==========

République algérienne démocratique et populaire
Ministère de l'enseignement supérieur et de la recherche scientifique

Université Des Frères Mentouri
Faculté des sciences exactes

==========

Département de mathématiques

N° d'ordre : 47/DS/2023
N° de série : 06/Math/2023

THÈSE

PRÉSENTÉE POUR L'OBTENTION
DU
DIPLÔME DE DOCTORAT EN SCIENCES
DE
MATHÉMATIQUES

« Quelques aspects d'inférence statistique fréquentielle : Méthode des ondelettes »

Par
DOUAS RYMA

OPTION
Probabilités Statistique

Devant le jury :

Président	M^{r}	M. Dalah	Professeur	Université Frères Mentouri Constantine 1
Directrice de thèse	M^{me}	S. Kharfouchi	Professeur	Université Salah Boubnider Constantine 3
Examinatrice	M^{me}	K. Kimouche	M.C.A	Université 20 Août 1955 Skikda
Co-directrice	M^{me}	I. Laroussi	M.C.A	Université Frères Mentouri Constantine 1

Soutenue le : ...

Remerciements

Je remercie infiniment mes directrices de thèse, madame I. Laroussi et S. kharfouchi pour toute l'aide est les moyens qu'elles ont mis à ma disposition pour avancer, je les salue particulièrement pour ses qualités humaines et leurs rigueur scientifique.

J'adresse mes remerciements sincères et chaleureux à monsieur M. Dalah qui me fait l'honneur de présider le jury de soutenance.

Mes sincères remerciements à K. Kimouche, pour l'honneur qu'elle ma fait d'examiner mon travail.

Je remercie grandement mes collègues F. Meghraoui , R. Gueres et K. Sahli pour toute l'aide qu'elles m'ont apportée avec autant de gentillesse et de générosité.

DEDICACE

A Mes chères mères, Leila, Rachida

Merci de m'avoir soutenu durant toutes ces années, ainsi que de votre confiance et de votre présence. Avec tout mon amour.

A Mon mari, Meheiddine

Merci pour votre soutien durant ces années d'études et pour vos encouragements.

A Mes sœurs, Mes frères adorées,

Merci de m'avoir encouragé durant ces années. Avec tout mon amour.

Je vous remercie grandement du temps et de l'attention que vous m'avez consacré.

Table des matières

INTRODUCTION

Cette étude a pour objet de mettre en lumière les méthodes des moindres carrés sur les bases des ondelettes. Il peut être utile de rappeler que l'analyse en ondelettes est née au début des années 80 a été introduite, il y a plusieurs décennies. D'abord par Gabor (1946) [16] dans le domaine pétrolier et dernièrement dans le domaine de l'analyse du signal par J. Morlet en faisant apparaître simultanément des informations temporelles et fréquentielles. D'autres auteurs, dans différents contextes, ont utilisés des outils similaires aux ondes de Morlet. Récemment, cette discipline se développe rapidement du fait qu'elle trouve un large champs d'application et même en statistique. Il est connu que le traitement du signal est régie par l'analyse de Fourier mais les ondelettes ont connu un essor fulgurant. Une première application des ondelettes est qui semble simple est la transformée de Fourier à fenêtre glissante ; appelée transformée de Gabor continue. En revanche, l'analyse par ondelettes basée sur la notion d'échelle au lieu du concept de fréquence est définie par l'utilisation d'ondelettes élémentaires. Ce procédé porte le nom d'analyse multi résolution. Cette dernière se présente comme la construction d'une ondelette de base nommée ondelette mère de sorte que les ondelettes élémentaires déduites par des translations et des dilatations constituent une base de l'espace $L_2(\mathbb{R})$. Elle permet l'écriture sous forme de combinaison linéaire lorsque on considère une fonction ou un signal de carré intégrable dans cette base. Des références incontournables et approfondies sur l'analyse multi- résolution est donnée dans Meyer (1989) [31] et Mallat [29].

Nous avons choisi cette méthode parce qu'elle nous permet d'analyser efficacement des signaux où se combinent des phénomènes d'échelles très différents. De nombreux scientifiques utilisaient déjà les ondelettes comme une alternative à l'analyse de Fourier traditionnelle qui a été introduite par le mathématicien Joseph Fourier (1822)[15], c'est donc une

méthode relativement récente. Dans la suite de ce travail, nous verrons comment les estimations de séries orthogonales utilisent les estimations d'un développement de séries pour reconstruire la fonction de régression. Nous nous intéresserons aux estimations orthogonales non linéaires ou une transformation non linéaire dite seuillage est appliquée aux coefficients estimés.

Il est utile de souligner que la clé de notre étude est la notion de "régression". Ce terme a été introduit par Francis Galton (1885)[17] à la suite d'une étude sur la taille des descendants de personnes de grande taille, qui décroit de génération en génération, vers une taille moyenne (donc leur taille régresse). La regression se définit comme étant l'ensemble des méthodes statistiques communément utilisées pour analyser la relation d'une variable par rapport à une ou plusieurs autres. Pendant, la régression d'une variable aléatoire Y sur le vecteur de variables aléatoires X était considérée comme étant la moyenne conditionnelle de Y sachant X. De nos jours, le terme de régression désigne tout élément de la distribution conditionnelle de Y sachant X, considérée comme une fonction de X. Donc, on considère le modèle $Y = m(X)+\varepsilon$ où ε est une variable aléatoire d'espérance nulle tout en étant indépendante de X. Ici, elle représente l'erreur commise pendant la sélection du modèle $m(x) = \mathbf{E}(Y|X = x)$. La fonction de régression la plus utilisée au début juste à la naissance de cette discipline est le modèle linéaire. Mais, dans le cas non linéaire. Il fallait utiliser une approche algorithmique itérative. D'autre fonctions peuvent être l'axe d'intérêt. Ainsi comme le quantile conditionnel de la distribution d'une variable aléatoire Y sachant le vecteur de variables aléatoires X, où l'on utilise un modèle de régression quantile. Aussi, si la forme fonctionnelle de la régression est inconnue, on peut utiliser un modèle de régression non paramétrique. Cette dernière est une forme d'analyse de régression dans laquelle la variable prédicatrice ou la fonction d'estimation, ne prend pas de forme prédéterminée, mais est construit selon les informations qui proviennent des données. Elle exige des tailles d'échantillons plus importantes que celles de la régression basée sur des modèles paramétriques parce que les données doivent fournir la structure du modèle ainsi que les estimations du modèle.

La méthode classique dans l'estimation est la méthode des moindres carrés qui à fait l'objet de plusieurs publications. Parmi les premières applications de cette méthode en 1805, fut celles au nom d'Adrien-Marie Legendre. Qui présentent une estimation par des polynômes et une estimation sur des bases de splines. Pour des publications sur les estimateurs par projection sur des bases orthonormées est plus récemment les bases des ondelettes et on peut se référer à Antoniadis et al. (1994) [1], Antoniadis (1996) [2], Donoho et Johnstone (1994, 1998) [9, 11] et Donoho

(1995) [10]. Parmi les travaux sur la consistance de ce type d'estimateur, nous citons, Vapnik et Chervonenkis (1971) [39], Vapnik (1982, 1998)[40] et Kohler (1997,1999)[25, 26]. Pour commencer, nous expliquons comment sont construits ces estimateurs dans le cas des données complètes ; autrement dit la variable expliquée est totalement observée. Puis, nous passons au cas où la variable réponse est censurée, ce qui veut dire que la valeur de cette variable peut être perdue au cours de l'étude. D'où l'utilisation de l'estimateur de Kaplan-Meier, inutile connu sous le nom de l'estimateur produit-limite et qui doit son nom à Edward L. Kaplan et Paul Meier. Il a été conçu pour estimer la fonction de survie à l'aide de données de durée de vie et est souvent utilisé pour mesurer le pourcentage de patients en vie pour une certaine durée après leur traitement en recherche médicale. Également utilisé en économie et en écologie. La fonction de survie admet une courbe de l'estimateur de Kaplan-Meier sous forme de série de marches horizontales de mesure décroissante qui est utilisé pour un échantillon suffisamment grand. Cette courbe approche très bien la fonction de survie réelle pour cette population. Un intérêt particulier de la courbe de Kaplan-Meier est que cette méthode peut prendre en compte certains types de données censurées, en particulier censurées à droite. Ce qui intervient lorsqu'un patient disparait d'une étude, c'est-à-dire qu'on ne dispose plus de ses données avant que l'événement attendu (par exemple le décès), soit observé. La méthode "plug in" qui consiste à insérer le précédent estimateur pour estimer la fonction de régression fut l'objet du travail de Carbonez et al. (1995)[5] où un estimateur à partitions de la fonction de régression est donné. En 2002, Kohler et al. [28] réutilisent l'idée de ce dernier travail, qui se traduit par l'étude d'un estimateur sans biais de la moyenne de Y tout en appliquant différentes méthodes d'estimation (à noyau, plus proches voisins, moindres carrés et spline de lissage). Ils montrent la consistance des estimateurs introduits en supposant l'indépendance du couple (X, Y) et de la variable de censure à droite noté R. Cette hypothèse n'est pas très difficile a retrouver dans la pratique, puisqu'elle est faisable lorsque la censure ne dépend pas des caractéristiques de l'individu sous étude. La deuxième supposition imposée est la continuité de la loi de R et que son support contient celui de Y. Il est possible de rencontrer, un autre cas de figure qui consiste dans la censure à gauche où l'observation de la variable d'intérêt et incomplète et pour laquelle on sait seulement qu'elle est inférieure à la valeur observée. En combinant ces deux types de censure sur un même échantillon, ils constituent l'apport principal de cette thèse. Nous étendons l'estimation de la fonction de régression au cas où la variable réponse est soumise à une censure mixte. Cela veut dire que Y est censurée à droite par une variable R (qui elle-même représente un temps de survie) et que le

minimum entre Y et R est censuré par une variable de censure à gauche. Aussi, toutes les variables latentes sont supposées indépendantes. C'est le modèle étudié dans l'article de Patilea et Rolin (2006)[36] dans lequel est proposé un estimateur produit limite de la fonction de survie fortement consistant. En 2010, Messaci [30] propose des estimateurs à poids (à noyau, à partitions et des plus proches voisins) fortement consistants de la fonction de régression pour Y censuré par ce dernier type de censure. L'étude de l'estimateur de la fonction de régression sur différentes classes de fonctions vérifiant des conditions de recouvrements s'est poursuivie dans l'article de Kebabi et al (2011)[24] dans lequel l'étude de convergence presque sûre à été établie.

Cette thèse contient cinq chapitres, le premier chapitre résume l'espace des ondelettes ainsi que leurs propriétés d'approximations exceptionnelle avec la présence de seuillage dur. Le deuxième chapitre est consacré à l'estimation de la fonction de régression dans le cas d'observation complète, ainsi que l'utilisation de la méthode de moindre carrées sur l'espace des ondelettes. En ce qui concerne le troisième chapitre, Il résume l'analyse de survie, la censure ainsi que l'étude de la consistance presque sûre des estimateurs de Kaplan-Meier, celui de Patilea et de Rolin. Le quatrième chapitre représente notre humble contribution dans la recherche, en présentant l'estimateur de la fonction de régression par la méthode des moindres carrées sur l'espace des ondelettes en présence de censure mixte. Il contient aussi l'étude de convergence p.s de cet estimateur. Finalement, un chapitre de simulation est rajouté pour appuyer l'efficacité de notre estimateur dans le domaine de l'application.

CHAPITRE 1

LES ONDELETTES

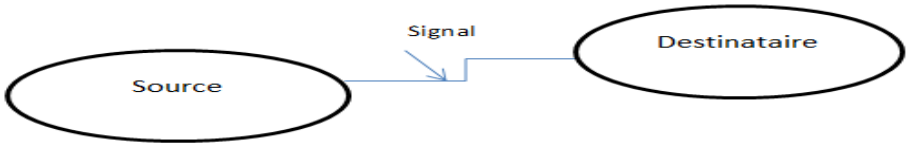
Ce chapitre présente l'analyse des ondelettes et leurs notions de base. Nous allons d'abord rappeler la transformée de Fourier puis introduire la transformation en ondelettes et les propriétés de ces fonctions analytiques (ondelettes). Enfin, nous discuterons de l'analyse multi-résolution est la méthode pour construire une base orthonormée pour l'espace L_2 par l'orthonormalisation des polynômes par morceaux.

1.1 Analyse de Fourier

Le traitement du signal a été longtemps dominé par la transformée de Fourier. Dans la suite. Nous représenterons et on donnerons les inconvénients d'une telle transformation.

1.1.1 Définitions des signaux

Définition 1 *Toute représentation graphique d'informations notée f est un signal, ce dernier est transféré de la source vers le destinataire.*



Par exemple un piano (source) produit des notes musicales (signaux) qui parviennent aux oreilles des personnes présentes (destinataires).

Définition 2 • *Un signal est dit analogique, s'il représente des informations en quantité "continue", il est défini sur* $\mathbb{R}$ *ou* $\mathbb{R}^2$*. C'est-à-dire, qu'on considère un signal* f *qui varie dans le temps (t continue) et on écrit* $f = f(t)$*.*

- *Un signal est numérique, s'il représente l'information par des suites de nombres. Il est défini sur les entiers* $\mathbb{Z}$*. C'est-à-dire qu'un signal* f *varie d'une manière discontinue et on écrit, pour* $(x_n)_{n\in\mathbb{Z}}$ *suite entière,* $f = f(x_n)_{n\in\mathbb{Z}}$*.*

Définition 3 *Le support d'un signal* f *défini sur* $\mathbb{R}$ *consiste à fermer l'ensemble des points auxquels la fonction ne s'annule pas*

$$supp(f) = \overline{\{t \in \mathbb{R}/f(t) \neq 0\}}.$$

- *Si* $f \in L_1(\mathbb{R}) \Longleftrightarrow \int_{-\infty}^{+\infty} \mid f(t) \mid dt < \infty$*, on obtient un signal stable.*
- *Si* $f \in L_2(\mathbb{R}) \Longleftrightarrow \int_{-\infty}^{+\infty} \mid f(t) \mid^2 dt < \infty$*, on obtient un signal d'énergie finie .*

On note $L_2([0,T])$ l'espace des fonctions périodiques et complexes de période T de carrés intégrables. Rappelons que le produit scalaire de deux fonctions f et g de $L_2([0,T])$ [1] est donné par

$$< f, g >= \int_0^T f(x)\overline{g(x)}dx. \tag{1.1}$$

Auquel on associe la norme $\|f\|_2^2 =< f, f >$ dans $L_2([0,T])$ et $\overline{g(x)}$ représente le conjugué d'un complexe.

1.1.2 Représentation d'un signal par la série de Fourier

On considère le signal $f : \mathbb{R} \longrightarrow \mathbb{C}$ périodique de période T. L'idée est de le lier à la décomposition de la forme

$$f(t) = \sum_{k=0}^{\infty} c_k \exp^{ik\omega t},$$

qui représente une série de Fourier où c_k sont des nombres complexes la période est donnée par $\omega = \frac{2\pi}{T}$.

Ainsi le problème est de donné, pour un entier n, une suite finie $(x_k)_{-n\leq k\leq n}$ telle que $\|f - p_n\|_2$ soit minimale. La définition suivante répond à cette idée.

Définition 4 *(Voir R. Douas [12]) On appelle polynôme trigonométrique de degré n et de période t, un polynôme de la forme*

$$p_n(t) = \sum_{k=-n}^{n} c_k e_k(t).$$

Où $e_k = \exp^{ik\omega t}$.

L'espace $L_2([0,T])$ contient l'espace des polynômes trigonométriques Γ_n de degré inférieur ou égale à n. Ceci permet à n'importe quelle fonction $f \in L_2([0,T])$ d'être approchée par le polynôme $p_n \in \Gamma_n$ pour n assez grand, d'où l'intérêt du théorème suivant.

1. $f \in L_2([0,T]) \Longleftrightarrow \int_0^T \mid f(t) \mid^2 dt < \infty$.

Théorème 1 *B. Torrésani [38]*

Il existe un unique polynôme trigonométrique f_n dans Γ_n tel que

$$\|f - f_n\|_2 = min\{\|f - p_n\|_2, p_n \in \Gamma_n\}.$$

Avec

$$c_k(f) = \frac{1}{T}\int_0^T f(t)e_k(t)dt.$$

Preuve 1 *Pour la démonstration se référer à Torrésani [38] et à l'ouvrage de C. Gasquet et P. Witomski [18].*

Remarque 1 *La famille de fonctions exponentielles e_k est une base orthonormée de $L_2([0,T])$.*

Pour le théorème ci dessous, on a besoin de la décomposition tiré de B. Torrésani [38] suivante

$$\|f - p_n\|_2^2 = \|f\|_2^2 + T\sum_{|k|>-n} |c_k(f)|^2.$$

Théorème 2 *B. Torrésani [38]*

Pour tout signal $f \in L_2([0,T])$, on a

$$\|f - f_n\|_2^2 = T\sum_{|k|>-n} |c_k(f)|^2 \longrightarrow 0, \text{ quand } n \longrightarrow +\infty.$$

Preuve 2 *Pour la démonstration ce référer à Torrésani [38] et l'ouvrage de C. Gasquet et P. Witomski [18].*

Théorème 3 *(Théorème de Dirichlet) C. Gasquet et P. Witcomski [18]*

Si f est une fonction périodique de période t et si f est de classe $\mathcal{C}^1$ par

morceaux sur $\mathbb{R}$ *alors la série de Fourier associée à* f *est convergente sur* $\mathbb{R}$ *et on a*

$$\frac{1}{2}(f(t_-) + f(t_+)) = a_0 + \sum_{n=1}^{+\infty}(a_n cos(n\omega t) + b_n sin(n\omega t)).$$

Où $f(t_-) = \lim\limits_{\substack{x<t \\ \to}} f(x)$ *et* $f(t_+) = \lim\limits_{\substack{x>t \\ \to}} f(x)$. *Et si de plus* f *est continue, alors on a*

$$f(t) = a_0 + \sum_{n=1}^{+\infty}(a_n cos(n\omega t) + b_n sin(n\omega t)).$$

Preuve 3 *Ce théorème est très connu pour la démonstration, on peut se référer à C. Gasquet et P. Witomski [18]*

1.1.3 Transformée de Fourier

La Transformation de Fourier a été publiée pour la première fois en 1822 par Joseph Fourier [15]. Elle convertit une fonction mathématique du domaine temporel au domaine fréquentiel, Cela nous permet d'obtenir d'autres propriétés de la fonction qui seraient autrement invisibles. Il existe plusieurs variantes de l'équation de la transformée de Fourier pour convertir $f(t)$ au domaine de fréquence.

Définition 5 *Soit* $f \in L_2(\mathbb{R})$, *sa transformée de Fourier est une fonction notée* $\widehat{f}$ *définie par*

$$\widehat{f}(\omega) = \int_{-\infty}^{+\infty} f(t) \exp(-i\omega t) dt,$$

pour toute valeur de ω.

Le théorème suivant permet d'avoir une version Le théorème inversée d'une fonction. Il représente une autre forme du théorème de Dirichlet pour les fonctions absolument intégrables.

Théorème 4 *(Voir B. Torrésani [38])*

Soit f une fonction de absolument intégrable. Sa transformée de Fourier $\widehat{f}$ est intégrable et si f est continue en t. Alors on obtient

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \widehat{f}(\omega) \exp(i\omega t) d\omega.$$

Preuve 4 *Pour la démonstration ce référer à B. Torrésani [38] et l'ouvrage de C. Gasquet et P. Witomski [18].*

Exemple 1 *Soit $f(t) = \mathbf{1}_{[-s,s]}(t)$, la fonction indicatrice de l'intervalle $[-s, s]$, donc elle est discontinue en $-s$ et $+s$. Sa transformée de Fourier est donnée par*

$$\widehat{f}(\omega) = \frac{2sin(s\omega)}{\omega}.$$

1.1.4 Les inconvénients de la transformée de Fourier et leur limitations

La transformée de Fourier ne nous permet pas une analyse du comportement local d'une fonction, cela représente un inconvénient majeur. Cequi est dû au fait qu'elles admettent (les fonctions analysantes) des supports infinis. Et il est connu qu'il est nécessaire de connaître l'ensemble des valeurs d'une fonction pour pouvoir calculer sa transformée de Fourier. Aussi, cette dernière ne permet pas d'avoir une localisation temporelle du contenu fréquentiel d'un signal. L'exemple de la Figure (1.1) représente un signal à deux signaux de fonctions consécutives. La transformée de Fourier de ce signal permet de retrouver ces deux fréquences, mais elle ne localise pas temporellement le changement de régime dans le signal. Le second inconvénient consiste dans la présence d'une coupure dans le signal qui affecte le comportement de la transformée de Fourier pour toutes les fréquences.

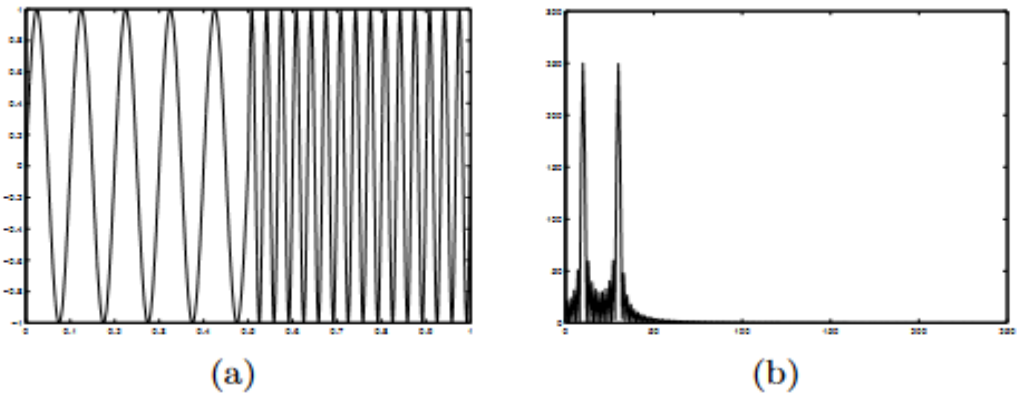


FIGURE 1.1 – Limitations de la transformée de Fourier

1.1.5 Comment dépasser ces limitations

Pour résoudre ce problème, on a ces deux points

- **Idée de fenêtre**
  La transformée de Fourier n'est pas idéale pour traiter un signal non stationnaire. D'où l'idée de couper le signal non stationnaire pour restituer des signaux stationnaires dans un intervalle de temps. C'est ce qu'on appelle une fenêtre.
- **Buts**
  — Déterminer les variations du signal.
  — Traiter un signal au fur et à mesure.

D'où l'introduction de représentations temps-fréquences. Par exemple la transformée de Fourier à fenêtre glissante dite de Gabor et qui représente une réponse aux questions suivantes. Comment localiser un signal en temps et en fréquence ? A quelle variation peut-on s'attendre pour (t, ω) ?

1.2 La transformée de Fourier à fenêtre glissante

Cette partie présente un type de transformée de Fourier bien connue qui n'est autre que la transformée en ondelettes ainsi que ses applications. L'avantage de cette transformée est qu'elle permet d'extraire à la fois les informations temporelles (spatiales) et fréquentielles d'un signal donné. La taille de la fenêtre accordable lui permet d'effectuer une analyse multi-résolution. Parmi les types de transformées en ondelettes, la transformée en

ondelettes de Gabor qui possède à la fois des propriétés mathématiques et biologiques. Toutes deux importantes et ont été fréquemment utilisées dans les recherches sur le traitement d'images.

Définition 6 *On définie l'énergie d'une fonction ou d'un signal f par la quantité*

$$\mathbf{E}_f = \int_{-\infty}^{+\infty} |f(t)|^2 dt.$$

Proposition 1 *B. Torrésani [38]*

Si le signal est de absolument intégrable, alors il est d'énergie finie. Dans ce cas on a

$$\mathbf{E}_f = \frac{1}{2\pi}\mathbf{E}_{\widehat{f}}, \tag{1.2}$$

où $\mathbf{E}_{\widehat{f}}$ est l'énergie de la transformée de Fourier $\widehat{f}$ de f.

Théorème 5 *(Le principe d'incertitude d'Heisenberg) (Voir C. K. Chui [6])*

Soit $f \in L_2(\mathbb{R})$ et telle que f et $\widehat{f}$ satisfont (1.2). Alors

$$\sigma_f \sigma_{\widehat{f}} \geq \frac{1}{2}.$$

Où

$$\sigma_{\widehat{f}}^2 = \frac{1}{\sqrt{\mathbf{E}_f}} \left\{ \int_{-\infty}^{+\infty} (t - t^*)^2 |f(t)|^2 dt \right\}^{\frac{1}{2}}$$

et

$$t^* = \frac{1}{\mathbf{E}_f} \int_{-\infty}^{+\infty} t|f(t)|^2 dt.$$

Preuve 5 *Pour la démonstration, consultez l'ouvrage de Chui [6] .*

Ce théorème montre la grandeur inversée des supports de f et de sa transformée de Fourier $\widehat{f}$. C'est-à-dire que si le support de f est assez grand, celui de $\widehat{f}$ doit être nécessairement assez petit et réciproquement. Ceci conduit à des erreurs dans la reconstruction de la fonction f à partir de sa transformée de Fourier. En 1946, Gabor a proposé d'utiliser une transformée de Fourier à fenêtre glissante pour résoudre le problème de location de temps de la transformée de Fourier. Cela consiste à calculer la transformée de Fourier sur une partie du signal choisi en utilisant une fenêtre temporelle bien positionnée. Puis des translations consécutives de cette fenêtre permettent d'étudier localement le courbure temps-fréquence du signal. La transformée de Gabor projette un signal sur les fonctions analysante du modèle comme suit

$$g_{b,\omega}(t) = e^{i\omega(t-b)}g(t-b), \omega \in \mathbb{R}$$

où g est une fenêtre et b un nombre réel fixé.

Définition 7 *Soit $g \in L_2(\mathbb{R})$ une fenêtre. On appelle transformée de Fourier à fenêtre glissante (ou transformée de Gabor continue) d'un signal $f \in L_2(\mathbb{R})$, l'application G_f définies sur $\mathbb{R}^2$ par*

$$G_f(\omega, b) = \int_{-\infty}^{+\infty} f(t)\overline{g_{b,\omega}}(t)dt.$$

où $\overline{g_{b,\omega}}$ est une fenêtre gaborettes.

Si la fenêtre g est d'énergie égale à 1 (cas de fenêtres gaussiennes par exemple), alors en intégrant la transformée de Gabor $G_f(\omega, b)$ et en utilisant le théorème de Fubini, on obtient

$$\int_{-\infty}^{+\infty} G_f(\omega, b)db = \widehat{f}(\omega).$$

Le Théorème suivant donne la formule d'inversion qui permet comme il a été présenter pour la transformée de Fourier, la connaissance de $G_f(\omega, b)$ pour toutes les valeurs de ω et b permet de reconstruire la fonction f.

Théorème 6 *(Formule d'inversion) (Voir B. Torrésani [38])*

Soit $g \in L_2(\mathbb{R})$. Alors tout signal $f \in L_2(\mathbb{R})$, on a en tout point de continuité

t de f :

$$f(t) = \frac{1}{2\pi\|g\|_2^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G_f(\omega, b) g_{\omega,b}(t) d\omega db.$$

Où $\|g\|_2^2$ *est la norme de la fonction* g.

Preuve 6 *Pour la démonstration se référer à B. Torrésani 1997 [38].*

1.3 La transformée en ondelettes

Comme pour le cas des transformées de Fourier, celles de Gabor admet des inconvénients comme la rigidité de la fenêtre de fréquence de temps. C'est-à-dire que la longueur de la fenêtre reste invariable. Lorsque le signal considéré est soumis à de fortes alternances, il y a un problème de fluctuation. En d'autres termes, la transformée de Gabor est imparfaite elle s'applique où juste convient, surtout à l'analyse simultanée des signaux de très haute fréquence et de très basse fréquence. Le passage aux ondelettes contourne le problème de fluctuation. À l'origine elles ont été proposée par J. Morlet [32], en se fondant sur le concept de d'échelle ou encore de résolution. Cela se précisé par l'utilisation d'ondelettes pour bien localiser, à la fois dans l'espace et dans le temps quelque soit l'intensité de la fluctuation. Les espaces spectraux sont créés de façon interchangeable par translation et dilatation. Les ondelettes sont identiques et ne varient que dans leur taille. Ils s'adaptent parfaitement et automatiquement à la forme et à la taille des éléments qu'ils recherchent. Elles sont très larges pour étudier les basses fréquences et très fines pour étudier des composantes plus provisoires. Cette procédure, développée par S. Mallat [29] et organisée par I. Daubechies [7], est appelée multi-résolution.

Définition 8 *(Voir R. Douas [12]) On appelle transformée en ondelette réelle d'une fonction* $f \in L_2(\mathbb{R})$ *associée à une ondelette analysante* ψ *la fonction*

$$\mathcal{C}_f(a, b) =< f, \psi_{a,b} >= \int_{-\infty}^{+\infty} f(t)\overline{\psi_{a,b}}(t)dt,$$

où ψ une ondelette et a,b deux nombres réels tels que $a > 0$ sont données par

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right).$$

Rappelons que $< .,. >$ est le produit scalaire dans $L_2(\mathbb{R})$ donné par la relation (1.1) sur l'espace $\mathbb{R}$.

- *$\psi_{a,b}$ est l'ondelette mère ψ translatée de b et dilatée de a (échelle) ou contractée si $a < 1$. Quand l'échelle a augmenté. Le support de la partie non nul augmente.*

- *$\frac{1}{\sqrt{a}}$ est le coefficient multiplicateur que permet d'avoir une formule de conservation de l'énergie du signal.*

Comme la figure (1.2) le montre bien, la forme d'ondelette ne change pas, elle est simplement étalée ou comprimée.

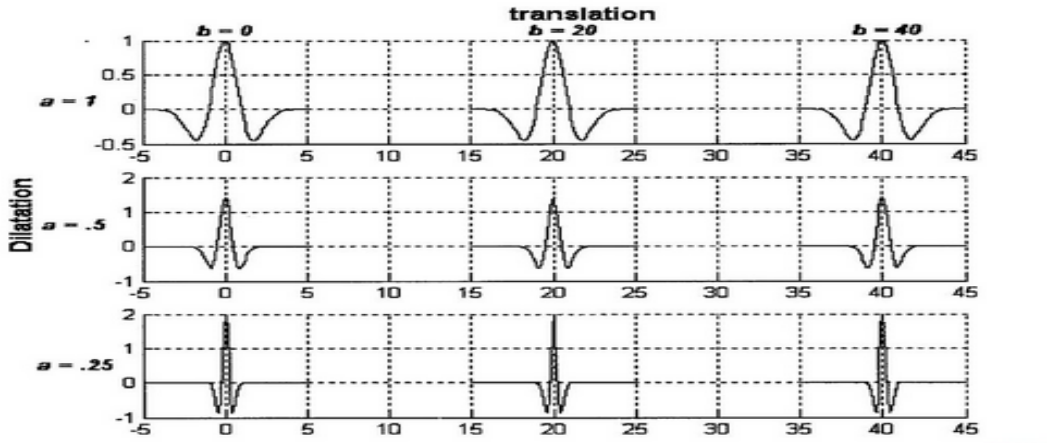


FIGURE 1.2 – Translation et dilatation des ondelettes

1.3.1 Analyse temps - échelle

L'énergie de la fonction d'ondelette est généralement de un. Des fonctions telles que le sinus et le cosinus ne peuvent pas être utilisées comme fonctions d'analyse, car elles ne vérifient pas cette condition en ayant une énergie infinie. Il existe une exigence implicite selon laquelle, même si elle a une énergie finie, elle doit avoir une certaine énergie, de sorte que l'intégration de la fonction doit être plus grande que zéro. La deuxième exigence est connue sous le nom de condition d'admissibilité, elle stipule que la transformée de Fourier de la fonction d'ondelettes ne peut pas avoir une composante de fréquence nulle.

Définition 9 *(Voir R.Douas [12])*

On appelle ondelette analysante ou ondelette mère, une fonction $\psi \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$, *d'énergie* $E_\psi = 1$ *vérifiant*

$$K_\psi = \int_{-\infty}^{+\infty} \frac{\|\widehat{\psi}(\omega)\|_2^2}{|\omega|} d\omega < \infty,$$

où $\widehat{\psi}$ *est la transformée de Fourier de* ψ.

Si la fonction ψ vérifie la condition suivante

$$\widehat{\psi}(0) = 0.$$

On dit qu'elle vérifie la condition d'admissibilité. Cette dernière est imposée aux fonctions de l'espace $L_2(\mathbb{R})$.

Définition 10 *Une fonction* ψ *admet* K *moments nuls si pour tout* $p = 0, ..., K-1$,

$$\int_{-\infty}^{+\infty} t^p \psi(t) dt = 0 \ et \ \int_{-\infty}^{+\infty} \mid t^K \psi(t) \mid dt < +\infty.$$

Dans un signal cette définition permet la détection la plus efficace des singularités. Ce qui signifie que si ψ a K moments nuls. Il est orthogonal à tout polynôme de degré $K-1$.
De plus, cette relation exprime le fait que $\widehat{\psi}$ est une fonction oscillante infiniment amortie. C'est l'origine du nom de l'ondelette. On remarque aussi qu'il y a une différence entre les propriétés de l'onde analytique et les propriétés de la fenêtre.

1.3.2 Analyse multirésolution et base d'ondelettes

On peut adapter la transformée en ondelettes dans le cas où le signal est à support discret. Cette approximation est notamment utilisée dans la compression de données numériques avec ou sans perte. Parmi les différentes méthodes possibles. On présente les ondelettes à travers l'idée de l'analyse multirésolution. Celle-ci consiste à définir une suite d'espaces emboîtés $\{(V_j),\ j \in Z\}$ de sorte qu' à chaque emboîtement, l'approximation résultante d'une fonction f de $L_2(\mathbb{R})$ sur cette espace est plus fiable. Pour être rigoureux, une analyse multirésolution, se définit comme suit

Définition 11 *Analyse multi-résolution (Voir R. Douas [12])*

On appelle analyse multi-résolution de $L_2(\mathbb{R})$ une suite croissante $\mathcal{M} = \{V_j\}_{j\in\mathbb{Z}}$ de sous espaces fermés de $L_2(\mathbb{R})$ telle que

1. *les sous espaces V_j sont emboîtés : $V_j \subset V_{j+1}$ pour tout $j \in \mathbb{Z}$.*
2. *$\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L_2(\mathbb{R})$ et $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$.*
3. *$\forall f \in L_2(\mathbb{R}), \forall j \in \mathbb{Z}, f \in V_j$ si et seulement si $x \mapsto f(2^{-j}x)$ appartient à V_{j+1}, c'est-à-dire, tous les espaces V_j sont obtenus par la dilatation des fonctions d'un espace unique (par exemple V_0).*
4. *Il existe ϕ, (appelée fonction d'échelle), telle que la famille $\{x \mapsto \phi(x-k)\}_{k\in\mathbb{Z}}\}$ soit une base orthonormée de V_0.*

   Cette propriété suppose l'existence d'une fonction qui permet de construire la base de V_0 par translation entière.

Définition 12 *(Base de Riesz)(Voir R. Douas [12])*

Une famille $(\phi_k : k \in \mathbb{Z}) \subset L_2(\mathbb{R})$ est une base de Riesz de $L_2(\mathbb{R})$ si

1. *$\forall f \in L_2(\mathbb{R}), \exists \alpha \in \mathfrak{l}^2(\mathbb{Z})$ [2], α étant unique tel que, $f = \sum_{k\in\mathbb{Z}} \alpha_k \phi_k$.*

2. $\mathfrak{l}^2(\mathbb{Z})$ est l'espace des suites complexes absolument sommables

2. $0 < A \leq B < +\infty$*, tels que pour tout* $f = \sum_{k\in\mathbb{Z}} \alpha_k \phi_k \in L_2(\mathbb{R})$*,*

$$A\|\alpha\|_{\mathbb{Z}}^2 \leq \sum_{k\in\mathbb{Z}} |\alpha_k|^2 \leq B\|\alpha\|_{\mathbb{Z}}^2,$$

où $\|\alpha\|_{\mathbb{Z}}$ *est la norme usuelle dans* $L_2(\mathbb{Z})$*. Aussi, on obtient que la norme de* $L_2(\mathbb{R})$ *appliquée à* f *donne le même résultat suivant*

$$\|f\|_2 = \left(\sum_{k\in\mathbb{Z}} |\alpha_k|^2\right)^{\frac{1}{2}}.$$

Par conséquent, cette définition dit qu'à chaque degré de résolution j, la famille de fonctions $\{\phi_{j,k} : x \mapsto 2^{j/2}\phi(2^j x - k)\}_{k\in\mathbb{Z}}$ forme une base orthonormée de l'espace V_j pour la norme L_2. Comme j appartient à V_0, qui est inclus dans V_1. Il peut être exprimé comme un groupe linéaire de $\{(\phi_{1,k})\}_{k\in\mathbb{Z}}$. Autrement dit, il existe une suite de réels $(c_k)_{k\in\mathbb{Z}}$ telle que

$$\forall x \in \mathbb{R}, \phi(x) = \sum_{k\in\mathbb{Z}} c_k \phi(2x - k).$$

Cette relation, dite relation à deux échelles permet de développer des algorithmes de décomposition ou de reconstruction rapide, dans un cadre d'analyse multi-boucles. La possibilité d'améliorer la connaissance d'une fonction en augmentant le niveau de résolution sans recalculer tous les coefficients associés est certainement très utile.

On introduit les espaces de détails retenus pour passer d'une résolution j à une résolution $j + 1$. Donc on ajoute des détails, compris dans l'espace W_j complémentaire de V_j dans V_{j+1}, pour tout $j \in \mathbb{Z}$. On a

$$V_{j+1} = V_j \bigoplus W_j.$$

Par cette définition, ainsi que par la définition (11) et pour tout n appartenant à $\mathbb{Z}$, l'espace $L_2(\mathbb{R})$ vérifie

$$L_2(\mathbb{R}) = \overline{V_j \bigoplus \bigoplus_{n=j}^{+\infty} W_n}.$$

Il existe une fonction ψ de sorte que $\{x \mapsto \psi(x - k)_{k\in\mathbb{Z}}\}$ soit une base orthonormée de W_0. La fonction ψ est alors appelée ondelette mère.

Comme nous l'avons vu précédemment, à tous les niveaux de la résolution $j \in \mathbb{Z}$, la famille $\{\psi_{j,k} : x \mapsto 2^{j/2}\psi(2^j x - k)\}_{k\in\mathbb{Z}}$ forme une base orthonormée de l'espace W_j. De même, une relation à deux échelles peut être établie comme W_0 est inclus dans V_1, il existe une suite de réels $(c_k)_{k\in\mathbb{Z}}$ telle que

$$\forall x \in \mathbb{R}, \psi(x) = \sum_{k\in\mathbb{Z}} c_k \phi(2x - k).$$

L'ondelette de Haar est l'ondelette dite la plus simple et qui représente un exemple d'analyse, créé par la fonction d'échelle $\phi = \mathbf{1}_{[0,1[}$ et l'ondelette $\psi = \mathbf{1}_{[1/2,1[} - \mathbf{1}_{[0,1/2[}$.

1.3.3 Approximations de fonctions

D'après ce qui à été dit plus haut, on conclut que pour tout $j' \in \mathbb{Z}$ et en utilisant la décomposition de l'espace $L_2(\mathbb{R})$, toute fonction f appartenant à $L_2(\mathbb{R})$ s'écrit comme suit

$$f = \sum_{k\in\mathbb{Z}} \alpha_{j',k}\phi_{j',k} + \sum_{j=j'}^{+\infty}\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k},$$

où $\alpha_{j',k} = \int f\phi_{j',k}$ et $\beta_{j,k} = \int f\psi_{j,k}$.

Définition 13 *Une analyse multirésolution orthogonale est dite d'ordre K si pour tout $j \in \mathbb{Z}$, le polynôme P de degré inférieur à $K-1$ appartient à V_j, peut s'écrire sous la forme suivante*

$$P = \sum_{k\in\mathbb{Z}} c_{j,k}\phi_{j,k}.$$

L'ordre d'une Analyse Multirésolution orthogonale est équivalent au nombre de moments nuls de l'ondelette associée. L'intérêt d'une décomposition multi-échelle est que, contrairement à une décomposition dans une base de Fourier, elle est localisée dans le temps et la fréquence. Les valeurs des coefficients de détails sont faibles lorsque la fonction est régulière, mais elles deviennent élevées dans le voisinage des points de discontinuité.

On sait que l'ondelette de Haar, a un inconvénient majeur en n'ayant qu'un seul moment nul. Il est donc généralement préférable de prendre des

ondelettes ayant un nombre élevé de moments nuls. Ainsi Daubechies (1992) a fourni des ondelettes au nombres des moments nuls supérieurs à un autre qui sera utilisé dans les applications. Une base des ondelettes peut être construite sur $L_2([0,1])$ par les fonctions périodiques ϕ et ψ.

Définition 14 *Soient $\tilde{\phi}_k$ et $\tilde{\psi}_k$ des ondelettes périodiques définies par*

$$\tilde{\phi}_k = \sum_{l\in\mathbb{Z}} \tilde{\phi}(x+l),$$

et

$$\tilde{\psi}_k = \sum_{l\in\mathbb{Z}} \tilde{\psi}(x+l).$$

Alors, le couple $(\tilde{\phi}, \tilde{\psi})$ engendre une analyse multirésolution orthonormée sur $[0,1]$.

Cette périodicité est un moyen évident de restreindre l'analyse multirésolution à un intervalle, mais son principal inconvénient est que les problèmes de discontinuités aux bords de l'intervalle en résultent.
Toute fonction $f \in L_2([0,1])$ peut alors se décomposer sous la forme

$$f = \sum_{k=0}^{2^{j'}-1} \alpha_{j',k}\tilde{\phi}_{j',k} + \sum_{j=j_0}^{+\infty}\sum_{k=0}^{2^j-1} \beta_{j,k}\tilde{\psi}_{j,k},$$

avec $\alpha_k = \int f\tilde{\phi}$ et $\beta_k = \int f\tilde{\psi}$, la restriction des indices k aux ensembles $\{0,1,....,2^j-1\}$ en raison de la périodicité des fonctions étudiées.

1.4 Exemples d'ondelettes

Il existe plusieurs types d'ondelettes dans la littérature. Le critère de sélection de la meilleure ondelette reste à déterminer. Puisque dire qu'une ondelette est meilleure que l'autre n'est pas possible car tout dépend de la fonction ou de l'application à déterminer. Dans certains cas, l'ondelette la plus simple dite de Haar sera optimale. Pour d'autres applications et après la présentation des différents types d'ondelettes. On abouti à dire que ce sera le pire des choix possibles.

1.4.1 L'ondelette de Haar

Cette ondelette est définie par

$$\psi(x) = \begin{cases} -1 & \text{si } x \in [0, \frac{1}{2}[\\ 1 & \text{si } x \in [\frac{1}{2}, 1[\\ 0 & \text{sinon} \end{cases} .$$

Et

$$\phi(x) = \begin{cases} 1 & \text{si } x \in [0, 1[\\ 0 & \text{sinon} \end{cases} .$$

Leurs dilatées et translatées sont déterminées par

$$\phi_{j,k}(x) = \sqrt{2^j}\phi(2^j x - k) \text{ et } \psi_{j,k}(x) = \sqrt{2^j}\psi(2^j x - k).$$

Pour $J \geqslant 0$. La famille de Haar $\{\phi_{j,k}\}_{0 \leq k < 2^J} \bigcup \{\psi_{j,k}\}_{j \geqslant J, 0 \leq k < 2^j}$ est une base orthonormée de $L_2([0,1])$.

D'après la figure (1.3), on voit bien que cette ondelette est une fonction

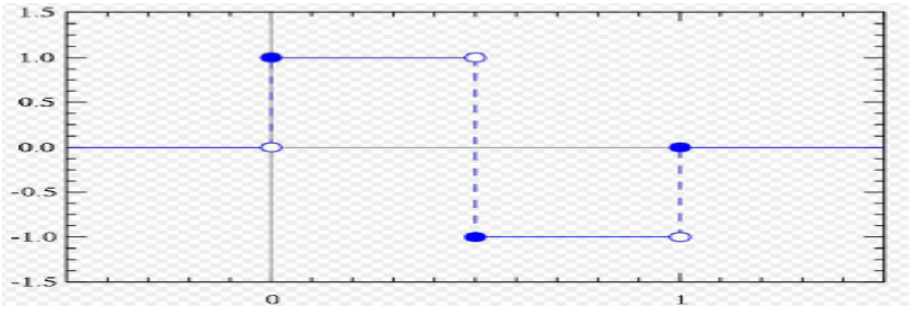


FIGURE 1.3 – Ondelettes Haar

étagée discontinue. D'où l'inconvénient de son utilisation.

1.4.2 Ondelette de Morlet

Cette ondelette est définie par

$$\psi(x) = \cos(5t) \exp^{\frac{x^2}{2}} .$$

La représentation graphique de cette fonction est donnée par la figure (1.4). Malheureusement, elle n'est pas normalisée et ne répond pas aux critères d'admissibilité. Cependant la valeur de $\widehat{\psi}(0)$ est de l'ordre de 10^{-5}, donc on peut la considérer comme presque nulle.

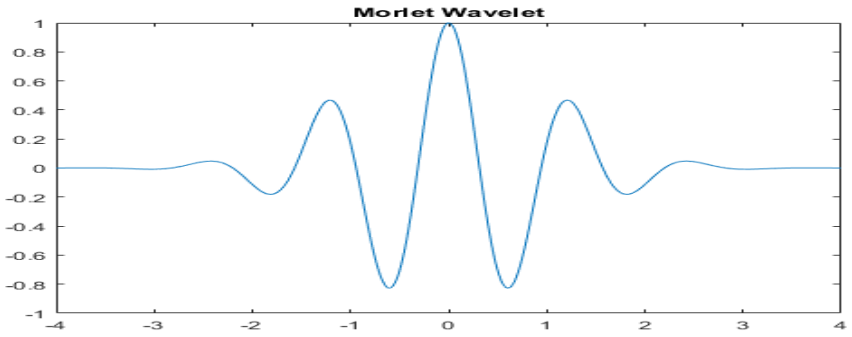


Figure 1.4 – Ondelettes Morlet

1.4.3 Ondelette de Meyer

Comme l'ondelette de Haar, l'ondelette de Meyer est définie à l'aide de

$$\psi(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \sin(\frac{\pi}{2}\phi(\frac{3|x|}{2\pi} - 1)) \exp^{\frac{x}{2}} & \text{si } \frac{2\pi}{3} < |x| < \frac{4\pi}{3} \\ \frac{1}{\sqrt{2\pi}} \cos(\frac{\pi}{2}\phi(\frac{3|x|}{2\pi} - 1)) \exp^{\frac{x}{2}} & \text{si } \frac{4\pi}{3} < |x| < \frac{8\pi}{3} \end{cases},$$

avec

$$\phi(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 < x < 1 \\ 1 & \text{si } x > 1 \end{cases}.$$

La représentation graphique de cette fonction est donnée par la figure (1.5).

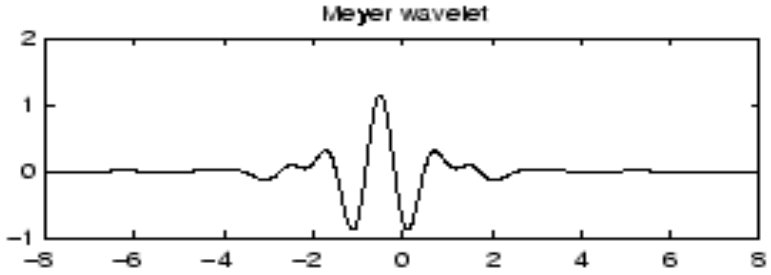


Figure 1.5 – Ondelettes Meyer

1.4.4 Ondelette de Daubechies

L'ondelette de Daubechies est la plus célèbre famille d'ondelettes orthonormales. Ses ondelettes sont généralement définies par le nombre de coefficients non nuls $c_k = \sqrt{2}h_k$ et des formules

$$\phi(2^{-(p+1)}x) = \sqrt{2}\sum_{l\in\mathbb{Z}} h_k \phi(2^{-p}x - l),$$

pour la fonction échelle et

$$\psi_j = \sqrt{2}\sum_{l\in\mathbb{Z}} (-1)^{1-l} h_{1-l}\phi(2^j - l),$$

pour l'ondelette mère.

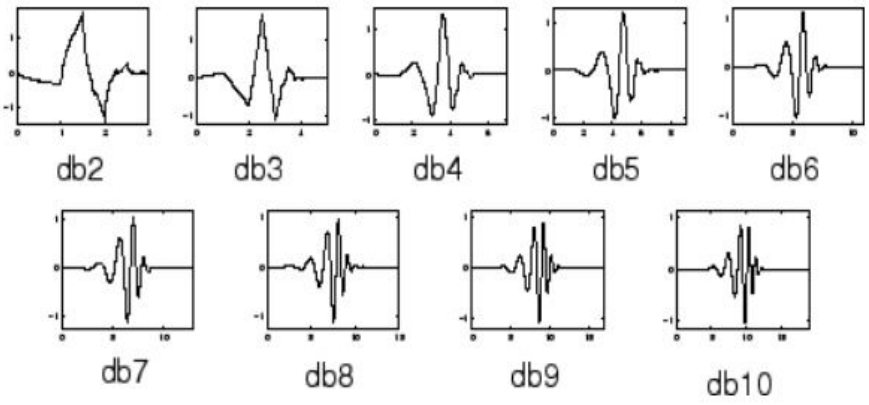


FIGURE 1.6 – Ondelettes Daubechies

1.4.5 Ondelette de Symmlet

Les symmlets sont des ondelettes presque symétriques proposées par Daubechies comme modifications de la famille db. Les propriétés des deux familles d'ondelettes sont similaires. Voici les fonctions d'ondelettes mères.

Dans la suite, nous nous s'intéresserons aux variables aléatoires ainsi que leurs mesures empirique (μ_n). Donc tout ce qui a été dit plus haut sera adapté aux espaces mesurables.

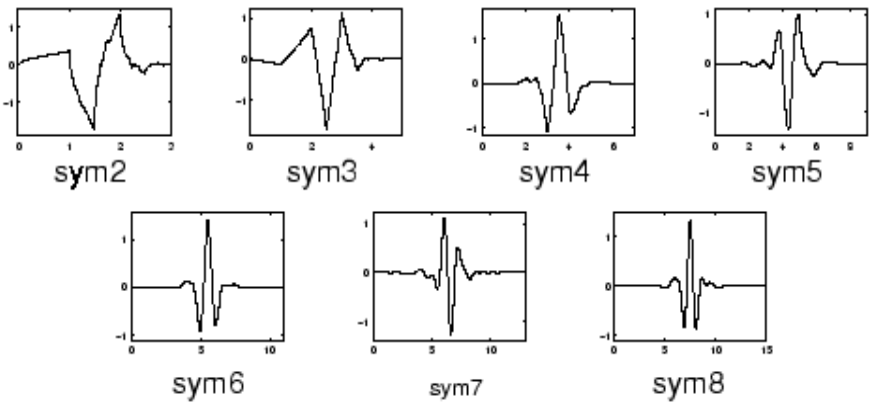


Figure 1.7 – Ondelettes Symmlet

1.5 Orthogonalisation empirique des polynômes par morceaux

Soient $X_1, ..., X_n$ des variables aléatoires. Notons leurs différentes valeurs par $x_1, ..., x_n$ et définissons un système orthonormé dans $L_2(\mu_n)$, en orthonormant des polynômes par morceaux, c'est-à-dire

$$n' \leq n, \quad \{x_1, ..., x_{n'}\} = \{x_1, ..., x_n\},$$

avec $x_1 < ... < x_{n'}$ et μ_n est la mesure empirique associer à $X_1, ..., X_n$.

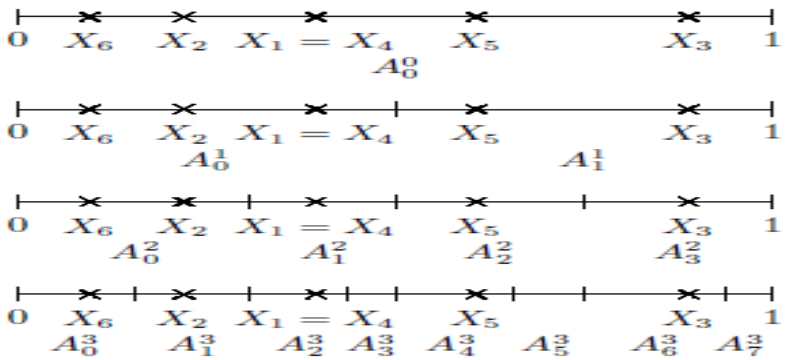


Figure 1.8 – Exemple pour la construction de $\mathcal{P}_l$ ($l \in \{0, 1, 2, 3\}$).

D'après la figure (1.8), nous commençons par définir les partitions $\mathcal{P}^l$ de $[0,1]$ ($l \in \mathbb{N}$) chaque $\mathcal{P}^l$ se compose de 2^l intervalles $A_0^l, ..., A_{2^l-1}^l$. Selon $x_1, ..., x_{n'}$ les partitions sont définie récursivement en fixant $A_0^0 = [0,1]$ et $\mathcal{P}^0 = \{A_0\}$.
Étant donné $\mathcal{P}^l = \{A_0^l, ..., A_{2^l-1}^l\}$ et $\mathcal{P}^{l+1} = \{A_0^{l+1}, ..., A_{2^l-1}^{l+1}\}$ qui est obtenue par la subdivision de chaque intervalle A_j^l en deux intervalles $A_{2j}^{l+1}, A_{2j+1}^{l+1}$, de sorte que chacun de ces deux intervalles contiennent presque le même nombre de $x_1, ..., x_{n'}$, c'est-à-dire

$$A_j^l = A_{2j}^{l+1} \cup A_{2j+1}^{l+1}, A_{2j}^{l+1} \cap A_{2j+1}^{l+1} = \emptyset$$

et

$$|card\{i : x_i \in A_{2j}^{l+1}\} - card\{i : x_i \in A_{2j+1}^{l+1}\}| \leq 1.$$

Cela est toujours possible parce que les $x_1, ..., x_{n'}$ sont distincts.
En utilisant ces partitions imbriquées $\mathcal{P}^0, \mathcal{P}^1, ...$ Nous définissons les espaces imbriqués de polynômes par morceaux, notés par $V_0^M, V_1^M, ..., où\ M \in \mathbb{N}$ représente le degré des polynômes.
Soit V_l^M l'ensemble de tous les polynômes par morceaux de degré inférieur ou égale à M par rapport à $\mathcal{P}^l$, c'est-à-dire

$$V_l^M = \left\{ f(x) = \sum_{j=0}^{2^l-1} \sum_{k=0}^{M} c_{j,k} x^k . \mathbf{1}_{A_j^l}(x) : c_{j,k} \in \mathbb{R} \right\}.$$

La somme par rapport à j dépend de la partition et celle de k dépend du dégré du polynôme. Il est évident que, $V_0^M \subseteq V_1^M \subseteq V_2^M \subseteq ....$ Nous allons construire une base orthonormée pour $({V_{\lceil \log_2(n) \rceil}^M}^3)$ dans $L_2(\mu_n)$.

Pour faire cela, nous décomposons d'abord V_{l+1}^M en une somme orthogonale d'espaces $U_{l+1,0}^M, ..., U_{l+1,2^l-1}^M$, c'est-à-dire que nous construisons les espaces orthogonales $U_{l+1,0}^M, ..., U_{l+1,2^l-1}^M$ avec la propriété que l'ensemble de toutes les fonctions de la forme $\sum_{j=0}^{2^l-1} f_j$ avec $f_j \in U_{l+1,j}^M$ soient égales à V_{l+1}^M. Donc tout $f \in V_{l+1}^M$ peut être écrite comme suit

$$f = \sum_{j=0}^{2^l-1} f_j \quad où \quad f_j = f\mathbf{1}_{A_j^l}\ et\ f \in V_{l+1}^M.$$

Du fait que, les supports des $f_0, ..., f_{2^l-1}$ sont tous disjoints ceci implique que ces fonctions sont orthogonales par rapport au produit scalaire. D'où,

$$V_{l+1}^M = \bigoplus_{j=0}^{2^l-1} U_{l+1,j}^M \quad avec \quad U_{l+1,j}^M = \{f\mathbf{1}_{A_j^l} : f \in V_{l+1}^M\}$$

3. $\lceil \log_2(n) \rceil$partie entière supérieure du nombre réel $\log_2(n) = \frac{\log(n)}{\log 2}$

est la décomposition orthogonale de V_{l+1}^M.

Maintenant, soit $\mathcal{B}_{l+1,j}^M$ la base orthonormée

$$V_l^M \cap U_{l+1,j}^M = \{f\mathbf{1}_{A_j^l} : f \in V_l^M\}$$

$$= \left\{\sum_{k=0}^M c_{j,k}x^k.\mathbf{1}_{A_j^l} : a_0, ..., a_M \in \mathbb{R}\right\}$$

sur

$$U_{l+1,j}^M = \{f\mathbf{1}_{A_j^l} : f \in V_{l+1}^M\},$$

c'est-à-dire que, $B_{l+1,j}^M$ est la base d'ensemble de toutes les fonctions de $U_{l+1,j}^M$ qui sont orthogonales sur $V_l^M \cap U_{l+1,j}^M$ une telle base orthonormée peut être calculée facilement. Supposer que g est un élément du complément orthogonal de $V_l^M \cap U_{l+1,j}^M$ sur $U_{l+1,j}^M$ donc $g \in U_{l+1,j}^M$ ce qui implique

$$g(x) = \sum_{k=0}^M c_k x^k \mathbf{1}_{A_{2j}^{l+1}}(x) + \sum_{k=0}^M c_k x^k \mathbf{1}_{A_{2j+1}^{l+1}}(x) \quad (x \in [0,1]) \tag{1.3}$$

avec $c_0, ..., c_M, b_0, ..., b_M \in \mathbb{R}$. En outre g est orthogonal à $V_l^M \cap U_{l+1}^M$, ce qui est équivalent à supposons que g est orthogonal à

$$1\mathbf{1}_{A_j^l},\ x\mathbf{1}_{A_j^l},\ x^M\mathbf{1}_{A_j^l},$$

par rapport à $<,>_n$. Cela conduit à un système d'équation linéaire homogène pour les coefficients $c_0, ..., c_M, b_0, ..., b_M$ de g. Par conséquent, toutes les fonctions du complément orthogonal de $V_l^M \cap U_{l+1}^M$ sur $U_{l+1,j}^M$ peuvent être calculées par la résolution d'un système d'équations linéaires et la base orthonormé de ce complément orthogonal est obtenue par l'orthonormalisation de la solution de ce système par rapport au produit scalaire $<,>_n$. Posons maintenant

$$B_{l+1}^M = B_{l+1,0}^M \cup ... \cup B_{l+1,2^l-1}^M.$$

Donc il est facile de voir que B_{l+1}^M est une base orthonormé du complément orthogonal de V_l^M sur V_{l+1}^M. On choisit arbitrairement une base orthonormé B_0^M pour V_0^M. Alors

$$B = B_0^M \cup ... \cup B_{\lceil \log_2(n) \rceil}^M \tag{1.4}$$

est une base orthonorméE de $V_{\lceil \log_2(n) \rceil}^M$ sur $L_2(\mu_n)$.

Soit $\mathcal{P}$ une partition arbitraire de $[0,1]$ constituée d'intervalles. La principale propriété du système orthonormé $\{f_j\}_{j=1,...,K}$ défini ci-dessus est que tout polynôme par morceaux de degré inférieur où égale à M par rapport à $\mathcal{P}$ peut être représenté dans $L_2(\mu_n)$ par une combinaison linéaire de pas plus que $card(\mathcal{P})$ des f_j. Pour mieux comprendre, le lemme suivant est donné.

Lemme 1 *(Voir L. Györfi [20])*

Soit $f_1, ..., f_K$ la famille de fonctions construite au dessus.

(a) *Chaque f_j est un polynôme par morceaux de degré inférieur où égale à M par rapport à une partition constituée de quatre intervalles ou moins.*

(b) *Soit $\mathcal{P}$ une partition finie de $[0,1]$ constituée d'intervalles, et soit f un polynôme par morceaux de degré inférieur où égale à M par rapport à cette partition $\mathcal{P}$. Alors il existe des coefficients $c_0, ..., c_K \in \mathbb{R}$ tel que*

$$f(X_i) = \sum_{j=1}^{K} c_j f_j(X_i) avec\ i = 1, \ldots, n$$

et

$$card\{j : c_j \neq 0\} \leq (M+1)(\lceil \log_2(n) \rceil + 1).card(\mathcal{P})$$
$$\leq 2(M+1)(log(n)+1)card(\mathcal{P}).$$

Tel que card représente le nombre d'élément d'un ensemble.

Preuve 7 *Pour la démonstration, Nous renvoyons à l'ouvrage de L. Györfi et al [20].*

CHAPITRE 2

RÉGRESSION NON PARAMÉTRIQUE PAR M.C SUR LES ONDELETTES

Dans le présent chapitre, et en premier lieu, nous exposerons l'idée utilisée pour estimer la fonction de régression, en générale. Ce qui revient à minimiser l'erreur d'approximation. Aussi, nous présenterons les critères utilisés pour choisir le meilleurs estimateurs pour des données complètes. nous présenterons ensuite le passage de l'estimation de la fonction de régression par la méthode des moindres carrées sur un espace quelconque. Et enfin, nous présenterons cette méthode pour la classe des fonctions ondelettes.

2.1 Analyse de régression et l'erreur L_2

Dans l'analyse de régression, un vecteur aléatoire (X, Y) à valeur $\mathbb{R}^d \times \mathbb{R}$ avec $\mathbf{E}(Y^2) < \infty$ est pris en compte et la dépendance de Y sur la valeur de X apporte des informations qui nous intéressent. Cette analyse se fait à l'aide de l'optimisation d'une distance entre les observations et une fonction objective ou théorique de cette étude. Plus précisément, le but est de trouver une fonction $f : \mathbb{R}^d \to \mathbb{R}$ telle que $f(X)$ est une $\ll$ bonne approximation $\gg$ de Y. Dans la suite, notre principal objectif est de minimiser l'erreur

moyenne quadratique où le risque L_2 donné par

$$\mathbf{E}|f(X) - Y|^2. \tag{2.1}$$

Il est connu que la fonction optimale est d'une fonction de régression $m : \mathbb{R}^d \to \mathbb{R}$, définies par l'espérance conditionnelle donnée par $m(x) = \mathbf{E}(Y|X = x)$. En effet, pour f une fonction arbitraire (mesurable) sachant que X est de distribution notée μ, on a

$$\begin{aligned}\mathbf{E}|f(X) - Y|^2 &= \mathbf{E}|f(X) - m(X) + m(X) - Y|^2 \\ &= \mathbf{E}|f(X) - m(X)|^2 + \mathbf{E}|m(X) - Y|^2.\end{aligned}$$

Nous obtenons,

$$\mathbf{E}|f(X) - Y|^2 = \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mu(dx) + \mathbf{E}|m(X) - Y|^2. \tag{2.2}$$

Ici, la deuxième équation découle de

$$\mathbf{E}[(f(X) - m(X))(m(X) - Y)] = \mathbf{E}\left[(f(X) - m(X)\mathbf{E}\left((m(X) - Y)|X\right)\right] = 0$$

Puisque l'intégrale à gauche de (2.2) est toujours non négative, (2.2) implique que la fonction de régression est le prédicteur optimal en vue de minimiser le risque L_2, donc on écrit

$$\mathbf{E}|m^*(X) - Y|^2 = \min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbf{E}|f(X) - Y|^2.$$

En outre, toute fonction f est un bon prédicteur dans le sens où son risque L_2 est proche de la valeur optimale, si et seulement si l'erreur L_2

$$\int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mu(dx)$$

est assez petite. Ceci motive à mesurer l'erreur causée par l'utilisation d'une fonction f au lieu de la fonction de régression par l'erreur L_2.

La fonction de régression est utilisée pour la prédiction, de sorte que son usage est commun dans de nombreux domaines de la vie et voila un exemple dans le domaine de la finance.

Exemple 2.1 : Gestion des prêts (Voir L. Györfi et al [20])
Une banque est intéressée par la prédiction du rendement Y d'un prêt accordé à un client. La banque dispose du profil X du client, ses antécédents de crédit, ses actifs, sa profession, son revenu, son âge, etc. Le rendement prévu influe sur la décision d'admettre ou de refuser un prêt ainsi que les conditions du prêt.

2.1.1 Concepts de l'estimation non paramétrique par MC et l'erreur L_2

Dans les applications, habituellement la distribution de (X, Y) (et donc aussi la fonction de régression) est inconnue. Mais il est souvent possible d'observer un échantillon de la distribution sous-jacente. Ceci mène au problème d'estimation de la régression. Ici $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ sont des vecteurs aléatoires indépendants et distribués de façon identique avec $E(Y^2) < \infty$. Soit D_n l'ensemble des données défini par

$$D_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}.$$

L'objectif est de construire une estimation

$$m_n(.) = m_n(., D_n) : \mathbb{R}^d \to \mathbb{R}$$

de la fonction de régression telle que l'erreur L_2

$$\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mu(dx)$$

est assez petite. Pour une introduction détaillée à la régression non paramétrique, nous renvoyons le lecteur à l'ouvrage de L. Györfi et al. (2002)[20]. En général, les estimations ne sont pas égales à la fonction de régression. Pour comparer différentes estimations, nous avons besoin d'un critère d'erreur qui mesure la différence entre la fonction de régression et l'estimateur m_n. Dans la littérature, plusieurs critères d'erreur distincts sont utilisés. Nous citons les plus utilises dans la littérature.

— Le premier critère est donné par

$$|m_n(x) - m(x)|.$$

Il représente l'erreur ponctuelle pour tout $x \in \mathbb{R}^d$.

— Le second critère est donné par

$$\int_C |m_n(x) - m(x)|^p \mu(dx).$$

Il représente l'erreur L_p, où C est un compact de $\mathbb{R}^d$. En général, la valeur utilisée dans la littérature est $p = 2$.

Un des points clés qu'on voudrait souligner est que la raison d'introduction de la fonction de régression conduit naturellement au critère d'erreur L_2 pour mesurer l'efficacité de l'estimateur de la fonction de régression. Rappelons que m minimise le risque L_2 dite approximation optimale. On

sait que pour montrer que l'estimateur m_n vérifie la relation (2.2), il doit s'écrire à l'aide des observations D_n comme suit

$$\mathbf{E}\{|m_n(X)-Y|^2|D_n\} = \int_{\mathbb{R}^d} |m_n(x)-m(x)|^2\mu(dx) + \mathbf{E}|m(X)-Y|^2. \quad (2.3)$$

Si $\int_{\mathbb{R}^d} |m_n(x)-m(x)|^2\mu(dx)$ tend vers zéro, alors le risque L_2 de l'estimateur m_n est proche de la valeur optimal.
Nous allons maintenant définir les modes de convergence des estimateurs de régression que nous étudierons. La première est la plus faible propriété qu'une estimation devrait avoir et que, â mesure que la taille de l'echantillon augmente, elle devrait converger vers la quantité estimée, c'est-à-dire que l'erreur de l'estimation devrait converger vers zéro pour une taille de l'èchantillon tendant vers l'infini. Les estimateurs qui admettent cette propriété sont dits cohérents. Pour mesurer l'erreur d'une estimation de régression, nous utiliserons l'erreur L_2 donnée par

$$\int_{\mathbb{R}^d} |m_n(x)-m(x)|^2\mu(dx).$$

L'estimation m_n dépend de la donnée D_n, donc l'erreur L_2 est une variable aléatoire.

Définition 15

Soit $\{m_n\}_{n\in\mathbb{N}}$ une suite de v. a. estimateurs de la fonction de régression. On dit que $\{m_n\}_{n\in\mathbb{N}}$ est faiblement consistante pour une certaine distribution de (X,Y), si

$$\lim_{n\to\infty} \mathbf{E}\int |m_n(x)-m(x)|^2\mu(dx) = 0.$$

Ce qui veut dire que la moyenne de l'erreur L_2 tend vers zéro.

Définition 16

Soit $\{m_n\}_{n\in\mathbb{N}}$ une suite de v. a. estimateurs de la fonction de régression. On dit que $\{m_n\}_{n\in\mathbb{N}}$ est fortement consistante pour une certaine distribution de (X,Y) si

$$\mathbf{P}\left\{\lim_{n\to\infty}\int |m_n(x)-m(x)|^2\mu(dx) = 0\right\} = 1.$$

Définition 17

Une suite de v. a. d'estimateurs de la fonction de régression $\{m_n\}$ *est appelée faiblement (resp. fortement) consistante universellement si elle est faiblement (resp. fortement) consistante pour toutes les distributions de* (X, Y) *avec* $\mathbf{E}(Y^2) < \infty$.

Dans le domaine non paramétrique, l'utilisation de cette dernière définition est importante du fait que la loi du couple (X, Y)est inconnue.

2.2 Estimation de la fonction de régression par la M.C

Dans cette partie, on utilise la méthode des moindres carrées pour construire une fonction de régression non paramétrique, également appelée modélisation globale. On étudie également la consistance de cet estimateur.

2.2.1 Construction de l'estimateur

L'idée de base pour construire une estimation de la fonction de régression lorsque nous devons réduire l'erreur L_2 de toutes les fonctions mesurables de R à R, et que la fonction de régression m dépend de la loi du couple (X, Y) est inconnue pour résoudre ce problème est due à l'idée d'estimer le risque de L_2 par le risque empirique qui s'écrit

$$\frac{1}{n}\sum_{i=1}^{n}|f(X_i) - Y_i|^2,$$

qui sera réduit au minimum. L'idée fondamentale de minimiser les risques empiriques en contribuant à toutes les fonctions mesurables est déraisonnable. Pour résoudre ce problème, nous choisissons la catégorie de fonction $\mathcal{F}_n$, qui dépend des données par la taille de l'échantillon. Donc, ces fonctions réduisent le risque empirique sur $\mathcal{F}_n$, et cela signifie que nous déterminons un estimateur des moindres carrés m_n par

$$\hat{m}_n(x) = \arg\min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(X_i) - Y_i|^2. \tag{2.4}$$

Dans la plus part des applications, $\mathcal{F}_n$ est définie comme l'ensemble des combinaisons linéaires de fonctions de base. Donc le fait qu'elle soit uniformément bornée veut dire ici, que les coefficients de la combinaison linéaire vérifiant des conditions plus faciles à utiliser. Dans le cas où, on n'a pas besoin de la condition $\ll$ uniformément bornée $\gg$, alors le calcul de la fonction qui minimise le risque L_2 est plus facile et on obtient comme estimateur la fonction suivante

$$\sum_{j=1}^{d} |Y_j - a_j f_{j,n}(X_i)|^2 = \inf_{(b_1,...,b_d)\in\mathbb{R}^d} \sum_{j=1}^{d} |Y_j - b_j f_{j,n}(X_i)|^2, \qquad (2.5)$$

où $(f_{j,n})$ sont des fonctions de base et $i = 1, ..., n, j = 1, ..., d$. Dans ce cas la solution de cette équation existe. Il reste à résoudre le système d'équations pour avoir la fonction qui minimise le risque empirique L_2. Un exemple qui assure l'existence de ce minimum se trouve dans L. Györfi et al [20].
Dans un théorème de Kohler et autres [28]. Il a été démontré qu'une classe de fonctions d'une forme série linéaire, donne un estimateur moindres carrés consistant d'une manière universellement forte sous certaines conditions et d'une manière universellement faible sous d'autres conditions.

2.2.2 Performances des estimateurs

Nous allons maintenant étudier la convergence de l'estimateur de la fonction de régression. Nous avons besoin d'un estimateur qui converge vers la quantité estimée lorsque la taille de l'échantillon devient suffisamment grande. D'une autre manière, l'erreur L_2 devrait tendre vers zéro lorsque l'estimateur de cette exigence est vérifié. On dit que l'estimation est « consistante ». Puisque $\hat{m}_n$ dépend de la fonction de régression, nous utiliserons l'erreur L_2, qui est une variable aléatoire, pour mesurer la différence de mesure dans l'estimation de la fonction de régression. Nous nous intéresserons aussi à la convergence en moyenne et presque sûre de cette variable vers zéro.

La classe de fonctions $\mathcal{F}_n$ affecte l'erreur d'estimation de deux manières. Premièrement, s'il n'est pas trop riche, le risque empirique se rapproche du risque L_2 de manière uniforme sur $\mathcal{F}_n$. Ainsi, l'erreur produite en minimisant le risque empirique devient faible. En revanche, comme $\hat{m}_n$ est dans $\mathcal{F}_n$, il ne peut pas dépasser en performance le meilleur choix dans $\mathcal{F}_n$. Ceci est formulé dans le lemme suivant.

Lemme 2 *(Voir L. Györfi et al [20])*

Soit $\hat{m}_n$ un estimateur vérifiant la relation (2.4)*, alors*

$$\int |\hat{m}_n(x) - m(x)|^2 \mu(dx)$$

$$\leq 2 \sup_{f \in \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 - \mathbf{E}\{(f(X) - Y)^2\} \right|$$

$$+ \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mu(dx).$$

Preuve 8 *D'après la formule* (2.3) *on a*

$$\begin{aligned} \int |\hat{m}_n(x) - m(x)|^2 \mu(dx) &= \mathbf{E}\{|\hat{m}_n(X) - Y|^2 | D_n\} - \mathbf{E}|m(X) - Y|^2 \\ &= \left(\mathbf{E}\{|\hat{m}_n(X) - Y|^2 | D_n\} - \inf_{f \in \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 \right) \\ &\quad + \left(\inf_{f \in \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 - \mathbf{E}|m(X) - Y|^2 \right). \end{aligned}$$

Selon l'équation (2.2), le deuxiéme terme du membre de droite de l'égalité précédente est

$$\inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mu(dx).$$

Il reste le premier terme. On a

$$\begin{aligned} &\mathbf{E}\{|\hat{m}_n(X) - Y|^2 | D_n\} - \inf_{f \in \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 \\ = &\sup_{f \in \mathcal{F}_n} \left\{ \mathbf{E}\{|\hat{m}_n(X) - Y|^2 | D_n\} - \frac{1}{n} \sum_{i=1}^n |\hat{m}_n(X_i) - Y_i|^2 \right. \\ &+ \frac{1}{n} \sum_{i=1}^n |\hat{m}_n(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 \\ &\left. + \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 - \mathbf{E}|f(X) - Y|^2 \right\}. \end{aligned}$$

De la définition de $\hat{m}_n$, il advient donc

$$\frac{1}{n} \sum_{i=1}^n |\hat{m}_n(X_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 \leq 0.$$

Donc

$$\begin{aligned}&\mathbf{E}\{|\hat{m}_n(X)-Y|^2|D_n\}-\inf_{f\in\mathcal{F}_n}\mathbf{E}|f(X)-Y|^2\\ \leq&\sup_{f\in\mathcal{F}_n}\left\{\mathbf{E}\{|\hat{m}_n(X)-Y|^2|D_n\}-\tfrac{1}{n}\textstyle\sum_{i=1}^n|\hat{m}_n(X_i)-Y_i|^2\right.\\ &\left.+\tfrac{1}{n}\textstyle\sum_{i=1}^n|f(X_i)-Y_i|^2-\mathbf{E}|f(X)-Y|^2\right\}.\end{aligned}$$

Il en résulte que

$$\begin{aligned}&\mathbf{E}\{|\hat{m}_n(X)-Y|^2|D_n\}-\inf_{f\in\mathcal{F}_n}\mathbf{E}|f(X)-Y|^2\\ \leq&\,2\sup_{f\in\mathcal{F}_n}\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n|f(X_i)-Y_i|^2-\mathbf{E}|f(X)-Y|^2\right|.\end{aligned}$$

Cette dernière quantité est l'erreur d'estimation qui représente la différence entre le risque L_2 de l'estimateur et le meilleur risque L_2 qui peut être obtenu dans la famille $\mathcal{F}_n$. La quantité

$$\inf_{f\in\mathcal{F}_n}\int|f(x)-m(x)|^2\mu(dx). \tag{2.6}$$

Elle se traduit comme étant l'erreur d'approximation.
Il suffit donc de montrer que les deux erreurs tendent vers zéro, pour obtenir un estimateur consistant.

Le choix d'une famille $\mathcal{F}_n$ pour avoir l'erreur d'approximation qui tend vers zéro est assez simple. Il se traduit par le fait que $\bigcup_n\mathcal{F}_n$ est dense dans L_2. Mais pour l'erreur d'estimation. Il est plus difficile de prouver la consistance de cette estimation. Donc, il est plus facile d'avoir une fonction bornée d'où le choix de tronquer notre estimateur. C'est-à-dire que pour $\hat{m}_n$ défini par

$$\hat{m}_n\in\mathcal{F}_n\text{ et }\frac{1}{n}\sum_{i=1}^n|\hat{m}_n(X_i)-Y_i|^2=\underset{f\in\mathcal{F}_n}{min}\frac{1}{n}\sum_{i=1}^n|f(X_i)-Y_i|^2.$$

On tronque $\hat{m}_n$ pour obtenir $\bar{m}_n$, pour $|Y|\leq B_n$ p.s., comme suit

$$\bar{m}_n(x)=\mathbb{T}_{B_n}\hat{m}_n(x).$$

Où B_n est une constante qui dépend de la taille de l'échantillon.

2.3 L'espace des ondelettes et l'estimateur M.C de la fonction de régression

Dans cette partie, nous nous s'intéressons à l'estimation des séries orthogonales en utilisant les estimations des coefficients d'un développement en série et cela pour la reconstitution de la fonction de régression. Plus particulièrement à l'estimation par des séries orthogonales non linéaires, où l'on applique une transformation non linéaire (seuil : terme déjà rencontré au chapitre précédent) aux coefficients estimés. Nous commençons notre étude d'estimateurs par séries orthogonales en donnant la motivation d'utilisation de ces estimations par ondelettes.

Par la suite, nous résumerons les notions vues au chapitre "Ondellettes" mais en adaptant nos écritures pour faciliter la lecture.

2.3.1 Estimations en séries orthogonales

On introduit les estimateurs de séries orthogonales dans le contexte de l'estimation de la fonction de régression avec un plan fixe équidistant. C'est le domaine où elles ont été appliquées avec le plus de succès. Là on obtient des données $(x_1, Y_1), ..., (x_n, Y_n)$ selon le modèle

$$Y_i = m(x_i) + \epsilon_i, \tag{2.7}$$

où les x_i sont des points fixes (non aléatoires) et équidistants sur $[0, 1]$, les variables ϵ_i sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d), centrées et $\mathbf{E}(\epsilon_1^2) < \infty$.
Supposons que $m \in L_2(\lambda)$ où λ désigne la mesure de Lebesgue sur $[0, 1]$ et soit $\{f_j\}_{j\in\mathbb{N}}$ une base orthonormée dans $L_2(\lambda)$, i.e.,

$$< f_j, f_k >_\lambda = \int f_j(x) f_k(x) d\lambda(x) = \delta_{j,k}, \ \ (j, k \in \mathbb{N}),$$

où $\delta_{j,k}$ représente le symbole delta de Kronecker.
Chaque fonction de $L_2(\lambda)$ peut être bien approchée arbitrairement par des combinaisons linéaires des $\{f_j\}_{j\in\mathbb{N}}$. Ensuite m peut être représenté par sa série de Fourier par apport à la base $\{f_j\}_{j\in\mathbb{N}}$

$$m = \sum_{j=1}^{\infty} c_j f_j \ où \ \ c_j =< m, f_j >_\lambda = \int_0^1 m(x) f_j(x) dx. \tag{2.8}$$

Dans l'estimation en série orthogonale, nous utilisons les estimations des coefficients du développement en série (2.8) pour reconstruire la fonction de régression.

Nous avons aussi, $x_1, ..., x_n$ qui sont équidistants dans $[0, 1]$. Les coefficients c_j peuvent être estimés par

$$\hat{c}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i f_j(x_i) \quad (j \in \mathbb{N}). \tag{2.9}$$

D'aprés la formule (2.7)

$$\hat{c}_j = \frac{1}{n}\sum_{i=1}^{n} m(x_i) f_j(x_i) + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f_j(x_i),$$

où, on voudrais bien que,

$$\frac{1}{n}\sum_{i=1}^{n} m(x_i) f_j(x_i) \approx \int_0^1 m(x) f_j(x) dx = c_j$$

et

$$\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f_j(x_i) \approx \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f_j(x_i)\right\} = 0.$$

La méthode traditionnelle est d'utiliser ces coefficients estimés pour construire une estimation $\hat{m}_n$ de m. De tronquer le développement en série à un indice $\tilde{K}$, puis à injecter (to plug in) les coefficients estimés $\hat{c}_j$ pour obtenir l'écriture de l'estimateur suivante

$$m_n^1 = \sum_{j=1}^{\tilde{K}} \hat{c}_j f_j. \tag{2.10}$$

Ici, nous essayons essaie de choisir $\tilde{K}$ tel que l'ensemble des fonctions $\{f_1, ..., f_{\tilde{K}}\}$ est le "meilleur" parmi tous les sous-ensembles $\{f_1\}, \{f_1, f_2\}$, $\{f_1, f_2, f_3, ...\}$ de $\{f_j\}_{j\in\mathbb{N}}$ en considèrant l'erreur de l'estimation (2.10). Ceci suppose implicitement que les informations les plus importantes sur m sont contenues dans les premiers coefficients $\tilde{K}$. Donoho et Johnstone (1994)[9] ont proposé un moyen de surmonter cette hypothèse. Il consiste à délimiter les coefficients estimés, par exemple pour utiliser tous les coefficients dont la valeur absolue est supérieure à un certain seuil θ_n (appelé seuillage dur). Cela conduit à des estimations de la forme

$$m_n^2 = \sum_{j=1}^{K} \eta_{\theta_n}(c_j) f_j, \tag{2.11}$$

où K est généralement beaucoup plus grand que $\tilde{K}$ dans (2.10), $\theta_n > 0$ est un seuil et

$$\eta_{\theta_n}(c) = \begin{cases} c & \text{si } |c| > \theta_n \\ 0 & \text{si } |c| \leq \theta_n \end{cases}. \tag{2.12}$$

Comme nous le verrons dans la sous section (2.3.3), nous essayerons de trouver le meilleur de tous les sous-ensembles de $\{f_1, ..., f_K\}$ au vu de l'estimation (2.10).

Le choix le plus populaire pour le système orthogonal $\{f_j\}_{j\in\mathbb{N}}$ sont les soi-disant systèmes d'ondelettes, où les f_j sont construites par la translation d'une ondelette père et par la translation et la dilatation d'une ondelette mère. Pour ces systèmes, le développement en série de la forme (2.8), donne quelques coefficients non nuls pour de nombreuses fonctions. Ceci avec le choix d'un sous-ensemble du système orthonormal par seuil dur.
Il conduit à des estimations qui atteignent un taux de convergence minimax presque optimal pour une variété d'espaces fonctionnels (par exemple, Hölder, Besov, etc.). En particulier, ces estimations peuvent s'adapter aux irrégularités locales (par exemple, les sauts de discontinuités) de la fonction de régression.
Motivées par le succès de ces estimations pour la régression par plan fixe, des estimations similaires ont également été appliquées pour la régression par plan aléatoire, où l'on a des données i.i.d. $(X_1, Y_1), ..., (X_n, Y_n)$. La difficulté à surmonter ici est de trouver une façon raisonnable d'estimer les coefficients c_j. Si X est uniformément réparti sur [0, 1], alors on peut utiliser la même estimation que pour les points équidistants et fixes $x_1, ..., x_n$ tel que

$$\hat{c}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i f_j(x_i),$$

puisque, dans ce cas,

$$\mathbf{E}\{\hat{c}_j\} = \mathbf{E}\{\mathbf{E}\{Y_1 f_j(X_1)|X_1\}\} = \mathbf{E}\{m(X_1)f_j(X_1)\} = c_j.$$

De toute évidence, il ne s'agit pas d'une estimation raisonnable si X n'est pas réparti uniformément sur $[0, 1]$. Dans ce cas, il a été suggéré dans la littérature d'utiliser les données $(X_1, Y_1), ..., (X_n, Y_n)$ pour construire de nouvelles données équidistantes $(x_1, Y_1), ..., (x_n, Y_n)$, où $x_1, ..., x_n$ sont équidistants dans $[0, 1]$ et Y_i est une estimation de $m(x_i)$, puis on applique (2.9) à ces nouvelles données. Les résultats concernant les taux de convergence de ces estimations n'ont été calculés que dans l'hypothèse où X a une densité par rapport à la mesure de Lebesgue-Borel, qui est délimitée

à l'infini par $[0,1]$. Si cette hypothèse est vraie, alors l'erreur L_2 peut être limitée par des temps constants

$$\int_{[0,1]} |\hat{m}_n(x) - m(x)|^2 dx.$$

Le dernier terme peut être exprimé comme la somme des carrés des coefficients de l'extension en série de $\hat{m}_n - m$ par rapport au système orthonormal dans $L_2(\lambda)$. Par conséquent, si l'on estime correctement les coefficients du développement en série de m, cela conduira automatiquement à des estimations avec une petite erreur L_2. Ce n'est plus vrai si μ n'est pas "proche" de la distribution uniforme. On ignore ensuite si une estimation presque correcte des coefficients conduit à une petite erreur L_2, parce que dans le terme $\int |\hat{m}_n(x) - m_(x)|^2 \mu(dx)$, on intégre par rapport à μ et non par rapport à λ.

Nous allons maintenant traiter le cas où la distribution d'échantillon est quelconque, Nous appliquerons l'estimation de séries orthogonales empiriques donnée ci-dessous.

2.3.2 Estimations en Séries Orthogonales Empiriques

Si μ n'est pas "proche" de la distribution uniforme, une approche naturelle consiste à estimer une expansion orthonormale de m en $L_2(\mu)$. De toute évidence, cela n'est pas possible car μ (c'est-à-dire que la distribution de X) est inconnue dans une application. Ce que nous ferons par la suite est d'utiliser un développement en série orthonormée de m dans $L_2(\mu_n)$ plutôt que dans $L_2(\lambda)$, où μ_n est la mesure empirique de $X_1, ..., X_n$, c'est-à-dire

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_A(X_i) \quad (A \subseteq \mathbb{R}).$$

On appelle les estimations résultantes " estimations empiriques de séries orthogonales ".

Pour $f, g : [0,1] \to \mathbb{R}$ définissons

$$< f, g >_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) g(X_i) \quad et \quad \|f\|_n^2 = < f, f >_n .$$

Nous décrirons une façon de construire un systéme orthonormé $\{f_j\}_{j=1,...,K}$ dans $\mathrm{L}_2(\mu_n)$, c-à-d des fonctions $f_1, ..., f_K$ qui satisfont

$$< f_j, f_k >_n = \delta_{j,k} \ (j, k = 1, ..., K).$$

Étant donné un systéme orthonormé, la meilleure approximation de m par rapport a $\|.\|_n$ par des fonctions du $span\{f_1, ..., f_K\}$ est donnée par

$$\sum_{j=1}^{K} c_j f_j \quad où \quad c_j =< m, f_j >_n = \frac{1}{n}\sum_{i=1}^{n} m(X_i) f_j(X_i). \tag{2.13}$$

Nous estimons les coefficients de (2.13) par

$$\hat{c}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i f_j(X_i), \tag{2.14}$$

Nous utiliserons " le seuillage dur" pour construire l'estimateur de m suivant

$$m_n^2 = \sum_{j=1}^{K} \eta_{\theta_n}(\hat{c}_j) f_j \tag{2.15}$$

où $\delta_n > 0$ est un seuillage et η_{δ_n} est défini par (2.12). Finalement nous tronquons l'estimateur à une hauteur β_n indépendante des données, c'est-à-dire que nous définissons ainsi

$$\bar{m}_n(x) = (T_{\beta_n}\hat{m}_n)(x) = \begin{cases} \beta_n & \text{si } \hat{m}_n(x) > \beta_n \\ \hat{m}_n(x) & \text{si } -\beta_n \leq \hat{m}_n(x) \leq \beta_n \\ -\beta_n & \text{si } \hat{m}_n(x) < -\beta_n \end{cases} \tag{2.16}$$

où $\beta_n > 0$ et $\beta_n \to \infty$ $(n \to \infty)$.

Finalement, nous injectons par la suite, cet espace dans l'écriture de l'estimateur de la fonction de régression par la méthode de moindres carrées.

2.3.3 Lien avec les estimations des moindres carrées

Donc en utilisant $\{f_j\}_{j=1,...,K}$ la famille de fonctions $f_j : \mathbb{R} \longrightarrow \mathbb{R}$. Pour $J \subseteq \{1, ..., K\}$ définissons $\mathcal{F}_{n,J}$ l'ensemble des combinaisons linéaires des fonctions $f_j (j \in J)$, c-à-d

$$\mathcal{F}_{n,J} = \left\{ \sum_{j \in J} c_j f_j : c_j \in \mathbb{R} \quad (j \in J) \right\}. \tag{2.17}$$

Rappelons que l'estimateur des moindres carrés $\hat{m}_{n,J}$ de m sur $\mathcal{F}_{n,J}$ est défini par

$$\hat{m}_{n,J} \in \mathcal{F}_{n,J} \quad et \quad \frac{1}{n}\sum_{i=1}^{n} |\hat{m}_{n,J}(X_i) - Y_i|^2 = \min_{f \in \mathcal{F}_{n,J}} \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{2.18}$$

En utilisant (2.17), $\hat{m}_{n,J}$ peut être réécrit comme suit

$$\hat{m}_{n,J} = \sum_{j \in J} c_j^* f_j,$$

pour des $c^* = \{c_j^*\}_{j \in J} \in \mathbb{R}^{|J|}$ satisfaisant

$$\frac{1}{n}\|\mathbf{B}c^* - \mathbf{Y}\|_2^2 = \underset{c \in \mathbb{R}^{|J|}}{min} \frac{1}{n}\|\mathbf{B}c - \mathbf{Y}\|_2^2, \tag{2.19}$$

où

$$\mathbf{B} = (f_j(X_i))_{1 \leq i \leq n, j \in J} \quad et \quad \mathbf{Y} = (Y_1, ..., Y_n)^T.$$

On peut montrer que la formule (2.19) est équivalente à

$$\frac{1}{n}\mathbf{B}^T\mathbf{B}c^* = \frac{1}{n}\mathbf{B}^T\mathbf{Y}, \tag{2.20}$$

Elle est dite équation normale du problème des moindres carrées.

Nous Considérons le cas où $\{f_j\}_{j=1,...,K}$ est une base orthonormé de $L_2(\mu_n)$ qui vérifie

$$\frac{1}{n}\mathbf{B}^T\mathbf{B} = (< f_j, f_k >_n)_{j,k \in J} = (\delta_{j,k \in J}).$$

Par conséquent la solution de (2.20) est donnée par

$$c_j^* = \frac{1}{n}\sum_{i=1}^{n} f_j(X_i)Y_i \quad (\forall j \in J). \tag{2.21}$$

Soit l'ensemble

$$\hat{J} = \{j \in \{1, ..., K\} : |c_j^*| > \theta_n\} \tag{2.22}$$

qui représente les coefficients touchés par le seuillage dure θ_n. Ensuite, l'estimation de série orthogonale ${m^2}_n$ définie par (2.15) satisfaisant $m_n^2 = {m^2}_{n,\hat{J}}$ et de mêmes propriétés que l'estimateur de m sur $\mathcal{F}_{n,\hat{J}}$. Alors le seuillage dur peut être considéré comme un moyen de choisir l'un des 2^K estimateurs des moindres carrés ${m^2}_{n,J}$ $(J \subseteq \{1, ..., K\})$.
Rappelons que nous avons montré dans le chapitre précédent que le systéme (1.3) forme une base orthonormé de l'espace L_2. Ainsi, dans la suite, nous étudions la consistance de notre estimateur induit par séries orthogonales.

2.3.4 Efficacité

Pour simplifier on considère le cas où $X \in [0,1]$ p.s. Il est facile de modifier la définition de notre estimateur de façon à obtenir un estimateur faiblement et fortement consistant universellement pour le cas uni-varié. Soient $\alpha \in (0, \frac{1}{2})$, les fonctions f_j et les coefficients $\hat{c}_j$ définies précédemment. Notons par $(\hat{c}_{(1)}, f_{(1)}), ..., (\hat{c}_{(K)}, f_{(K)})$ la permutation de $(\hat{c}_1, f_1), ..., (\hat{c}_K, f_K)$ avec

$$|\hat{c}_{(1)}| \geq |\hat{c}_{(2)}| \geq ... \geq |\hat{c}_{(K)}|. \tag{2.23}$$

Définissons l'estimateur m_n^3 par

$$m_n^3 = \sum_{j=1}^{min\{K, \lfloor n^{1-\alpha} \rfloor\}} \eta_{\theta_n}(\hat{c}_{(j)}) f_{(j)}. \tag{2.24}$$

Cela garantit que m_n^3 soit une combinaison linéaire d'au plus $n^{1-\alpha}$ des fonctions f_j. Et comme il a été déjà donné, on peut montrer que (2.24) implique que

$$m_n^3 = m_{n,J^*}^3 \quad avec \quad J^* \subseteq \{1, ..., K\}, \tag{2.25}$$

où, J^* satisfaisant $|J^*| \leq n^{1-\alpha}$. Et

$$\frac{1}{n} \sum_{i=1}^{n} |m_{n,J^*}^3(X_i) - Y_i|^2 + pen_n(J^*)$$

$$= \min_{\substack{J \subseteq \{1,...,K\} \\ |J| \leqslant n^{1-\alpha}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} |m_{n,J}^3(X_i) - Y_i|^2 + pen_n(J) \right\}, \tag{2.26}$$

avec, $pen_n(J) = c_n \frac{|J|}{n}$ $(J \subseteq \{1, ...K\})$ et $c_n > 0$.

Toutes ces modifications sur l'estimateur vont nous permettre de montrer le théoréme suivant.

Théorème 7 *(cf théorème 18.1 dans L. Györfi et al. [20])*

Soient $M \in \mathbb{N}$ fixé et m_n l'estimateur de m défini par (2.14),(2.16), (2.23), (2.24), avec $B_n = \log(n)$ et $\theta_n \leq \frac{1}{(\ log(n)+1)^2}$. Alors

$$\int \left| m_n^2(x) - m(x) \right|^2 \mu(dx) \underset{n \to \infty}{\to} 0 \ p.s \ .$$

Preuve 9 *Soit $L > 0$, posons $Y_L = T_L Y$, $Y_{1,L} = T_L Y_1, ..., Y_{n,L} = T_L Y_n$. Soit $\mathcal{F}_n$ l'ensemble de tous les polynômes par morceaux de degré inférieur ou égale à M par rapport à une partition de $[0,1]$ constituée d'au plus $4n^{1-\alpha}$ intervalles, Soit G_M l'ensemble des polynômes de degré inférieur ou égale à M, soit $\mathcal{P}_n$ une partition équidistante de $[0,1]$ dans $\lceil log(n) \rceil$ (partie entière supérieure du nombre réel $log(n)$) intervalles, et notons $G_M \circ \mathcal{P}_n$ l'ensemble de tous les polynômes par morceaux de degré inférieur ou égale à M par rapport à $\mathcal{P}_n$.*

Nous avons également besoin des notations suivantes

$$\mathcal{L}_n^{**} = \mathrm{T}_{\log n}(\mathcal{F}_n).$$

$$\mathcal{F}_n^{**} = \{\forall f \in G_M \circ \mathcal{P}_n, \|f\|_\infty \ \leq log(n)\}.$$

Dans la première étape de la preuve, nous montrons que l'affirmation découle de

$$\inf_{f \in \mathcal{F}_n^{**}} \int |f(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty). \tag{2.27}$$

Et pour tout $L > 0$,

$$\sup_{f \in \mathcal{L}_n^{**}} \left| \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_{i,L}|^2 - \mathbf{E}\{|f(X) - Y_L|^2\} \right| \to 0 \ (n \to \infty) \ p.s. \tag{2.28}$$

Dans la deuxième et troisième étape, nous allons prouver (2.27) et (2.28), respectivement.

Supposons donc temporairement que (2.27) et (2.28) sont vrais. Comme

$$\int |m_n^2(x) - m(x)|^2 \mu(dx) = \mathbf{E}\{|m_n^2(X) - Y|^2 | D_n\} - \mathbf{E}\{|m(x) - Y|^2,$$

il suffit de montrer

$$\{\mathbf{E}\{|m_n^2(X)-Y|^2|D_n\}\}^{\frac{1}{2}} - \{\mathbf{E}\{|m(x)-Y|^2\}^{\frac{1}{2}} \to 0 \quad p.s. \tag{2.29}$$

Nous utilisons la décomposition

$$\begin{aligned} 0 &\leq \{\mathbf{E}\{|m_n^2(X)-Y|^2|D_n\}\}^{\frac{1}{2}} - \{\mathbf{E}\{|m(x)-Y|^2\}^{\frac{1}{2}} \\ &= \left(\{\mathbf{E}\{|m_n^2(X)-Y|^2|D_n\}\}^{\frac{1}{2}} - \inf_{f\in\mathcal{F}_n^{**}}\{\mathbf{E}|f(X)-Y|^2\}^{\frac{1}{2}}\right) \\ &\quad + \left(\inf_{f\in\mathcal{F}_n^{**}}\{\mathbf{E}|f(X)-Y|^2\}^{\frac{1}{2}} - \{\mathbf{E}|m(x)-Y|^2\}^{\frac{1}{2}}\right). \end{aligned} \tag{2.30}$$

Il résulte de (2.27) et de l'inégalité du triangle que le second terme de (2.30) converge vers zéro. Donc pour (2.29) il suffit de montre que

$$\limsup_{n\to\infty}\left(\{\mathbf{E}\{|m_n^2(X)-Y|^2|D_n\}\}^{\frac{1}{2}} - \inf_{f\in\mathcal{F}_n^{**}}\{\mathbf{E}|f(X)-Y_L|^2\}^{\frac{1}{2}}\right) \leq 0 \ p.s. \tag{2.31}$$

Dans la suite on considère d'une manière arbitraire $L>0$ et sans perte de généralité que $log(n) > L$. Alors

$$\begin{aligned} &\{\mathbf{E}\{|m_n^2(X)-Y|^2|D_n\}\}^{\frac{1}{2}} - \inf_{f\in\mathcal{F}_n^{**}}\{\mathbf{E}|f(X)-Y|^2\}^{\frac{1}{2}} \\ &= \sup_{f\in\mathcal{F}_n^{**}}\{\mathbf{E}\{|m_n^2(X)-Y|^2|D_n\}\}^{\frac{1}{2}} - \{\mathbf{E}|f(X)-Y|^2\}^{\frac{1}{2}} \\ &\leq \sup_{f\in\mathcal{F}_n^{**}}\Big\{\{\mathbf{E}\{|m_n^2(X)-Y|^2|D_n\}\}^{\frac{1}{2}} - \{\mathbf{E}\{|m_n^2(X)-Y_L|^2|D_n\}\}^{\frac{1}{2}} \\ &\quad + \{\mathbf{E}\{|m_n^2(X)-Y_L|^2|D_n\}\}^{\frac{1}{2}} - \left\{\frac{1}{n}\sum_{i=1}^n |m_n^2(X_i)-Y_{i,L}|^2\right\}^{\frac{1}{2}} \\ &\quad + \left\{\frac{1}{n}\sum_{i=1}^n |m_n^2(X_i)-Y_{i,L}|^2\right\}^{\frac{1}{2}} - \left\{\frac{1}{n}\sum_{i=1}^n |m_n^3(X_i)-Y_{i,L}|^2\right\}^{\frac{1}{2}} \end{aligned}$$

$$+\left\{\frac{1}{n}\sum_{i=1}^{n}|m_n^3(X_i)-Y_{i,L}|^2\right\}^{\frac{1}{2}}-\left\{\frac{1}{n}\sum_{i=1}^{n}|m_n^3(X_i)-Y_i|^2\right\}^{\frac{1}{2}}$$

$$+\left\{\frac{1}{n}\sum_{i=1}^{n}|m_n^3(X_i)-Y_i|^2\right\}^{\frac{1}{2}}-\left\{\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2\right\}^{\frac{1}{2}}$$

$$+\left\{\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2\right\}^{\frac{1}{2}}-\left\{\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_{i,L}|^2\right\}^{\frac{1}{2}}$$

$$+\left\{\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_{i,L}|^2\right\}^{\frac{1}{2}}-\{\mathbf{E}|f(X)-Y_L|^2\}^{\frac{1}{2}}$$

$$+\{\mathbf{E}|f(X)-Y_L|^2\}^{\frac{1}{2}}-\{\mathbf{E}|f(X)-Y|^2\}^{\frac{1}{2}}\Big\}.$$

Nous donnons maintenant des limites supérieures aux termes de chaque ligne du coté droit de la dernière inégalité.

Le deuxième et septième terme sont délimités ci-dessus par

$$\sup_{f\in\mathcal{L}_n^{**}}\left|\left\{\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_{i,L}|^2\right\}^{\frac{1}{2}}-\{\mathbf{E}|f(X)-Y_L|^2\}^{\frac{1}{2}}\right|.$$

Remarquons que $m_n^2 = T_{log(n)}m_n^3$ *et* $m_n^3 \in \mathcal{F}_{n,J^*} \subseteq \mathcal{F}_n$. *Pour le troisième terme, on voit que si* $x, y \in \mathbb{R}$ *avec* $|y| \leq \log(n)$ *et* $z = T_{log(n)}x$, *alors* $|z-y| \leq |x-y|$. *Par conséquent, le troisiéme terme n'est pas supérieur à zéro.*

Ensuite, nous allons borner le cinquième terme. Posons $f \in \mathcal{G}_M \circ \mathcal{P}_n$. *Par définition de* $\mathcal{P}_n$ *et le lemme (1), il existe* $\bar{J} \subseteq \{1, ..., n\}$ *et* $\bar{f} \in \mathcal{F}_n$, *tels que*

$$f(X_i) = \bar{f}(X_i)\ (i = 1, ..., n) \quad et \quad card\bar{J} \leq 2(M+1)(\log(n)+1)^2.$$

Ceci, en plus (2.26), implique ce qui suit

$$\frac{1}{n}\sum_{i=1}^{n}|m_n^3(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}|m_n^3(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|\tilde{f}(X_i)-Y_i|^2$$

$$\leq pen_n(\tilde{J})$$

$$\leq n\theta_n^2\frac{2(M+1)(\log(n)+1)^2}{n}.$$

En utilisant ces limites supérieures et l'inégalité triangulaire pour les termes restants, on obtient

$$\{\mathbf{E}\{|m_n^2(X)-Y|^2|D_n\}\}^{\frac{1}{2}}-\inf_{f\in\mathcal{F}_n^{**}}\{\mathbf{E}|f(X)-Y_L|^2\}^{\frac{1}{2}}$$

$$\leq 2\{\mathbf{E}|Y-Y_L|^2\}^{\frac{1}{2}}+2.\left\{\frac{1}{n}\sum_{i=1}^{n}|Y_i-Y_{i,L}|^2\right\}^{\frac{1}{2}}+2(M+1)\theta_n^2(\log(n)+1)^2$$

$$+2\sup_{f\in\mathcal{L}_n^{**}}\left|\left\{\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_{i,L}|^2\right\}^{\frac{1}{2}}-\{\mathbf{E}|f(X)-Y_L|^2\}^{\frac{1}{2}}\right|.$$

D'après (2.28), $\theta_n \leq \frac{1}{(\log(n)+1)^2}$ *et la loi forte des grands nombres, il en résulte que*

$$\limsup_{n\to\infty}\left(\{\mathbf{E}\{|m_n^2(X)-Y|^2|D_n\}\}^{\frac{1}{2}}-\inf_{f\in\mathcal{F}_n^{**}}\{\mathbf{E}|f(X)-Y_L|^2\}^{\frac{1}{2}}\right)$$

$$\leq 4\{\mathbf{E}|Y-Y_L|^2\}^{\frac{1}{2}}\quad p.s.$$

Avec $L\to\infty$ *on obtient l'assertion.*

Dans la seconde étape, nous montrons (2.27). Puisque la fonction m peut être approchée arbitrairement dans $L_2(\mu)$ *par des fonctions continuellement différentiables. Nous pouvons supposer que m est continuellement*

différentiable. Pour chaque $A \in \mathcal{P}_n$. Nous choisissons quelques $x_A \in A$ et posons $f^* = \sum_{A\in\mathcal{P}_n} m(x_A)\mathbf{1}_A$. Alors $f \in \mathcal{G}_M \circ \mathcal{P}_n$ et pour n suffisamment grand (c'est-à-dire pour n tel que $||m||_\infty \leq \log(n)$), nous obtenons

$$\inf_{f\in\mathcal{F}_n^{**}} \int |f(X) - m(x)|^2 \mu(dx) \leq \sup_{x\in[0,1]} |f^* - m(x)|^2$$

$$\leq \frac{c}{\log^2(n)} \to 0 \ (n \to \infty),$$

où c est une constante qui dépend de la premiére dérivée de m.

Dans la troisième étape, nous allons montrer (2.28). Posons

$$\mathcal{H}_n := \{h : \mathbb{R}\times\mathbb{R} \to \mathbb{R} : h(x,y) = |f(x) - T_L y|^2 \ ((x,y) \in \mathbb{R}^d \times \mathbb{R}) \text{ pour } f \in \mathcal{L}_n^{**}\}.$$

Pour $h \in \mathcal{H}_n$, on a $0 \leq h(x,y) \leq 4\log(n)^2$ $((x,y) \in \mathbb{R}^d \times \mathbb{R})$. En utilisant la notion de nombre couvrant et le théoréme 9.1 de Györfi and al[20], on conclut

$$\mathbf{P}\left\{\sup_{f\in\mathcal{L}_n^{**}} \left|\frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_{i,L}|^2 - \mathbf{E}|f(X) - Y_L|^2\right| > t\right\}$$

$$= \mathbf{P}\left\{\sup_{h\in\mathcal{H}_n} \left|\frac{1}{n}\sum_{i=1}^n h(X_i, Y_i) - \mathbf{E}h(X,Y)\right| > t\right\}$$

$$\leq 8\mathbf{E}\left\{\mathcal{N}_1\left(\frac{t}{8}, \mathcal{H}_n, (X,Y)_1^n\right)\right\} \exp\left(-\frac{nt^2}{2048\log(n)^4}\right). \tag{2.32}$$

Nous allons ensuite borner le nombre couvrant défini par (2.32). Observons d'abord que si

$$h_j(x,y) = |f_j(x) - T_{\log(n)y}|^2 \ ((x,y) \in \mathbb{R}\times\mathbb{R}),$$

pour certaines fonctions f_j *liées en valeur absolue par* $\log(n)$ $(j = 1, 2)$, *alors*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|h_1(X_i,Y_i)-h_2(X_i,Y_i)| \\
&=\frac{1}{n}\sum_{i=1}^{n}\left|f_1(X_i)-T_{\log(n)}Y_i+f_2(X_i)-T_{\log(n)Y_i}\right||f_1(X_i)-f_2(X_i)| \\
&\leq 4\log(n)\tfrac{1}{n}\textstyle\sum_{i=1}^{n}|f_1(X_i)-f_2(X_i)|\,.
\end{aligned}$$

Ainsi

$$\mathcal{N}_1\left(\frac{t}{8},\mathcal{H}_n,(X,Y)_1^n\right)\leq\mathcal{N}_1\left(\frac{t}{32\log(n)},\mathcal{L}_n^{**},X_1^n\right). \tag{2.33}$$

En utilisant la notion de dimension VC, les nombres de partitionnement, le théorème 9.4 et le lemme 13.1 dans Györfi et al [20], on obtient

$$\begin{aligned}
&\mathcal{N}_1\left(\frac{t}{32\log(n)},\mathcal{L}_n^{**},X_1^n\right) \\
&\leq\Delta_n(\mathcal{P})\left\{\sup_{z_1,\ldots,z_l\in X_1^n,l\leq n}\mathcal{N}_1\left(\frac{1}{32\log(n)},T_{\log(n)}\mathcal{G}_M,z_1^l\right)\right\}^{4n^{1-\alpha}} \\
&\leq\Delta_n(\mathcal{P})\left\{3\frac{6e\log(n)}{\frac{t}{32\log(n)}}\right\}^{2V_{T_{\log(n)}\mathcal{G}_M^+}}4n^{1-\alpha} \\
&=\Delta_n(\mathcal{P})\left\{\frac{576e\log^2(n)}{t}\right\}^{2V_{T_{\log(n)}\mathcal{G}_M^+}}4n^{1-\alpha},
\end{aligned} \tag{2.34}$$

où $\mathcal{P}$ *est l'ensemble de toutes les partitions de* $[0,1]$ *composées d'au plus* $4n^{1-\alpha}$ *intervalles. Ce qui implique*

$$\begin{aligned}
\Delta_n(\mathcal{P})&\leq(n+4n^{1-\alpha})^{4n^{1-\alpha}} \\
&\leq(5n)^{4n^{1-\alpha}}.
\end{aligned} \tag{2.35}$$

De plus, nous pouvons facilement conclure, de la définition de la dimension VC, que

$$V_{T_{\log(n)}\mathcal{G}_M^+} \leq V_{\mathcal{G}_M^+},$$

qui, avec le théorème 9.5(Györfi et al [20]), implique

$$V_{T_{\log(n)}\mathcal{G}_M^+} \leq M + 2. \tag{2.36}$$

Il découle de (2.32) et (2.33) que,

$$\mathbf{P}\left\{ \sup_{f\in\mathcal{L}_n^{**}} \left| \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_{i,L}|^2 - \mathbb{E}|f(X) - Y_L|^2 \right| > t \right\}$$

$$\leq 8(5n)^{4n^{1-\alpha}} \left(\frac{576e\log^2(n)}{t} \right)^{8(M+2)n^{1-\alpha}} \exp\left(-\frac{nt^2}{2048\log^4(n)} \right),$$

à partir de laquelle on obtient l'affirmation par une application du lemme Borel-Cantelli, le résultat s'ensuit.

CHAPITRE 3

ANALYSE DE SURVIE ET DONNÉES CENSURÉES

Ce chapitre représente une synthèse sur l'analyse de survie, en vue de l'utilisation des données censurées. Nous,nous intéresserons spécialement au passage entre l'estimation de la fonction de régression par les méthodes des moindres carrés pour des données complètes aux données censurées. C'est l'introduction sur les données de survie et la censure avec ces types distincts ainsi que l'estimation de la fonction de répartition et de son inverse dans un modèle de censure mixte ce qui est essentielle à la construction de l'estimateur de notre travail.

3.1 La survie et le phénomène de censure

L'analyse de survie est une analyse de temps-à-événement. C'est-à-dire lorsque le résultat d'intérêt est le temps jusqu'à ce qu'un événement se produise. Les exemples de délais sont le temps qui s'écoule avant l'infection, la réapparition d'une maladie ou le rétablissement la santé en médecine, la durée du chômage en économie, le temps qui s'écoule jusqu'à la défaillance d'une partie de la machine ou bien la durée de vie des ampoules en génie. Ainsi, l'analyse de survie fait partie des études de fiabilité dans ce cas. Elle

est habituellement utilisée pour étudier la durée de vie des composants industriels. Pour la fiabilité, les temps de survie sont habituellement appelés temps de défaillance car la variable d'intérêt est le temps pendant lequel un composant fonctionne correctement avant qu'il ne tombe en panne.
L'analyse de survie consiste en des méthodes paramétriques, semi-paramétriques et non paramétriques. Nous pouvons les utiliser pour estimer les mesures les plus couramment utilisées dans les études de survie, les fonctions de survie et de danger, les comparer à de différents groupes et évaluer la relation entre les variables prédictives et le temps de survie dans la littérature. Certaines distributions de probabilités statistiques décrivent assez bien les temps de survie et les distributions les plus utilisées sont les distributions exponentielles, Weibull, lognormal.

Un concept important dans l'analyse de survie est la censure. Les temps de survie de certains individus peuvent ne pas être complètement observés pour différentes raisons. Dans les sciences de la vie, cela peut se produire lorsque l'étude de survie (p. ex., l'essai clinique) prend fin avant que les périodes de survie de tous les individus ne puissent être observées. Ou lorsqu'une personne abandonne l'étude, ou pour des études à long terme, lorsque le patient est perdu pour un suivi. Dans le contexte industriel, tous les composants ne sont pas tombés en panne avant la fin de l'étude de fiabilité. Dans de tels cas, l'individu survit au-delà du temps de l'étude et le temps de survie exact est inconnu. Ceci est appelé censure à droite.

Au cours d'une étude de survie, on observe un échec de l'individu au temps Y, où l'observation sur cet individu cesse au temps R. Ensuite, l'observation est $\min(Y, R)$ aussi une variable indicatrice δ indique si l'individu est censuré ou non. Le calcul des fonctions de répartition ou de survie ou tout autre doivent être ajustées pour tenir compte de la censure.
Parmi les types de censure : La censure de type I, le chercheur détermine le moment de la censure comme la fin de l'étude. Quant au second type, l'étude s'arrête une fois qu'un certain nombre d'événements préalablement identifiés par l'expérimentateur se produisent. Une autre forme de censure aléatoire est que la censure est hors du contrôle des chercheurs.

3.2 Estimation de la fonction de survie

L'objet d'intérêt principal est la fonction de survie, appelée conventionnellement S, qui est définie comme

$$S(t) = \mathbf{P}(Y > t),$$

où t est le temps, Y est une variable aléatoire indiquant l'heure du décès par exemple. Autrement dit, la fonction de survie est la probabilité que l'heure du décès soit postérieure à une certaine période spécifiée t.
Généralement, on suppose $S(0) = 1$, bien qu'il pourrait être inférieur à 1 s'il y a une possibilité de mort ou d'échec à l'instant $t = 0$.

La fonction de survie est non croissante : $S(u) \leq S(t)$ si $u \geq t$. Cette propriété suit directement parce que $Y > u$ implique $Y > t$ pour $u \geq t$. Cela traduit l'idée selon laquelle la survie à un âge avancé n'est possible que si tous les jeunes sont atteints.
On suppose habituellement que la fonction de survie s'approche de zéro $S(t) \longrightarrow 0$ lorsque $t \longrightarrow \infty$).

Dans la sous section qui va suivre on s'intéresse à la construction de l'estimateur de Kaplan et Meier (1958), pour faciliter la compréhension, on donne à la fin un petit exemple.

3.2.1 La méthode d'estimation de Kaplan-Meier

En se basant sur le fait que l'observation est censurée n'est pas liée à la cause de l'échec, Kaplan-Meier ont supposé que le temps de censure est indépendant du temps de survie. Ils ont déduit un estimateur de la fonction de survie très efficaces pour le cas de données censuré à droite et l'ont nommé estimateur produit limité vue son écriture.
Soit $(Y_1, \ldots, Y_n)$, n variables aléatoires indépendantes qui représentent les durées d'intérêt, de fonction de répartition F_Y et les temps de censure $R_1, \ldots, R_n$ qui sont indépendantes des fonction de répartition F_R. Si le modèle est censuré à droite, on observe pas Y_i mais la plus petite des deux valeurs $Z_i = \min(Y_i, R_i)$, ainsi que l'indicateur de censure δ_i équivaut à 1 si la durée d'intérêt est observée et 0 si elle est censurée, i.e. $\delta_i = \mathbf{1}_{\{Y_i \leq R_i\}}$.
Pour cela la fonction de répartition F_Y est estimée par l'estimateur fourni par Kaplan- Meier [23], donné pour $z < Z_{(n)}$ où $Z_{(n)} = \max\{Z_1, \ldots, Z_n\}$ par

$$F_n(z) = 1 - \prod_{i:Z_i \leq z} \left(\frac{N_n(Z_i) - 1}{N_n(Z_i)} \right)^{\delta_i},$$

avec $N_n(x) = \sum_{i=1}^{n} \mathbf{1}_{\{Z_i \geq x\}}$ qui représente le nombre des données qui sont supérieure à x. Si $z \geq Z_{(n)}$, il existe plusieurs accords pour sélectionner $F_n(z)$. Il est défini comme $F_n(Z_{(n)})$, ce qui implique un résultat indésirable que F_n peut ne pas être une fonction de répartition à l'instant où $Z_{(n)}$ est une donnée censurée. Cet estimateur présente des caractéristiques assez semblables à celles de la fonction de répartition empirique, comme la conver-

gence uniforme presque sûr de Stute, Winter et al [37, 43] et la normalité asymptotique de Breslow et al [4, 19].

Exemple 2 *(F.Bouhajera [3])*

Un ensemble de données contenant les dossiers médicaux de 290 patients souffrants d'insuffisance rénale, soit 100 femmes et 190 hommes âgés de 40 à 95 ans. On s'intéresse à l'analyse du temps de survie de ces patients par rapport â l'âge. L'estimateur de Kaplan-Meyer (voir les figures (3.1) (b) et (c)) a été utilisée pour étudier un modèle général de survie qui montrait des taux de mortalité élevés au début, puis une décente graduelle jusqu'â la fin de l'étude.

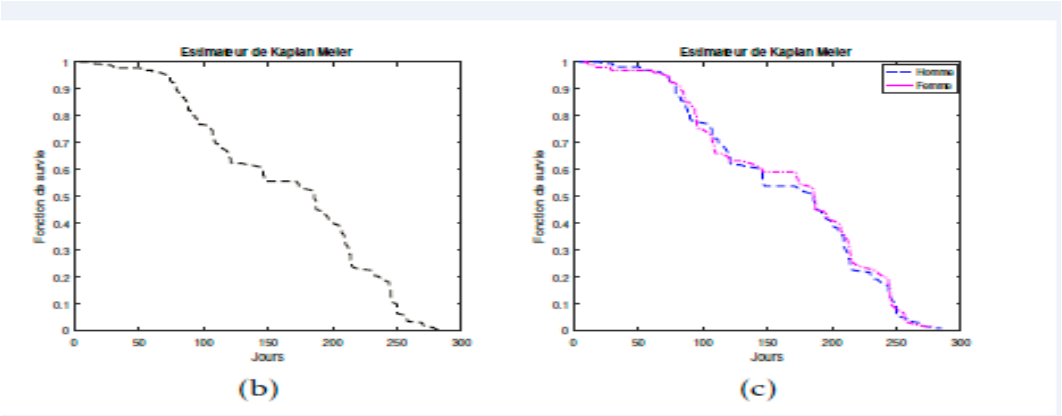


Figure 3.1 – (b) Fonctions de survie pour les patients insuffisants rénaux. (c) Fonctions de survie pour les hommes et les femmes souffrant d'insuffisance rénale.

Dans ce qui suit, nous, nous sommes intéressés à injecter l'estimation de Kaplan-Meyer dans le cas de données censurées mixtes parce que cette nouvelle estimation va interférer avec l'expression de nos estimateurs.

3.3 Estimation de la fonction de survie dans un modèle de censure mixte

Quand la censure à droite et la censure à gauche sont présentes sur un même échantillon conjointement, on se retrouve avec un modèle de censure mixte.

3.3.1 Définition de la censure mixte

Notons par R la variable de censure à droite et L la variable de censure à gauche. Ce modèle est noté modèle I dans Patilea and Rolin (2006) [36], où on étudie un échantillon du couple (Z, A) avec la variable observée $Z = \max(\min(Y, R), L)$ et l'indicateur de censure est donné par

$$A = \begin{cases} 0 & \text{si} \quad L < Y \leq R \\ 1 & \text{si} \quad L < R < Y \\ 2 & \text{si} \quad \min(Y, R) \leq L \end{cases} .$$

Pour ce type de données Kebabi laroussi et Messaci (2011) [24] donnent un estimateur de la fonction de régression par la même méthode en se basant sur des données censurées mixtes. Elles obtiennent sa convergence presque sûre.

3.3.2 Fonction de survie et son estimateur en présence de censure mixte

Soient Y, L et R des variables aléatoires positives indépendantes, de fonctions de répartition F_Y, F_L et F_R, et des fonctions de survie respectivement S_Y, S_L et S_R, où Y c'est la durée d'intérêt L la durée de censure à gauche et R la durée de censure à droite. Dans le modèle I de Patiléa et Rolin [36], rappelons qu'au lieu d'étudier un échantillon de Y nous disposons seulement d'un échantillon du couple (Z, A) où $Z = \max(\min(Y, R), L)$ et

$$A = \begin{cases} 0 & \text{si } L < Y \leq R \\ 1 & \text{si } L < R < Y \\ 2 & \text{si } \min(Y, R) \leq L \end{cases} .$$

Nous pouvons définir la fonction de répartition F_Z de Z, comme suit : $\sum_{k=0}^{2} F_Z^{(k)}(t)$ où

$$F_Z^{(k)}(t) = \mathbf{P}(Z \leq t, A = k), \quad \text{pour } k = 0, 1, 2.$$

Lorsque $R^-(t)$ la limite de R à gauche de t existe, pour toute application R de $\mathbb{R}$ dans $\mathbb{R}$, ces fonctions s'écrivent

$$\begin{aligned}
F_Z^{(0)}(t) &= \int_0^t F_L^-(u) S_R^-(u)\, dF_X(u), \\
F_Z^{(1)}(t) &= \int_0^t F_L^-(u) S_X(u)\, dF_R(u), \\
F_Z^{(2)}(t) &= \int_0^t \{1 - S_X(u) S_R(u)\}\, dF_L(u).
\end{aligned}$$

Nous considérons que $Y = \min(X, R)$ et L dans un modèle de censure à gauche sont d'estimer la fonction de répartition de Y. Ensuite, nous utiliserons pour estimer la fonction de répartition de la variable d'intérêt X en utilisant un modèle de censure à droite.
Pour obtenir l'estimateur de la fonction de survie S_X. En remplaçant les fonctions de $F_Z^{(0)}$, $F_Z^{(1)}$ et $F_Z^{(2)}$ par leurs estimateurs empiriques, comme suit

$$\hat{S}_n(Z'_j) = 1 - F_n(Z'_j) = \prod_{1 \leq l \leq j} \left\{ 1 - \frac{D_{0l}}{U_{l-1} - N_{l-1}} \right\},$$

où les valeurs $(Z'_j)_{\mathbf{1} \leq j \leq M}$ sont distinctes des Z_i dans l'ordre croissant, et

$$D_{kj} = \sum_{\mathbf{1} \leq i \leq n} \mathbf{1}_{\{Z_i = Z'_j, A_i = k\}}, \quad N_j = \sum_{\mathbf{1} \leq i \leq n} \mathbf{1}_{\{Z_i \leq Z'_j\}},$$

$$U_{j-1} = n \prod_{j \leq l \leq M} \left\{ 1 - \frac{D_{2l}}{N_l} \right\}$$

pour $0 \leq l \leq 2$ et $1 \leq j \leq M$.

Si $L \equiv 0$, c'est-à-dire pas de censure à gauche, $\hat{S}_n$ devient l'estimateur de Kaplan-Meier qui lui devient le complément à 1 de la fonction de répartition empirique si $R \equiv \infty$ (pas de censure).

Patiléa et Rolin [36] qui ont présenté cet estimateur et démontré la convergence uniforme presque sûr et la convergence comme un processus vers un gaussien dans les conditions d'identifiabilité du modèle.
Notons que $\hat{S}_n$ plus interfère avec la place dans les expressions d'estimées de la fonction de régression, avec un devoir de faire une exigence nécessaire

et suffisante pour s'annulé.

$$\hat{S}_n(Z'_j) = \prod_{1\leq l\leq j} \left\{1 - \frac{D_{1l}}{U_{l-1} - N_{l-1}}\right\}. \tag{3.1}$$

Cet estimateur est proposé par Patilea et Rolin [36]. En inversant le temps dans l'estimateur de Kaplan-Meir. Nous pouvons actualiser l'estimateur de F_n de F_L (cas de la censure à gauche) qui fournissait par

$$\hat{F}_n(Z'_j) = \prod_{j<l\leq M} \left\{1 - \frac{1_{\{A_j=2\}}}{l}\right\}. \tag{3.2}$$

En utilisant les deux hypothèses H_1 et H_5, on a

$$\sup_{t\in\mathbb{R}^+} \left|\hat{S}_n(t) - S_R(t)\right| \underset{n\longrightarrow\infty}{\longrightarrow} 0 \text{ p.s.} \tag{3.3}$$

Voir Patilea et Rolin [36]. On a aussi

$$\sup_{t\in\mathbb{R}^+} \left|\hat{F}_n(t) - F_L(t)\right| \underset{n\longrightarrow\infty}{\longrightarrow} 0 \text{ p.s.} \tag{3.4}$$

D'après l'hypothèse H_3. Nous trouvons que

$$S_R(T) > 0 \text{ et } F_L(I) > 0. \tag{3.5}$$

Et à l'aide des équations (3.3)–(3.5) et pour n assez grand. Nous déduisons que

$$\hat{S}_n(T) > 0 \text{ et } \hat{F}_n(I) > 0 \text{ p.s.} \tag{3.6}$$

Lemme 3 i) *Une exigence nécessaire et adéquate* $\hat{S}_n(Z'_{k_0}) = 0$ *pour la première fois et reste nul est que* $D_{0,k_0} \neq 0$, $D_{1,k_0} = 0$ *et* $\forall j > k_0, D_{0,j} = D_{1,j} = 0$ *si* $k_0 \neq M$.

ii) $\hat{S}_n$ *s'annule pour la première fois en* Z'_M *si et seulement si* $D_{0,M} \neq 0$ *et* $D_{1,M} = 0$.

Preuve 10 i) *La première étape est de montrer que pour tout* $k : 0 \leq k \leq M-1$, *on a*

$$U_k \geq N_k. \tag{3.7}$$

On sait que

$$\begin{aligned} U_k &= n\left(\frac{N_{k+1} - D_{2,(k+1)}}{N_{k+1}}\right)\left(\frac{N_{k+2} - D_{2,(k+2)}}{N_{k+2}}\right)\cdots\left(\frac{N_M - D_{2,M}}{N_M}\right) \\ &= \left(\frac{N_k + D_{0,(k+1)} + D_{1,(k+1)}}{N_{k+1}}\right)\cdots\left(\frac{N_{M-2} + D_{0,(M-1)} + D_{1,(M-1)}}{N_{M-1}}\right) \\ &\quad \times (N_{M-1} + D_{0,M} + D_{1,M}) \\ &\geq \frac{N_k}{N_{k+1}} \times \cdots \times \frac{N_{M-2}}{N_{M-1}} \times N_{M-1} = N_k. \end{aligned}$$

S'il existe j *tel que* $j > k$ *avec* $D_{1,j} \neq 0$ *ou* $D_{0,j} \neq 0$ *alors* $U_k > N_k$.

ii) *Soit* k_0 *le premier* k *pour lequel* $\hat{S}_n(Z'_k) = 0$ *tel que* $D_{0,k} = U_{k-1} - N_{k-1}$.

On a

$$D_{0,k_0} \neq 0 \text{ et } D_{0,k_0} = U_{k_0-1} - N_{k_0-1}. \tag{3.8}$$

Par ailleurs

$$U_{k_0-1} = n\left(1 - \frac{D_{2,k_0}}{N_{k_0}}\right)\cdots\left(1 - \frac{D_{2,M}}{N_M}\right) = \left(1 - \frac{D_{2,k_0}}{N_{k_0}}\right)U_{k_0}. \tag{3.9}$$

D'après (3.8) *et* (3.9), *il vient*

$$\begin{aligned} D_{0,k_0} + N_{k_0-1} &= \left(\frac{N_{k_0} - D_{2,k_0}}{N_{k_0}}\right)U_{k_0} \\ &= \left(\frac{N_{k_0-1} + D_{0,k_0} + D_{1,k_0}}{N_{k_0}}\right)U_{k_0}. \end{aligned}$$

Et selon (3.7), *nous devons avoir* $D_{1,k_0} = 0$, *donc* $U_{k_0=N_{k_0}}$ *alors* $D_{0,k_0} \neq 0$, $D_{1,k_0} = 0$ *et* $\forall j > k_0, D_{1,j} = D_{0,j} = 0$, *(au-delâ de* k_0,

ce qui montre que l'exigence énoncée est nécessaire. Supposons que $D_{1,k_0} = 0$, $D_{0,k_0} \neq 0$ *et* $\forall j > k_0, D_{1,j} = D_{0,j} = 0$, *et montrons que* $\hat{S}_n(Z'_{k_0}) = 0$. *On a*

$$\begin{aligned} U_{k_0-1} &= n\left(\frac{N_{k_0} - D_{2,k_0}}{N_{k_0}}\right)\left(\frac{N_{k_0+1} - D_{2,(k_0+1)}}{N_{k_0+1}}\right)\cdots\left(\frac{N_M - D_{2,M}}{N_M}\right) \\ &= n\left(\frac{N_{k_0-1} + D_{0,k_0}}{N_{k_0}}\right)\left(\frac{N_{k_0}}{N_{k_0+1}}\right)\cdots\left(\frac{N_{M-1}}{N_M}\right) \\ &= N_{k_0-1} + D_{0,k_0}, \end{aligned}$$

avec $D_{0,k_0} \neq 0$, *où* $\hat{S}_n(Z'_{k_0}) = 0$.

Autrement dit, si la première donnée est censurée à gauche le dénominateur sera indéfini et de même si la dernière donnée est censurée à droite.

CHAPITRE 4

RÉGRESSION NON PARAMÉTRIQUE PAR M.C DANS UN MODÈLE DE CENSURE MIXTE SUR LES ONDELETTES

Ce chapitre représente une combinaison et une synthèse de tout ce qui à été dit dans les chapitres précédents. A, cause de la lourdeur des notations et de leur masse importante, nous avons trouvé intéressant de réécrire les notations utilisées précédemment pour mieux facilité la lecture. C'est-à-dire que nous avons introduit un estimateur de la fonction de régression des moindres carrés pour Y censuré à droite par R et $\min(Y, R)$ censurée à gauche par L comme il a été défini dans le chapitre 3. Il est basé sur des idées dérivées du contexte d'estimations par ondelettes et est construit par un seuillage dur des estimateurs des coefficients d'un développement en série de la fonction de régression. Nous avons établi la convergence en norme L_2 et cela, en donnant suffisamment de critères pour la cohérence de cet estimateur. Le résultat montre que notre estimateur est capable de s'adapter à la régularité locale de la fonction de régression et de la distribution associées.

4.1 Principe de l'estimation et hypothèses

Soit Y un vecteur aléatoire prenant des valeurs dans $\mathbb{R}$ et soit X une variable aléatoire de $\mathbb{R}^d$. Posons R et L des variables aléatoires positives de censure. Plus précisément, l'objectif est de trouver une fonction $f : \mathbb{R}^d \to \mathbb{R}$ telle que $f(X)$ soit une "bonne approximation" de Y.

4.1.1 Modèle

Nous introduisons des estimations orthogonales en série de $m(x) = \mathbf{E}(Y|X = x)$ en fonction de l'échantillon formé d'observations iid $\mathcal{D}_n = \{X_i,\ Z_i = \max(\min(Y_i, R_i), L_i),\ A_i\}$ de même loi que (X, Z, A) ou $Z = \max(\min(Y, R), L)$ et

$$A = \begin{cases} 0 & \text{si} \quad L < Y < R \\ 1 & \text{si} \quad L < R \leq Y \\ 2 & \text{si} \quad \min(Y, R) \leq L \end{cases} .$$

En effet, soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction mesurable. Nous dénoterons la distribution de X par μ aurons alors

$$\begin{aligned} \mathbf{E}|f(X) - Y|^2 &= \mathbf{E}|f(X) - m(X) + m(X) - Y|^2 \\ &= \mathbf{E}|f(X) - m(X)|^2 + \mathbf{E}|m(X) - Y|^2. \end{aligned}$$

Finalement, nous obtient

$$\mathbf{E}|f(X) - Y|^2 = \mathbf{E}|m(X) - Y|^2 + \int |f(x) - m(x)|^2 \mu(dx).$$

Par la suite, nous noterons la fonction de répartition de la variable aléatoire V par F_V et sa fonction de survie par $S_V = 1 - F_V$ et

$$T_V = \sup\{t : F_V(t) < 1\} \text{ et } I_V = \inf\{t : F_V(t) \neq 0\},$$

où T_V et I_V sont les points terminaux du support de la variable V.
Supposons que les variables X, Y, R et L vérifient les hypothèses suivantes

H_1 : Y, R et L sont indépendantes.
H_2 : (L, R) est indépendant de (X, Y).
H_3 : $\exists T < T_R$ et $I > I_L$ tel que, $\forall n \in \mathbb{N}, \forall i (1 \leq i \leq n) : A_i = 0 \Rightarrow I \leq Z_i \leq T$ p.s.
H_4 : F_L est continue sur $]0, \infty[$,
H_5 : $T_R \leq T_Y \leq T_L < \infty$ est $I_Y \leq I_L < I_R$.

H_1 est une hypothèse implicite au modèle de Patilea et Rolin [36]. L'hypothèse H_3 nous semble acceptable car $I \leq Z_i \leq T$ lorsque $A_i = 0$. L'hypothèse H_5 garantit notamment que le modèle est identifiable.
Soit h une application de $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, nous proposons comme estimateur sans biais de $\mathbf{E}(h(X,Y))$ la quantité

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{h(X_i,Z_i)}{S_R(Z_i)F_L(Z_i)}. \tag{4.1}$$

Le problème est que les fonctions S_R et F_L sont généralement inconnues, nous les remplacerons respectivement par leurs estimateurs.
Nous poserons $(Z'_j)_{1\leq j\leq M},\ (M \leq n)$ les valeurs distinctes de Z_i classées par ordre croissant.

4.1.2 Estimation et propriétés

Posons

$$D_{kj} = \sum_{i=1}^{n} 1_{\{Z_i=Z'_j, A_i=k\}}, \text{ et } \ N_j = \sum_{i=1}^{n} 1_{\{Z_i \leq Z'_j\}},$$

Patilea et Rolin [36] proposent d'estimer S_R par

$$\hat{S}_n(t) = \prod_{j/Z'_j \leq t} \left\{1 - \frac{D_{1j}}{U_{j-1} - N_{j-1}}\right\} \text{ et } U_{j-1} = n \prod_{j\leq l\leq M} \left\{1 - \frac{D_{2l}}{N_l}\right\}. \tag{4.2}$$

Nous pouvons en déduire l'estimateur $\hat{F}_n$ à partir de F_L, si en inversant le temps dans l'estimateur de Kaplan-Meir (cas de la censure à gauche) donné par

$$\hat{F}_n(t) = \prod_{j/Z'_j > t} \left\{1 - \frac{1_{\{A_j=2\}}}{j}\right\}. \tag{4.3}$$

Sous les hypothèses H_1 et H_5. Il a été prouvé par Patilea et Rolin [36] que

$$\sup_{t\in\mathbb{R}^+} \left|\hat{S}_n(t) - S_R(t)\right| \underset{n\longrightarrow\infty}{\longrightarrow} 0 \text{ p.s}, \tag{4.4}$$

et

$$\sup_{t\in\mathbb{R}^+} \left|\hat{F}_n(t) - F_L(t)\right| \underset{n\longrightarrow\infty}{\longrightarrow} 0 \text{ p.s.} \tag{4.5}$$

Remarquons que l'hypothèse H_3 implique que

$$S_R(T) > 0 \text{ et } F_L(I) > 0. \tag{4.6}$$

Sous les équations (4.4),(4.5) et (4.6) ; en déduit que, pour n est assez grand

$$\hat{S}_n(T) > 0 \text{ et } \hat{F}_n(I) > 0 \text{ p.s.}$$

Si Y est entièrement observé, l'estimateur de la fonction de régression par les moindres carrés est obtenu en réduisant le risque empirique de L_2 et est donné par

$$\arg\min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2,$$

où $\mathcal{F}_n$ est une classe de fonctions dépendantes de la taille de l'échantillon n. Ainsi, selon la relation h et après avoir estimé S_R et F_L, l'estimateur des moindres carrés de $m(x)$ est donné par

$$\widetilde{m}_n = \arg\min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}1_{\{A_i=0\}}\frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)}\left(\frac{0}{0}:=0\right). \tag{4.7}$$

$\mathcal{F}_n$ est la classe des fonctions ondelettes données par la suite. Nous savons que $\hat{S}_n(Z_i)$ ne s'annule pas dans l'expression de $\widetilde{m}_n$ si $A_i = 0$, Il est facile de vérifier que $\hat{F}_n(Z_i)$ ne s'annule pas non plus si $A_i = 0$. Mais comme Y est bornée, Nous allons l'imposer à notre estimateur et pour cela nous réintroduisons la notation d'application de troncature. Pour $0 \le t < \infty$ et $x \in \mathbb{R}$, définissons

$$T_{[0,t]}(x) = \begin{cases} t & \text{si } x > t \\ x & \text{si } 0 \le x \le t \\ 0 & \text{si } x < 0 \end{cases}.$$

Ainsi pour $f : \mathbb{R}^d \to \mathbb{R}$, définissons $(T_{[0,t]}f)(x) = T_{[0,t]}(f(x))$. Nous pouvons aussi réutiliser le fait que cette application vérifie la relation suivante.

$$\forall b > a, \quad |T_{[0,b]}(x) - T_{[0,a]}(x)| \le (b-a). \tag{4.8}$$

Y étant bornée pour $M_n = max(Z_1,\ldots,Z_n)$ avec $M_n \underset{n\to+\infty}{\to} T_L$ p.s. Nous proposons enfin comme estimateur de $m(x)$

$$m_n(x) = T_{[0,M_n]}(\tilde{m}_n(x)). \tag{4.9}$$

4.1.3 Bases d'ondelettes

Nous introduisons l'estimation orthogonale en série dans le contexte de l'estimation de la fonction de régression avec un plan fixe et équidistant, qui

est le domaine dans lequel elles ont été appliquées avec le plus de succès. Soit les données $(x_1, Y_1), ..., (x_n, Y_n)$ selon le modèle $Y_i = m(x_i) + \varepsilon_i$ où x_i sont des points fixes (non aléatoires) et équidistants $[0, 1]$. Les variables ε_i sont des variables aléatoires indépendantes et identiquement distribuées (iid), centrées et $\mathbf{E}(\epsilon_i^2) < \infty$. m une fonction de régression $m : [0, 1] \to \mathbb{R}$. Supposons que $m \in L_2(\mu)$ où μ est la mesure de Lebesgue sur $[0, 1]$ et que $(f_j)_{j\in\mathbb{N}}$ soit une base orthonormale dans $L_2(\mu)$, c'est-à-dire

$$\langle f_j, f_k\rangle = \int f_j(x)f_k(x)\mu(dx) = \begin{cases} 1 & \text{si} \quad j = k \\ o & \text{si} \quad j \neq k \end{cases} .$$

Chaque fonction de $L_2(\mu)$ peut être approximée arbitrairement par des combinaisons linéaires de $(f_j)_{j\in\mathbb{N}}$. Alors m représentée par son développement en série par rapport à $(f_j)_{j\in\mathbb{N}}$

$$m = \sum_{j=1}^{\infty} c_j f_j \text{ où } c_j = \langle m, f_j\rangle_{L_2(\mu)} = \int m(x)f_j(x)\mu(dx). \tag{4.10}$$

Dans l'estimation en série orthogonale. Nous utilisons des estimations de coefficients de croissance en série (4.1) pour reconstruire la fonction de régression.
Dans le modèle $Y_i = m(x_i) + \varepsilon_i$, où $x_1, \ldots, x_n$ sont équidistants dans $[0, 1]$. Les coefficients c_j peuvent être estimés par

$$\hat{c}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i f_j(x_i), j \in \mathbb{N}. \tag{4.11}$$

La méthode traditionnelle d'utilisation de ces coefficients estimés pour construire une estimation m_n de m est de tronquer le développement en série à un indice K et d'injecter les coefficients estimés.

$$m_n^1 = \sum_{j=1}^{\tilde{K}} \hat{c}_j f_j. \tag{4.12}$$

Ici, nous essayons de choisir $\tilde{K}$ tel que l'ensemble des fonctions $\{f_1, ..., f_{\tilde{K}}\}$ est le "meilleur" parmi tous les sous-ensembles $\{f_1\}, \{f_1, f_2\}, \{f_1, f_2.....\}$ de $\{f_j\}_{j\in\mathbb{N}}$ en considérons l'erreur de l'estimation (4.7). Cela suppose implicitement que les informations les plus importantes sur m sont contenues dans les premiers coefficients $\tilde{K}$ du développement en série (4.1).
Donoho et John-stone [9] ont proposé un moyen de surmonter cette hypothèse. Cela consiste à seuiller les coefficients estimés. Par exemple, on

utilise tous les coefficients dont la valeur absolue est supérieure à un seuil θ_n (revoir l'indice appelé seuillage dur). Cela conduit à des estimations de la forme

$$m_n^2 = \sum_{j=1}^{K} \eta_{\theta_n}(\hat{c}_j) f_j,$$

où K est généralement beaucoup plus grand que $\tilde{K}$ dans (4.12), $\theta_n > 0$ est un seuil et

$$\eta_{\theta_n}(\hat{c_j}) = \begin{cases} \hat{c_j} & \text{si} \quad |\hat{c_j}| > \theta_n \\ o & \text{si} \quad |\hat{c_j}| \leq \theta_n \end{cases} .$$

Nous tronquons l'estimation à une hauteur indépendante des données B_n, c'est-à -dire que nous définissons

$$\bar{m}_n(x) = \mathrm{T}_{B_n}\tilde{m}_n(x) = \begin{cases} B_n & \text{si } \tilde{m}_n(x) > B_n \\ \tilde{m}_n(x) & \text{si } -B_n \leq \tilde{m}_n(x) \leq B_n \\ -B_n & \text{si } \tilde{m}_n(x) < -B_n \end{cases}, \tag{4.13}$$

où $B_n > 0$ et $B_n \to \infty$ $(n \to \infty)$.
Dans cette partie nous étudions la consistance de notre estimateur de séries orthogonales. Pour simplifier nous allons considérer le cas où $X \in [0;1]$ p.s. Il est facile de modifier la définition de notre estimateur de façon à obtenir un estimateur faiblement et fortement consistant universellement pour l'univarié X. Pour montrer la consistance forte de notre estimateur nous avons besoin d'appliquer des modifications sur sa définition. Soit $\alpha \in (0, \frac{1}{2})$. Soient les fonctions f_j et les coefficients $\hat{c}_j$ définies dans les sections (1.5) et (4.1). Notons par $(\hat{c}_{(1)}; f_{(1)}), ..., (\hat{c}_{(K)}; f_{(K)})$ la permutation de $(\hat{c_1}, f_1), ..., (\hat{c_K}, f_K)$ avec

$$|\hat{c}_1| \geq |\hat{c}_2| \geq ..... \geq |\hat{c}_k|. \tag{4.14}$$

Définissons l'estimateur m_n^3 par

$$m_n^3 = \sum_{j=1}^{\min\{k, n^{\lfloor 1-\alpha \rfloor}\}} \eta_{\theta_n}(\hat{c}_j) f_j. \tag{4.15}$$

Cela garantit que m_n^3 et une combinaison linéaire de non plus que $n^{1-\alpha}$ des fonctions f_j. Et comme dans $\mathbf{E}|f(X) - Y|^2 = \mathbf{E}|m(X) - Y|^2 + \int |f(x) - m(x)|^2 \mu(dx)$ on peut montrer que

$$m_n^3 = m_{n,J^*}^3 \text{avec } J^* \subseteq \{1, ..., K\} \text{ où } J^* \text{ satisfaisant } card J^* \leq n^{1-\alpha}.$$

Enfin, Nous combinons la notation des deux estimations pour obtenir comme estimation de $\tilde{m}_n$ les formules suivantes m_n^3 et m_n avec $T_L \leq B_n = \log(n)$. Nous aurons également besoin des notations suivantes

$$\mathcal{L}_n^* = \mathrm{T}_{T_L}(\mathcal{F}_n),$$

et

$$\mathcal{F}_n^* = \{g : \exists f \in G_M \circ P_n, g = \mathrm{T}_{[0,T_L]} f\}.$$

Où $\mathcal{F}_n$ l'ensemble de tous les polynômes par morceaux de degré inférieur ou égal à M par rapport à une partition de $[0,1]$ constituée d'au plus $4n^{1-\alpha}$ intervalles et G_M l'ensemble des polynômes de degré inférieur ou égale à M, nous aurons $\mathcal{P}_n$ une partition équidistante de $[0,1]$ dans $\lceil log(n) \rceil$ intervalles et notée $G_M \circ \mathcal{P}_n$ l'ensemble de tous les polynômes par morceaux de degré inférieur ou égale à M par rapport à $\mathcal{P}_n$.
L'ensemble des graphes de fonctions dans G_M et dont V_G^+ désigne la dimension V.C. Nous sommes maintenant en mesure d'énoncer le résultat suivant relatif à la convergence de l'estimateur m_n proposé. Référence peut être faite au supplément pour quelques définitions et résultats, sur la théorie de Vapnik-Chervonenkis utilisée dans ce travail.

4.2 Résultats

Théorème 8 *(R. Douas et al [13])*

Sous les hypothèses H_1 - H_5*, soit* $M \in \mathbb{N}$ *fixé, et soit* m_n *l'estimateur de* m *défini par (4.11), (4.13) ,(4.14),(4.15), avec* $T_L \leq B_n = \log(n)$ *et* $\theta_n \leq \frac{1}{(\ log(n)+1)^2}$ *. Then*

$$\int |m_n(x) - m(x)|^2 \mu(dx) \underset{n\to\infty}{\to} 0 \ p.s \ .$$

Lemme 4 *Nous définissons la quantité* $\bar{m}_n(x) = \mathbb{T}_{[0,T_L]}(\tilde{m}_n(x))$ *et d'aprés*

les relations (4.2), (4.3), on a

$$\begin{aligned} & \int |\bar{m}_n(x) - m(x)|^2 \mu(dx) \qquad (4.16) \\ \leq \; & 2 \sup_{f \in \mathcal{L}_n^*} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \; E\,|f(X) - Y|^2 \right| \\ + \; & n\theta_n^2 2(M+1) \frac{(log(n)+1)^2}{n} \\ + \; & \inf_{f \in \mathcal{F}_n^*} \int |f(x) - m(x)|^2 \mu(dx). \end{aligned}$$

Le lemme précédent sera utilisé pour établir notre résultat principal.

4.2.1 Preuve

En premier lieu, le théorème est prouvé si et seulement si

$$\int |m_n(x) - m(x)|^2 \mu(dx) \underset{n\to\infty}{\to} 0 \text{ p.s } . \Longleftrightarrow \int |\bar{m}_n(x) - m(x)|^2 \mu(dx) \underset{n\to\infty}{\to} 0 \text{ p.s } .$$

En effet, selon la relation (4.8), on a $|m_n(x) - \bar{m}_n(x)|^2 \leq |T_L - M_n|$ ce qui implique que $\int_{\mathbb{R}^d} |m_n(x) - \bar{m}_n(x)|^2 \leq (T_L - M_n)^2 \to 0$ p.s. puisque par H_5 nous avons $\lim_{n\to+\infty} M_n = T_L$ p.s, (voir K. Kebabi et al [24]). D'abord, nous démontrons le lemme 4, et enfin, nous démonterons le théorème.

Preuve lemme 4
D'abord nous savons que

$$\begin{aligned} \int |\bar{m}_n(x) - m(x)|^2 \mu(dx) = & \left\{ \mathbf{E}(|\bar{m}_n(X) - Y|^2 \,|\mathcal{D}_n) - \inf_{f \in \mathcal{F}_n^*} \mathbf{E}\,|f(X) - Y|^2 \right\} \\ & + \left\{ \inf_{f \in \mathcal{F}_n^*} \mathbf{E}\,|f(X) - Y|^2 - \mathbf{E}\,|m(X) - Y|^2 \right\}. \end{aligned}$$

De plus, la fonction de régression vérifie

$$\inf_{f \in \mathcal{F}_n^*} E\,|f(X) - Y|^2 - \; E\,|m(x) - Y|^2 = \inf_{f \in \mathcal{F}_n^*} \int |f(x) - m(x)|^2 \mu(dx). \qquad (4.17)$$

D'un autre coté

$$
\begin{aligned}
& E\left(|\bar{m}_n(X)-Y|^2\,|\mathcal{D}_n\right) - \inf_{f\in\mathcal{F}_n^*} E\,|f(X)-Y|^2 \\
= & \sup_{f\in\mathcal{F}_n^*}\left\{ E\left(|\bar{m}_n(X)-Y|^2\,|\mathcal{D}_n\right) - E\left(|f(X)-Y|^2\,|\mathcal{D}_n\right)\right\} \\
= & \sup_{f\in\mathcal{F}_n^*}\Bigg\{ E\left(|\bar{m}_n(X)-Y|^2\,|\mathcal{D}_n\right) - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\frac{|\bar{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \\
& + \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\frac{|\bar{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\frac{|\tilde{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \\
& + \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\frac{|\tilde{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \\
& + \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - E\,|f(X)-Y|^2\Bigg\} \le \sum_{i=1}^{4} Q_{n,i},
\end{aligned}
$$

où les $Q_{n,i}$ sont expliqués ci-dessous pour tous les $i, 1 \le i \le 4$.

— Comme $\tilde{m} \in \mathcal{F}_n$,$\bar{m}_n \in \mathcal{F}_n^*$ et $\mathcal{F}_n^* \subset \mathcal{L}_n^*$, il est clair que

$$
\begin{aligned}
Q_{n,1} = & \sup_{f\in\mathcal{F}_n^*}\left\{ E\left(|\bar{m}_n(X)-Y|^2\,|\mathcal{D}_n\right) - \tfrac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\tfrac{|\bar{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)}\right\} \\
\le & \sup_{f\in\mathcal{L}_n^*}\left|\tfrac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\tfrac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - E\,|f(X)-Y|^2\right|,
\end{aligned}
$$

et

$$
\begin{aligned}
Q_{n,4} = & \sup_{f\in\mathcal{F}_n^*}\left\{\left|\tfrac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\tfrac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - E\,|f(X)-Y|^2\right|\right\} \\
\le & \sup_{f\in\mathcal{L}_n^*}\left|\tfrac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\tfrac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - E\,|f(X)-Y|^2\right|.
\end{aligned}
$$

— Comme $\bar{m}_n(X_i) \le T_L$ et $Z_i \le T_L$ p.s. nous obtenons $1_{\{A_i=0\}}|\tilde{m}_n(X_i) - Z_i| \ge 1_{\{A_i=0\}}|\bar{m}_n(X_i) - Z_i|$, ce qui implique

$$
Q_{n,2} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\frac{|\bar{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}}\frac{|\tilde{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \le 0.
$$

— Comme $\mathcal{F}_n^* \subset \mathcal{F}_n^{**}$ du fait que $T_L \le \log(n)$ et fixer un f dans $G_M \circ \mathcal{P}_n$. Selon la définition de $\mathcal{P}_n$, et le lemme 18.1 dans L. Györfi et al

[20] existe $\bar{J} \subset \{1, ..., n\}$ et $\bar{f} \in \mathcal{F}_{n,\bar{J}}$, tel que $f(X_i) = \bar{f}(X_i)$ et $card\bar{J} \leq 2(M+1)(log(n)+1)^2$, ce qui implique

$$\begin{aligned}
&\tfrac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \tfrac{|\tilde{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \tfrac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \tfrac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \\
&= \tfrac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \tfrac{|\tilde{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \tfrac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \tfrac{\left|\bar{f}(X_i)-Z_i\right|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \\
&\leq n\theta_n^2 2(M+1)\tfrac{(log(n)+1)^2}{n}.
\end{aligned}$$

D'après la définition de $\tilde{m}$, il est évident que

$$\begin{aligned}
Q_{n,3} = &\sup_{f\in\mathcal{F}_n^*} \left\{ \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|\tilde{m}_n(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \right\} \\
\leq \quad & n\theta_n^2 2(M+1)\frac{(log(n)+1)^2}{n}.
\end{aligned}$$

L'inégalité (4.16) est donc démontrée.

Preuve 11 *Il reste à prouver que les trois termes du deuxième membre de l'equation (4.16) tendent vers zéro de façon presque sûre lorsque $n \to \infty$. Pour cela, nous allons procéder en trois étapes. Dans la première étape, nous montrons que*

$$\lim_{n\to\infty} \sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \; E\,|f(X)-Y|^2 \right| = 0 \; p.s.$$

A cette fin, utilisons les inégalités suivantes

$$\begin{aligned}
&\sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \; E\,|f(X)-Y|^2 \right| \\
\leq &\sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)\hat{F}_n(Z_i)} \right| \\
+ &\sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)} \right| \\
+ &\sup_{f\in\mathcal{L}_n^*} \left| \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)} - E\,|f(X)-Y|^2 \right| \leq \sum_{i=1}^{3} Q_{n,i}^*.
\end{aligned}$$

Comme $f \in \mathcal{L}_n^*$ *implique que* $0 \leq f(x) \leq B_n$, *nous obtenons avec les relations (4.4)-(4.5)*

$$\begin{aligned} Q_{n,1}^* = & \sup_{f \in \mathcal{L}_n^*} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n} \sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{S_R(Z_i)\hat{F}_n(Z_i)} \right| \\ \leq & \frac{T_L^2}{\hat{S}_n(T) S_R(T) \hat{F}_n(I)} \sup_{t \in \mathbb{R}^+} \left| \hat{S}_n(t) - S_R(t) \right| \underset{n \to \infty}{\to} 0, \ p.s. \end{aligned}$$

et

$$\begin{aligned} Q_{n,2}^* = & \sup_{f \in \mathcal{L}_n^*} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{S_R(Z_i)\hat{F}_n(Z_i)} - \frac{1}{n} \sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{S_R(Z_i)F_L(Z_i)} \right| \\ \leq & \frac{T_L^2}{F_L(I) S_R(T) \hat{F}_n(I)} \sup_{t \in \mathbb{R}^+} \left| \hat{F}_n(t) - F_L(t) \right| \underset{n \to \infty}{\to} 0 \ p.s. \end{aligned}$$

Introduisons les notations suivantes

$V = (X, Z, 1_A), V_1 = (X_1, Z_1, 1_{A_1}), \ldots, V_n = (X_n, Z_n, 1_{A_n})$ *n vecteurs aléatoires i.i.d de même répartition que* V.

Posons

$$\begin{aligned} \mathcal{H}_n = \big\{ & h : \mathbb{R}^d \times [0, B_n] \times \{0,1\} \to \mathbb{R}^+ : \exists f \in \mathcal{L}_n^* \text{ tel que,} \\ & h(x, z, 1_A) = \frac{1_A \left| f(x) - z \right|^2}{S_R(z) F_L(z)} \\ & \forall \, (x, z, 1_A) \in \mathbb{R}^d \times [0, T_L] \times \{0,1\} \big\}. \end{aligned}$$

Les fonctions de $\mathcal{H}_n$ *sont positives et limitées par* $\frac{T_L^2}{S_R(T)F_L(I)}$ *et*

$$\mathbf{E}h(V) = \mathbf{E}\left(\frac{1_A \left| f(X) - Z \right|^2}{S_R(Z) F_L(Z)} \right) = \mathbf{E}\left[\mathbf{E}\left(\frac{1_A \left| f(X) - Z \right|^2}{S_R(Z) F_L(Z)} \mid X, Y \right) \right] = \mathbf{E}\left(|f(X) - Z|^2 \right).$$

Sous H_1,H_2 et H_4. De plus on a

$$Q^*_{n,3} = \sup_{f\in\mathcal{L}^*_n}\left|\frac{1}{n}\sum_{i=1}^{n}1_{\{A_i=0\}}\frac{|f(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)} - \ E\,|f(X)-Y|^2\right|$$
$$= \sup_{f\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n}h(V)-\mathbf{E}h(V)\right|.$$

Pour tous h_1 and $h_2 \in \mathcal{H}_n$, soient f_1 et f_2 leurs fonctions correspondantes dans $\mathcal{L}^*_n$ alors

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}|h_1(V_i)-h_2(V_i)| \\
&= \frac{1}{n}\sum_{i=1}^{n}\left|1_{\{A_i=0\}}\frac{|f_1(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)} - 1_{\{A_i=0\}}\frac{|f_2(X_i)-Z_i|^2}{S_R(Z_i)F_L(Z_i)}\right| \\
\leq\ & \frac{1}{S_R(T)F_L(I)}\frac{1}{n}\sum_{i=1}^{n}|(f_1(X_i)+f_2(X_i)-2Z_i)(f_1(X_i)-f_2(X_i))| \\
\leq\ & \frac{2T_L}{S_R(T)F_L(I)}\frac{1}{n}\sum_{i=1}^{n}|f_1(X_i)-f_2(X_i)|,
\end{aligned}$$

ce qui implique que $\mathcal{N}(\varepsilon,\mathcal{H}_n,V_1^n) \leq \mathcal{N}\left(\varepsilon\frac{S_R(T)F_L(I)}{2T_L},\mathcal{L}^*_n,X_1^n\right)$,

où $\mathcal{N}(\varepsilon,\mathcal{F}_n,Z_1^n)$ dénote le nombre recouvrant. Par l'application du théorème 9.1 dans L. Györfi et al [20]. Nous obtenons pour tout $\delta>0$

$$\begin{aligned}
&p\left\{\sup_{f\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n}h(V_i)-\mathbf{E}h(V)\right|>\delta\right\} \\
\leq\ & 8\ E\left\{\mathcal{N}\left(\delta\frac{S_R(T)F_L(I)}{16T_L},\mathcal{L}^*_n,X_1^n\right)\right\}\exp\left(-\frac{n\delta^2S_R^2(T)F_L^2(I)}{128T_L^4}\right),
\end{aligned}$$

Selon le théorème 9.4 et 9.5 et le lemme 13.1 dans L. Gyöfi et al [20], on obtient

$$\begin{aligned}
&p\left\{\sup_{f\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n}h(V_i)-\mathbf{E}h(V)\right|>\delta\right\} \\
\leq\ & 8(5n)^{4n^{1-\alpha}}\left(-\frac{288eB_n^2}{\delta\left(S_R(T)F_L(I)\right)^4}\right)^{2(M+2)n^{1-\alpha}}\exp\left(-\frac{n\delta^2S_R^2(T)F_L^2(I)}{128T_L^4}\right).
\end{aligned}$$

La relation combinée avec le $V_{\mathrm{T}_{\log n} G_M^+} \leq V_{G_M^+}$ *du théorème permet d'appliquer le lemme de Borel cantelli pour arriver à*

$$\sup_{f \in \mathcal{L}_n^*} \left| \frac{1}{n} \sum_{i=1}^{n} 1_{\{A_i=0\}} \frac{|f(X_i) - Z_i|^2}{S_R(Z_i) F_L(Z_i)} - \mathbf{E} \left| f(X) - Y \right|^2 \right| \underset{n \to \infty}{\to} 0 \ p.s.$$

Dans la deuxième étape, on obtient

$$n\theta_n^2 2(M+1) \frac{(log(n)+1)^2}{n} \underset{n \to \infty}{\to} 0 \ p.s \ parce \ que \ \theta_n \leq \frac{1}{(\log(n)+1)^2}.$$

Dans la troisième étape, nous prouvons que

$$\inf_{f \in \mathcal{F}_n^*} \int_{\mathbb{R}^d} \left| f(x) - m(x) \right|^2 \mu(dx) \underset{n \to \infty}{\to} 0 \ p.s.$$

Étant donné que m *peut être approximativement fermé arbitrairement dans* $L_2(\mu)$ *par des fonctions continûment différentiables que nous assumons sans perte de généralité.*

$$\begin{aligned} &\inf_{\forall f \in G_M \circ P_n, \|f\|_\infty \ \leq log(n)} \int_{\mathbb{R}^d} \left| f(x) - m(x) \right|^2 \mu(dx) \\ &\leq \sup_{x \in [0,1]} \left| f^*(X) - m(x) \right|^2 \leq \frac{c}{(log(n))^2} \to 0, \end{aligned}$$

où c *est constant en fonction de la première dérivé de* m*.*

$$\int_{\mathbb{R}^d} \left| m_n(x) - m(x) \right|^2 \mu(dx) \underset{n \to \infty}{\to} 0 \ p.s \ \Leftrightarrow \int_{R^d} \left| \bar{m}_n(x) - m(x) \right|^2 \mu(dx) \underset{n \to \infty}{\to} 0 \ p.s.$$

Ce qui achève cette démonstration.

CHAPITRE 5

SIMULATION

Ce chapitre présente des exemples de simulations d'estimation de modèles linéaires et non linéaires de fonctions de régression pour différentes bases induites par différents types d'ondelettes pour le cas de données complètes. Nous comparerons ensuite ces résultats et effectuons un seuillage dur pour sélectionner le meilleur estimateur. Une deuxième partie est consacrée aux données censurées. L'étude se fera dans le cas non linéaire avec bruit et sans bruit.

5.1 Les données complètes

Dans le cas de données complètes, nous simulons le cas linéaire et non linéaire pour différentes tailles d'échantillons ainsi qu'une partie de seuillage dur en utilisant la base de Haar et Daubechies. Pour les tableaux des erreurs. Nous rajoutons la base de Symmlet.

5.1.1 Estimation de la fonction de régression dans un modèle linéaire

Considérons le modèle linéaire suivant $Y = m(X) + \epsilon$, où $m(X) = 3X + 4$. X suit la loi uniforme sur l'intervalle $[0, 1]$ et $\epsilon \sim \mathcal{N}(0, (0.2)^2)$, pour un échantillon de taille n=$2^8, 2^9$ et 2^{10}.

Les exemples suivants illustrent les principes de l'analyse multi-résolution et les décompositions basées sur Haar et Daubechies de cette fonction pour $n = 2^8$.

Différentes bases

Les figures (5.1, 5.2) représentent l'estimateur de la fonction de régression en s'appuyant sur la base de Haar pour des niveaux $L = 1 : 8$.

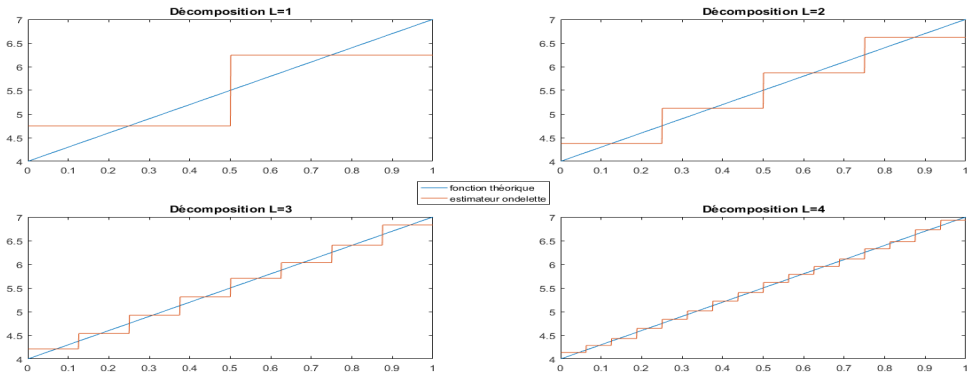


Figure 5.1 – Décomposition de fonction sur la base de Haar à L niveaux (L=1 :4).

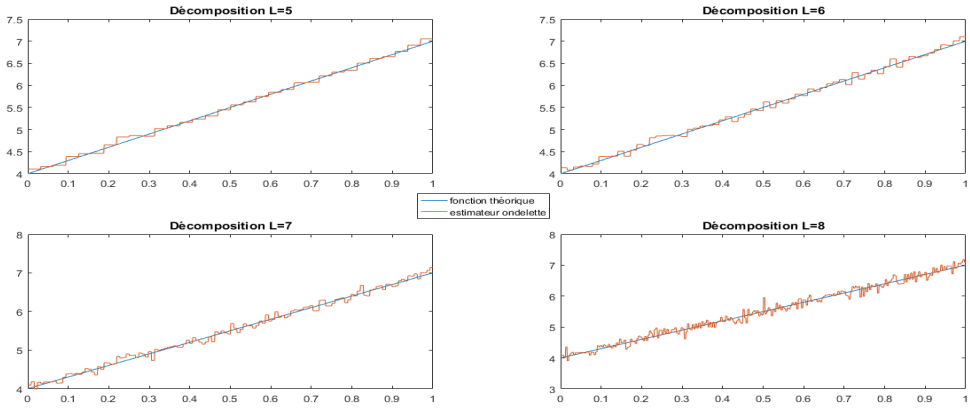


Figure 5.2 – Décomposition de fonction sur la base de Haar à L niveaux (L=5 :8).

On remarque visuellement pour les différents niveaux, qu'il y a une bonne approximation à partir du niveau $L = 5$.

Les figures (5.3, 5.4) représentent l'estimateur de la fonction de régression en s'appuyant sur la base de Daubechies pour des niveaux $L = 1 : 8$. Nous

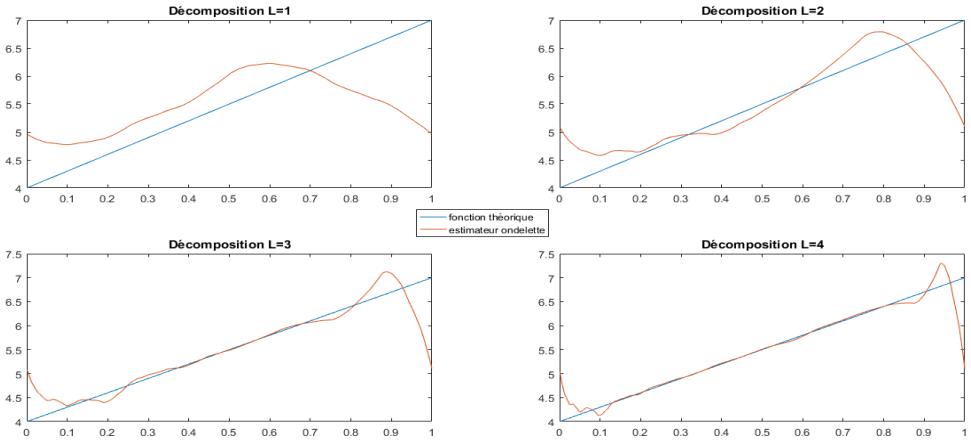


FIGURE 5.3 – Décomposition de fonction sur la base de Daubechies à L niveaux (L=1 :4).

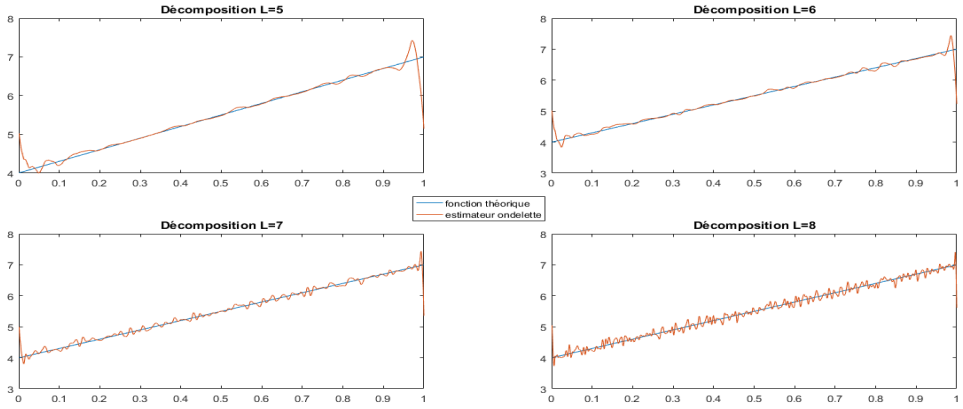


FIGURE 5.4 – Décomposition de fonction sur la base de Daubechies à L niveaux (L=5 :8).

remarquons là aussi une bonne approximation a partir du niveau $L = 5$.

n	Haar	Daubechies	Symmlet
2^8	0.0424	0.1015	0.1077
2^9	0.0411	0.0985	0.1076
2^{10}	0.0404	0.0944	0.0889

Table 5.1 – Tableau des erreurs à niveau L=4.

Pour pouvoir choisir le meilleur niveaux parmi ces 8 niveaux, nous calculerons l'erreur quadratique moyenne pour les bases déjà rencontrés ainsi que la base de Symmlet. Après avoir conclu que le niveau 4 donne les plus petites erreurs pour toutes les bases et pour différentes tailles d'échantillons. Nous résumons les résultats dans le tableau (5.1) pour choisir la meilleure base.

Dans ce qui a été dit plus haut, nous concluons que notre estimateur des moindres carrés s'approche très bien des données simulées et la meilleure base est celle de Haar pour le modèle linéaire ce qui confirme les résultats obtenus dans la littérature.

Le seuillage dur est une partie intégrante de l'estimateur proposé dans ce travail, d'où la partie qui suit où $L = 4$ pour obtenir le nombre de coefficients avec une erreur minimale.

Seuillage dur

Comme dans la sous section précédente, nous commencerons par le cas où la base utilisée est celle de Haar. Pour la sélection, nous utiliserons 4 valeurs de seuillages entre 0.01 et 2 noté sur les figures (5.5 et 5.6) En pratique, différents seuils sont mis en œuvre. Puis le meilleur est sélectionné en comparant les erreurs résultantes. Le meilleur seuillage est celui qui correspond à la valeur d'erreur la plus basse.

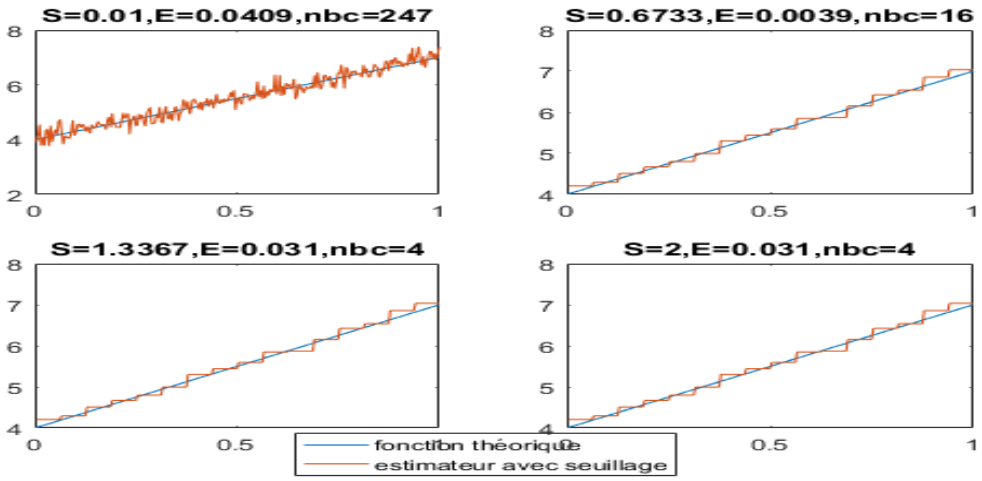


FIGURE 5.5 – Résultats d'application du seuillage dur sur la base de Haar d'un modèle linéaire.

Nous remarquons d'après la figure (5.5) que l'erreur notée E sur la figure diminue jusqu'à ce que la valeur minimale soit atteinte. Ce qui correspond à une valeur de $E = 0.0039$ et immédiatement après l'erreur augmente. Cela indique que la valeur optimale du seuil est de $S = 0,6733$ et que le nombre de coefficients notés nbc correspondant est de 16.

En nous basant sur la même idée, nous faisons un seuillage dur sur les coefficients d'ondelettes "Daubechies 8", on obtient

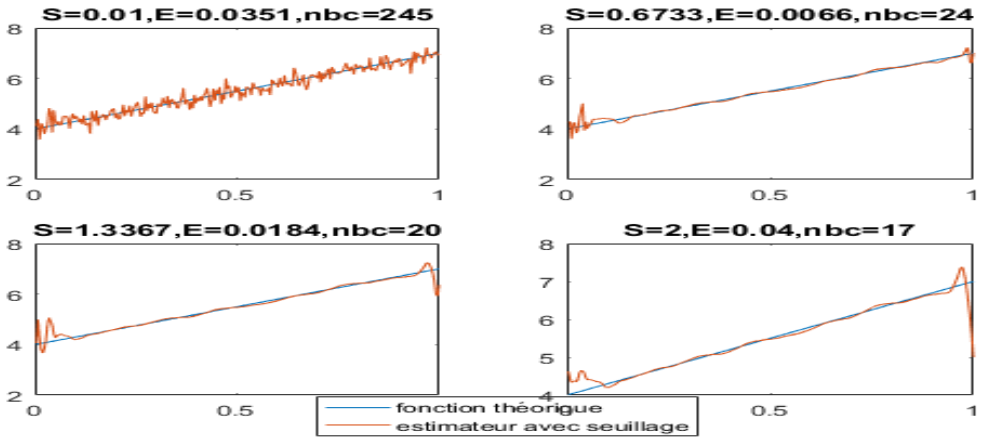


FIGURE 5.6 – Résultats d'application du seuillage dur sur la base de Daubechies d'un modèle linéaire.

Nous constatons d'après la figure (5.6) que le seuil optimal est $S = 0,6733$. Le nombre de coefficients est 24 et l'erreur est de E=0,0066.
En comparant les résultats obtenus par la base de Haar et Daubechies les figures (5.5) et (5.6), nous obtenons la même conclusions donc la meilleure base utilisée dans cet exemple est celle de Haar. C'est parce que l'estimateur admet le plus petit nombre de coefficients d'ondelettes avec de petites valeurs d'erreur correspondantes.

Par la suite, nous reprenons les mêmes étapes que dans la section précédente pour les modèles non linéaires.

5.1.2 Estimation de la fonction de régression dans un modèle non linéaire

Considérons le modèle non linéaire suivant $Y = m(X) + \epsilon$, où $m(X) = (-3X+4)sin(4\pi X)+5sin(8\pi X)+e^{2X}$, X suit la loi uniforme sur l'intervalle $[0, 1]$ et $\epsilon \sim \mathcal{N}(0, (0.2)^2)$, sur un échantillon de taille n=2^8. Cet exemple illustre les principes d'analyse multirésolution et de décomposition de ces fonctions non linéaires basées sur Haar et Daubechies.

Différentes bases

Les figures (5.7,5.8) représentent l'estimateur de la fonction de régression basée sur Haar pour les niveaux $L = 1 : 8$.

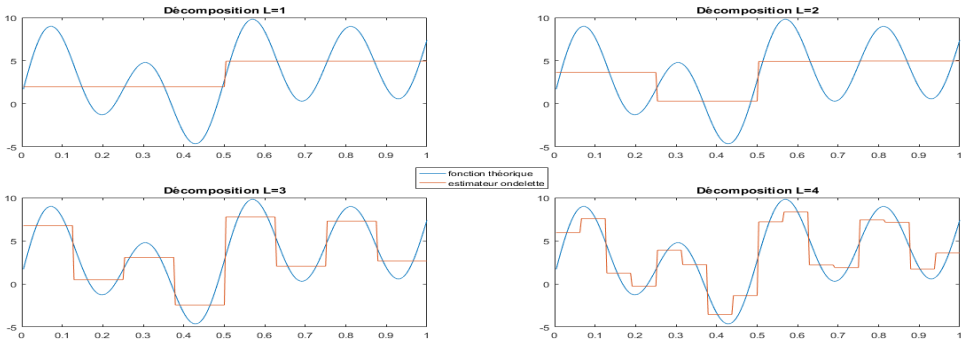


Figure 5.7 – Décomposition de fonction sur la base de Haar à L niveaux (L=1 :4).

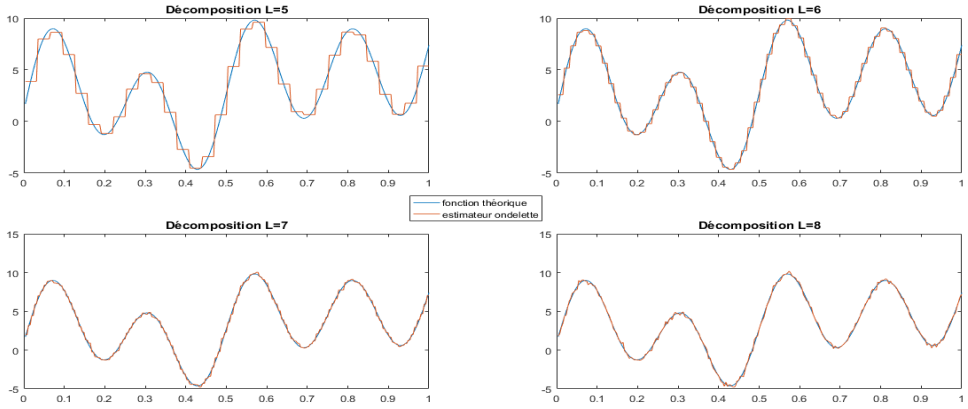


FIGURE 5.8 – Décomposition de fonction sur la base de Haar à L niveaux (L=5 :8).

Nous remarquons que pour les différents niveaux, un bon niveau est d'environ $L = 4$.

Les figures (5.9, 5.10) représentent l'estimateur de la fonction de régression sur la base de Daubechies pour des niveaux $L = 1 : 8$.
D'après les figures (5.9,5.10), nous remarquons que le meilleur niveau d'ap-

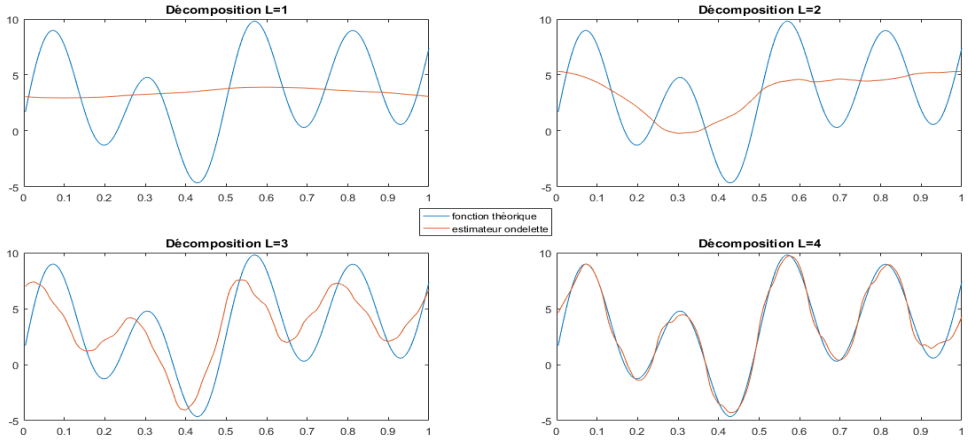


FIGURE 5.9 – Décomposition de fonction sur la base de Daubechies à L niveaux (L=1 :4).

proximation est $L = 4$.

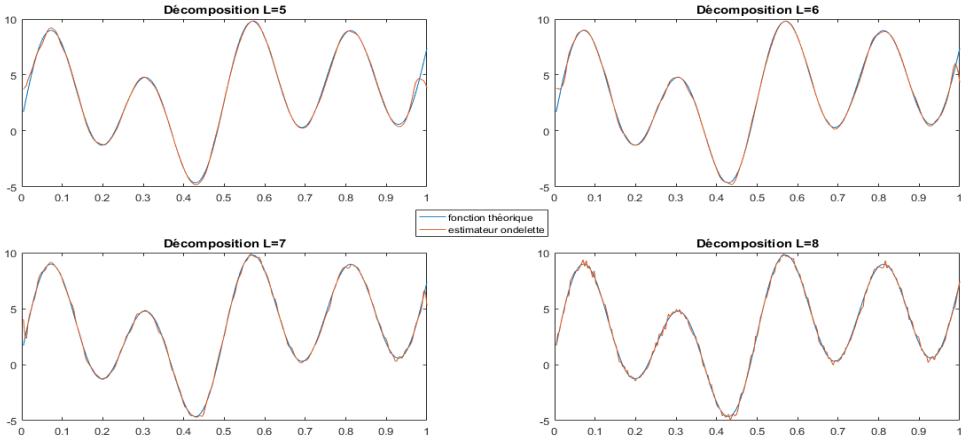


Figure 5.10 – Décomposition de fonction sur la base de Daubechies à L niveaux (L=5 :8).

n	Haar	Daubechies	Symmlet
2^8	2.5989	0.4599	0.3646
2^9	2.5960	0.4454	0.3331
2^{10}	2.5910	0.4370	0.2498

Table 5.2 – Tableau des erreurs à niveau L=4.

Le tableau (5.2) présente différentes bases d'ondelettes et tailles d'échantillon qui résument les moyennes des écarts carrés entre la courbe théorique et la courbe de l'estimateur. nous constatons une erreur minimale pour la décomposition à niveau L=4 et surtout pour la base "Symmlet". Notons que le choix de la base influence l'efficacité de notre estimateur.

Seuillage dur

Comme dans l'exemple précédent. Commençons par utiliser la base de Haar. Lors des sélections, nous choisirons 4 valeurs du seuillages et nous obtenons ce qui suit.

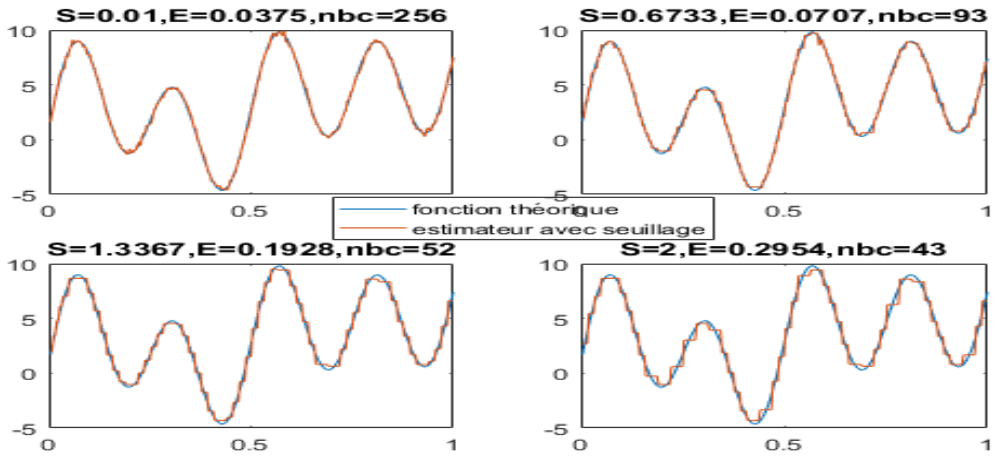


FIGURE 5.11 – Résultats d'application du seuillage dur sur la base de Haar d'un modèle non linéaire.

Nous concluons que le seuil optimal est $S = 0,01$ parce qu'il a la valeur d'erreur la plus basse de $E = 0,0375$, après quoi la valeur d'erreur augmente.

Sur la base de l'idée précédente, nous effectuons un seuillage dur sur les coefficients d'ondelettes "Daubechies 8" obtenues.

Nous constatons que le seuil optimal est $S = 0,6733$, le nombre de coef-

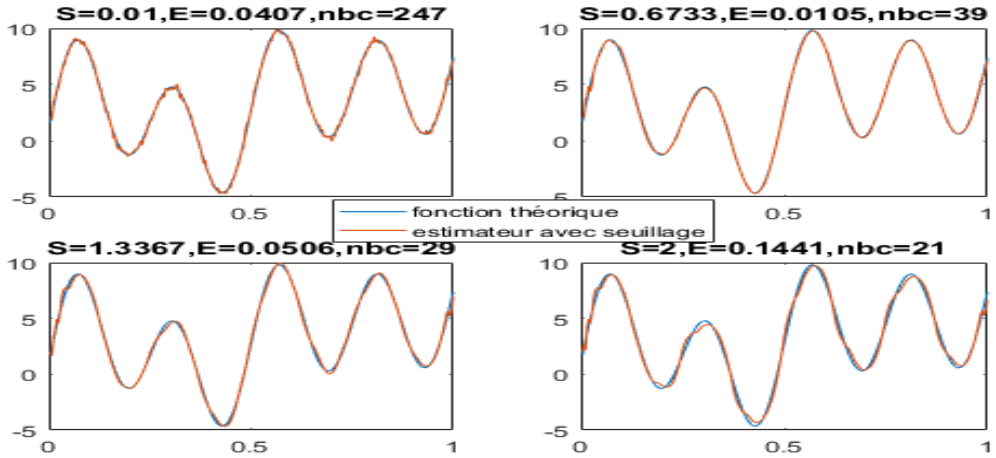


FIGURE 5.12 – Résultats d'application du seuillage dur sur la base de Daubechies d'un modèle non linéaire.

ficients est 39 et l'erreur est $E = 0,0105$. D'après les figures 5.11 et 5.12.

Nous concluons que la meilleure base utilisée dans cet exemple est celle de Daubechies. Cela permet de prendre en charge le nombre minimum de coefficients d'ondelettes avec de petites valeurs d'erreur.
Grâce aux deux exemples précédents de modèles linéaires et non linéaires, il est facile de constater que les deux règles sont efficaces. La différence réside dans l'application d'un seuillage dur, qui réduit le nombre de coefficients d'ondelettes.
Cela signifie que lorsque vous choisissez une base d'ondelettes, il est optimisé que sa forme corresponde étroitement à celle du signal.

Maintenant, nous étudions un exemple de simulation d'estimation de la fonction de régression du modèle de censure mixte, dans le cas non linéaire de fluctuations bien surveillées. Nous fondons notre discussion sur l'erreur ϵ.

5.2 Les données censurées

En présence de la censure mixte, l'estimateur des moindres carrés, s'écrit comme suit pour $M_n = \max_{1\leq i\leq n} Z_i$

$$m_n(x) = \mathbb{T}_{[0,M_n]}(\tilde{m}_n(x)),$$

où

$$\widetilde{m}_n = \arg\min_{f\in\mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} \frac{1_{\{A_i=0\}}}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \left|f(X_i) - Z_i\right|^2 \left(\frac{0}{0} := 0\right). \tag{5.1}$$

On pose

$$\tau_i(Z_i) = \frac{1_{\{A_i=0\}}}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)}.$$

On obtient

$$\widetilde{m}_n = \arg\min_{c_j\in\mathcal{R}_n} \frac{1}{n}\sum_{i=1}^{n} \tau_i(Z_i) \left(\sum_{j=1}^{\min\{K,\lfloor n^{1-\alpha}\rfloor\}} c_j f_j(X_i) - Z_i\right)^2.$$

Après, nous calculerons les dérivées partielles de $\widetilde{m}_n$ par rapport à $(c_k, \forall\, k = 1, \ldots, \min\{K, \lfloor n^{1-\alpha} \rfloor\})$, nous trouvons

$$\begin{aligned}\frac{\partial Q}{\partial c_k} &= \frac{2}{n}\sum_{i=1}^{n}\tau_i(Z_i)f_k(X_i)\left(\sum_{j=1}^{\min\{K,\lfloor n^{1-\alpha}\rfloor\}} c_j f_j(X_i) - Z_i\right) \\ &= \frac{2}{n}\sum_{i=1}^{n}\tau_i(Z_i)\left\{f_k(X_i)\sum_{j=1}^{\min\{K,\lfloor n^{1-\alpha}\rfloor\}} c_j f_j(X_i) - f_k(X_i)Z_i\right\}.\end{aligned}$$

Et comme (f_j) est une base orthonormale, on obtient

$$\frac{\partial Q}{\partial c_k} = \frac{2c_k}{n}\sum_{i=1}^{n}\tau_i(Z_i) - \frac{2}{n}\sum_{i=1}^{n}\tau_i(Z_i)f_k(X_i)Z_i.$$

En annulant les dérivées partielles. Les estimateurs des coefficients s'écrivent sous la forme suivante

$$\hat{c}_k = \frac{\frac{1}{n}\sum_{i=1}^{n}\tau_i(Z_i)f_k(X_i)Z_i}{\frac{1}{n}\sum_{i=1}^{n}\tau_i(Z_i)}.$$

Pour étudier notre simulation, nous choisissons comme famille de fonctions $\mathcal{F}_n$ la classe des ondelettes (par exemple Daubechies, Symmly et Haar).

5.2.1 Estimation de la fonction de régression dans un modèle non linéaire sans bruit avec seuillage dur

Considérons le modèle non linéaire suivant

$$m(X) = (-3X + 4)sin(4\pi X) + 5sin(8\pi X) + e^{2X}.$$

Où X suit la loi uniforme sur l'intervalle $[0, 1]$. Les variables de censure sont $D \simeq \mathcal{N}(10, (3)^2)$ et $G \simeq \mathcal{N}(-0.8, (3)^2))$ sur un échantillon de taille n=2^9. Cet exemple démontre les principes d'analyse multi-résolution et de décomposition de ces fonctions non linéaires basées sur Daubechies, Haar et Symmlet.

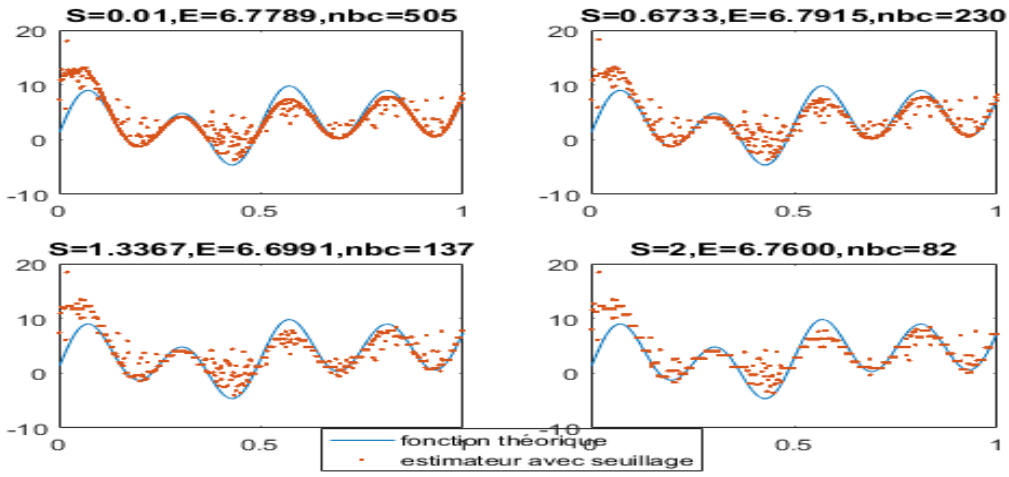


Figure 5.13 – $m(X) = (-3X + 4)sin(4\pi X) + 5sin(8\pi X) + e^{2X}$, avec un taux de 21% censure à gauche et 9% à droite sur la base Haar.

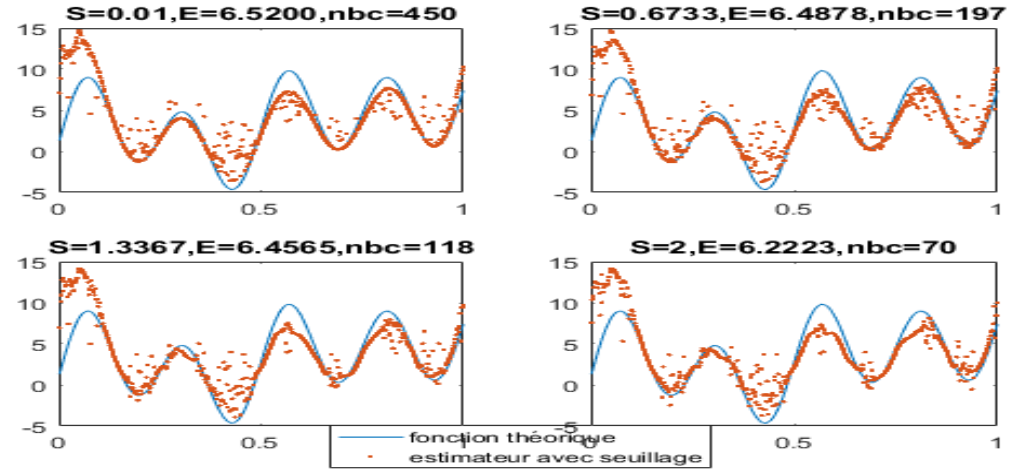


Figure 5.14 – $m(X) = (-3X + 4)sin(4\pi X) + 5sin(8\pi X) + e^{2X}$, un taux de censure à gauche et à droite de 20% et 10% respectivement sur la base de Daubechies .

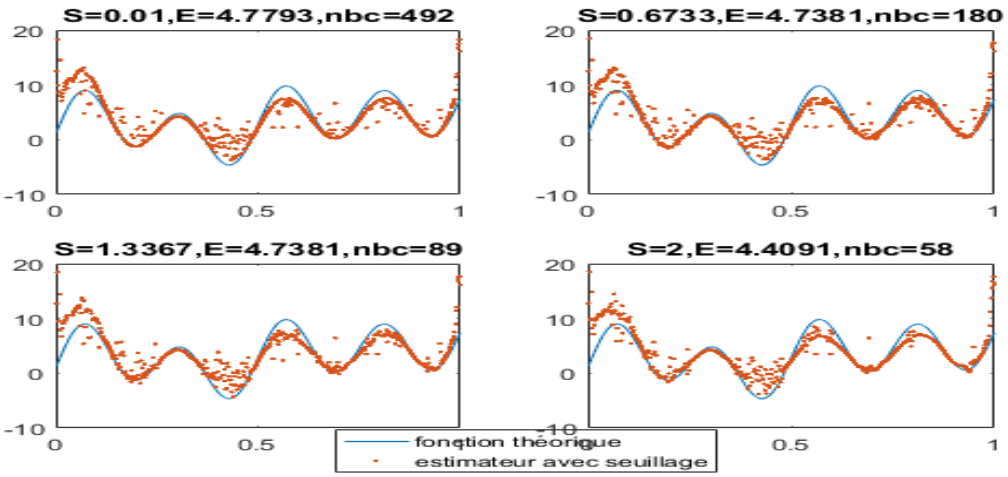


FIGURE 5.15 – $m(X) = (-3X+4)sin(4\pi X) + 5sin(8\pi X) + e^{2X}$, avec un taux de 19% censure à gauche et 12%à droite sur la base Symmlet.

Dans les figures 5.5,5.6 et 5.15 précédentes d'un modèle non linéaire, nous remarquons que la meilleure base utilisée est celle de Symmlet avec une erreur $E = 4.4091$ et un seuillage $S = 2.000$ où le nombre de coefficients est 58. En outre, nous constatons que plus le seuillage est élevé, plus la valeur d'erreur est diminuée.

A travers les trois exemples précédents d'un modèle non linéaire, nous pouvons constater facilement que les deux bases sont efficaces. La différence se voit très bien lors de l'application du seuillage dur qui consiste à diminuer le nombre de coefficients d'ondelettes.

5.2.2 Estimation de la fonction de régression dans un modèle non linéaire bruité avec seuillage dur

Reprenons l'exemple $m(X) = (-3X+4)sin(4\pi X)+5sin(8\pi X)+e^{2X}+\varepsilon$ où X suit la loi uniforme sur l'intervalle $[0,1]$ et $\varepsilon \simeq \mathcal{N}(0,(0,2)^2)$.Les variables de censure sont $D \simeq \mathcal{N}(10,(3)^2)$ et $G \simeq \mathcal{N}(-0.8,(3)^2)$ sur un échantillon de taille n=2^9.

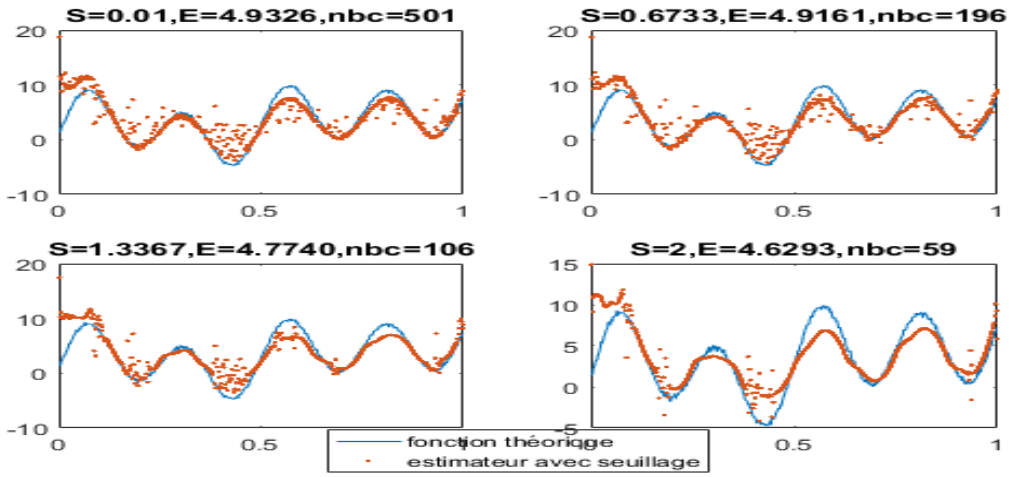


Figure 5.16 – $m(X) = (-3X+4)\sin(4\pi X) + 5\sin(8\pi X) + \exp^{2X} + \varepsilon$, avec un taux de 19% censure à gauche et 9% à droite sur la base de Daubechies .

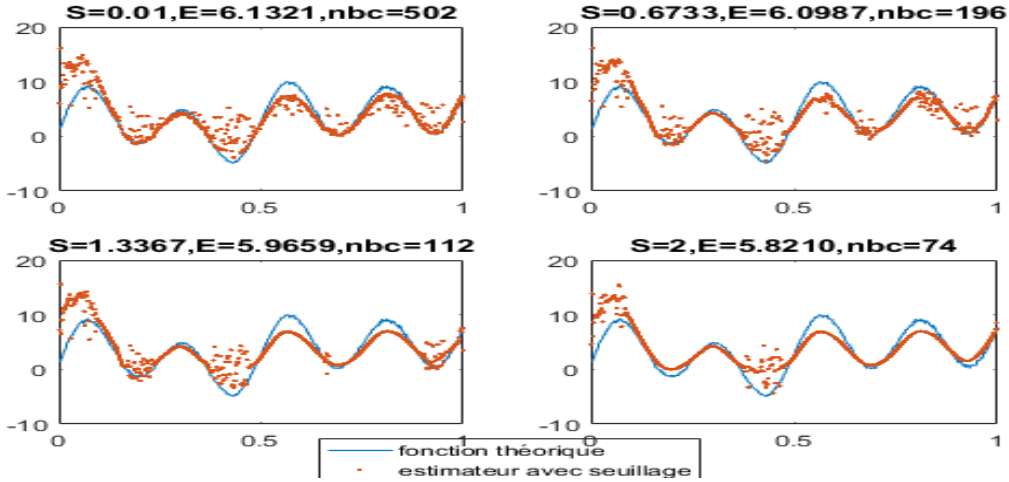


Figure 5.17 – $m(X) = (-3X+4)sin(4\pi X) + 5sin(8\pi X) + e^{2X} + \varepsilon$, avec un taux de 18% censure à gauche et 9%à droite sur la base Symmlet.

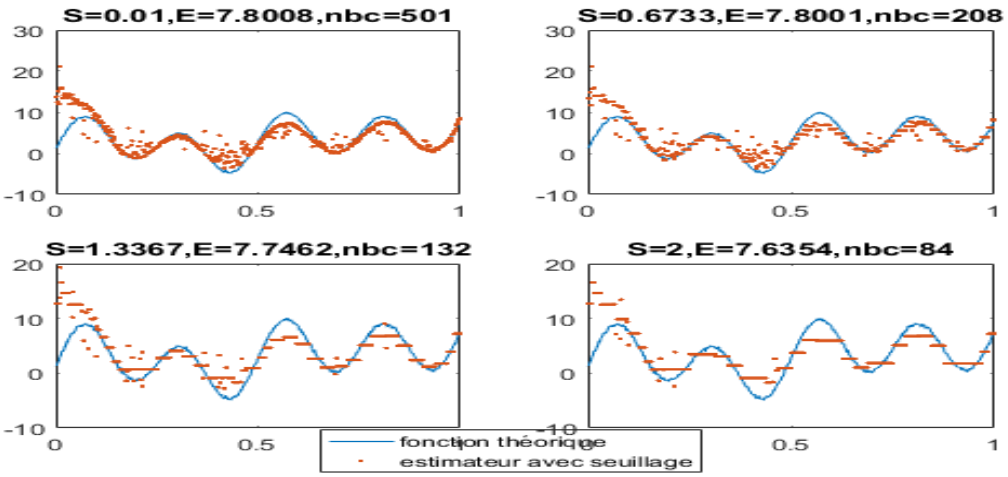


FIGURE 5.18 – $m(X) = (-3X + 4)sin(4\pi X) + 5sin(8\pi X) + e^{2X} + \varepsilon$, avec un taux de 20% censure à gauche et 11% à droite sur la base Haar.

Nous constatons que le seuil optimal est $S = 2$. Le nombre de coefficients est 59 et l'erreur est $E = 4.6293$. D'après les figures 5.17 et 5.18,5.18 nous concluons que la meilleure base utilisée dans cet exemple est celle de Daubechies.

5.2.3 Comparaison de l'estimateur avec bruit et sans bruit

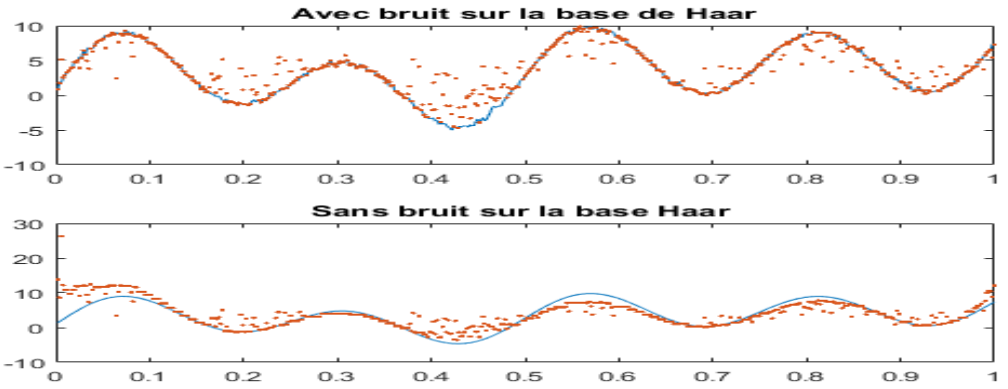


FIGURE 5.19 – $m(X) = (-3X + 4)sin(4\pi X) + 5sin(8\pi X) + e^{2X} + \varepsilon$, avec un taux 20% de censure à gauche et 10% à droite sur la base Haar.

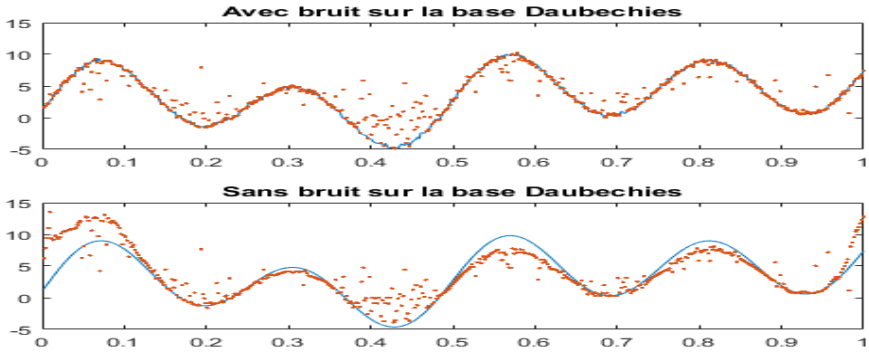


Figure 5.20 – $m(X) = (-3X+4)sin(4\pi X)+5sin(8\pi X)+e^{2X}+\varepsilon$, avec un taux de 20% censure à gauche et 10% à droite de sur la base Daubechies.

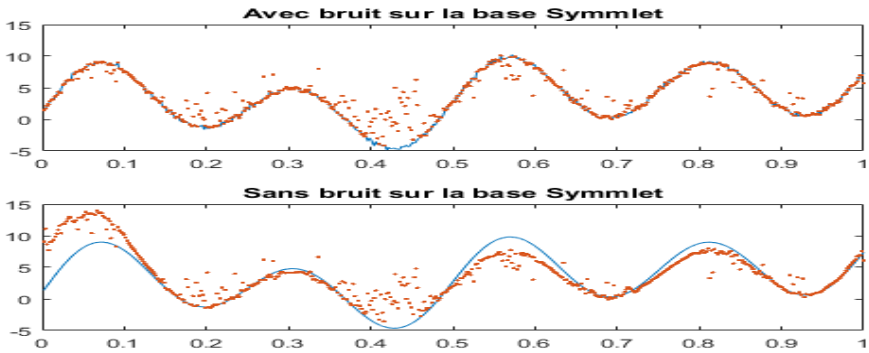


Figure 5.21 – $m(X) = (-3X+4)sin(4\pi X)+5sin(8\pi X)+e^{2X}+\varepsilon$, avec un taux de 20% censure à gauche et 10% à droite sur la base Symmlet.

Notez que la meilleure base d'ondelettes utilisée dans les figures 5.19,5.20 et 5.21 est celle de Daubechies. Le meilleur seuillage est $S = 2$ où le nombre de coefficient est 81 avec une erreur $E = 2.7191$. Nous concluons également dans le cas du bruit le pourcentage de données complètes est inférieur par rapport au données complètes dans le cas sans bruit.

Perspectives

Pour compléter cette thèse, nous fournissons une liste de quelques points qui pourraient faire l'objet de travaux futurs.

- Étude de l'estimateur de la fonction de régression par la méthode des S-splines dans un modèle de censure mixte ainsi que les taux de convergence.
- Étude d'autres modèles de censure généralisée sur l’espace des ondelettes complexe.
- Étude de l'estimateur de la fonction de régression par la méthode des moindres carrées dans un modèle de censure mixte avec différents modes de convergences.

BIBLIOGRAPHIE

[1] A. Antoniodis, G. Grégoire, I. W. McKeague, Wavelet methods for curve estimation. *J. Am. Statist. Ass.*, **89**, (1994) 1340–1353.

[2] A. Antoniadis, Smoothing noisy data with traped coifletsseries. *Scand. J. Statist.*, **23**, (1996), 313–330.

[3] F. Bouhajera *Estimation non paramétrique de la fonction de régression pour des données censurées*. Hal (2020).

[4] N. Breslow et J. Crowley, A large sample study of the life table and product limit estimates under random censorship. *The Annals of statistics* (1974) .

[5] A. Carbonez, L. Györfi and E. C. Van der Meulen, Partitioningestimates of a regression function under random censoring. *Statistics and Decisions*, **13**, (1995) 21–37.

[6] C. K. CHUI. *An introduction to wavelets. San Diego*. Academic Press (1992).

[7] I. Daubechies. Ten lectures on wavelets. *CBMS - NFS Regional Series in applied Mathematics* 61(1992).

[8] L. Devroye, L.Györfi, G. Lugosi Devroye, *A probabilist theory of pattern Recognition*, Springer Verlag, (1996).

[9] D. Donoho and I. M. Johnstone, Ideal spatial adaptation by wavelet shrinkage. *Biometrika*,81 (1994) 425 - 455.

[10] D. Donoho, Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, 90 1200-1224.

[11] D. Donoho and I. M. Johnstone, Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(1998) 879-921.

[12] R. Douas , *Application des ondelettes à l'analyse des séries chronologique* 269/Mat. (2011).

[13] R. Douas, I. Laroussi, S. Kharfouchi, Incomplete Least Squared Regression Function Estimator Based on Wavelets, *Journal of siberian federal university. mathematics and physics*, (2023) 16(2)-204-2015.

[14] N. R. Draper and H. Smith, *Applied Regression Analysis*, 2nd ed. Wiley, New York, (1981).

[15] J. B. J. Fourier. *Théorie analytique de la chaleur*(1822).

[16] D. Gabor. *Theory of Communication. J. IEE. London 93 pp.* 429-457 (1946).

[17] F. Galton. Regression towards mediocrity and the stability. *Types Studies in History and Philosophy of Science* volume 86, April (2021), Pages 6-19.

[18] C. Gasquet et P. Witomski . *Analyse de Fourier et applications.* Dunod (1990).

[19] R. Gill, Large sample behaviour of the product limit estimator on the whole line. *The annals of statistics*(1983).

[20] L. Györfi, M. Kohler, A. Krzyżak, H. Walk, *A Distribution Free theory of Non parametric Regression.* Springer-Verlag New York, Inc. (2002).

[21] D. J. Hart, *Non parametric Smoothing and Lack-of-Fit Tests.* Springer-Verlag, New York, (1997).

[22] T. Hastie and R. J. Tibshirani . *Generalized Additive Models.* Chapman and Hall, London, U. K, (1990).

[23] E. L. Kaplan, P. Meier, Nom parametric estimation from incomplete observations, *J. Amer. Statist. Assoc.* 53, (1958) 457-481.

[24] K. Kebabi, I. Laroussi, F. Messaci, Least squares estimators of the regression function with twice censored data, *Statist. Probab. Lett.* 81, (2011) 1588-1593.

[25] M. Kohler, On the universal consistency of a least squares spline regression estimator. *Math. Methods Statist.*, (1997) (6) 349-364.

[26] M. Kohler, Universally consistent regression function estimation using hierarchical b-splines. *J. Multivariate Anal.*, (1999) (67) 138–164.

[27] M. Kohler, A. Krzyżak, Nonparametric regression estimation using penalized least squares,*IEEE Trans. Inform. Theory.* 47, (2001) 3054-3058.

[28] M. Kohler, K. Màthé, M. Pintér,Prediction from randomly right censored data,*J. Multivariate Anal.* 80, (2002) 73-100.

[29] S. Mallat . Multiresolution approximation and wavelets. *Trans. AMS.* 615, pp. 69-88 (1989).

[30] F. Messaci, Local averaging estimates of the regression function with twice censored data,*Statist. Probab. Lett.*, 80, (2010) 1508-1511.

[31] Y. Meyer, *Ondelettes et oprateurs I. Hermann.*(1989)

[32] J. Morlet, A Decomposition of hardy into square integrable wavelets of constant shape. *SIAM J.Math Anal.* (1980)

[33] D.Morales, L. Pardo, V. Quesada, Bayesian survival estimation for incomplete data when the life distribution is proportionally related to the censoring time distribution. *Comm. Statist. Theory Methods*, 20, (1991) 831-850. MR1131189.

[34] E. A. Nadaraya, On estimating regression. *Theory of Probability and Its Applications*, 9 (1) (1964) 141-142.

[35] A. Nobel, Histogram Regression Estimation Using Data-dependent Partitions. *Ann. Statist.* 24,(1996) 1084-1105.

[36] V. Patilea, J. M. Rolin, Product-limit estimators of the survival function with twice censored data, *Ann. Statist.* 34, No 2, (2006) 925-938.

[37] W. Stute et J.-L. Wang, The strong law under random censorship. *The Annals of statistics*(1993) .

[38] B. Torrésani. *Ondelettes - Analyse temps fréquences et signaux non stationnaires.* Montpelliers 97.

[39] V. N. Vapnik et A. Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16 (1971) 264-280.

[40] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data.* Springer-Verlag, New York (1982).

[41] V. N. Vapnik, *Statistical Learning Theory.* Wiley, New York (1998).

[42] G. S. Watson, Smooth regression analysis. *Sankhya Series A*, **26** (1964) 359–372.

[43] B. Winter et L. Rejtő, Glivenko-Cantelli theorems for the product limit estimate. *Problems of Control and Information Theory* (1978).

بعض جوانب الاستدلال الترددي الإحصائي بطريقة الموجات

ملخص

ضمن هذه الأطروحة سنهتم بطريقة لتقدير دالة الانحدار دون وسائط . المسماة بطريقة المربعات الصغرى على فئة الموجات.

مساهمتنا تكمن في تمديد هذه الطريقة إلى متغير الإجابة خاضع إلى حجب مزدوج حيث بينا أن المقدر المدروس يتقارب بشكل شبه مؤكد نحو القيمة المثلى.

في النهاية نقدم دراسة محاكاة الهدف منها التأكيد على جودة المقدر المقدم من جهة والمقارنة بينهما من جهة اخرى.

الكلمات المفتاحية : دالة الانحدار ، المربعات الصغرى ،المعطيات الخاضعة لحجب مزدوج، الموجات.

Quelques aspects d'inférence statistique fréquentielle Méthode des ondelettes

Résumé

Dans ce travail, nous nous intéressons à la méthode non paramétrique de la fonction de régression sur la classe des ondelettes. Notre apport se situe dans le fait que nous avons étendu cette méthode au cas où la variable réponse est soumise à une censure mixte. Nous avons montré que l'estimateur introduit convergent presque sûrement vers la valeur optimale.

Mots clés: Fonction de régression, Moindres carrés, Ondelettes, Censure mixte.

Some aspects of statistical frequency inference Wavelets method

Summary

In this work, we are interested in the nonparametric method of the regression function on the wavelets class. Our contribution lies in the fact that we have extended this method in case the response variable is subject to mixed censorship. We have shown that the introduced estimator almost certainly converges towards the optimal value.

Keywords: Function of regression, Least squares, Wavelets, Twice censure data.