est fermée.

Les auteurs contournent cette difficulté en montrant que les espaces $\sum_{k\leq p} \mathcal{M}_k$ sont fermés sans avoir à invoquer un argument de compacité des projections, et en utilisant le lemme suivant, plus particulièrement (3) :

Lemme (Lemme S.7 (11))
soit $\mathbb{H}$ de Hilbert et $\mathcal{M}_1, \ldots, \mathcal{M}_d$ des sous-espaces vectoriels fermés de $\mathbb{H}$. Sont équivalents :

1. $\sum\limits_{k\leq p} H_k$ est un sous-espace vectoriel fermé de $\mathbb{H}$
2. $\left\| \underset{k\leq p}{\circ} (I - P_k) \right\|_{\mathcal{L}(\overline{\sum\limits_{k\leq p} H_k})} < 1$
3. $\exists c > 0 \quad \forall h \in \mathbb{H} \quad h = \sum_k h_k$ avec $h \in \mathbb{H}_k$ et $\sum_k \|h_k\|^2 \leq c\|h\|^2$

7. Un opérateur est dit compact si il envoie toute partie bornée sur une partie d'adhérence compacte pour la topologie induite par la norme

Chapitre 3

Conclusion

L'étude de la méthodologie liée à la régression non paramétrique pour le modèle additif ainsi que le modèle semi-paramétrique permet de découvrir des outils qui peuvent s'avérer utiles à la fois pour la compréhension de la théorie de l'estimation non paramétrique mais aussi pour la mise en place d'algorithmes d'estimation.

Parmis les idées les plus intéressantes, on peut citer l'utilisation du domaine fréquentiel et de la transformée de Fourier pour la déconvolution de l'estimation. Une des idées les plus marquantes doit être le point de vue de la régression non paramétrique du modèle additif comme une projection de la fonction m sur un sous espace des fonctions additives, à la fois très géométrique et pour autant clé et non purement esthétique dans l'étude de la convergence. Enfin le cheminement qui mène à la considération du noyau normalisé de déconvolution pour le modèle semi-paramétrique comme l'assemblage de briques fondatrices qui proviennent des multiples éléments méthodologiques abordés dès le début de ce rapport est très satisfaisant.

Le stage a aussi permis de rendre compte de la difficulté de généraliser des concepts à des espaces plus généraux tels que les espaces de Banach, et les éléments auxquels il faut faire attention lorsque l'on manipule des données fonctionnelles avec un exemple concret de difficulté qui a dû être contournée pour pouvoir appliquer la méthodologie bien connue dans le cadre des réels.

Annexe A

Évaluation du rapport de stage

ENSAI École nationale de la statistique et de l'analyse de l'information

FICHE D'EVALUATION DU STAGE D'APPLICATION EN STATISTIQUE 2è année

Nom et prénom de l'élève	**Hugo Brunet**
Organisme d'accueil	CREST (ENSAO)
Nom du correcteur	VIMOND Myriam

Critères d'appréciation		**Commentaire**
Description de l'environnement professionnel		Bonne présentation
Description des travaux menés :		Le rapport est un rapport méthodologique sur les modèles additifs en semi-paramétrique.
	Présentation de la problématique	
	Description des données	
	Analyses descriptives	
	Interprétation des résultats	
Présentation de la méthodologie		Effort appréciable de l'étudiant pour présenter les méthodes de manière rigoureuse. Les procédures sont dans l'ensemble maitrisées.
Analyse critique/prolongements		Le rapport fait de manière régulière des allers-retours avec la bibliographie existante démontrant que Hugo a bien investi la littérature.

Présentation (qualité rédactionnelle, lisibilité)	Le résumé n'est pas un résumé (il y a confusion avec l'introduction) et prend comme référence l'ENSAI alors qu'il devrait se placer dans un contexte plus général. Les notations mathématiques sont globalement bien définies (avec un glossaire), par contre il y a des problèmes avec la ponctuation. Bon usage de la bibliographie. Attention au style parlé.

Commentaire général

Le travail effectué est conséquent, la méthodologie étudiée est de niveau 3A et elle est correctement maîtrisée par Hugo. Une étude de Monte Carlo pour illustrer la méthodologie aurait été attendue. Un effort supplémentaire sur la rédaction serait attendu pour l'obtention d'une note supérieure.

NOTE	15

Bibliographie

(1) Jean-Marc Azaïs and Jean-Claude Fort. Remark on the finite-dimensional character of certain results of functional statistics. Comptes Rendus. Mathématique, 351(3-4) :139–141, 2013.

(2) Richard C Bradley. Basic properties of strong mixing conditions. a survey and some open questions. 2005.

(3) Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33 :1877–1901, 2020.

(4) Sungho Cho, Jeong Min Jeon, Dongwoo Kim, Kyusang Yu, and Byeong U Park. Partially linear additive regression with a general hilbertian response. Journal of the American Statistical Association, pages 1–15, 2023.

(5) StackExchange Contributors. n-ball volume and surface with n $\rightarrow \infty$. `https://math.stackexchange.com/questions/584774/n-ball-volume-and-surface-with-n-rightarrow-infty`, 2023. Accessed : 2023-11-08.

(6) Aurore Delaigle and Peter Hall. Defining probability density for a distribution of random functions. The Annals of Statistics, pages 1171–1193, 2010.

(7) Jean François Le Gall. Intégration, probabilités et processus aléatoires. (télécharger), 2006.

(8) Kyunghee Han and Byeong U. Park. Smooth backfitting for errors-in-variables additive models. The Annals of Statistics, 46(5) :2216–2250, 2018.

(9) T.J. Hastie and R.J. Tibshirani. Generalized Additive Models. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1990.

(10) Joel Horowitz, Jussi Klemelä, and Enno Mammen. Optimal estimation in additive regression models. Bernoulli, 12(2) :271–298, 2006.

(11) Jeong Min Jeon and Byeong U Park. Additive regression with hilbertian responses. 2020.

(12) Jeong Min Jeon and Germain Van Bever. Additive regression with general imperfect variables. arXiv preprint arXiv :2212.05745, 2022.

(13) KAIST. Lecture 8 : High-dimensional space. `https://alinlab.kaist.ac.kr/resource/2021_AI503_Lec8.pdf`, 2021. Accessed : 2023-11-08.

(14) Alec Koppel, G Warnell, E Stump, P Stone, and Alejandro Ribeiro. Breaking bellman's curse of dimensionality : Efficient kernel gradient temporal difference. arXiv preprint arXiv :1709.04221, 2017.

(15) Eun Ryung Lee, Kyunghee Han, and Byeong U. Park. Estimation of errors-in-variables partially linear additive models. Statistica Sinica, 28(5) :2353–2373, 2018.

(16) Young Kyung Lee, Enno Mammen, and Byeong U Park. Hilbertian additive regression with parametric help. Journal of Nonparametric Statistics, pages 1–20, 2023.

(17) Enno Mammen, Oliver Linton, and Janni Nielsen. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. The Annals of Statistics, 27(5) :1443–1490, 1999.

(18) Raymond A. Ryan. Introduction to Tensor Products of Banach Spaces. Springer London, London, 2002. The Radon-Nikodým Property : pages 93–126 – ISBN : 978-1-4471-3903-4.

(19) Leonie Selk and Jan Gertheiss. Nonparametric regression and classification with functional, categorical, and mixed covariates. Advances in Data Analysis and Classification, 17(2) :519–543, 2023.

(20) Charles J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. The Annals of Statistics, 10(4) :1040 – 1053, 1982.