Université de Kasdi Merbah

Ouargla (Algérie)

THÈSE DE DOCTORAT

en Sciences : 2024/2025

Présentée en vue de l'obtention du diplôme
de Doctorat en Mathématiques

Intitulée :

La Convergence Presque Complète en Statistique Semi-paramétrique Fonctionnelle

Spécialité : Mathématiques Appliquées
Option : Probabilités et Statistiques

Présenté par :
Agoune Rachid

Soutenue publiquement le 24/05/2025 devant le jury :

Président :	M. Meflah Mabrouk	Prof, Université de Ouargla, Algérie
Directeur de thèse :	M. Merahi Fateh	MCA, Université de Batna 2, Algérie
Co-encadreur :	M. Saouli Mustapha	MCA, Université de Ouargla, Algérie
Examinateur :	M. Houas Amrane	MCA, Université de Biskra, Algérie
Examinateur :	M. Amara Abdelkader	MCA, Université de Ouargla, Algérie
Examinateur :	M. Benhadid Ayache	MCA, Université de Batna 2, Algérie

Résumé

Cette thèse explore plusieurs méthodes d'estimation pour les données fonctionnelles dans le cadre des modèles de régression fonctionnelle, en mettant l'accent sur les techniques basées sur le noyau. Notre contribution réside dans la reformulation de l'estimateur proposé par Ferraty et Vieu [23], ce qui nous permet d'estimer la fonction de distribution de la variable réponse en utilisant un seul noyau. Ce noyau se concentre sur la partie fonctionnelle, permettant la simulation de la fonction de répartition cumulative conditionnelle dérivée de la fonction de régression et sa comparaison avec la fonction de distribution théorique. Dans la première partie, nous examinons l'estimation de l'opérateur de régression dans un modèle fonctionnel à indice unique en utilisant des méthodes basées sur le noyau. Nous appliquons plusieurs techniques, y compris la validation croisée Leave-One-Out (LOO), pour sélectionner le paramètre de lissage optimal. Les résultats démontrent la convergence presque sûre de l'estimateur par noyau, validant ses fondements théoriques et son efficacité pratique. La deuxième partie se concentre sur l'estimation de la fonction de répartition cumulative conditionnelle pour les données fonctionnelles. Nous utilisons une approche de sélection de la largeur de fenêtre et validons les résultats en étudiant la convergence presque complète de l'estimateur de la fonction de répartition conditionnelle. À travers des simulations, nous confirmons la robustesse et les capacités prédictives de la méthode par noyau proposée, tout en soulignant l'importance de choisir des paramètres de lissage appropriés. Dans la troisième partie, nous abordons les défis liés à la gestion de deux paramètres de lissage dans les estimateurs à double noyau utilisés pour l'estimation de la distribution conditionnelle. En reformulant l'approche, nous montrons les avantages de l'utilisation d'un seul noyau, simplifiant ainsi le processus d'estimation. Nous recommandons l'utilisation de la validation croisée pour optimiser la sélection de ces paramètres.
Mots-clés : Modèle de regression semi-paramétrique, Modèle à indice fonctionnel, Convergence presque complète, Variables aléatoires fonctionnelles.

Abstract

This thesis explores several estimation methods for functional data within the framework of functional regression models, with an emphasis on kernel-based techniques. Our contribution lies in the reformulation of the estimator proposed by Ferraty and Vieu [23], allowing us to estimate the distribution function of the response variable using a single kernel. This kernel focuses on the functional part, enabling the simulation of the conditional cumulative distribution function derived from the regression function, and its comparison with the theoretical distribution function. In the first part, we examine the estimation of the regression operator in a single-index functional model using kernel methods. We apply several techniques, including Leave-One-Out (LOO) cross-validation, to select the optimal smoothing parameter. The results demonstrate the almost complete convergence of the kernel estimator, validating its theoretical foundations and practical efficacy. The second part focuses on the estimation of the conditional cumulative distribution function for functional data. We use a bandwidth selection approach and validate the results by studying the almost complete convergence of the conditional cumulative distribution estimator. Through simulations, we confirm the robustness and predictive capabilities of the proposed kernel method, while emphasizing the importance of choosing appropriate smoothing parameters. In the third part, we address the challenges associated with managing two smoothing parameters in the double-kernel estimators used for conditional distribution estimation. By reformulating the approach, we demonstrate the practical benefits of using a single kernel, simplifying the estimation process. We recommend the use of cross-validation to optimize the selection of these parameters, further improving the flexibility and efficiency of the estimator for prediction.

Keywords : Semi-parametric regression model, Functional single index model, Almost complete convergence, Functional random variables.

ملخص

هذه الأطروحة تستكشف عدة طرق تقدير للبيانات الوظيفية في إطار نماذج الانحدار الوظيفي، مع التركيز على التقنيات القائمة على النواة. تكمن مساهمتنا في إعادة صياغة المُقدِّر الذي اقترحه فيراتي وفييوFerraty et Vieu، مما يسمح لنا بتقدير دالة توزيع المتغير التابع باستخدام نواة واحدة. تركز هذه النواة على الجزء الوظيفي، مما يُتيح محاكاة دالة التوزيع التراكمي الشرطي المشتقة من دالة الانحدار ومقارنتها بالدالة النظرية للتوزيع .في الجزء الأول، نستعرض تقدير معامل الانحدار في نموذج وظيفي بمؤشر واحد باستخدام طرق قائمة على النواة. نطبّق عدة تقنيات، بما في ذلك التحقق المتقاطع "اترك واحدًا خارجًا(LOO) "، لاختيار أفضل معامل للتنعيم. تظهر النتائج تقاربًا شبه مؤكد للمُقدِّر القائم على النواة، مما يؤكد على أسسها النظرية وكفاءتها العملية .يركز الجزء الثاني على تقدير دالة التوزيع التراكمي الشرطي للبيانات الوظيفية. نستخدم نهجًا لاختيار عرض النافذة ونتحقق من النتائج بدراسة التقارب شبه المؤكد للمُقدِّر الخاص بدالة التوزيع الشرطي. من خلال المحاكاة، نؤكد على متانة وقدرات التنبؤ للطريقة المقترحة القائمة على النواة، مع تسليط الضوء على أهمية اختيار المعاملات المناسبة للتنعيم. كما نسلط الضوء على الطرق البديلة، مثل التحقق المتقاطعLOO ، لتحسين دقة التقدير . في الجزء الثالث، نتناول التحديات المرتبطة بإدارة معاملين للتنعيم في المُقدِّرات ذات النواتين المستخدمة لتقدير التوزيع الشرطي. من خلال إعادة صياغة النهج، نُبرز الفوائد العملية لاستخدام نواة واحدة، مما يُبسِّط عملية التقدير. نوصي باستخدام التحقق المتقاطع لتحسين اختيار هذه المعاملات، مما يزيد من مرونة وكفاءة المُقدِّر في التنبؤ .بشكل عام، تُقدِّم هذه الأطروحة مساهمة هامة في مجال تحليل البيانات الوظيفية الآخذ في التوسع. وتوفر أعمالنا رؤى حول أفضل الممارسات لتقدير معامل الانحدار والتوزيعات الشرطية، لا سيما فيما يتعلق باختيار المعاملات واستخدام النواة. تفتح النتائج آفاقًا جديدة للبحث المستقبلي، خصوصًا في تحسين تقنيات اختيار المعاملات وتطبيق هذه الطرق على البيانات الواقعية.

الكلمات المفتاحية : نموذج شبه معلمي، نموذج بمؤشر وظيفي، التقارب شبه التام، المتغيرات العشوائية الوظيفية

Dédicace

À ma mère,
Ta tendresse, ton courage et ta foi inébranlable ont été, tout au long de ma vie, une source constante de force et d'inspiration. C'est à toi, avec une reconnaissance infinie, que je dédie ce travail.
À ma petite famille — mon épouse et mes enfants —

Merci pour votre amour, votre patience et votre présence rassurante, qui m'ont soutenu dans les moments de doute et de fatigue. Vous êtes mon équilibre et ma joie de vivre.

À mes frères et sœurs,

Pour votre affection, vos encouragements et vos pensées bienveillantes, toujours ressentis malgré la distance ou le silence.

Ce travail est aussi le vôtre. Il porte la trace de tout ce que vous m'avez donné.

Remerciements

Je tiens à exprimer ma profonde gratitude à mes encadreurs dévoués. Tout d'abord, au Professeur Baheddi, qui, malgré les défis imposés par la pandémie de COVID-19, a fait preuve d'une persévérance et d'un engagement exemplaires. Son soutien et sa sagesse ont été des lumières dans les moments difficiles de ce parcours académique.

Je souhaite également remercier sincèrement M. Fateh Merahi, qui a généreusement accepté de prendre le relais dans des circonstances éprouvantes. Son expertise et son dévouement ont grandement contribué à la réussite de ce travail.

Je tiens à exprimer ma gratitude au président du jury, M. Mabrouk Meflah, Professeur à l'Université de Ouargla, pour avoir accepté de présider ce jury, ainsi qu'au co-encadreur M. Mustapha Saouli, et aux examinateurs : M. Abdelkader Amara, Professeur à l'Université de Ouargla, M. Amrane Houas de l'Université de Biskra, et M. Ayache Benhadid de l'Université de Batna 2. Je les remercie d'avoir accepté d'évaluer cette thèse et pour leurs commentaires et suggestions précieux, qui ont été d'une grande aide dans l'amélioration de ce travail.

Un remerciement tout particulier à ma famille, qui a été mon pilier de soutien inébranlable tout au long de cette aventure. Votre amour, vos encouragements et votre compréhension m'ont permis de surmonter les hauts et les bas de cette recherche.

À tous ceux qui ont contribué, de près ou de loin, à la réalisation de cette thèse, je vous adresse mes plus sincères remerciements. Vos efforts ont été essentiels à sa réussite.

Table des matières

Abréviations et Symboles

Les notations suivantes seront utilisées dans les chapitres de cette thèse.

Abréviations

- **v.a.f.** : Variable aléatoire fonctionnelle
- **p.co.** : Presque complète
- **p.s.** : presque sûrement.
- **p.** : probabilité
- **FDA** : Functional Data Analysis (Analyse des données fonctionnelles)
- **i.i.d.** : Indépendant et identiquement distribué
- **GAM** : Generalized Additive Model (Modèle additif généralisé)
- **FDC** : Fonction de distribution cumulative conditionnelle
- **MSE** : Erreur quadratique moyenne (Mean Squared Error)
- **kNN** : k plus proches voisins (k Nearest Neighbors)

Symboles mathématiques

- $\mathcal{N}_x$: Voisinage de x
- $\mathcal{H}$: Espace de Hilbert
- $\mathcal{S}$: Sous-ensemble compact de $\mathbb{R}$
- Y : Variable réponse
- $E(Y|X)$: Espérance conditionnelle de Y sachant X
- ϵ : Terme d'erreur avec $E(\epsilon|X) = 0$
- P : Probabilité
- $B(x, \epsilon)$: Boule de centre x et de rayon ϵ
- h : Paramètre de bande (positif)
- K : Fonction noyau (de type I ou II)

- g : Paramètre de lissage
- K_0 : Noyau de type 0
- $\sigma_m(x)$: Fonction continue en x
- r : Opérateur de régression
- F_Y^X : Fonction de répartition conditionnelle de Y sachant X.
- $R_\lambda(y \mid x)$: fonction de répartition conditionnelle (approche semi-paramétrique)
- $d(x_1, x_2)$: Distance entre x_1 et x_2
- $\hat{r}(x)$: Estimateur de régression en x
- $\hat{F}_Y^X(y)$: Estimateur de la fonction de répartition conditionnelle
- $\Omega_i(x)$: Variable aléatoire associée à x
- X_i : $i^{\text{ème}}$ observation de la variable fonctionnelle X
- $\Gamma_i(y)$: Variable aléatoire associée au point y
- c_1, c_2, c_3, c_4 : Constantes positives
- EK : Espérance du noyau K
- φ_x : Probabilité de petite boule associée à x
- λ : Indice fonctionnel

Table des figures

Liste des tableaux

Introduction

La statistique fonctionnelle a connu un très important développement ces dernières années. Cette branche de statistique vise à étudier des données qui, de par leurs structures et le fait qu'elles soient collectées sur des grilles très fines, comme la fonction du temps par exemple. Le besoin de considérer ce type de données, maintenant couramment rencontré sous le nom de données fonctionnelles dans la littérature, est avant tout un besoin pratique. Compte tenu des capacités actuelles des appareils de mesure et de stockage informatique, les situations pouvant fournir de telles données sont multiples et issues de domaines variés. Cependant au-delà de cet aspect pratique, il est nécessaire de donner un cadre théorique pour l'étude de ces données. Bien que la statistique fonctionnelle ait les mêmes objectifs que les autres branches de la statistique (analyse des données, inférence,...), les données qui ont cette particularité prennent leurs valeurs dans des espaces fonctionnels, et les méthodes usuelles de la statistique multivariée sont ici mises en défaut. En effet, la principale source de difficultés, tant d'un point de vue théorique que pratique, provient du fait que les observations de ce type de variables fonctionnelles sont supposées appartenir à un espace de dimension infinie. Ainsi, l'intérêt de ce travail réside dans l'apport de solutions à ce problème de dimension infinie, dans les deux cadres indépendant et mélangeant, en mettant en place un cadre théorique suffisamment général. Les outils utilisés pour les développements théoriques sont de natures variées. En effet, ils relèvent de l'analyse fonctionnelle, mais aussi d'outils probabilistes tels que les inégalités exponentielles pour des sommes de variables aléatoires. Ce travail a pour objet une familiarisation avec les méthodes statistiques qui permettent de tester la validité des théories économiques (et éventuellement celles d'autres sciences sociales). L'objectif est d'aider les utilisateurs de ces méthodes à sélectionner les techniques les plus appropriées pour résoudre un problème spécifique, à interpréter correctement les résultats obtenus lors de leur application et à vérifier la validité des hypothèses sur lesquelles reposent leurs performances optimales. Un intérêt particulier sera porté aux exemples concrets d'application. Ces exemples seront choisis dans diverses disciplines de la science économique (macro-économie, micro-économie, économie du travail, économie publique, économie internationale) mais également dans les domaines extérieures à l'économie (criminologie, agronomie, écologie, pédagogie et démographie). Dans cette thèse, nous proposons d'apporter une contribution à l'étude des données fonctionnelles dans le contexte où la variable fonctionnelle sert à expliquer un phénomène représenté par une autre variable. Le premier problème qui va nous intéresser est celui d'une fonction de régression et de sa fonction de répartition conditionnelle correspondante dans le cas où la variable explicative est fonctionnelle. C'est un sujet sur lequel la littérature est très conséquente. Le document est organisé en quatre chapitres principaux :

- **Chapitre 1 : Outils probabilistes** : Ce premier chapitre présente les concepts probabilistes et statistiques fondamentaux qui serviront de base pour les développements ultérieurs. Il comprend notamment des rappels sur les processus stochastiques, les inégalités de concentration, et d'autres outils théoriques indispensables pour l'analyse des modèles de régression fonctionnelle.

- **Chapitre 2 : Estimation de la fonction de distribution conditionnelle sur un modèle de régression non paramétrique** : Dans ce chapitre, nous étudions l'estimation de la fonction de distribution cumulative conditionnelle pour des modèles de régression fonctionnelle avec réponses réelles. L'objectif est de développer des méthodes d'estimation non paramétriques qui soient à la fois flexibles et précises dans ce contexte fonctionnel.

- **Chapitre 3 : Choix de la bande passante pour l'estimateur de la régression fonctionnelle semi-paramétrique** : Ce chapitre se concentre sur l'estimation dans un cadre semi-paramétrique en utilisant la technique de validation croisée "Leave-One-Out" pour la sélection du paramètre de lissage dans la régression à noyau. La méthodologie est appliquée aux données fonctionnelles, et les propriétés théoriques de l'estimateur proposé sont examinées.
- **Chapitre 4 : Estimation de la fonction de distribution conditionnelle sur le modèle de régression à indice fonctionnelle** : Le dernier chapitre porte sur l'estimation de la fonction de distribution conditionnelle dans un cadre semi-paramétrique. Nous analysons en détail les propriétés de l'estimateur et explorons diverses applications pratiques pour illustrer son utilité.

Chapitre 1

Outils probabilistes

On va présenter en bref quelques outils probabilistes. Parmi ces outils, ceux qui sont reformulés dans des nouveaux types afin de les rendre simplement applicables pour les modèles non paramétriques fonctionnels. Ces nouvelles formulations seront également utiles pour toute personne intéressée pour le développement de nouvelles avancées sur l'étude asymptotique en statistique fonctionnelle semi-paramétrique. Le fil conducteur de cette thèse consiste à présenter des récents développements dans le cas des variables fonctionnelles. Cependant, l'obtention des résultats asymptotiques nécessite l'utilisation des outils de probabilité de base pour les variables aléatoires réelles et de nombreux résultats présentés ci-dessous concernent les variables aléatoires réelles. La première section de ce chapitre traite de la notion de convergence presque complète et met l'accent sur le lien entre ce mode de convergence et d'autres modes standards (tels que la convergence presque sûre ou la convergence en probabilité). L'étude des propriétés de la convergence presque complète repose principalement sur certaines inégalités exponentielles pour des sommes de variables aléatoires. La deuxième section rappelle certaines de ces inégalités ayant une forme adaptée au type des développements théoriques. La dernière section de ce chapitre sera consacrée aux conditions d'indépendance des variables aléatoires (réelles ou fonctionnelles). Plus précisément, dans la dernière partie de cette section, nous présenterons quelques inégalités pour la somme des variables aléatoires réelles indépendantes. En outre, comme pour la deuxième section, nous avons choisi, parmi la large littérature de ces inégalités, celles ayant une forme adaptée au cadre de cette thèse. Certaines de ces inégalités ont été reformulées dans de nouveaux types, afin de rendre leur application plus facile.

Il est impossible, dans cette thèse, de donner les preuves de tous ces outils probabilistes, et nous nous référerons principalement à la littérature existante.

Tout au long de ce chapitre, $(X_n)n \in \mathbb{N}$ et $(Yn)n \in \mathbb{N}$ sont deux suites de variables aléatoires réelles, (un) est une suite de nombres réels positifs. On note que $(\zeta n)n \in \mathbb{Z}$ est une suite de variables aléatoires (non nécessairement réelles), et que $(T_n)n \in \mathbb{Z}$ est une suite stationnaire de variables aléatoires réelles. On note aussi que $(Zn)n \in \mathbb{N}$ est une suite de variables aléatoires réelles indépendantes et centrées, et que $(Wn)_{n\in\mathbb{N}}$ est une suite stationnaire de variables aléatoires dépendantes et centrées.

1.1 La convergence presque complète

Le mode de convergence presque complète implique les autres modes standards de convergence. Par conséquent, en raison de ces deux avantages [18], il est devenu tout à fait habituel, pour les modèles non paramétriques et semi-paramétriques fonctionnels, d'exprimer les résultats asymptotiques à l'aide de la notion de convergence presque com-

plète. Cette convergence a été introduite depuis longtemps dans [34]. Dans cette section, nous avons décidé de rappeler quelques définitions et propriétés de base au sujet de cette notion sans démonstration (on peut consulter toutes les preuves dans Ferraty et al [24]).

Définition 1. *On dit que la suite* $(X_n)_{n\in\mathbb{N}}$ *converge presque complètement vers la variable aléatoire réelle si et seulement si :*

$$\forall \epsilon > 0, \sum_{n\in\mathbb{N}} P\left(|X_n - X| > \epsilon\right) < \infty, \tag{1.1}$$

et on la note

$$\lim_{n\to\infty} X_n = X, p.co.$$

Définition 2. *On dit que la suite* $(X_n)_{n\in\mathbb{N}}$ *converge en probabilité vers la variable aléatoire réelle* X*, si et seulement si*

$$\forall \epsilon > 0, \lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0, \tag{1.2}$$

et on écrit :

$$\lim_{n\to\infty} X_n = X, p.$$

Définition 3. *On dit que la suite* $(X_n)_{n\in\mathbb{N}}$ *converge presque sur vers la variable aléatoire réelle* X*, si et seulement si :*

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1. \tag{1.3}$$

On peut trouver dans des livres de probabilité élémentaire une présentation plus générale des divers liens entre ces modes de convergence (on peut consulter [13] par exemple). Les preuves de ces propriétés peuvent également être trouvées dans [13].

Proposition 4. *Si* $\lim_{n\to\infty} X_n = X, p.co$*, alors on a :*

1. $\lim_{n\to\infty} X_n = X, p.$
2. $\lim_{n\to\infty} X_n = X, p.s.$

Selon nos connaissances, il n'existe pas de notion équivalente à la notion de convergence complète. Le but de la définition suivante est de préciser cette notion.

Définition 5. *On dit que la vitesse de convergence presque complète de la suite* $(X_n)_{n\in\mathbb{N}}$ *vers* X *est d'ordre* u_n *si et seulement si :*

$$\exists \epsilon_0 > 0, \sum_{n\in\mathbb{N}} P\left(|X_n - X| > \epsilon_0 u_n\right) < \infty, \tag{1.4}$$

et on écrit :

$$X_n - X = O\left(u_n\right), p.co.$$

Nous pouvons trouver d'autres points de vue au sujet de manière à mesurer un tel type de vitesse de convergence (voir par exemple [35] et [36]). Cette nouvelle définition est à double intérêt. D'une part, elle mérite de donner une définition précise et formelle qui est intéressante d'un point de vue probabiliste puisqu'elle implique la vitesse de convergence en probabilité O_p et la vitesse de convergence presque sûre. D'autre part, elle est intéressante d'un point de vue statistique puisqu'elle facilite la démonstration des résultats de convergence en O_p et $O_{p.s}$.

Proposition 6. *Supposons que* $X_n - X = O(u_n), p.co,$ *alors :*

i) $X_n - X = O(u_n), p.$

ii) $X_n - X = O(u_n), p.s.$

Maintenant, on va présenter dans les propositions 7 et 8 quelques règles de calcul élémentaires concernant le mode de convergence presque complète.

Proposition 7. *Supposons que* $\lim_{n\to\infty} u_n = 0$, $\lim_{n\to\infty} X_n = l_x, p.co$ *et* $\lim_{n\to\infty} Y_n = l_y, p.co$ *où* l_x *et* l_y *sont deux nombres réels déterminés :*

i) *On a :*

1. $\lim_{n\to\infty} X_n + Y_n = l_x + l_y, p.co.$

2. $\lim_{n\to\infty} X_n Y_n = l_x l_y, p.co.$

3. $\lim_{n\to\infty} \frac{1}{Y_n} = \frac{1}{l_y}, p.co,$ *avec* $l_y \neq 0.$

ii) *Si* $X_n - l_x = O(u_n), p.co$ *et* $Y_n - l_y = O(u_n), p.co.$ *on a :*

1. $(X_n + Y_n) - (l_x + l_y) = O(u_n), p.co.$

2. $(X_n Y_n) - (l_x l_y) = O(u_n), p.co.$

3. $\lim_{n\to\infty} \frac{1}{Y_n} - \frac{1}{l_y} = O(u_n), p.co,$ *avec* $l_y \neq 0.$

Proposition 8. *Supposons que* $\lim_{n\to\infty} u_n = 0, X_n = O(u_n), p.co,$ *et* $\lim_{n\to\infty} Y_n = l_y, p.co,$ *où* l_y *est un nombre réel déterminé :*

i) $X_n Y_n = O(u_n), p.co.$

ii) $\frac{X_n}{Y_n} = O(u_n), p.co,$ *avec* $l_y \neq 0.$

1.2 Les inégalités exponentielles pour les v.a.r indépendantes

Dans cette section, on considère que $Z_1, Z_2, ..., Z_n$ sont des variables aléatoires réelles centrées et indépendantes. Comme nous pouvons voir tout au long de ce travail, l'énoncé des propriétés de convergence presque complète nécessite de trouver une borne supérieure pour la probabilité de la somme des v.a.r tel que :

$$P\left(\left|\sum_{i=1}^{n} Z_i\right| > \epsilon\right),$$

où ϵ est un réel positif qui décroit avec n. Dans ce contexte, il existe des outils probabilistes puissants appelés inégualités exponentielles et la littérature fait état de plusieurs versions de ces inégalités qui diffèrent selon les différentes hypothèses vérifiées par les variables Z_i. Nous nous concentrons ici sur les inégalités exponentielles de type Bernstein, ce choix est dû au fait que la forme de l'inégalité de type Bernstein est facile et plus adaptée aux développements théoriques de statitstique fonctionnelle qui vont être énoncés tout au long de cette thèse. D'autres formes de ces inégalités peuvent être trouvées dans [27], [44].

Proposition 9. *Soit* $A_n = a_1^2 + a_2^2 + ...a_n^2$ *et supposons que :* $\forall m \geq 2, |EZ_i^m| \leq (m!/2)\, a_i^2 b^{m-2}$ *alors on a :*

$$\forall \epsilon \geq 0, P\left(\left|\sum_{i=1}^{n} Z_i\right| > \epsilon A_n\right) \leq 2\exp\left(-\frac{\epsilon^2}{2a^2\left(\frac{\epsilon b}{A_n}\right)}\right). \tag{1.5}$$

La preuve de la proposition 9 ci dessus est donnée dans [52]. Notons que cette inégalité est énoncée pour les variables aléatoires réelles non identiquement distribuées. Notons également que chaque variable Z_i peut dépendre de n. C'est pour ces deux raisons que nous introduisons le corollaire 10 qui est plus utilisé que la proposition générale 9 car il est plus adapté au cadre de notre travail.

Corollaire 10.

i) *Si* $\forall m > 1, \exists C_m > 0, E\,|Z_1^m| \leq C_m a$, *on a :*

$$\forall \epsilon > 0, P\left(\left|\sum_{i=1}^{n} Z_i\right| > \epsilon n\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{2a^2(1+a^2)}\right). \tag{1.6}$$

ii) *Supposons que les variables aléatoires sont dépendantes de n (c.à.d* $Z_i = Z_{in}$*). Si*

$\forall m > 1, \exists C_m > 0, E\,|Z_1^m| \leq C_m a^{2(m-1)}$, *et si* $u_n = \frac{a_n^2 \log n}{n}$ *vérifie :* $\lim_{n\longrightarrow\infty} u_n = 0$ *alors :*

$$\frac{1}{n}\sum_{i=1}^{n} Z_i = O(\sqrt{u_n})\ p.co. \tag{1.7}$$

Notons aussi que toutes les inégalités précédentes sont fournies pour des variables aléatoires non bornées utilisées dans la régression fonctionnelle non paramétrique. Naturellement, elles s'appliquent directement pour les variables bornées, comme c'est le cas pour la densité conditionnelle fonctionnelle. C'est pour cette raison qu'on va présenter une nouvelle version du corollaire précédent, qui est directement adapté aux variables bornées.

Corollaire 11. *i) Si* $\exists M < \infty, |Z_1| < M$ *et en faisant la notation* $\sigma^2 = E(Z_1^2)$ *on a :*

$$\forall \epsilon > 0, P\left(\left|\sum_{i=1}^{n} Z_i\right| > \epsilon n\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{2\sigma^2(1+\epsilon\frac{M}{\sigma^2})}\right), \tag{1.8}$$

ii) Supposons que les variables aléatoires sont dépendantes de n (c.à.d $Z_i = Z_{in}$*) telles que :* $\exists M = M_n < \infty, |Z_1| < M$, *et en faisant la notation* $\sigma^2 = E(Z_1^2)$ *et si* $u_n = \frac{a_n^2 \log n}{n}$ *vérifie* $\lim_{n\longrightarrow\infty} u_n = 0$ *et si* $\frac{M}{\sigma_n^2} < C < \infty$ *alors :*

$$\frac{1}{n}\sum_{i=1}^{n} Z_i = O(\sqrt{u_n}),\ p.co.$$

1.3 Présentation de différentes situations des modèles fonctionnels

La statistique fonctionnelle est une branche de la statistique à l'analyse des données fonctionnelles ([47],[46]), [24] et[9]. La modélisation des données observées dans le temps

d'une manière continu, ceci apparait normale grace aux appareils informatiques nous permet de survenus ces données à travers le temps.

On appelle modèle fonctionnel, tout modèle prenant en compte au moins une variable aléatoire fonctionnelle (une v.a.f est une variable qui prend ces valeurs dans un espace de dimension infini). Il ya deux grandes familles de modèles fonctionnels les modèles paramétriques et non paramétriques. La condition que nous utiliserons pour caractériser les 2 modèles est la forme $\Phi \in \mathsf{C}$ où C est une classe de fonctions. Le modèle est dit paramétrique si C est indéxable par un nombre fini de paramètres appartenant à un ensemble E. Le modèle fonctionnel peut être défini sous la forme suivante :

$$Y = \Phi(x) + \epsilon,$$

avec ϵ une v.a.r centrée et $\Phi \in \mathsf{C}$.
Si

$$\mathsf{C} = \left\{ \Phi : E \to \mathbb{R} / \exists \beta > 0, \exists c > 0, \forall (x_1, x_2) \in E^2, |\Phi(x_1) - \Phi(x_2)| \leq c \left(d(x_1, x_2)\right)^{\beta} \right\}.$$

Dans ce cas une telle famille n'est pas indéxable par un nombre fini de paramètres de E, on parle ainsi de modèle fonctionnel non paramétrique de régression. Il existe d'autre modèle entre le paramétrique et le non paramétrique c'est notamment le semi paramétrique. Il convient de préciser la terminologie inhérente à ces modèles. Dans ce but nous considérons $(X_i, Y_i)_{i=1,\ldots,n}$ n couples i.i.d selon la loi de (X, Y) avec X v.a.f. Selon la définition d'un modèle fonctionnel de régression, plusieurs situations peuvent être distinguées.

- La variable aléatoire Y est une réponse scalaire ou fonctionnelle.
- L'échantillon est formé de couples indépendamment ou dépendamment distribuées.
- Le modèle fonctionnel est paramétrique, non paramétrique ou semi-paramétrique.

La littérature scientifique aborde ces différents contextes, en fournissant des bases théoriques et des méthodes spécifiques adaptées à chaque situation.

1.3.1 Première situation : Modèle fonctionnel paramétrique

La réponse est scalaire et l'échantillon est formé de couples indépendamment distribués. De nombreux statisticiens ont développé des applications adaptées aux variables aléatoires fonctionnelles quand la variable réponse est scalaire. Seulement pendant très longtemps la nature fonctionnelle de ces données n'a été que partiellement prise en compte. En effet les premiers travaux se sont focalisés directement sur la version discrétisée des variables aléatoires fonctionnelles observées, ce qui revient à considérer un nombre fini de régresseurs. Frank et al. [26] proposent une importante bibliographie lié à des applications dans le domaine de chimie quantitative. Qu'il s'agisse de régression partial (partial least squares regression) ([42]) ou de regression sur composante principale [43]. Toutes ces méthodes ont pour but d'adapter la regression linéaire classique (ordinary least squares regression) lorsque le nombre de regresseur est très important devant le nombre d'individus ou bien encore quand les régresseurs sont très colérés, ces deux cas pathologique pouvant se porduire simultanément. Il faut attendre la disscussion de hatsie et Mallaws en 1993 [33] pour voir apparaitre la version fonctionnelle du modèle de régression linéaire classique. Un premier travail synthétique dans ce cadre est fourni par [45], lequel propose une méthode PLS au cadre fonctionnel. Par suite [15] ont mis en place le cadre théorique permettant l'obtention de premiers résultats asymptotiques dans le modèle fonctionnel de régression linéaire, et ils ont introduit et étudié en 2003 les propriétés asymptotiques

d'estimateurs (lisses). Enfin [16] mis en place des tests pour les modèles fonctionnels. Le modèle fonctionnel de la régression linéaire est défini comme suit : Soit X la v.a.f à valeurs dans l'espace H des fonctions définies sur C (un compact de $\mathbb{R}$) et de carrée intégrable (on parle alors de variables aléatoires hilbertiennes). On a donc $X = \{X(t), t \in C\}$. Soit $(X_i, Y_i)_{i=1,\ldots,n}$ n couples de variables aléatoires identiques et indépendamment distribuées selon le couple (X, Y) où Y est une variable aléatoire réelle. Ils se sont intéressés alors au modèle fonctionnel de régression :

$$Y = \int_C \Phi(t)X(t)dt + \epsilon,$$

où $\Phi \in H$ et ϵ est une variable aléatoire réelle centrée de variance σ^2 et telle que $E(\varepsilon|X) = 0$. Bien entendu, un tel modèle peut être reformuler de manière équivalente en munissant H du produit scalaire $\langle u, v\rangle = \int_C \Phi(t)X(t)dt$ est défini par :

$$Y = \langle \Phi, X\rangle + \epsilon, \tag{1.9}$$

ou encore

$$Y = \Phi(X) + \epsilon. \tag{1.10}$$

Dans le cadre de ce modèle, ils ont proposé différents estimateurs. Deux concernant directement l'opérateur linéaire continu Φ alors q'un autre s'intéresse uniquement à son représentant dans H à savoir Φ.

1.3.2 Deuxième situation : modèle fonctionnel paramétrique

La réponse est fonctionnelle et l'échantillon est formé de couples indépendamment distribués. Les premiers travaux théoriques dans les modèles fonctionnels de régression sont dus à Bosq [10] dans le cas des variables aléatoires fonctionnels indépendantes (processus autorégressifs hilbertiens). Il s'agit de modèles fonctionnels paramétriques où la variable réponse est aussi fonctionnelle. de nombreux développements asymptotiques concernant ces processus linéaires ont été réalisés par ([11] et [20]).

1.3.3 Troisième situatuion : Modèle fonctionnel paramétrique

La réponse est fonctionnelle et l'échantillon est formé de couples dépendamment distribués. Le papier concernant ce type de modèle a été réalisé par [19]. Ils ont fournis des premiers résultats théoriques dans le cadre de la régression linéaire fonctionnelle d'une variable aléatoire fonctionnelle sur une fonction déterministe (fixed design).

1.3.4 Quatrième situatuion : modèle fonctionnel non paramétrique

La réponse est scalaire et l'échantillon est formé de couples indépendamment distribués. Dans ce cadre, les premiers résultats, tant d'un point de vue théorique que pratique sont fournis par [23] qui ont considéré un modèle de la forme :

$$Y = \Phi(x) + \epsilon, \tag{1.11}$$

où Y est une variable aléatoire réelle et X une variable aléatoire non nécessairement réelle mais à valeurs dans un espace vectoriel semi normé $(H, \|.\|)$ et ϵ est une variable aléatoire réelle centrée et indépendante de X . Dans ce cadre ils ont cherché à estimer l'opérateur fonctionnel

$$\Phi(x) = E\left(Y|X = x\right), \text{avec} \tag{1.12}$$

$$Y = \Phi(x) + \epsilon = E(Y/X = x) + \epsilon,$$

à partir de n observation$(X_i, Y_i)_{i=1,....,n}$ indépendante du couple (X, Y). L'estimateur est défini par :

$$\hat{\Phi}_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{||X_i - x||}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{||X_i - x||}{h_n}\right)},$$

où h_n est une suite de nombres positifs. La convergence presque complète de cet estimateur est établi quand il est continu. Les vitesses sont précisées par le théorème 12 pour des opérateurs vérifiant quelques hypothèses et un ensemble de conditions de régularité.

Théorème 12. *Si Φ est continu en x alors on a :*

$$\lim_{n\to\infty} \Phi_n(x) = \Phi(x),\ p.co, \tag{1.13}$$

et

$$\Phi_n(x) - \Phi(x) = O\left(\frac{\log n}{n}\right)^{\frac{\gamma(x)}{2\gamma(x)+\alpha(x)}},\ p.co. \tag{1.14}$$

1.3.5 Cinqième situatuion : modèle fonctionnel non paramétrique

La réponse est scalaire et l'échantillon est formé de couples dépendamment distribués. Dans ce cadre, les premiers développements théoriques et pratiques ont été produit par [23] : Prédiction et variables aléatoires fonctionnelles dépendantes ; théorie et application. Ces méthodes ont été appliquées à des données de pollution dans un travail récent dû à [37].

1.3.6 Sixième situatuion : Modèle fonctionnel non paramétrique

La réponse est fonctionnelle et l'échantillon est formé de couples dépendamment distribués. On peut citer [12] qui étudièrent une prédiction non paramétrique d'une variable aléatoire fonctionnelle dans le cadre de processus markovien strictement stationnaire. Dans le contexte de variables aléatoires fonctionnelles dépendantes Besse et al [6] ont mis en œuvre une méthode non paramétrique de prédiction (réponse fonctionnelle) sur des données climatiques.

1.3.7 Septième situatuion : Modèle fonctionnel semi- paramétrique

L'échantillon est formé de couples identiquement distribués. Parmi les études rentrant dans ce sujet celle de Seron et al. [51] dans le cadre d'une extension de la méthode SIR aux variables aléatoires fonctionnelles et dont une partie des outils utilisés est similaire à l'approche théorique de cardot [15], en plus de quelques contributions concernant le modèle à indice fonctionnel simple.

1.4 La méthode du noyau

En ce qui concerne les estimateurs de type noyau, il existe une littérature abondante à laquelle Gérard collomb apporta une contribution déterminante comme en témoigne la compilation de ses travaux dans [17]. Pour les aspects théoriques ont peut trouver de bons compléments et un large éventuel de références dans [7], le manuscrit de [32] étant plus utile concernant les aspects appliqués des techniques du noyau et les problèmes

d'implémentation de ces techniques d'estimation. Les estimateurs de type noyau, introduits indépendamment par Nadaraya et Watson(1964), sont une des techniques les plus populaires d'estimation sous des modèles de régression de type non paramétrique. Pour comprendre les idées qui ont amené à l'introduction de ces estimateurs, il faut remonter au régressogramme de Tuckey(1961) défini comme suit :

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} Y_i I_{(X_i \in B_j)}}{\sum_{i=1}^{n} I_{(X_i \in B_j)}}, \forall x, \tag{1.15}$$

où B_j pour $j = 1, ..., J$ est une partition du support fixé de X a priori. Cet estimateur primitif présente comme inconvénient d'avoir à choisir à la fois le nombre de J découpages et la position exacte des bornes des intervalles B_j . Afin de résoudre ce problème un nouvel estimateur a été construit en remplaçant la discrétisation a priori en intervalles B_j par un seul intervalle mais qui varie de manière continue. Concrètement, cela donne l'estimateur de la fenêtre mobile défini comme suit :

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} Y_i I_{(X_i \in [x-h, x+h])}}{\sum_{i=1}^{n} I_{(X_i \in [x-h, x+h])}}, \forall x. \tag{1.16}$$

1.4.1 Les differents types de noyau

Une fonction K de $\mathbb{R}$ dans $\mathbb{R}^+$ telle que $\int K = 1$ est dite noyau de type I s'il existe deux constantes réelles c_1 et c_2 satisfaisant :

$$0 < c_1 < c_2 < \infty \ ,$$

telle que :

$$c_1 1_{[0,1]} \leq K \leq c_2 1_{[0,1]}.$$

Elle est dite noyau de type II si son support est $[0, 1]$ et si sa dérivée K' existe sur $[0, 1]$ et satisfait pour deux constantes réelles c_3 et c_4 :

$$-\infty < c_3 < c_4 < 0,$$

tel que :

$$c_3 \leq K' \leq c_4.$$

Une fonction K_0 de $\mathbb{R}$ dans $\mathbb{R}^+$ telle que $\int K_0 = 1$ est appelée noyau de type 0 si son support compact est $[-1, 1]$ et telle que $\forall t \in]0, 1[\, , K_0(u) > 0$. Les noyaux de type I sont souvent utilisés dans des méthodes d'estimation non-paramétriques, comme l'estimation de densité ou de régression par noyau, où ils jouent un rôle crucial dans le lissage des données par une pondération locale uniforme. Les noyaux de type II sont d'avantage utilisés dans les modèles qui nécessitent une différenciation des comportements directionnels, comme c'est le cas dans les modèles fonctionnels et semi-paramétriques. Ferraty et Vieu [24] ont utilisé ces noyaux, notamment pour l'aspect théorique, dans le cadre de l'étude asymptotique des estimateurs. Ces travaux théoriques permettent de démontrer les propriétés de convergence des estimateurs non-paramétriques basés sur ces noyaux, garantissant ainsi la validité des méthodes dans des contextes où les données sont de nature fonctionnelle. Les noyaux de type 0, en particulier, sont couramment employés pour leur support compact, permettant une pondération locale limitée aux observations les plus pertinentes. L'aspect asymptotique étudié par Ferraty [24] concerne notamment la convergence presque complète des estimateurs, des propriétés cruciales pour assurer la performance et la robustesse des méthodes dans des applications réelles.

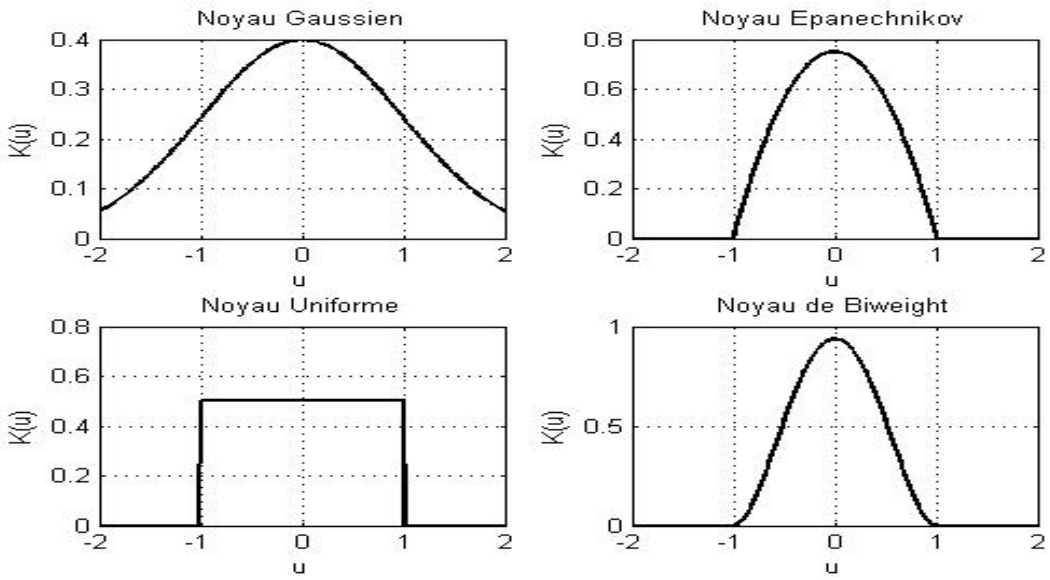

Fig. 1.1 : Quatres noyaux courrament utilisés.

1.4.2 Noyaux courrament utilisés

Prmi les noyaux courrament utilisés on cite les quatres noyaux suivants :

1. Le noyau gaussien est défini par :

$$K(u) = \frac{1}{\sqrt{2\pi}} exp(\frac{-u^2}{2}), \text{où } u \in \mathbb{R}$$

.

2. Le noyau d'Épanechnikov s'exprime comme suit :

$$\frac{3}{4}(1-u^2)1_{\{|u|\leq 1\}}.$$

3. Le noyau uniforme est défini par :

$$\frac{1}{2}1_{\{|u|\leq 1\}}.$$

4. Le noyau de biweight (ou quadratique) est une fonction de type polynomial de degré 2, définie par :

$$\frac{15}{16}(1-u^2)^2 1_{\{|u|\leq 1\}}.$$

La figure 1.1 ci-dessus illustre la représentation graphique de ces quatres noyaux.

1.5 Diverses approches au problème de prédiction

1.5.1 Définitions des prédicteurs

Les prédicteurs jouent un rôle crucial dans l'analyse de données fonctionnelles, où l'objectif est de prédire une variable scalaire Y en fonction d'une variable fonctionnelle X, prise dans un espace semi métrique $\mathcal{F}$. Ferraty et al. [24] ont proposé plusieurs approches

de prédiction non-paramétrique pour capturer la relation entre ces deux variables dans différents contextes. Soit l'échantillon $(X_i, Y_i)_{i=1,....,n}$ identiquement distribués et indépendantes, à valeurs dans $\mathcal{F} \times \mathbb{R}$ où $\mathcal{F}$ est un espace semi-métrique. Soit x (resp y) un élément fixé de $\mathcal{F}$ (resp $\mathbb{R}$), soit S un sous espace de $\mathbb{R}$ fixé. Etant donné x, on appelle $\hat{y}$ la valeur de prédiction du scalaire réponse.

La régression r de Y en x est définie par :

$$r(x) = E(Y|X = x). \tag{1.17}$$

La fonction de répartition conditionnelle de Y sachant x est définie par :

$$\forall y \in \mathbb{R} : F_Y^X(x, y) = P(Y \leq y|X = x). \tag{1.18}$$

En outre, supposons que $F_X^Y(x, y)$ est différentiable en y donc on peut écrire :

$$\forall y \in \mathbb{R} : f_Y^X(x, y) = \frac{\partial}{\partial y} F_X^Y(x, y). \tag{1.19}$$

$f_Y^X(x, y)$ est la valeur de la fonction de densité conditionnelle en (x, y). Il s'avère clair que chaque opérateur non linéaire cité donne une information sur la relation entre X et Y. Le premier opérateur nous permet de poser :

$$\hat{y} = \hat{r}(x), \tag{1.20}$$

de sorte que $\hat{r}(x)$ est l'estimateur de $r(x)$. Puisque la médiane conditionnelle $m(x)$ de F_X^Y est définie par :

$$m(x) = \inf \left\{ y \in \mathbb{R}, F_X^Y(x, y) \geq \frac{1}{2} \right\}, \tag{1.21}$$

alors le deuxième opérateur nous permet d'utiliser le prédicteur :

$$\hat{y} = \hat{m}(x), \tag{1.22}$$

où $\hat{m}(x)$ est l'estimateur de $m(x)$. Le mode conditionnel $\theta(x)$ de $f_Y^X(x, y)$ est définie par :

$$\theta(x) = \arg\sup_{x \in S} f_Y^X(x, y), \tag{1.23}$$

alors le troisième opérateur nous permet d'utiliser le pré dicteur :

$$\hat{y} = \hat{\theta}(x), \tag{1.24}$$

où $\hat{\theta}(x)$ est l'estimateur de $\theta(x)$. Les quantiles sont définés pour $\alpha \in]0, 1[$ par :

$$t_\alpha(x) = \inf \left\{ y \in \mathbb{R}, F_X^Y(x, y) \geq \alpha \right\}, \tag{1.25}$$

donc l'estimateur $\hat{t}_\alpha(x)$ se construit de l'intervalle :

$$\left[\hat{t}_\alpha(x), \hat{t}_{1-\alpha}(x) \right] \text{ pour } \alpha \in \left] 0, \frac{1}{2} \right[. \tag{1.26}$$

1.5.2 Présentation des modèles associés aux prédicteurs

Ferraty et al.[24] ont étudié plusieurs modèles non paramétriques qui permettent d'estimer ces prédicteurs dans différents cadres fonctionnels. Deux grandes catégories de modèles sont souvent considérées : les modèles sous condition de continuité et ceux basés sur la propriété de Lipschitz. Dans le cadre de la régression non-paramétrique, un modèle courant repose sur l'hypothèse r appartient à l'espace de fonctions continues sur $\mathcal{F}$, noté C_F^0, défini comme suit :

$$r \in C_F^0, \tag{1.27}$$

où

$$C_F^0 = \left\{ f : \mathcal{F} \to \mathbb{R}, \lim_{d(x,x')\to 0} f(x) = f(x') \right\},$$

ou $\exists \beta > 0$ such that :

$$r \in Lip_{\mathcal{F},\beta}, \tag{1.28}$$

avec

$$Lip_{\mathcal{F},\beta} = \left\{ f : \mathcal{F} \to \mathbb{R}, \exists C \in \mathbb{R}+^*, \forall x' \in \mathcal{F}, |f(x) - f(x')| < C d(x, x^{'}) \right\}.$$

Pour l'opérateur de la médiane conditionnelle, qui est basé sur l'opérateur F_X^Y (opérateur non-linéaire de $\mathcal{F} \times \mathbb{R}$ dans $\mathbb{R}$), on suppose que :

$$F_X^Y \in S_{cfd}^x, \tag{1.29}$$

avec

$$S_{cfd}^x = \{ f : \mathcal{F} \times \mathbb{R} \to \mathbb{R}, f(x, .) \text{ est strictement croissante} \}.$$

En effet $F_X^Y \in S_{cfd}^x$ assure l'existence et l'unicité de la médiane conditionnelle qui est définit comme suit :

$$m(x) = F_Y^{x^{-1}}(y, x). \tag{1.30}$$

On considère les deux ensembles suivants :

$$C_{\mathcal{F}\times\mathbb{R}}^0 = \left\{ \begin{array}{l} f : \mathcal{F} \times \mathbb{R} \to \mathbb{R}, \forall x' \in N_x, \\ \lim\limits_{d(x,x')\to 0} f(x, y) - f(x', y) = 0, \\ \forall y' \in \mathbb{R}, \lim\limits_{|y'-y|\to 0} f(x, y^{'}) - f(x, y) = 0. \end{array} \right\} \tag{1.31}$$

$$lip_{\mathcal{F}\times\mathbb{R},\beta} = \left\{ \begin{array}{l} f : \mathcal{F} \times \mathbb{R} \to \mathbb{R}, \\ \forall (x_1, x_2) \in N_x^2, \forall (y_1, y_2) \in \mathbb{R}^2, \\ |f(x_1, y_1) - f(x_2, y_2)| \leq C \left(d^{\beta}(x_1, x_2) + |y_1 - y_2|^{\beta} \right) \end{array} \right\} \tag{1.32}$$

Nous pouvons donc définir deux modèles non-paramétriques fonctionels de l'opérateur de la fonction de distribution fonctionelle comme suit :

$$F_Y^X \in C_{\mathcal{F}\times\mathbb{R}}^0 \cap S_{cdf}^x, \tag{1.33}$$

et $\exists \beta > 0$ tel que :

$$F_Y^X \in Lip_{\mathcal{F}\times\mathbb{R},\beta} \cap S_{cdf}^x. \tag{1.34}$$

La prédiction concernant le mode conditionnel repose sur l'opérateur de la fonction de densité conditionnelle f_Y^X (opérateur non linéaire de $\mathcal{F} \times \mathbb{R}$ dans $\mathbb{R}$).
On introduit l'ensemble suivant :

$$S_{dens}^X = \left\{ \begin{array}{l} f : \mathcal{F} \times \mathbb{R} \to \mathbb{R}, \\ \exists \xi > 0, \exists! y_0 \in S, f(x, .) \text{ est strictement} \\ \text{croissante sur } (y_0 - \xi, y_0). \end{array} \right\} \tag{1.35}$$

Alors, si $f(x,y) \in S^X_{dens}$ selon le problème de maximisation de $f(x,y)$ sur S, elle admet une solution unique y_0. Donc le mode conditionnel peut être défini de la façon suivante :

$$\theta(x) = \arg\sup_{y \in S} f_Y^X \ (x,y)\,. \tag{1.36}$$

$f(x,y) \in S^X_{dens}$ assure l'unicité de $\theta(x)$ et nous avons deux modèles non-paramétriques pour la prédiction du mode conditionnelle. Soit

$$f_Y^X \in C^0_{\mathcal{F}\times\mathbb{R}} \cap S^x_{dens}\ , \tag{1.37}$$

ou

$$f_Y^X \in Lip_{\mathcal{F}\times\mathbb{R},\beta} \cap S^x_{dens}\ . \tag{1.38}$$

Remarque 13. *Les modèles 1.25 , 1.31 et 1.35 s'appellent modèles non-paramétriques fonctionnels de type continuité, ils nous permettent d'etudier la convergence presque complète alors que 1.26, 1.32 et 1.36 sont de type Lipschitz et nous permettent d'etudier la vitesse de convergence presque complète.*

Chapitre 2

Estimation de la fonction de distribution conditionnelle sur un modèle de régression non paramétrique

2.1 Introduction

Ce chapitre est tiré de l'article publié intitulé *Estimation in nonparametric functional-on regression models with real responses* paru dans le journal *Utilitas Mathematica* [3]. L'analyse de données fonctionnelles est un domaine des statistiques qui traite des observations qui sont des fonctions ou des courbes, plutôt que des scalaires ou des vecteurs. Les données fonctionnelles apparaissent dans de nombreux domaines d'application, tels que la médecine, la biologie, la finance, l'écologie, etc. Il convient de noter que la modélisation des variables fonctionnelles gagne en popularité, comme en témoignent les monographies de James. O. Ramsay et Bernard. W. Silverman [46] et Ferraty et Vieu [24] sur l'analyse fonctionnelle des données (FDA). L'un des problèmes fondamentaux de l'analyse des données fonctionnelles est la régression fonctionnelle, qui consiste à étudier la relation entre une variable de réponse fonctionnelle et une ou plusieurs variables explicatives, qui peuvent également être fonctionnelles ou scalaires. L'estimation de la fonction de distribution cumulative conditionnelle de la variable réponse repose sur la fonction de régression fonctionnelle, compte tenu des variables explicatives, comme discuté par Ahmed Ait Saidi et Mecheri Kheira [49]. La fonction de distribution conditionnelle contient toutes les informations sur la distribution de la variable réponse, et permet de calculer des quantités d'intérêt, telles que la moyenne conditionnelle, la variance conditionnelle, les quantiles conditionnels, etc., comme discuté par Ali Gannoun, Jérôme Saracco, et Keming Yu [28] et Manteiga et Vieu. [41].
L'estimation de la fonction de distribution conditionnelle est donc une étape essentielle de la régression fonctionnelle. Dans ce chapitre, nous nous intéressons à la méthode du noyau pour l'estimation de la fonction de distribution conditionnelle et sa régression pour les variables fonctionnelles. La méthode du noyau est une technique non paramétrique qui utilise une fonction de pondération, appelée noyau, pour lisser les données et obtenir une estimation de la fonction de distribution conditionnelle, comme détaillé par Gao et Gijbels [29] et Ferraty et Vieu. [25].
La méthode du noyau présente plusieurs avantages, tels que sa simplicité, sa flexibilité et sa robustesse. Elle permet également de traiter des cas complexes, tels que la régression fonctionnelle multivariée, la régression fonctionnelle avec des données censurées ou la régression fonctionnelle avec des données hétérogènes, comme l'expliquent Li et Hsing [39].

Nous présentons dans ce chapitre la théorie, la pratique et les applications de la méthode du noyau pour l'estimation de la fonction de distribution conditionnelle et sa régression pour des variables aléatoires fonctionnelles (c'est-à-dire prenant leurs valeurs dans un espace de dimension infinie). Nous décrivons le principe de la méthode, les choix du noyau et de la bande passante, les propriétés asymptotiques et les critères de performance. Il est indiqué qu'il existe d'autres méthodes appropriées pour estimer la régression fonctionnelle non paramétrique, telles que la régression fonctionnelle linéaire ou la régression spline. Nous illustrons la méthode avec des données simulées, et nous discutons de ses limites et de ses perspectives de recherche futures. Nous montrons que la méthode du noyau est simple, flexible et robuste, et qu'elle peut traiter des cas complexes de régression fonctionnelle. Citons comme exemple, pour le cas indépendant et identiquement distribué (i.i.d.) des données, sous certaines conditions, certains résultats asymptotiques tels que la convergence ponctuelle presque complète et la convergence uniforme presque complète de l'estimateur du noyau avec leurs vitesses sont établis. Notre intérêt vient principalement du fait que de nombreux chercheurs se sont penchés sur l'aspect théorique de l'estimation de la fonction de distribution cumulative conditionnelle sous le modèle de régression à indice fonctionnel. les travaux de Ait Saidi et Kheira [49] ont établi le cadre théorique de la fonction de distribution conditionnelle pour un échantillon de données indépendantes, où la variable réponse est scalaire et la variable explicative prend ses valeurs dans un espace vectoriel semi-normé de dimension infinie. D'une part, Delsol [22] a conduit des simulations pour l'estimation non paramétrique de la moyenne conditionnelle par méthode à noyau, utilisant un échantillon de données fonctionnelles indépendantes et identiquement distribuées (i.i.d.). D'autre part, Ait Saidi et al. [4] ont étudié la convergence de l'estimateur semi-paramétrique de la moyenne conditionnelle fonctionnelle à partir de données fonctionnelles i.i.d. simulées. Enfin, Ait Saidi et Kheira [49] ont développé l'aspect théorique de la fonction de répartition conditionnelle dans le modèle à indice fonctionnel, cadre sur lequel s'appuie l'étude pratique présentée dans ce chapitre de notre thèse. Le manuscrit de [1] propose un estimateur linéaire local de la fonction de risque conditionnel pour les modèles d'indice en cas de données manquantes au hasard. La méthode consiste à utiliser une approche basée sur le noyau pour estimer la fonction de risque à chaque point du domaine et à incorporer les données manquantes à l'aide d'une pondération de probabilité inverse. L'article de [50] aborde la régression non paramétrique et la classification avec des covariables fonctionnelles, catégorielles et mixtes. La méthode proposée repose sur l'utilisation d'un cadre de modèle additif généralisé (GAM) pour modéliser la relation entre la variable réponse et les covariables, qui peuvent inclure des données fonctionnelles, catégorielles ou mixtes. Par conséquent, l'estimation de la fonction de distribution conditionnelle dans le modèle à indice fonctionnel pour le cas d'indépendance des données simulées devrait être une question importante. Le chapitre est structuré comme suit : dans la section 2, nous présentons notre modèle et notre estimateur, ainsi que les concepts clés, les hypothèses et les notations. La section 3 est consacrée à une étude de simulation. La section 4 présente les résultats et la discussion. Enfin, nous résumons nos résultats et présentons nos conclusions dans la dernière section.

2.2 Modèle et estimation non paramétrique

2.2.1 Construction de l'estimateur non paramétrique

Soit (X, Y) un couple de variables aléatoires prenant leurs valeurs dans $E \times \mathbb{R}$, où (E, d) est un espace semi-métrique (c'est-à-dire que X est une variable aléatoire fonctionnelle $(v.a.f)$ et d une semi-métrique). On considère l'échantillon $(X_i, Y_i)_{i=1,\dots,n}$ de n

couples indépendants identiquement distribués (i.i.d) comme le couple (X, Y). Soit (x, y) un élément fixe de $(E, \mathbb{R})$, soit $N_x \subset E$ un voisinage de x et S un sous-ensemble compact fixe de $\mathbb{R}$. Étant donné x, désignons par $\hat{y}$ une valeur prédite pour la réponse scalaire. L'opérateur de régression (non linéaire) r de Y sur X est défini par

$$r(x) = E(Y|X = x). \tag{2.1}$$

et la fonction de distribution cumulative conditionnelle (FCD) de Y étant donné X est définie par :

$$\forall y \in \mathbb{R}, F_Y^X(x, y) = P(Y \leq y|X = x). \tag{2.2}$$

Nous considérons les conditions introduites par Ferraty et al. [24] sur les deux opérateurs. Nous définissons par $\hat{r}(x)$ l'estimateur à noyau de $r(x)$ tel que :

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} Y_i K(\frac{1}{h} d(X_i, x))}{\sum_{i=1}^{n} K(\frac{1}{h} d(X_i, x))}, \tag{2.3}$$

et par $\hat{F}_Y^X(x, y)$ l'estimateur à noyau de $F_Y^X(x, y)$ tel que :

$$\forall y \in \mathbb{R} : \hat{F}_Y^X(x, y) = \frac{\sum_{i=1}^{n} G(\frac{1}{g}(y - Y_i)) K(\frac{1}{h} d(X_i, x))}{\sum_{i=1}^{n} K(\frac{1}{h} d(X_i, x))}. \tag{2.4}$$

La fonction K est un noyau de type I ou de type II et lé noyau G est défini par : $\forall t \in \mathbb{R}$, $G(t) = \int_{-\infty}^{t} K_0(v)dv$ où K_0 est un noyau de type 0 et $h = h(n)$ $(resp.g = g(n))$ est une suite de nombres réels positifs qui tend vers zéro lorsque n tend vers $+\infty$. Cette estimation prolonge, de manière différente, les travaux de Roussas [48] dans le cas réel et de Ferraty et al. [24] dans le cas fonctionnel.

2.2.2 Convergence presque complète ponctuelle en estimation non paramétrique

Nous allons rappeler quelques résultats théoriques développés en détail dans [24], ainsi que leurs démonstrations, sur lesquels nous nous appuierons pour établir la partie théorique de ce chapitre. Soit $\mathcal{N}_x \subset \mathcal{H}$ un voisinage de x et $\mathcal{S}$ un sous-ensemble compact fixé de $\mathbb{R}$. Afin d'établir la convergence presque complète (p.co.) de nos estimateurs, nous présentons quelques conditions de régularité nécessaires à la formulation de ce résultat, introduites par Ferraty et al. [24].

(H_1) Le modèle de la fonction du lien :

$$Y = E(Y|X) + \epsilon, \text{ où } E(\epsilon|X) = 0. \tag{2.5}$$

(H_2) La probabilité que la variable fonctionnelle soit dans une petite boule est non nulle,

$$\forall \epsilon > 0, \ P(X \in B(x, \epsilon)) = \varphi_x(\epsilon) > 0. \tag{2.6}$$

$$(H_3) \begin{cases} \text{h est une suite de nombres positifs telle que :} \\ \lim_{n\to+\infty} h = 0 \text{ et } \lim_{n\to+\infty} \frac{1}{\varphi_x(h)} \cdot \frac{\log n}{n} = 0. \\ K \text{ est un noyau de type I ou de type II et} \\ \exists C > 0, \exists \epsilon > 0, \forall \epsilon < \epsilon_0, \int_0^1 \varepsilon\varphi_x(u)du > C\varepsilon\varphi_x(\varepsilon). \end{cases} \tag{2.7}$$

(H_4) g est une suite positive telle que : $\lim_{n\to\infty} g = 0$, et K_0 est de type 0.
(H_5) Nous considérons une variable de réponse scalaire Y telle que :

$$\forall m \geq 2,\ E\left(|Y^m||X = x\right) < \sigma_m(x) < \infty \ \text{ avec } \sigma_m(.) \text{ continue en } \ x. \tag{2.8}$$

(H_6) L'opérateur de régression r (qui est un opérateur non linéaire de E dans $\mathbb{R}$) vérifie :

$$r : E \to \mathbb{R}, \lim_{d(x_1,x_2)\to 0} r(x_1) = r(x_2). \tag{2.9}$$

(H_7) L'opérateur F_Y^X (qui est un opérateur non linéaire de $E \times \mathbb{R}$ vers $\mathbb{R}$) vérifie :

$$F : E \times \mathbb{R} \to \mathbb{R}, \forall x' \in \mathcal{N}_x, \lim_{d(x',x)\to 0} F_Y^X(x', y) = F_Y^X(x, y), \tag{2.10}$$

et

$$\forall y' \in \mathbb{R}, \lim_{|y'-y|\to 0} F_Y^X(x, y') = F_Y^X(x, y). \tag{2.11}$$

Théorème 14. *Sous les hypothèses* $(H_1), (H_2), (H_3), (H_5)$ *et* (H_6) *nous avons :*

$$\lim_{n\to\infty} \hat{r}(x) = r(x),\ p.co. \tag{2.12}$$

Théorème 15. *Sous l'hypothèse* $(H_1), (H_2), (H_3), (H_4)$ *et* (H_7) *on a pour tout nombre réel fixé* y

$$\lim_{n\to\infty} \hat{F}_Y^X(y) = F_Y^X(y),\ p.co. \tag{2.13}$$

Nous examinons maintenant la convergence presque complète ainsi que la vitesse de convergence de l'estimateur de régression (respectivement de répartition conditionnelle).

Démonstration des théorèmes 14 et 15. Tout d'abord on rappelle que

$$\Omega_i(x) = \frac{K\left(\frac{d(X_i,x)}{h}\right)}{EK\left(\frac{d(X_i,x)}{h}\right)}, \text{ pour } i = 1, ..., n. \tag{2.14}$$

La variable Ω_i est définie lorsque $EK\left(\frac{d(X_i,x)}{h}\right)$ est différent de zéro. Ainsi, nous présentons ici une démonstration détaillée du résultat 2.15 issu directement du lemme de Ferraty [24], qui permet de définir rigoureusement la variable aléatoire Ω_i. Ce résultat est crucial pour la suite de notre travail, car il servira de base, sous forme du modèle semi-paramétrique, dans les deux chapitres suivants.

$$c_1\varphi_x(h) \leq EK\left(\frac{d(X,x)}{h}\right) \leq c_2\varphi_x(h). \tag{2.15}$$

Passons maintenant à la démonstration du 2.15. Si K est un noyau de type I, on a :

$$c_1 1_{[0,1]} \leq K \leq c_2 1_{[0,1]} \text{ où } c_1, c_2 \in \mathbb{R}_+^*, \tag{2.16}$$

ce qui implique :

$$c_1 1_{B(x,h)}(X) \leq K\left(\frac{d(X,x)}{h}\right) \leq c_2 1_{B(x,h)}(X), \tag{2.17}$$

ce qui implique directement :

$$c_1\varphi_x(h) \leq EK\left(\frac{d(X,x)}{h}\right) \leq c_2\varphi_x(h). \tag{2.18}$$

Si K est un noyau de type II, on a :

$$EK\left(\frac{d(X,x)}{h}\right) = \int_0^1 K\left(\frac{d(X,x)}{h}\right) dP\left(\frac{d(X,x)}{h} < 1\right). \tag{2.19}$$

Puisque K' existe, nous pouvons écrire : $K(t) = K(0) + \int_0^t K'(u)\,du$, ce que cela implique :

$$\begin{aligned} EK\left(\frac{d(X,x)}{h}\right) &= \int_0^1 K(0) dP\left(\frac{d(X,x)}{h} < 1\right) + \int_0^1 \left(\int_0^t K'(u)\,du\right) dP\left(\frac{d(X,x)}{h} < 1\right) \\ &= K(0)\varphi_x(h) + \int_0^1 \left(\int_0^1 K'(u) 1_{[u,1]}(t)\,du\right) dP\left(\frac{d(X,x)}{h} < 1\right). \end{aligned}$$

En appliquant le théorème de Fubini, nous obtenons :

$$EK\left(\frac{d(X,x)}{h}\right) = K(0)\varphi_x(h) + \int_0^1 K'(u) P\left(u \leq \frac{d(X,x)}{h} \leq 1\right) du. \tag{2.20}$$

En utilisant le fait que $K(1) = 0$ nous pouvons écrire :

$$EK\left(\frac{d(X,x)}{h}\right) = -\int_0^1 K'(v)\varphi_x(hv)\,dv. \tag{2.21}$$

En utilisant l'hypothèse (H_3) avec $h < \epsilon$, on arrive à :

$$EK\left(\frac{d(X,x)}{h}\right) \geq c_4\varphi_x(h). \tag{2.22}$$

Puisque K est borné et de support $[0,1]$ en utilisant le résultat de l'équation (2.15) on obtient :

$$EK\left(\frac{d(X,x)}{h}\right) \leq c_3\varphi_x(h). \tag{2.23}$$

Pour aborder la preuve des deux théorèmes mentionnés, nous aurons besoin de la décomposition 2.24 ci-dessous afin de démontrer la convergence presque complète de la régression, ainsi que de la décomposition 2.25 pour établir la convergence de sa fonction de distribution conditionnelle. Nous utiliserons, à cet effet, les lemmes 16, 17 et 18 présentés ci-après.

$$\hat{r}(x) - r(x) = \frac{1}{\hat{r}_1(x)}\left[(\hat{r}(x) - E\hat{r}_2(x) - (r(x) - E\hat{r}_2(x))\right] - \frac{r(x)}{\hat{r}_1(x)}(\hat{r}_1(x) - 1), \tag{2.24}$$

et

$$\begin{aligned} \hat{R}_Y^X(x,y) - R_Y^X(x,y) = \frac{1}{\hat{r}_1(x)}&\left[\left(\hat{R}_Y^X(x,y) - E\hat{r}_3(x,y) - (R_Y^X(x,y) - E\hat{r}_3(x,y)\right)\right] \\ &- \frac{R_Y^X(x,y)}{\hat{r}_1(x)}(\hat{r}_1(x) - 1), \end{aligned} \tag{2.25}$$

où

$$\hat{r}_1(x) = \frac{1}{n}\sum_{i=1}^{n}\Omega_i(x) \quad \text{et} \quad \hat{r}_3(x,y) = \hat{r}_1(x)\,\hat{R}_Y^X(x,y) = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i(y)\Omega_i(x),$$

$$\hat{r}_1(x) = \frac{1}{n}\sum_{i=1}^{n}\Omega_i(x) \quad \text{et} \quad \hat{r}_2(x) = \frac{1}{n}\sum_{i=1}^{n}Y_i\Omega_i(x),$$

avec $\Omega_i(x) = \dfrac{K\left(\dfrac{d(X_i,x)}{h}\right)}{EK\left(\dfrac{d(X_i,x)}{h}\right)}$ et $\Gamma_i(y) = G\left(\dfrac{y-Y_i}{g}\right)$.

La démonstration des deux théorèmes 14 et 15 est une conséquence directe des lemmes 16, 17 et 18 suivants : □

Lemme 16.

1. *Sous les hypothèses du théorème du 14, lorsque n tend vers l'infini, on a*

$$\hat{r}_1(x) - 1 = O_{p.co.}\left(\sqrt{\frac{\log n}{n\varphi_x(h)}}\right). \tag{2.26}$$

2. *Sous les hypothèses du théorème du 14, lorsque n tend vers l'infini, on a*

$$E\hat{r}_2(x) - \hat{r}_2(x) = O_{p.co.}\left(\sqrt{\frac{\log n}{n\varphi_x(h)}}\right). \tag{2.27}$$

Lemme 17. *Sous les conditions (H_3) et (H_6), lorsque n tend vers l'infini, on a*

$$\lim_{n\to\infty} E\hat{r}_2(x) = r(x),\ p.co. \tag{2.28}$$

Lemme 18. *Sous les hypothèses du théorème 15, lorsque n tend vers l'infini, on a*

1.

$$\lim_{n\to\infty} E\hat{r}_3(x,y) = R_Y^X(x,y),\ p.co. \tag{2.29}$$

2.

$$\lim_{n\to\infty} \hat{r}_3(x,y) - E\hat{r}_3(x,y) = 0,\ p.co. \tag{2.30}$$

Le dénominateur des décompositions 2.24 et 2.25 est traité en utilisant la première partie du lemme 16 et la première partie de la proposition A.6 (voir [24]). Parallèlement, les numérateurs de la décomposition 2.24 sont traités en appliquant les lemmes 16 et 17 ci-dessus, et les numérateurs de la décomposition 2.25 sont traités en utilisant la première partie du lemme 16 et le lemme 18. Pour plus de détails sur les preuves, on peut se référer à [24].

2.3 Simulation : approche non paramétrique

Dans cette section, nous nous intéressons à l'étude appliquée de l'étude théorique réalisée par [49] pour estimer la fonction de distribution cumulative conditionnelle sous le modèle de régression non paramétrique dans les conditions de régularité citées dans la section 2, où nous adoptons les données de simulation qui ont servi à estimer la régression

fonctionnelle semi-paramétrique d'une part, puis nous adoptons les données de simulation proposées par [22] pour estimer la régression fonctionnelle non paramétrique d'autre part.

Pour illustrer les concepts cités dans la section 2 par simulation, considérons deux exemples. Dans le premier exemple, nous générons un échantillon de courbes en utilisant des techniques de simulation de données. Nous construisons un ensemble de 100 courbes, chacune définie par une fonction.
On construit un échantillon de $n = 100$ courbes comme suit :

$$x_i(t_j) = \cos(w_i + \pi(2t_j - 1)),$$

où $0 < t_1 < t_2 < ... < t_{100} = 1$ sont des points équidistants, w_i étant des observations indépendantes uniformément distribuées sur $\left[0, \frac{\pi}{4}\right]$. La figure 2.1 donne une idée de leur forme.

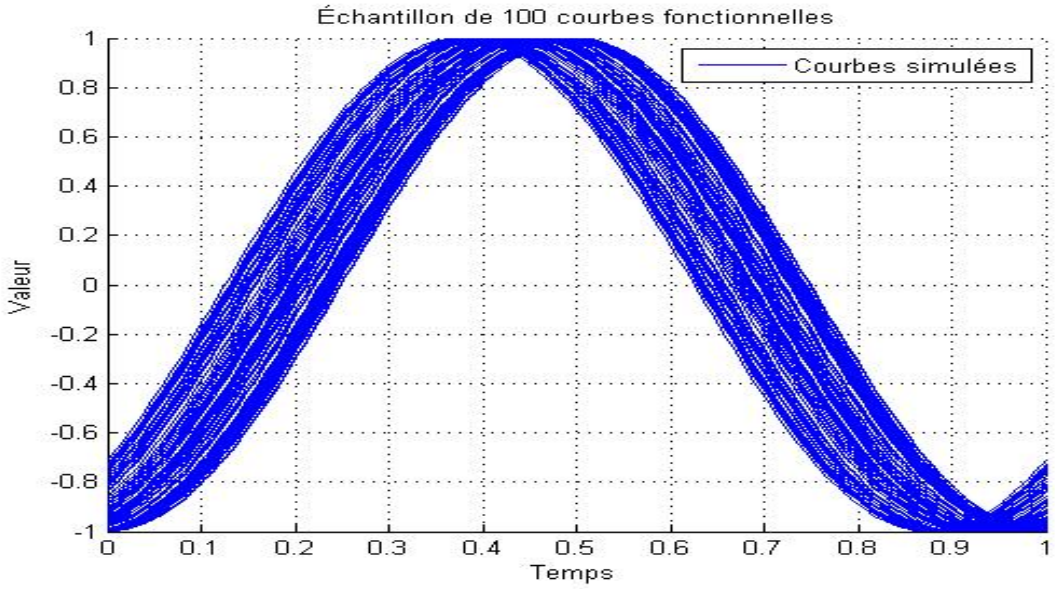

Fig. 2.1 : Un échantillon de données fonctionnelles $n = 100$.

On simule un modèle de régression non paramétrique comme suit :

1. On prend une fonction de lien r définie de $C^1(\mathbb{R})$ dans $\mathbb{R}$ par

$$r(f) = \int_{\frac{1}{4}}^{\frac{3}{4}} \left(f'(t)\right)^2 dt.$$

2. Générer indépendamment $\varepsilon_1, ..., \varepsilon_{100}$ à partir d'une loi normale centrée de variance égale à $0,05$.

3. Simuler les réponses correspondantes $Y_i = r(X_i) + \varepsilon_i$.

4. Simuler les estimateurs associés à la fonction de lien r ainsi que leur fonction de distribution cumulative conditionnelle correspondante R.

5. En plus, nous avons utilisé les valeurs de l'erreur quadratique moyenne (MSE) pour évaluer la qualité de toute estimation de courbe pour la régression.

$$MSE = \frac{1}{100}\sum_{i=1}^{n}\left(\hat{r}(x_i) - r(x_i)\right)^2.$$

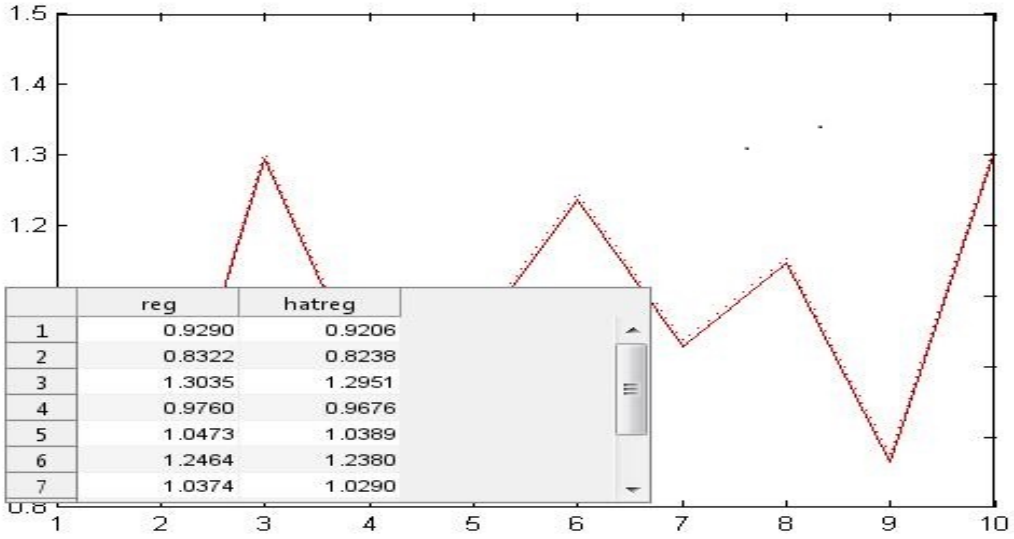

Fig. 2.2 : Régression et son estimateur pour 10 courbes.

n	h	$r(x)$	$\widehat{r}(x)$	MSE
10	.001	113.9808 -2.5103e+03 -2.3740e+03 -231.4304 1.6084e+03 7.1522e+03 -940.1715 -1.1542e+04 318.8997 -34.3421	113.9593 -2.5103e+03 -2.3740e+03 -231.4787 1.6084e+03 7.1521e+03 -940.2087 -1.1542e+04 318.8587 -34.2866	0.02

Tab. 2.1 : Tableau de comparaison de la régression et de son estimateur avec MSE.

Dans le deuxième exemple, nous comparons l'estimateur correspondant de la distribution cumulative conditionnelle à la vraie fonction tout en conservant les mêmes données et le même modèle de régression étudiés dans le premier exemple comme suit

1. La vraie distribution conditionnelle est définie comme une fonction de distribution cumulative (CDF) normale centrée autour de $0,5$ avec un écart type de $0,1$.

2. Simuler l'estimateur correspondant à la fonction de distribution cumulative conditionnelle (CDF).

3. Simuler la moyenne et l'écart type du (CDF) estimé.

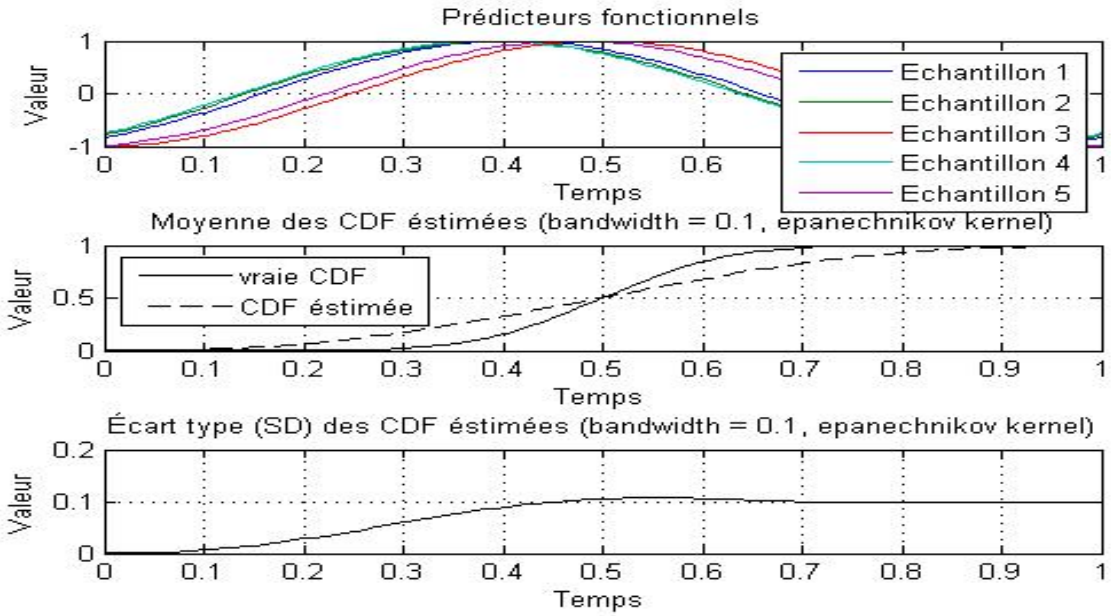

Fig. 2.3 : Moyenne et écart type de l'estimateur.

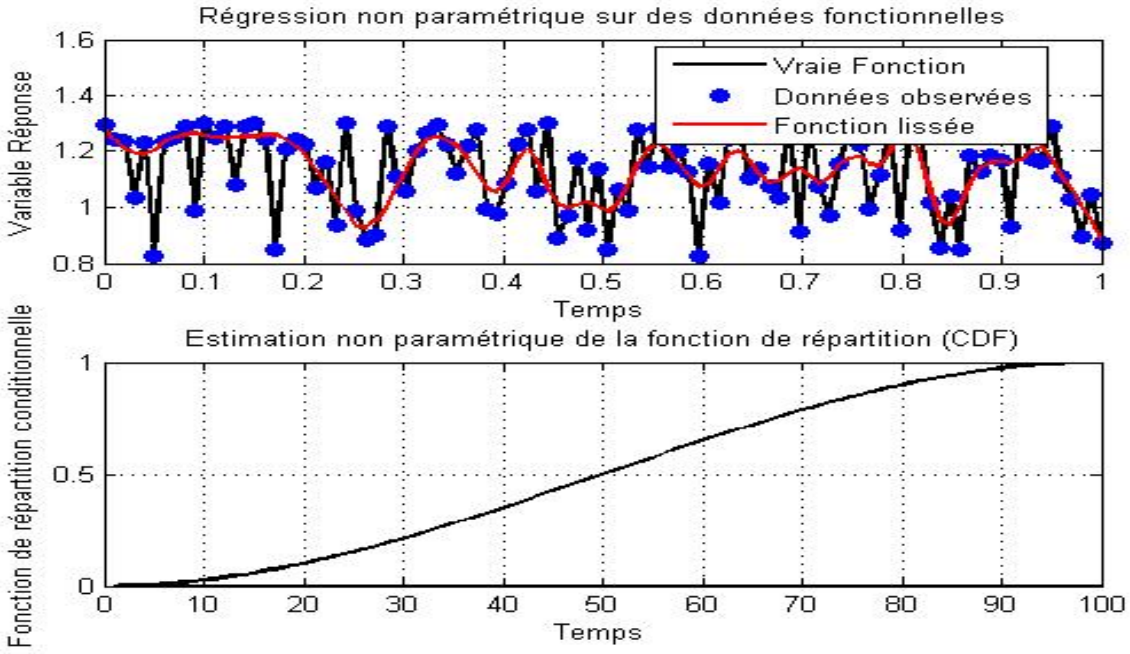

Fig. 2.4 : Estimateurs de régression et de répartition conditionnelle .

2.4 Discussion

Notre approche globale vise à affirmer la convergence complète de la fonction de distribution conditionnelle pour les variables aléatoires fonctionnelles. En intégrant les techniques de simulation empiriques de [22] et [4], ainsi que le cadre théorique fourni par [49], nous cherchons à mieux comprendre les propriétés de convergence et le comportement de la fonction de distribution cumulative. Le défi rencontré lors de la simulation concerne l'absence de la fonction de distribution cumulative conditionnelle théorique, cruciale pour comparer les résultats simulés. En plus de la difficulté associée à la sélection des deux paramètres de lissage et des deux noyaux intégrés à l'estimateur, cet écart théorique représente une limitation importante. Pour répondre à cette contrainte, nous avons adopté une approche de balayage pour estimer la fonction de distribution cumulative conditionnelle. Les résultats de la simulation sont présentés dans le tableau 2.1 et la figure 2.2 pour le premier exemple, ainsi que dans les figures 2.3 et 2.4 pour le second. Ils mettent en évidence la nécessité de poursuivre les recherches afin de développer des méthodes plus robustes et plus complètes dans ce domaine.

2.5 Conclusion

En conclusion, notre étude s'est focalisée sur l'estimation de la fonction de répartition cumulative pour des données fonctionnelles à l'aide de la méthode du noyau. Pour la sélection de la largeur de fenêtre, nous avons opté pour une approche par balayage. Nous avons validé nos résultats en vérifiant la convergence presque complète théorique de l'estimateur de la fonction de répartition conditionnelle, renforcée par des simulations qui ont servi d'outil prédictif, confirmant ainsi la robustesse et l'applicabilité de notre approche. Nos résultats soulignent également l'importance de choisir avec soin les paramètres de lissage dans la méthode du noyau. Bien que nous ayons employé une approche par balayage, courante dans la littérature, il est important de noter que des alternatives, telles que la méthode Leave-One-Out, pourraient aussi être envisagées. L'adoption de techniques de sélection de paramètres plus avancées pourrait accroître la précision de nos estimations et constitue une piste intéressante pour les recherches futures. Ainsi, nos travaux apportent une contribution significative au domaine en expansion de l'analyse des données fonctionnelles et fournissent des recommandations pour les meilleures pratiques en matière d'estimation des fonctions de répartition conditionnelles, notamment dans le cadre de la prédiction de variables fonctionnelles.

Chapitre 3

Choix de la bande passante pour l'estimateur de la régression fonctionnelle

3.1 Introduction

Ce chapitre est basé sur un article publié dans le Journal of Applied Probability and Statistics [2]. Il présente une approche semi-paramétrique pour l'estimation de la régression fonctionnelle, et les résultats obtenus sont discutés ici. En statistiques, de nombreuses situations visent à étudier la corrélation entre deux variables afin de pouvoir prédire les valeurs de l'une à partir de la connaissance de l'autre. Le problème de la prédiction a été largement étudié dans la littérature, notamment lorsque les valeurs des deux variables sont dans un espace de dimension finie. De même, le problème de prédiction persiste lorsque les deux variables sont fonctionnelles, c'est-à-dire qu'elles prennent leurs valeurs dans un espace de dimension infinie. Notre objectif est d'examiner ce problème, en particulier lorsque la variable endogène est fonctionnelle et que la variable exogène est à valeurs réelles. Depuis la publication de l'étude fondatrice sur l'analyse des données fonctionnelles (Functional Data Analysis, FDA) par Ramsay et Silverman [46], l'application des techniques de la FDA à des problèmes impliquant des variables fonctionnelles et à valeurs réelles est devenue de plus en plus courante. Ces dernières années, la modélisation des variables fonctionnelles a suscité un intérêt croissant. La vitesse de convergence uniforme des estimateurs non paramétriques avec des variables fonctionnelles a été étudiée par Ferraty et al. [25]. L'étude asymptotique des estimateurs de certains opérateurs dans des modèles non paramétriques pour des variables fonctionnelles a été abordée dans [24]. Ces auteurs ont examiné l'estimation non paramétrique de certaines fonctions à partir de l'estimation de l'espérance conditionnelle d'une variable de réponse scalaire Y, sachant que la variable aléatoire X prend ses valeurs dans un espace semi-métrique, telle que la régression conditionnelle, la densité conditionnelle, la fonction de répartition conditionnelle, le mode conditionnel, et la médiane conditionnelle. Notre travail se consacre à l'étude pratique du traitement des données fonctionnelles, basé sur une approche de réduction de dimension par le biais de la modélisation à indice unique. Nous étudions ainsi le lien entre la variable explicative et la variable de réponse à travers un indice fonctionnel. L'approche à indice unique est une technique couramment utilisée dans les modèles semi-paramétriques, qui offrent un compromis entre les modèles paramétriques et non paramétriques. Cette approche a suscité un intérêt considérable dans le domaine de l'économétrie, et de nombreuses études dans la littérature exploitent le modèle à indice lorsque la variable explicative est fonctionnelle. De plus, cette approche a été appliquée à des variables explicatives fonctionnelles, comme on le voit dans les travaux de [30], ce qui démontre sa polyvalence et sa capacité à gérer une large gamme de types de données.

Par exemple, les travaux de Ling [40] sur l'estimation de la densité conditionnelle dans un modèle à indice fonctionnel unique pour des données fonctionnelles α-mélangeantes illustrent cette méthode. Celle-ci consiste à estimer la densité conditionnelle de la variable de réponse, compte tenu du prédicteur fonctionnel, ce qui est utile dans diverses applications telles que la prédiction et la classification. Par ailleurs, l'article [49] propose une méthode pour estimer la fonction de répartition conditionnelle dans un modèle à indice fonctionnel unique en utilisant la régression locale linéaire pour estimer la fonction de répartition conditionnelle en chaque point du domaine. L'article [4] traite des estimations validées par la méthode de validation croisée dans un modèle à indice fonctionnel unique, il vise à choisir le paramètre de lissage optimal pour estimer les fonctions de régression liées aux variables fonctionnelles. L'article [14] présente un modèle de régression à indice unique pour des séries temporelles fonctionnelles quasi-associées, il propose de modéliser la réponse fonctionnelle en fonction d'une seule variable indice, tout en intégrant une fonction de covariance quasi-associée pour prendre en compte la dépendance temporelle. Les travaux de [1] proposent un estimateur linéaire local de la fonction de risque conditionnelle pour les modèles à indice, dans le cas de données manquantes de manière aléatoire. Cette méthode utilise une approche par noyau pour estimer la fonction de risque en chaque point du domaine, en intégrant les données manquantes via la pondération par probabilité inverse. Les travaux de [50] discutent de la régression non paramétrique et de la classification avec des covariables fonctionnelles, catégorielles, et mixtes, en utilisant un cadre de modèle additif généralisé (GAM) pour modéliser la relation entre la réponse et les covariables, lesquelles peuvent inclure des données fonctionnelles, catégorielles et mixtes.

La principale contribution de ce chapitre réside dans l'étude pratique de l'estimateur de la fonction de régression pour des variables fonctionnelles dans le cadre du modèle à indice fonctionnel unique, déjà étudié théoriquement par des chercheurs, à travers des exemples de simulation illustratifs. En pratique, nous appliquons la méthode du noyau pour estimer la fonction de régression de Y en fonction de X, en utilisant spécifiquement la technique de validation croisée Leave-One-Out (LOO) pour sélectionner le paramètre de lissage. Nous illustrons la méthode proposée à travers des exemples de simulation. Le reste de ce chapitre est structuré comme suit :
Dans la section 2, nous présentons notre modèle et notre estimateur, en introduisant les concepts clés ainsi que leurs hypothèses et notations correspondantes. La section 3 rappelle la théorie de la validation croisée. La section 4 est consacrée à l'étude des simulations. Les preuves des résultats théoriques sont reportées en annexe A. Enfin, nous présentons nos conclusions dans la dernière section.

3.2 Modèle et estimation semi-paramétrique de régression

Soit (X, Y) un couple de variables aléatoires prenant ses valeurs dans $\mathcal{H} \times \mathbb{R}$, où $\mathcal{H}$ est un espace de Hilbert réel séparable de norme $||.||$ engendré par un produit scalaire $\langle ., . \rangle$. Considérons maintenant l'échantillon $(X_i, Y_i)_{i=1,\dots,n}$ de n paires indépendantes identiquement distribuées comme la paire (X, Y). Supposons que la fonction de régression de Y étant donné X possède une structure à indice unique. Une telle structure suppose que l'explication de Y à partir de X se fait par un indice fonctionnel fixe λ dans $\mathcal{H}$. Plus précisément, la fonction de régression de Y étant donné $X = x$ est notée par :

$$r_\lambda(x) = E(Y | \langle \lambda, x \rangle). \tag{3.1}$$

L'indice fonctionnel λ agit comme un filtre permettant d'extraire la partie de X expliquant la réponse Y et représente une direction fonctionnelle révélant une explication pertinente

de la variable réponse. Concernant l'identifiabilité de ce modèle, nous considérons les hypothèses introduites par Ferraty et al. [24] sur l'opérateur de régression. Autrement dit, la fonction de régression r_λ est différentiable par rapport à x et λ, telle que $\langle \lambda, e_1 \rangle = 1$, où e_1 est le premier vecteur d'une base orthonormée de $\mathcal{H}$. Alors, nous avons :

$$\forall x \in \mathcal{H},\ r_1\left(\langle \lambda_1, x \rangle\right) = r_2\left(\langle \lambda_2, x \rangle\right) \Rightarrow \lambda_1 = \lambda_2 \text{ et } r_1 = r_2. \tag{3.2}$$

On considère la semi-métrique d_λ, associée à l'indice unique fonctionnel $\lambda \in \Gamma_{\mathcal{H}} \subset \mathcal{H}$ défini par $d_\lambda(x_1, x_2) = |\langle x_1 x_2, \lambda \rangle|$. On note $r_\lambda(x)$ la fonction de régression de Y étant donné $\langle \lambda, x \rangle$ et on définit par $\hat{r}_\lambda(x)$ l'estimateur à noyau de $r_\lambda(x)$ tel que :

$$\hat{r}_\lambda(x) = \frac{\sum_{i=1}^{n} Y_i K(h^{-1} d_\lambda(X_i, x))}{\sum_{i=1}^{n} K(h^{-1} d_\lambda(X_i, x))}. \tag{3.3}$$

La fonction K est un noyau de type I ou de type II et $h = h(n)$ est une suite de nombres réels positifs qui tend vers zéro lorsque n tend vers $+\infty$. Cette estimation prolonge, de manière différente, les travaux de Roussas [48] dans le cas réel et de Ferraty et al. [23] dans le cas fonctionnel.

3.2.1 La convergence presque complète ponctuelle en estimation semi-paramétrique de régression

Afin d'établir la convergence presque complète (p.co.) de notre estimateur nous donnons quelques conditions de régularité nécessaires à l'énonciation de ce résultat introduit par [24]. Soit x un élément fixe de $\mathcal{H}$.

(H_1) Le modèle de la fonction de lien r_λ.

$$Y = r_\lambda(X) + \epsilon \text{ avec } E(\epsilon | X) = 0. \tag{3.4}$$

(H_2) Le modèle est basé sur une condition de type continuité,

$$r_\lambda \in C_E^0 = \{f : E \to \mathbb{R} / \lim_{d_\lambda(x,x') \to 0} f(x) = f(x')\}. \tag{3.5}$$

(H_3) Le modèle est basé sur une condition de type Lipschitz, $\exists \alpha > 0$ tel que :

$$r_\lambda \in Lip_{E,\alpha} = \{f : E \to \mathbb{R}, \exists c > 0, \forall x, x' \in E, |f(x) - f(x')| < c d^\alpha(x, x')\}. \tag{3.6}$$

(H_4) La probabilité de la variable fonctionnelle sur la petite boule est non nulle,

$$\forall \epsilon > 0,\ P(X \in B_\lambda(x, \epsilon)) = \varphi_{x,\lambda}(\epsilon) > 0. \tag{3.7}$$

(H_5)

$$\begin{cases} \text{h est une suite de nombres positifs telle que} \\ \lim_{n \to +\infty} h = 0 \ et \ \lim_{n \to +\infty} \dfrac{\log n}{n \varphi_{x,\lambda}(h)} = 0. \\ K \text{ est un noyau de type } I \text{ ou de type } II \text{ et} \\ \exists C > 0, \exists \epsilon > 0, \forall \epsilon < \epsilon_0, \int_0^1 \varepsilon \varphi_{x,\lambda}(u)\, du > C \varepsilon \varphi_{x,\lambda}(\varepsilon). \end{cases} \tag{3.8}$$

(H_6) Pour la variable réponse Y, on considère

$$\forall p \geq 0 : E(Y^p | X = x) < \sigma_p(x) < \infty \text{ avec } \sigma_p(.) \quad \text{continue en x.} \tag{3.9}$$

Théorème 19 (Ferraty et al.[24]). *Sous les hypothèses* $(H_2), (H_4), (H_5)$ *et* (H_6) *nous avons :*

$$\lim_{n\to+\infty} \hat{r}_\lambda(x) = r_\lambda(x), \ p.co. \tag{3.10}$$

Rappelons que l'on dit qu'une suite $(Z_n)_{n\in\mathbb{N}}$ de variables aléatoires réelles est presque complètement convergente (p.co.) vers une variable aléatoire réelle Z si et seulement si

$$\forall \epsilon > 0, \sum_{n\geq 1} P(|Z_n - Z| > \epsilon) < \infty. \tag{3.11}$$

Ensuite, la vitesse de convergence presque complète de $(Z_n)_{n\in\mathbb{N}}$ vers Z est de l'ordre de u_n si et seulement si

$$\exists \epsilon_0 > 0, \sum_{n\geq 1} P(|Z_n - Z| > \epsilon_0 u_n) < \infty, \tag{3.12}$$

et on écrit $Z_n - Z = O_{p.co}(u_n)$.

Démonstration du théorème 19. On prend

$$\Omega_{i,\lambda} = \frac{K\left(\frac{d_\lambda(X_i,x)}{h}\right)}{EK\left(\frac{d_\lambda(X_i,x)}{h}\right)}, \text{ pour } i = 1, ..., n. \tag{3.13}$$

La variable $\Omega_{i,\lambda}$ est définie si $EK\left(\frac{d_\lambda(X_i,x)}{h}\right)$ est différent de zéro. Alors montrons que : $EK\left(\frac{d_\lambda(X,x)}{h}\right) > 0$.
Si K est un noyau de type I, on a :

$$c_1 1_{[0,1]} \leq K \leq c_2 1_{[0,1]} \text{ où } c_1, c_2 \in \mathbb{R}_+^*, \tag{3.14}$$

ce qui implique :

$$c_1 1_{B_\lambda(x,h)}(X) \leq K\left(\frac{d_\lambda(X,x)}{h}\right) \leq c_2 1_{B_\lambda(x,h)}(X), \tag{3.15}$$

ce qui implique directement :

$$c_1 \varphi_{x,\lambda}(h) \leq EK\left(\frac{d_\lambda(X,x)}{h}\right) \leq c_2 \varphi_{x,\lambda}(h). \tag{3.16}$$

Si K est un noyau de type II, on a :

$$EK\left(\frac{d_\lambda(X,x)}{h}\right) = \int_0^1 K\left(\frac{d_\lambda(X,x)}{h}\right) dP\left(\frac{d_\lambda(X,x)}{h} < 1\right). \tag{3.17}$$

Puisque K' existe, nous pouvons écrire : $K(t) = K(0) + \int_0^t K'(u)\,du$, ce qui implique :

$$EK\left(\frac{d_\lambda(X,x)}{h}\right) = \int_0^1 K(0) dP\left(\frac{d_\lambda(X,x)}{h} < 1\right) + \int_0^1 \left(\int_0^t K'(u)\,du\right) dP\left(\frac{d_\lambda(X,x)}{h} < 1\right) \tag{3.18}$$

$$= K(0)\varphi_{x,\lambda}(h) + \int_0^1 \left(\int_0^1 K'(u) 1_{[u,1]}(t)\,du\right) dP\left(\frac{d_\lambda(X,x)}{h} < 1\right). \tag{3.19}$$

En appliquant le théorème de Fubini, on obtient :

$$EK\left(\frac{d_\lambda(X,x)}{h}\right) = K(0)\varphi_{x,\lambda}(h) + \int_0^1 K'(u)\, P\left(u \leq \frac{d_\lambda(X,x)}{h} \leq 1\right) du, \tag{3.20}$$

en utilisant le fait que $K(1) = 0$, nous pouvons écrire :

$$EK\left(\frac{d_\lambda(X,x)}{h}\right) = -\int_0^1 K'(v)\, \varphi_{x,\lambda}(hv)\, dv. \tag{3.21}$$

En utilisant l'hypothèse (H_3) avec $h < \epsilon$ on arrive à :

$$EK\left(\frac{d_\lambda(X,x)}{h}\right) \geq c_4 \varphi_{x,\lambda}(h). \tag{3.22}$$

Puisque K est borné et de support $[0,1]$ en utilisant le résultat (3.17), on obtient :

$$EK\left(\frac{d_\lambda(X,x)}{h}\right) \leq c_3 \varphi_{x,\lambda}(h). \tag{3.23}$$

Pour traiter la preuve du théorème 19 nous avons besoin de la décomposition 3.24 et des lemmes 20 et 21 suivants :

$$\hat{r}_\lambda(x) - r_\lambda(x) = \frac{1}{\hat{r}_{1,\lambda}(x)}\left[(\hat{r}_{2,\lambda}(x) - E\hat{r}_{2,\lambda}(x) - (r_\lambda(x) - E\hat{r}_{2,\lambda}(x))\right] - \frac{r_\lambda(x)}{\hat{r}_{1,\lambda}(x)}(\hat{r}_{1,\lambda}(x) - 1), \tag{3.24}$$

où

$$\hat{r}_{1,\lambda}(x) = \frac{1}{n}\sum_{i=1}^n \Omega_{i,\lambda} \text{ et } \hat{r}_{2,\lambda}(x) = \frac{1}{n}\sum_{i=1}^n Y_i \Omega_{i,\lambda}.$$

Lemme 20 (Ferraty et al.[24]). *Sous les hypothèses (H_2) et (H_5), on a :*

$$E\hat{r}_{2,\lambda}(x) \to r_\lambda(x) \text{ quand } n \to +\infty. \tag{3.25}$$

Lemme 21 (Ferraty et al.[24]).

1. *Sous les hypothèses $(H_4), (H_5)$ et (H_6) nous avons :*

$$E\hat{r}_{2,\lambda}(x) - \hat{r}_{2,\lambda}(x) = O\left(\sqrt{\frac{\log n}{n\varphi_{x,\lambda}(h)}}\right). \tag{3.26}$$

2. *Sous les hypothèses (H_4) et (H_5) nous avons :*

$$\hat{r}_{1,\lambda}(x) - 1 = O\left(\sqrt{\frac{\log n}{n\varphi_{x,\lambda}(h)}}\right). \tag{3.27}$$

Les numérateurs de la décomposition 3.24 sont traités en faisant les deux lemmes 20 et 21 ci-dessus, tandis que le dénominateur est traité en utilisant la deuxième partie du lemme 21 et la première partie de la proposition $A.6$.(voir [24]). □

3.2.2 La vitesse de convergence

Cette section indique la vitesse de convergence presque complète ponctuelle. Pour cela, nous allons considérer un modèle de type Lipschitz pour r_λ. Comme nous le verrons, cela nous permettra d'énoncer précisément le comportement du biais et d'en déduire la vitesse de convergence.

Théorème 22 (Ferraty et al.[24]). *Sous le modèle de type Lipschitz* (H_3) *et avec la condition de probabilité* (H_4)*, si l'estimateur vérifie* (H_5) *et si la variable de réponse* Y *vérifie* (H_6)*, alors on :*

$$\hat{r}_\lambda(x) - r_\lambda(x) = O\left(h^\alpha\right) + O_{p.co}\left(\frac{\log n}{n\varphi_{x,\lambda}(h)}\right). \tag{3.28}$$

Démonstration du théorème 22. En utilisant le lemme 21 conjointement avec le lemme 23 ci-dessous, la preuve du résultat de ce théorème suit étape par étape la démonstration du théorème 19. □

Lemme 23. (Ferraty et al.[24]). *Sous le modèle de type Lipschitz* (H_3) *et avec la condition de probabilité* (H_4) *on a :*

$$Er_{2,\lambda}(x) - r_\lambda(x) = O(h^\alpha). \tag{3.29}$$

3.3 Quelques techniques de validation croisée

Cette section se concentre sur la théorie de la technique de validation croisée qui est elle-même très générale et s'applique à de nombreuses procédures d'estimation. La méthode de lissage par noyau est une technique utilisée pour estimer la fonction de régression à partir de données brutes. Cependant, une partie essentielle de cette méthode consiste à choisir un paramètre de lissage approprié, également appelé bande passante, qui détermine la largeur de la fenêtre utilisée pour estimer la fonction de régression. Le choix de la bande passante optimale peut souvent être difficile, mais la technique du leave-one-out (LOO) peut aider à résoudre ce problème. Ici, nous l'appliquerons pour le choix de la fenêtre h de l'estimateur par la méthode du noyau, pour ce faire : On se donne une grille de valeurs H de fenêtres, parmi lesquelles on souhaite choisir une fenêtre optimale $\hat{h}$ en se basant uniquement sur les données. Le principe général est de diviser l'échantillon en un ensemble d'apprentissage et un ensemble de validation. Les estimateurs sont construits à partir de l'ensemble d'apprentissage, puis l'ensemble de validation est utilisé pour estimer leur risque de prédiction. Les schémas les plus populaires sont :

Hold-out CV : on divise l'échantillon en deux parties I_1 et I_2 (I_1 et I_2 sont deux ensembles disjoints de $\{1, 2, ..., n\}$). Les estimateurs $(\hat{r}_h^{I_1})_{h\in H}$ sont calculés à partir de $(X_i; Y_i)_{i\in I_1}$. On calcule ensuite les estimateurs des risques associés pour $n_2 = Card(I_2)$

$$\hat{R}_h = \frac{1}{n_2}\sum_{i\in I_2}\left(Y_i - \hat{r}_h^{I_1}(X_i)\right)^2. \tag{3.30}$$

V -fold CV : les données sont divisées en V ensembles disjoints $I_1, ..., I_V$. Chacun des V sous-ensembles est utilisé à son tour comme ensemble de validation, le reste étant donc utilisé pour l'apprentissage : on calcule, pour chaque $j \in \{1, 2, ..., V\}$, l'ensemble des estimateurs $(\hat{r}_h^{-I_j})_{h\in H}$ réalisés avec $(X_i; Y_i)_{i\in I_j}$. Puis le risque de prédiction pour une fenêtre h est estimé par

$$\hat{R}_h = \frac{1}{V}\sum_{j=1}^{v}\frac{1}{n_j}\sum_{i\in I_j}\left(Y_i - \hat{r}_h^{-I_j}(X_i)\right)^2, \tag{3.31}$$

où on note $n_j = Card(I_j)$ et $\hat{r}_h^{-I_j}(x) = \dfrac{\sum\limits_{i\in I_j} Y_i K\left(\frac{\langle\lambda, X_i - x\rangle}{h}\right)}{\sum\limits_{i\in I_j} K\left(\frac{\langle\lambda, X_i - x\rangle}{h}\right)}$.

En pratique, on choisit souvent $V = 5$ ou $V = 10$.

Leave-one out : est un cas particulier de V -fold CV avec $V = n$.

Dans tous les cas nous choisissons :

$$\hat{h} = \arg\min_{h\in H} \hat{R}_h,$$

et l'estimateur final est donné par :

$$\hat{r} = \hat{r}_{n,\hat{h}},$$

le principe du cas particulier "leave-one-out" est que pour chaque valeur h de la grille de valeurs H et pour chaque $i \in \{1, 2, ..., n\}$ on construit un estimateur $\hat{r}_h^{-i}$ en utilisant toutes les observations sauf la ième. La ième observation est ensuite utilisée pour mesurer la performance de $\hat{r}_h^{-i}$ par $\left(Y_i - \hat{r}_h^{-i}(X_i)\right)^2$. On pose donc :

$$\hat{R} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{r}_h^{-i}(X_i)\right)^2. \tag{3.32}$$

On minimise R pour trouver $\hat{h}$.

3.4 Étude de simulation

Dans cette section, nous analysons le comportement de notre estimateur dans le cadre du modèle d'indice fonctionnel sous les conditions de régularité énoncées à la section 3.2. Nous présentons plusieurs simulations pour évaluer les performances de l'estimateur sur des échantillons de taille finie. De plus, nous comparons les performances de l'estimateur proposé à celles de méthodes concurrentes, notamment l'estimateur par Spline, l'estimateur de régression polynomiale locale, ainsi que l'estimateur des k plus proches voisins.

3.4.1 Exemples

Exemple 24. On construit un échantillon de n = 100 courbes comme suit :

$$x_i(t_j) = a_i \cos(2\pi t_j) + b_i \sin(4\pi t_j) + 2c_i(t_j - 0.25)(t_j - 0.5), \tag{3.33}$$

où $0 < t_1 < t_2 < ... < t_{100} = 1$ sont des points équidistants, a_i, b_i et c_i étant des observations indépendantes uniformément distribuées sur $[0, 1]$. La figure 3.1 (en haut) donne une idée de leur forme.

Simulation de modèle

On simule le modèle à indice fonctionnel comme suit :

1. on prend $\lambda(t) = tsin\left(\dfrac{1}{t^4 + 0.05}\right)$,
2. on prend une fonction de lien r_λ définie de $C^1(\mathbb{R})$ dans $\mathbb{R}$ par :

$$r(\langle\lambda, x\rangle) = 100 * (\langle\lambda, x\rangle - 0.05)^3, \tag{3.34}$$

3. on calcule les produits scalaires $\langle\lambda, x_1\rangle, ..., \langle\lambda, x_{100}\rangle$,

4. générer indépendamment $\varepsilon_1, ..., \varepsilon_{100}$ à partir d'une loi centrée de variance égale à 0,05.

5. simuler les réponses correspondantes $Y_i = r\left(\langle \lambda, x_i \rangle\right) + \varepsilon_i$.

Choix du paramètre de lissage

La méthode de lissage par noyau est une technique utilisée pour estimer la fonction de régression à partir de données brutes. Cependant, une partie essentielle de cette méthode consiste à choisir un paramètre de lissage approprié, également appelé bande passante, qui détermine la largeur de la fenêtre utilisée pour estimer la fonction de régression. Le choix de la bande passante optimale peut souvent être difficile, mais la technique du leave-one-out (LOO) peut aider à résoudre ce problème.

1. Divisez les données en deux ensembles : un ensemble d'apprentissage et un ensemble de validation. L'ensemble d'apprentissage est constitué des données excluant celles de validation, tandis que l'ensemble de validation est composé exclusivement des données réservées à la validation.

2. Estimez la fonction de régression pour chaque fonction de validation à l'aide des données restantes (ensemble d'apprentissage). Explorez une gamme de valeurs pour le paramètre de lissage et répétez cette estimation pour chaque point de validation.

3. Calculez l'erreur de prédiction pour chaque fonction de validation et pour chaque valeur du paramètre de lissage. L'erreur de prédiction est généralement mesurée par la somme des carrés des écarts entre les valeurs prédites et les valeurs observées.

De plus, nous avons utilisé les valeurs de l'erreur quadratique moyenne MSE pour évaluer la qualité de toute estimation de courbe.

$$MSE = \frac{1}{100} \sum_{i=1}^{n} \left(\hat{r}_\lambda\left(x_i\right) - r_\lambda\left(x_i\right)\right)^2. \tag{3.35}$$

La simulation de la régression sous le modèle de régression non paramétrique a été étudiée par Delsol Laurent et peut être trouvée dans [22], tandis que sous le modèle semi-paramétrique nous pouvons nous référer aux travaux réalisés par Ait Saidi dans [4]. Notre contribution est de valider la notion de convergence presque complète de l'estimateur de la fonction de régression sous un modèle semi-paramétrique plus précisément sous le modèle d'indice fonctionnel où nous avons choisi le paramètre de lissage par validation croisée LOO. Les résultats de simulation sont résumés dans le tableau 3.1 et la figure 3.1.

n					10					
h					0.1					
$r_\theta \cdot 10^{-3}$	-0.0891	0.0000	-0.0477	-0.0001	-0.0455	0.0114	-0.0114	-0.2510	-0.2374	0.0231
$\hat{r}_\theta \cdot 10^{-3}$	-0.0890	0.0000	-0.0001	-0.0478	-0.0001	0.0455	-0.0114	-0.2511	-0.2374	0.0232

Tab. 3.1 : Paramètre de lissage optimal (Exemple 24).

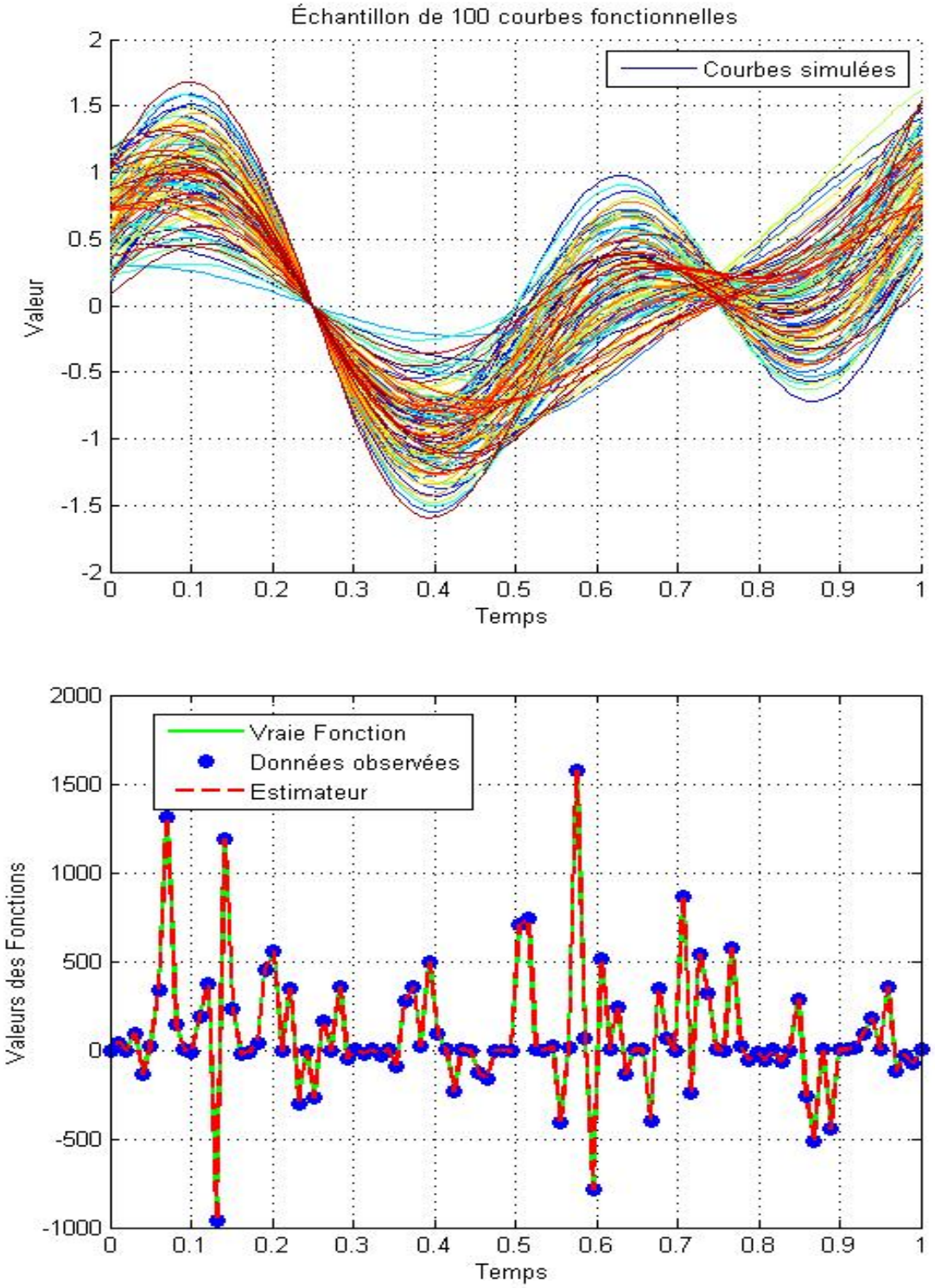

Fig. 3.1 : (En haut) : 100 courbes simulées. (En bas) : Courbes de régression, estimateur et données observées.

Exemple 25. Dans cet exemple, nous conservons la même structure générale de traitement (organigramme) utilisée pour l'exemple 24. Cependant, nous modifions uniquement les données simulées à la première étape. Nous supposons que la variable fonctionnelle X est définie sur $[0, 1]$ comme suit :

$$X(t) = a\cos\left(w + \pi\left(2t - 1\right)\right), \tag{3.36}$$

où a et w sont deux variables aléatoires indépendantes, avec a suivant la loi uniforme sur $[0, 1]$ et w suivant la loi uniforme sur $\left[0, \frac{\pi}{4}\right]$. On considère une fonction de lien r définie de $C^1(\mathbb{R})$ dans $\mathbb{R}$ par :

$$r\left(\langle\theta, x\rangle\right) = 1 + exp\left(\langle\theta, x\rangle\right), \tag{3.37}$$

Le paramètre de lissage sélectionné h, les courbes simulées ainsi que le tracé de l'estimateur sont présenté respectivement dans le tableau 3.2 et la figure 3.2.

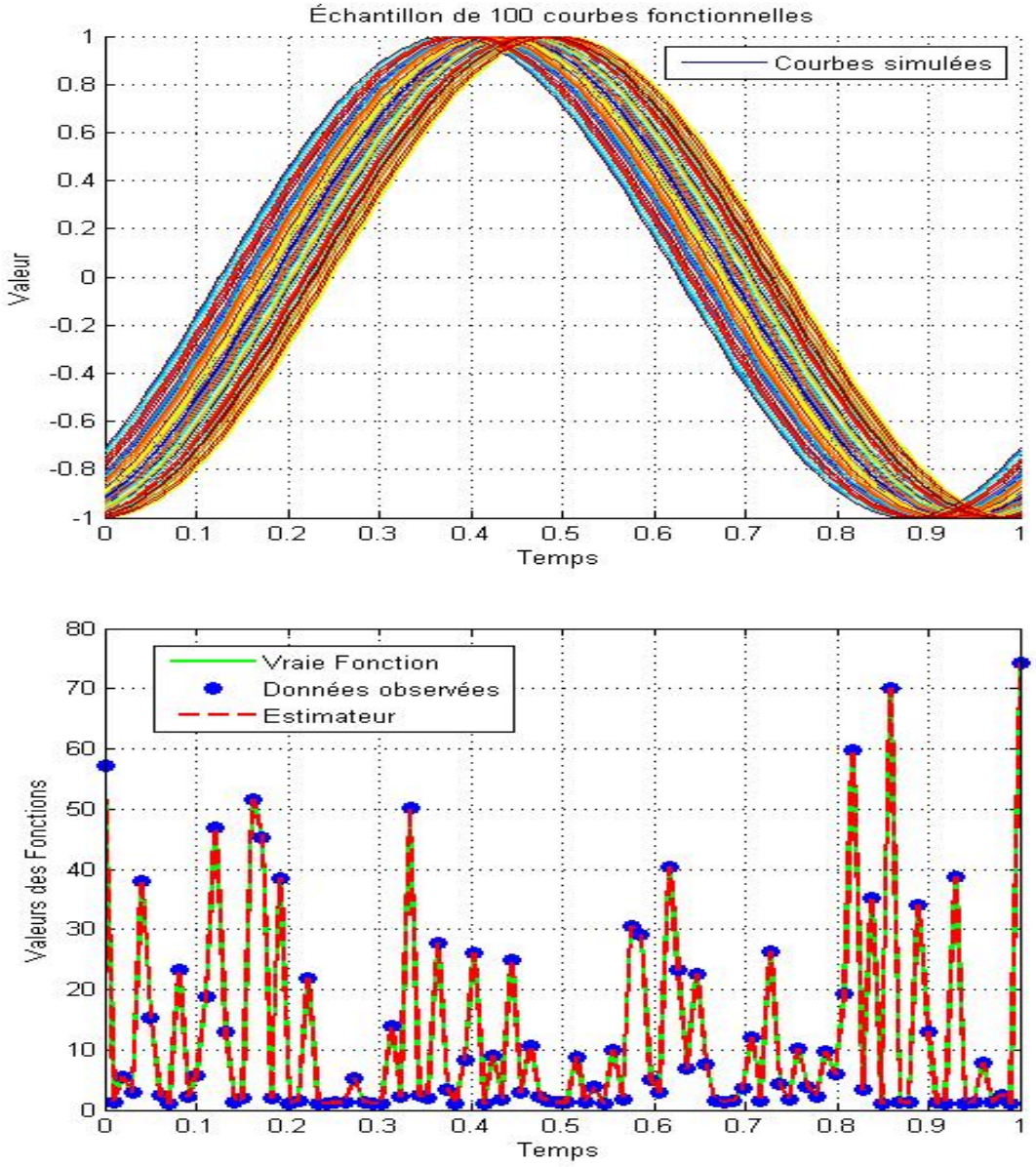

Fig. 3.2 : (En haut) : 100 courbes. (En bas) : Courbes de régression, estimateur et données observées.

n	10
h	0.1000

r_θ	57.1665	1.2478	5.4583	2.7635	37.9632	15.3719	2.4045	1.0501	23.2627	2.2826
$\hat{r}_\theta$	57.2097	1.2526	5.4157	2.8071	37.9413	15.3504	2.3494	1.0700	23.2145	2.2911

Tab. 3.2 : Paramètre de lissage optimal (Exemple 25).

Exemple 26. Dans cet exemple, nous supposons que la covariable fonctionnelle X est définie sur $[0, 1]$ qui est générée par l'équation suivante :

$$X(t) = a\cos(b + \pi wt) + c\sin(d + \pi wt) + (1 - a)\sin(\pi wt), \tag{3.38}$$

où $a, c \rightsquigarrow Bernoulli\,(0.5)$ et b, d, w sont des variables aléatoires indépendantes suivants la loi normale centrée et réduite. Nous prenons $\theta(t) = 0.15t\sin t$ et une fonction de lien r définie de $C^1(\mathbb{R})$ dans $\mathbb{R}$ par :

$$r(\langle\theta, x\rangle) = \frac{1}{1 + \langle\theta, x\rangle}, \tag{3.39}$$

nous calculons les produits scalaires $\langle\theta, x_1\rangle, ..., \langle\theta, x_{100}\rangle$, et générer indépendamment $\varepsilon_1, ..., \varepsilon_{100}$ à partir d'une loi normale centrée de variance égale à $0, 1$ fois. Nous simulons les réponses correspondantes $Y_i = r(\langle\theta, x_i\rangle) + \varepsilon_i$. Les résultats de la simulation sont résumés dans la figure 3.3, la figure 3.4 et le tableau 3.3.

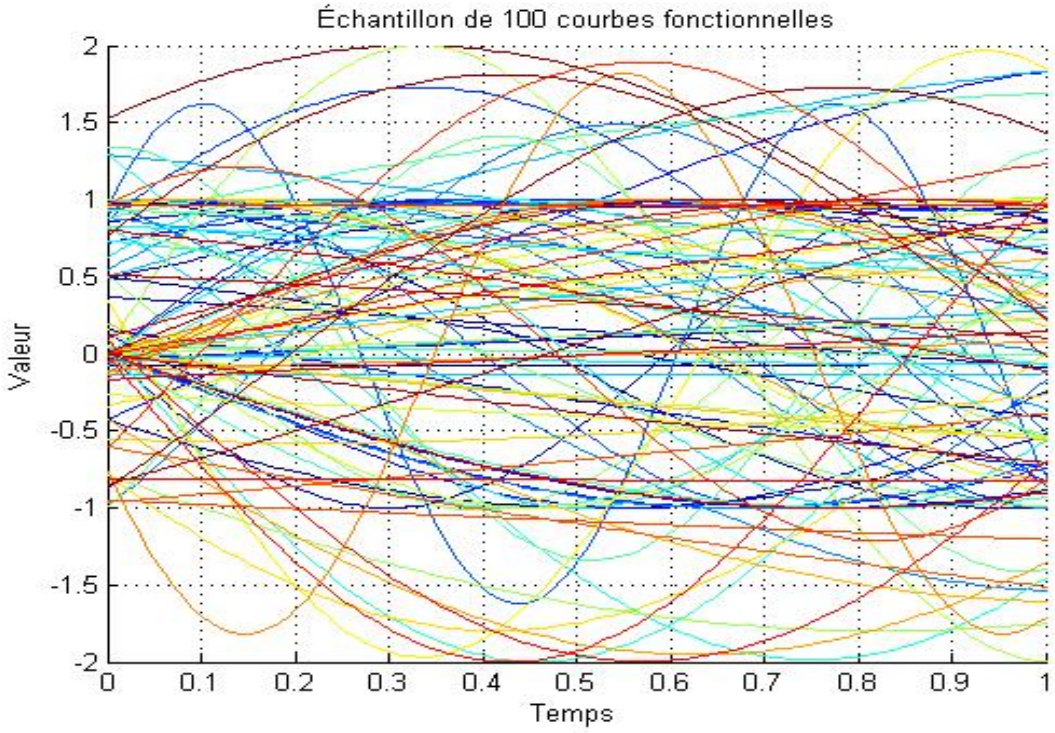

Fig. 3.3 : 100 courbes simulées.

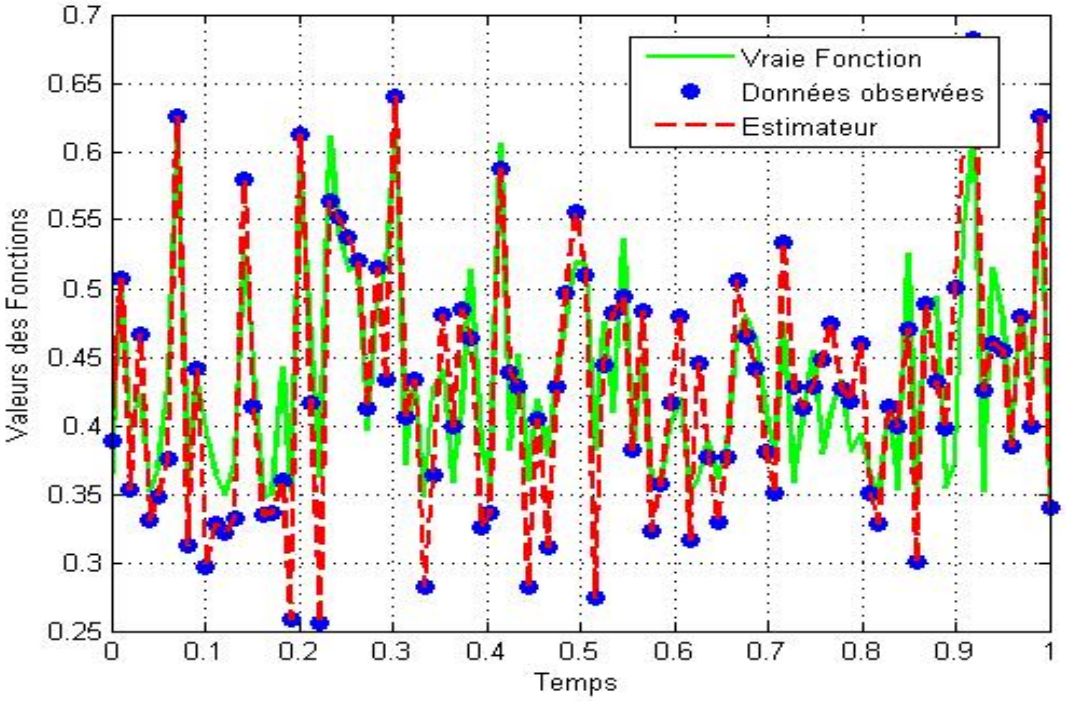

Fig. 3.4 : Courbes de régression, estimateur et données observées.

n				10						
h				0.1000						
r_θ	0.3457	0.5022	0.3965	0.4231	0.3523	0.3696	0.4305	0.6062	0.3612	0.4335
$\hat{r}_\theta$	0.3889	0.5069	0.3539	0.4667	0.3304	0.3482	0.3753	0.6260	0.3129	0.4419

Tab. 3.3 : Paramètre de lissage optimal (Exemple 26).

3.4.2 Erreur absolue moyenne, erreur quadratique moyenne et normalité asymptotique.

Dans cette sous-section, nous commençons par calculer l'erreur absolue moyenne $\overline{MAE}$ et l'erreur quadratique moyenne $\overline{RMSE}$ qui sont définies respectivement par :

$$\overline{MAE} = \frac{1}{ns}\sum_{s=1}^{ns} |\overline{\hat{r}}_\lambda^{(s)} - \overline{r}_\lambda|, \tag{3.40}$$

et

$$\overline{RMSE} = \sqrt{\frac{1}{ns}\sum_{s=1}^{ns} |\overline{\hat{r}}_\lambda^{(s)} - \overline{r}_\lambda|^2}, \tag{3.41}$$

où $\overline{\hat{r}}_\lambda^{(s)} = \frac{1}{n}\sum_{i=1}^{n} \hat{r}_\lambda^{(s)}(x_i)$ et $\overline{r}_\lambda = \frac{1}{n}\sum_{i=1}^{n} r_\lambda(x_i)$ avec ns et n désignent respectivement le nombre d'échantillons et le nombre de points. Nous résumons les valeurs obtenues de $\overline{MAE}$ et $\overline{RMSE}$ dans le tableau 3.4 comme suit :

Errors	$\overline{MAE}$	$\overline{RMSE}$
Example 4.1	0.0401	0.0050
Example 4.2	0.0399	0.0050
Example 4.3	0.0198	0.0155

Tab. 3.4 : Erreur absolue moyenne (MAE), Erreur quadratique moyenne ($RMSE$) de l'estimateur pour h optimal.

La courbe de la fonction d'erreur absolue moyenne :

$$MAE(x) = \frac{1}{ns}\sum_{s=1}^{ns} |\hat{r}_\lambda^{(s)}(x) - r_\lambda(x),$$

et de la fonction d'erreur quadratique moyenne :

$$RMSE(x) = \sqrt{\frac{1}{ns}\sum_{s=1}^{ns}\left(\hat{r}_\lambda^{(s)}(x) - r_\lambda(x)\right)^2},$$

sont illustrés dans la figure 3.5 ci-dessous.

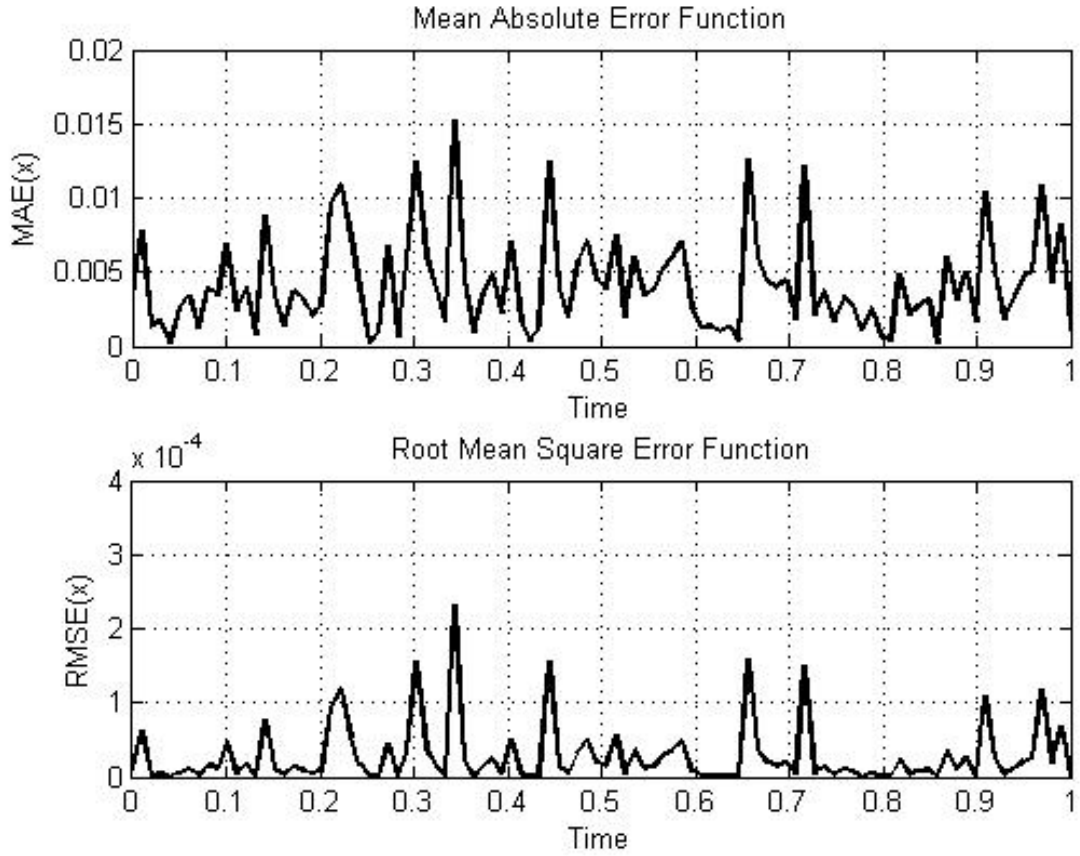

Fig. 3.5 : (panneau supérieur) : $MAE(x)$. (panneau inférieur) : $RMSE(x)$.

La courbe de la densité asymptotique ainsi que le diagramme en boîte (basé sur 500 répétitions) de $\left(\overline{\hat{r}}_\lambda^{(s)}\right)_{s=1,\ldots,ns}$ pour l'estimateur proposé NW à partir de 200 observations du modèle de l'Exemple 24 sont présentés dans la Figure 3.6 ci-dessous. L'intervalle de confiance (IC) et la normalité Q-Q $\overline{\hat{r}}^{(s)}$ sont illustrés dans les figures 3.7, 3.8 respectivement.

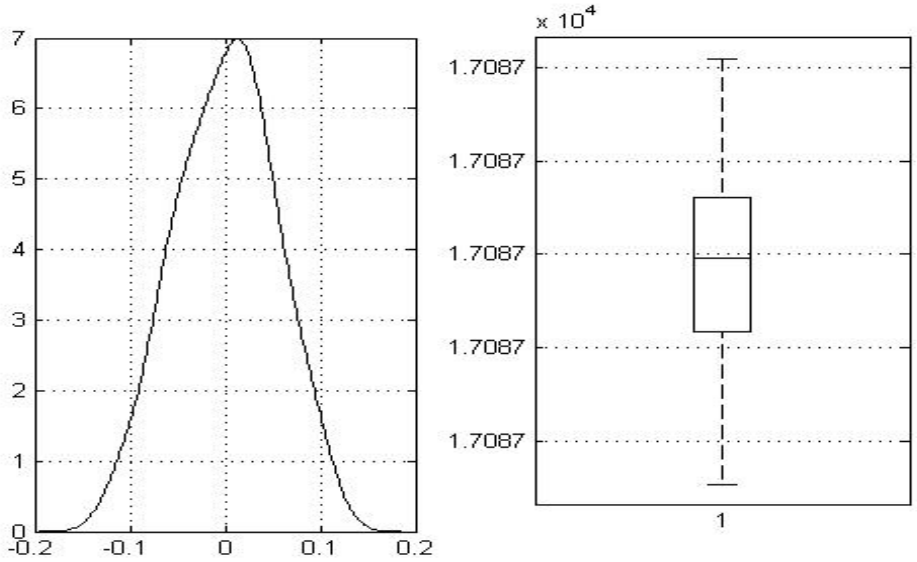

Fig. 3.6 : (gauche) : Distribution asymptotique de $\sqrt{ns}\left(\overline{\hat{r}}_\lambda^{(s)} - \overline{r}_\lambda\right)$. (droite) : Diagrammes en boites de $\left(\overline{\hat{r}}_\lambda^{(s)}\right)_{s=1,\ldots,ns}$.

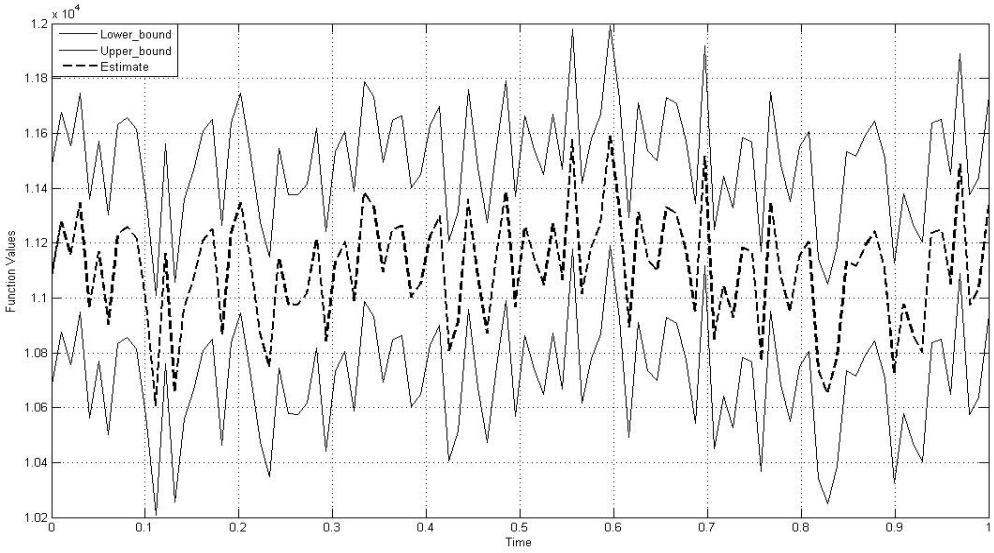

Fig. 3.7 : Intrevalle de confiance de $\overline{\hat{r}}^{(s)}$.

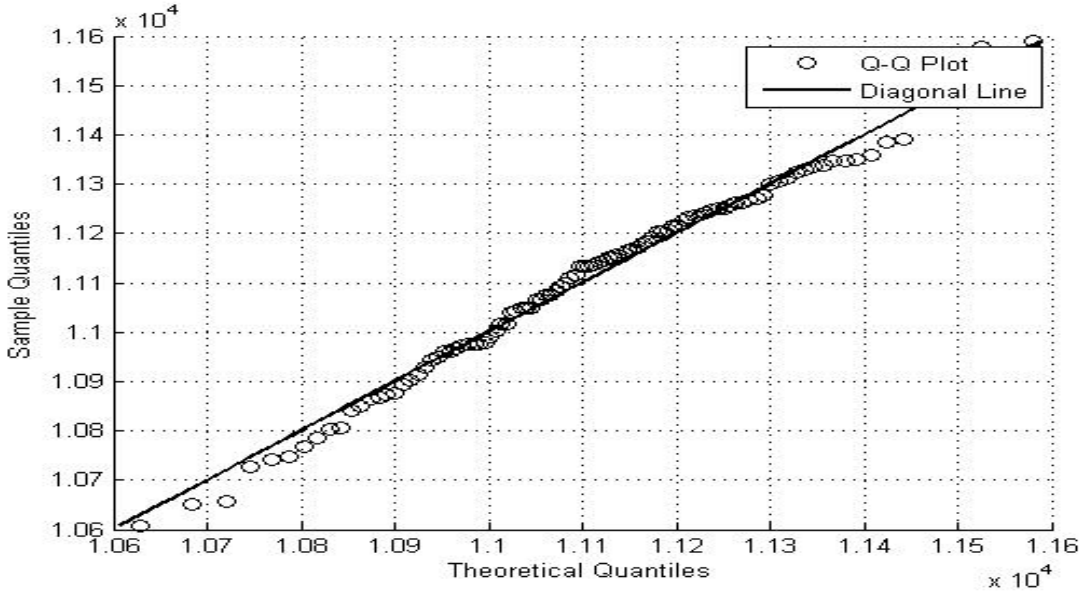

Fig. 3.8 : Diagramme quantile-quantile de $\left(\overline{\hat{r}}_\lambda^{(s)} - \overline{r}_\lambda\right)$ à partir de 500 réplications du modèle de l'exemple 24.

3.4.3 Comparisons

Nous comparons également les performances en échantillon fini de l'estimateur proposé N-W (Nadaraya-Watson) à celles de l'estimateur Smoothing Spline, de l'estimateur LPR (régression polynomiale locale) et de l'estimateur K-NN (K-plus proches voisins). Comme on peut le voir dans le tableau 3.5, l'erreur absolue moyenne et l'erreur quadratique moyenne pour $\left(\overline{\hat{r}}_\lambda^{(s)}\right)_{s=1,\ldots,ns}$ données par l'estimateur proposé NW indiquent que celui-ci a un taux de convergence plus rapide que celui $\left(\overline{\hat{r}}_\lambda^{(s)}\right)_{s=1,\ldots,ns}$ donné par les estimateurs Spline, LPR et KNN.

Methods	NW		Spline		LPR		KNN	
Errors	$\overline{MAE}$	$\overline{RMSE}$	$\overline{MAE}$	$\overline{RMSE}$	$\overline{MAE}$	$\overline{RMSE}$	$\overline{MAE}$	$\overline{RMSE}$
Example 4.1	0.0401	0.0050	0.0401	0.0050	0.4157	0.0224	571.5613	53.4566
Example 4.2	0.0399	0.0050	0.0401	0.0050	0.0389	0.0053	0.0170	0.0066
Example 4.3	0.0198	0.0155	0.0714	0.0148	0.0714	0.0148	0.0190	0.0161

Tab. 3.5 : Erreur absolue moyenne, erreur quadratique moyenne de NW, méthodes Splines, LPR et KNN pour h optimal avec $n = 100$ et $ns = 500$.

Les courbes des densités asymptotiques, le diagramme en boîte et le diagramme quantile-quantile récapitulatif de la composante $\left(\overline{\hat{r}}_\lambda^{(s)}\right)_{s=1,\ldots,ns}$ selon les méthodes NW, Spline, LPR et KNN sont présentés dans la figure 3.9, la figure 3.10 et la figure 3.11 ci-dessous.

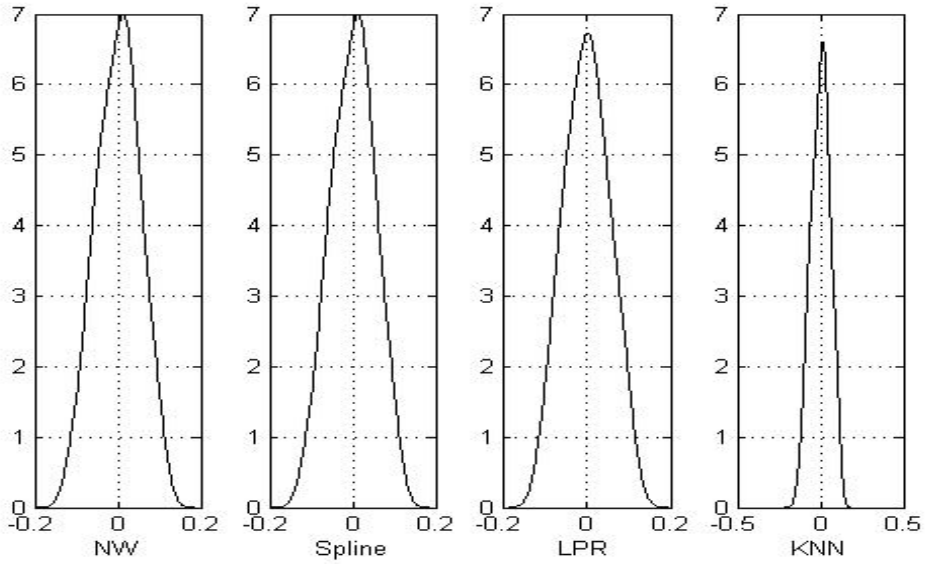

Fig. 3.9 : Distribution asymptotique de $\sqrt{ns}\left(\overline{\hat{r}}_{\lambda}^{(s)} - \overline{r}_{\lambda}\right)$ selon : NW, $Spline$, LPR et KNN pour l'Exemple 24.

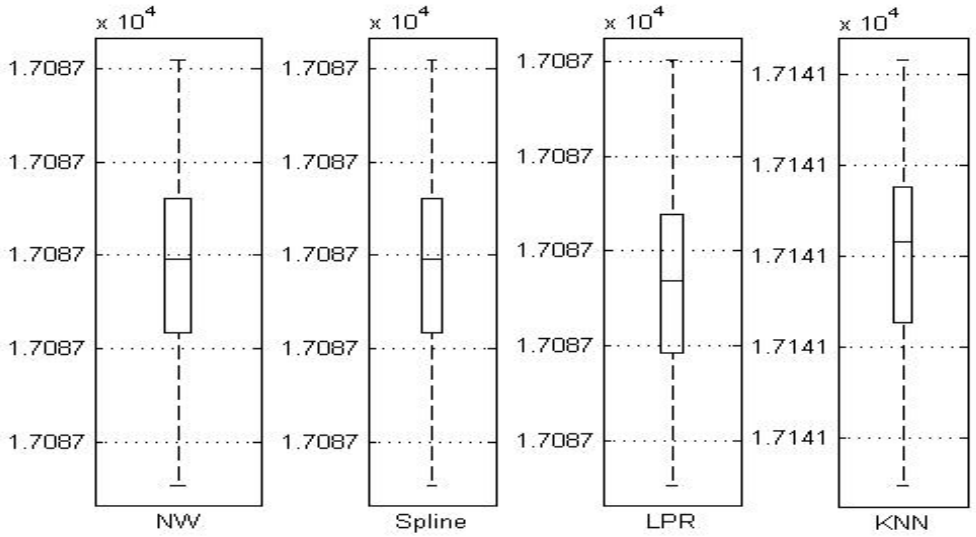

Fig. 3.10 : Les diagrammes en boites de $\left(\overline{\hat{r}}_{\lambda}^{(s)}\right)_{s=1,\ldots,ns}$ selon NW, Spline, LPR et KNN pour l'exemple 24.

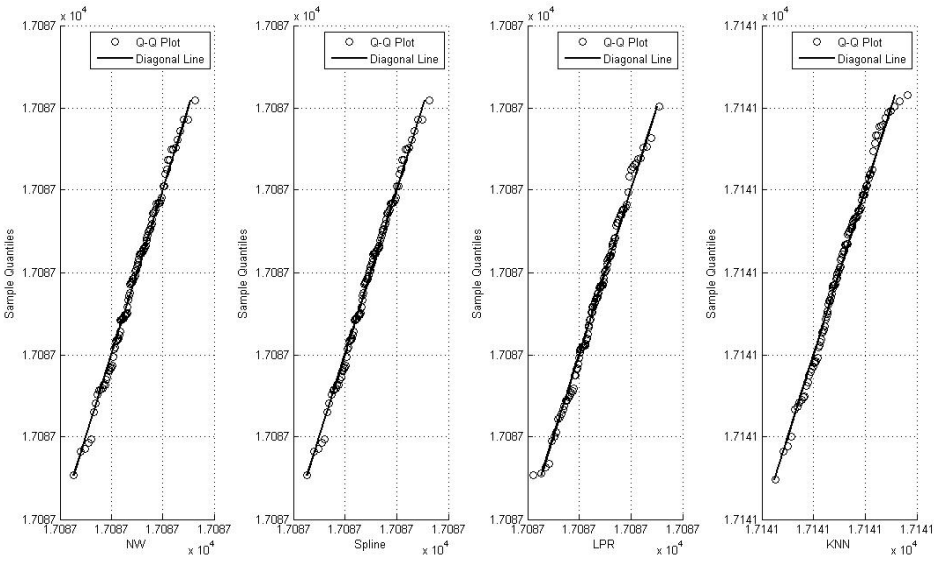

Fig. 3.11 : Graphique quantile-quantile de $\left(\overline{\hat{r}}_{\lambda}^{(s)}\right)_{s=1,\ldots,ns}$ selon NW, Spline, LPR et KNN pour l'exemple 24.

3.5 Analyse des données réelles

Dans cette section, nous passons de l'exploration théorique à l'application pratique en examinant un cas réel de spectrométrie, un outil moderne et précieux pour analyser la composition chimique des substances, comme l'a noté [26]. Les chimiométristes ont développé leurs propres techniques basées sur le raisonnement heuristique et des idées intuitives pour analyser ces données. Les deux méthodes les plus importantes sont les moindres carrés partiels (voir [53] et [42] pour plus de détails) et la régression en composantes principales ([43]). La chimiométrie sert de base au développement de méthodologies fonctionnelles non paramétriques.

3.5.1 Description des données spectrométriques

Les données originales ont d'abord été étudiées par [8] en utilisant une approche de réseau neuronal. Elles proviennent d'un problème de contrôle qualité dans l'industrie alimentaire et peuvent être consultées sur http ://lib.stat.cmu.edu/datasets/tecator. Cet ensemble de données concerne un échantillon de viande finement hachée. La figure 3 montre quelques unités des données spectrométriques originales. Cette figure représente l'absorbance en fonction de la longueur d'onde pour 10 morceaux de viande sélectionnés au hasard. Plus précisément, pour chaque échantillon de viande, les données consistent en un spectre d'absorbance à 100 canaux. Rappelant que l'absorbance est le $-log10$ de la transmittance mesurée par le spectromètre, et que les données ont été enregistrées sur un analyseur d'aliments et d'aliments pour animaux infractés Tecator fonctionnant dans la gamme de longueurs d'onde de 850 à 1050 nanomètres (nm) par le principe de transmission proche infrarouge (NIR). Une unité apparaît clairement comme une courbe discrétisée. En raison de la finesse de la grille (englobant la discrétisation), chaque sujet peut être considéré comme une courbe continue. [38] a déclaré que dans les données chimiométriques, les spectres observés sont des observations pratiquement fonctionnelles.

Ainsi, chaque analyse spectrométrique peut être résumée par une courbe continue donnant l'absorbance observée en fonction de la longueur d'onde. L'ensemble des données de courbes, c'est-à-dire l'ensemble de données (continues), est présenté dans la figure 3.13. Pour plus de détails, nous pouvons nous référer à Ferraty [24].

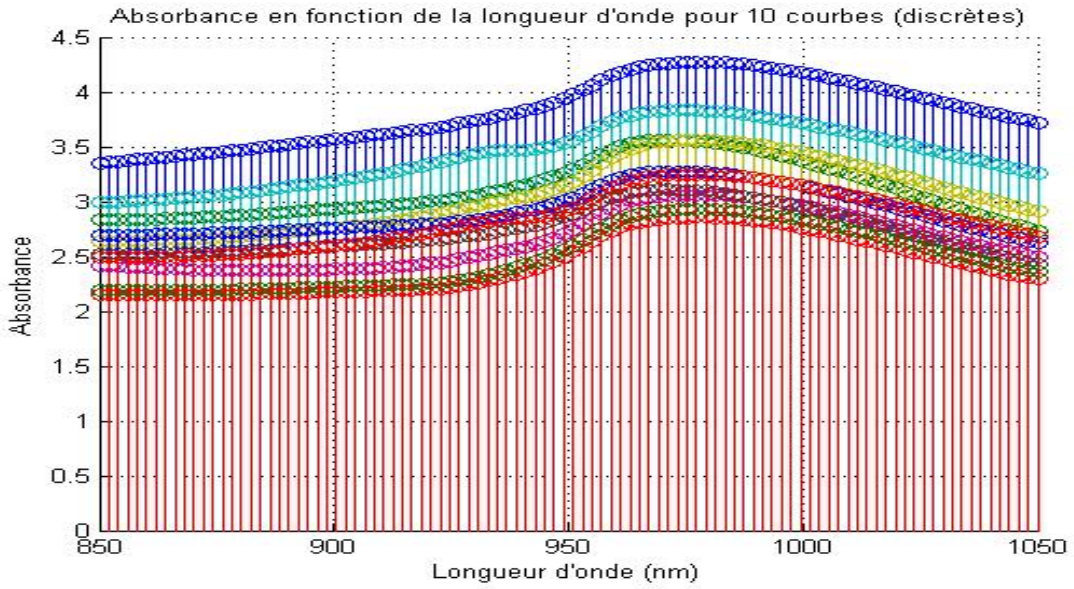

Fig. 3.12 : 10 courbes de données (discrétisées).

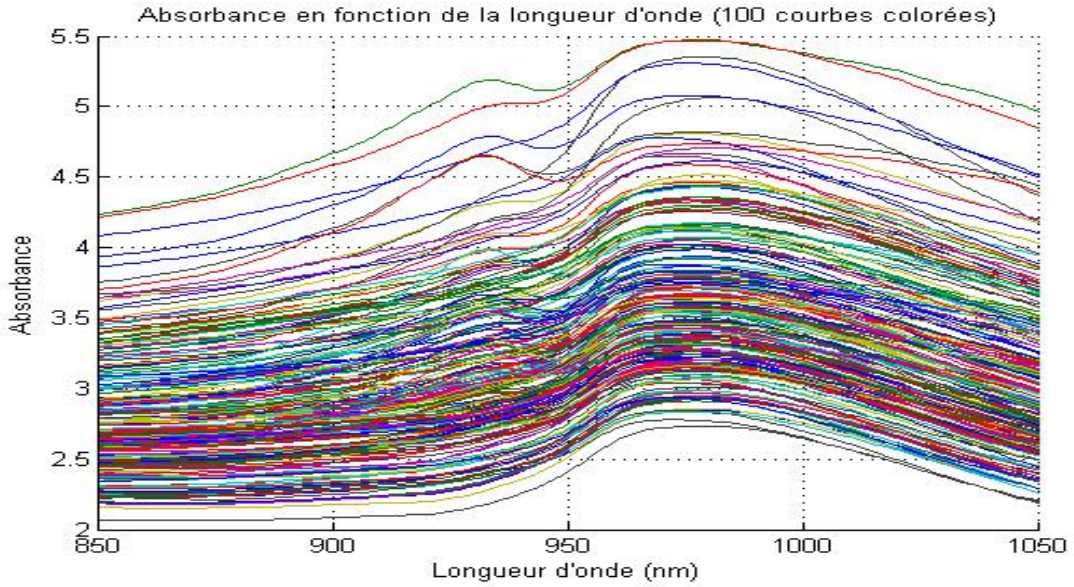

Fig. 3.13 : Échantillon de 100 courbes de données.

L'absorbance en fonction de la longueur d'onde pour 10 courbes discrétisées, ainsi que le jeu de données continu correspondant, sont illustrés respectivement dans les figures 3.12, et 3.13.

Le Tableau 3.6 résume les données réelles Y et leur estimation Y_{est}.

n				10					
h				0.01					
Y 22.5000	40.1000	8.4000	5.9000	25.5000	42.7000	10.6000	19.9000	19.9000	46.0000
Y_{est} 22.5000	40.1000	8.4000	5.9000	25.5000	42.7000	10.6000	19.9000	19.9000	46.0000

Tab. 3.6 : Paramètre de lissage optimal pour les données réelles.

La Figure 3.14 présente l'échantillon des données réelles et leur estimation sous le modèle de régression fonctionnelle.

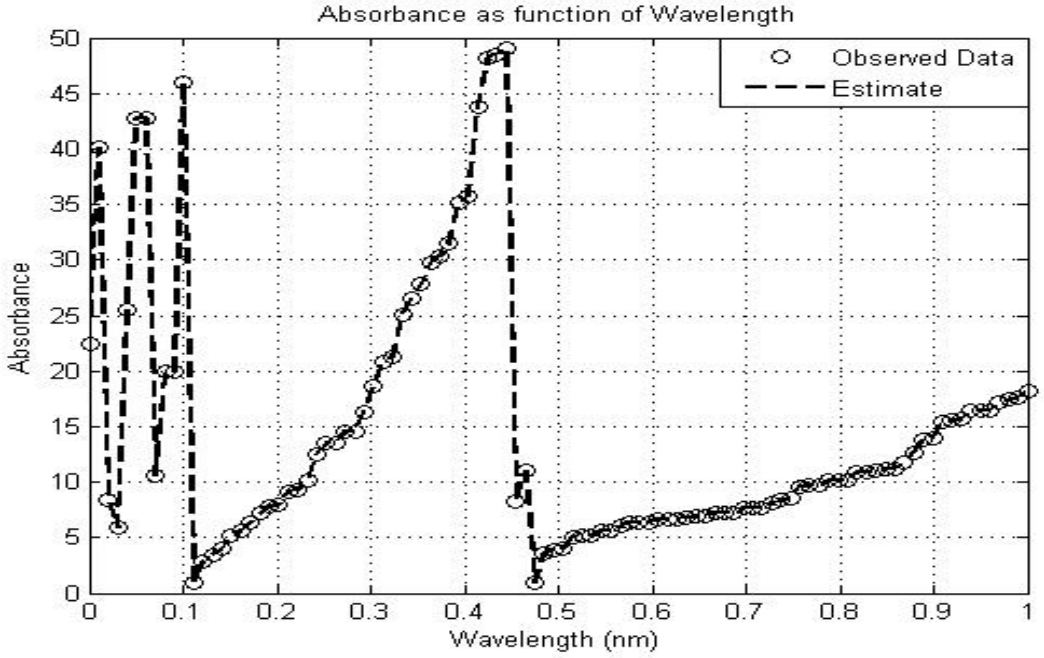

Fig. 3.14

3.6 Conclusion

En conclusion, ce chapitre s'est concentré sur l'estimation de l'opérateur de régression dans le modèle à indice fonctionnel unique pour les données fonctionnelles en appliquant la méthode du noyau. Grâce à l'utilisation du paramètre de lissage, nous avons exploré plusieurs techniques, dont la technique de validation croisée LOO, pour obtenir des résultats optimaux. Notre approche a consisté à valider nos résultats en vérifiant la notion théorique de résultats limits ou asymptotiques, qui a démontré la convergence presque complète de l'estimateur de régression par la méthode du noyau. Nous avons constaté que la méthode LOO fournit un bon estimateur de la régression, grâce à la sélection du

paramètre de lissage optimal. Nos résultats contribuent au corpus croissant de recherches en analyse de données fonctionnelles et donnent un aperçu des meilleures pratiques pour estimer l'opérateur de régression dans ce domaine.

3.7 Annexe A

Dans cette annexe, nous présentons les démonstrations détaillées des lemmes utilisées dans la preuve des deux théorèmes exposés dans le chapitre précédent. Les démonstrations sont volontairement placées ici afin de ne pas alourdir le corps principal du texte, tout en permettant au lecteur intéressé de suivre rigoureusement le raisonnement mathématique sous-jacent.

Démonstration du lemme 21. Le modèle (H_2) nous permet d'écrire :

$$\begin{aligned} E\hat{r}_{2,\lambda}(x) - r_\lambda(x) &= E(Y_1\Delta_{1,\lambda}) - r_\lambda(x) \\ &= E(E(Y_1\Delta_{1,\lambda}|X_1)) - r_\lambda(x) \\ &= E(r_\lambda(X_1)\Delta_{1,\lambda}) - r_\lambda(x) \\ &= E((r_\lambda(X_1)) - r_\lambda(x))\Delta_{1,\lambda}, \end{aligned}$$

puisque le support du noyau K est $[0,1]$ nous avons obtenu :

$$|r_\lambda(X_1) - r_\lambda(x)|\Delta_{1,\theta} \leq \sup_{x' \in B_\lambda(x,h)} |r_\lambda(x') - r_\lambda(x)|\Delta_{1,\lambda}.$$

En supposant la continuité de r_λ, nous arrivons au résultat revendiqué.

□

Démonstration du Lemme 23. On prend pour $i = 1, ..., n$, $K_{i,\lambda} = K\left(h^{-1}d_\lambda\left(x, X_i\right)\right)$. La preuve de ce résultat est basée sur l'application de l'inégalité exponentielle de Bernstein donnée par le corollaire $A.9 - i$ (voir Ferraty et al. [24]). Puisque :

$$|\hat{r}_{2,\lambda}\left(x\right) - E\hat{r}_{2,\lambda}\left(x\right)| = \frac{1}{n}|\sum_{i=1}^{n}\left(Y_i\Delta_{i,\lambda} - E\left(Y_i\Delta_{i,\lambda}\right)\right)|.$$

Nous prouvons qu'il existe $\epsilon_0 > 0$ tel que :

$$P\left(\frac{1}{n}|\sum_{i=1}^{n}\left(Y_i\Delta_{i,\lambda} - E\left(Y_i\Delta_{i,\lambda}\right)\right)| > \epsilon_0\sqrt{\frac{\log n}{n\varphi_{x,\lambda}\left(h\right)}}\right) < \infty,$$

puis nous avons appliqué l'inégalité exponentielle donnée par le corollaire $A.8 - ii$ dans l'annexe A [24] pour une variable $T_i = Y_i\Delta_{i,\lambda} - E\left(Y_1\Delta_{1,\lambda}\right)$.
Tout d'abord nous prouvons :

$$\exists C > 0, \forall m \geq 2 : |E\left(Y_1\Delta_{1,\lambda} - E\left(Y_1\Delta_{1,\lambda}\right)\right)^m| \leq C\left(\varphi_{x,\lambda}\left(h\right)\right)^{1-m}.$$

Ensuite, nous prouvons pour $m \geq 2$,

$$E\left(|Y_1|^m\Delta_{1,\lambda}^m\right) = O\left(\varphi_{x,\lambda}\left(h\right)\right)^{1-m}, \tag{3.42}$$

aussi puisque nous avons : $\Delta_{1,\lambda}^m = \dfrac{K_{1,\lambda}^m}{(EK_{1,\lambda})^m}$, alors :

$$\begin{aligned}
E\left(|Y_1|^m \Delta_{1,\lambda}^m\right) &= \frac{1}{(EK_{1,\lambda})^m} E\left(|Y_1|^m K_{1,\lambda}^m\right) \\
&= \frac{1}{(EK_{1,\lambda})^m} E\left(E\left(|Y_1|^m / X\right) K_{1,\lambda}^m\right) \\
&= \frac{1}{(EK_{1,\lambda})^m} E\left(\sigma_m(X) K_{1,\lambda}^m\right) \\
&= \frac{1}{(EK_{1,\lambda})^m} E\left((\sigma_m(X) - \sigma_m(x)) K_{1,\lambda}^m\right) + \sigma_m(x) \frac{EK_{1,\lambda}^m}{(EK_{1,\lambda})^m} \\
&= E\left((\sigma_m(X) - \sigma_m(x)) \Delta_{1,\lambda}^m\right) + \sigma_m(x) E\Delta_{1,\lambda}^m,
\end{aligned}$$

ensuite :

$$\begin{aligned}
|E\left(|Y_1^m| \Delta_{1,\lambda}^m\right)| &\leq E\left(|\sigma_m(X) - \sigma_m(x)| \Delta_{1,\lambda}^m\right) + \sigma_m(x) E\Delta_{1,\lambda}^m \\
&\leq \left(\sup_{x' \in B_\lambda(x,h)} |\sigma_m(X) - \sigma_m(x)|\right) E\Delta_{1,\lambda}^m + \sigma_m(x) E\Delta_{1,\lambda}^m.
\end{aligned}$$

Puisque $0 < \int K^m < \infty$ donc si K est de type I (respectivement de type II) alors $\frac{K^m}{\int K^m}$ est de type I (respectivement de type II). Par conséquent, en utilisant le résultat 3.15 ou 3.23 nous arrivons à :

$$C_1 \varphi_{x,\lambda}(h) \leq EK_1^m \leq C_2 \varphi_{x,\lambda}(h). \tag{3.43}$$

En utilisant 3.43 et les résultats de l'inéquation 3.15 ou 3.23, nous pouvons écrire :

$$\forall\ m \geq 2,\ C_1 \varphi_{x,\lambda}(h)^{1-m} \leq E\Delta_{1,\lambda}^m \leq C_2 \varphi_{x,\lambda}(h)^{1-m},$$

alors :

$$E\left(|Y_1^m| \Delta_{1,\lambda}^m\right) = O\left(\varphi_{x,\lambda}(h)^{1-m}\right).$$

On a :

$$(Y_1\Delta_{1,\lambda} - EY_1\Delta_{1,\lambda})^m = \sum_{k=0}^m c_{m,k} (Y_1\Delta_{1,\lambda})^k (EY_1\Delta_{1,\lambda})^{m-k} (-1)^{m-k},$$

où $c_{m,k} = \dfrac{m!}{k!(m-k)!}$, ce qui implique :

$$\begin{aligned}
E\left(|Y_1\Delta_{1,\lambda} - EY_1\Delta_{1,\lambda}|^m\right) &\leq \sum_{k=0}^m c_{m,k} E|Y_1\Delta_{1,\lambda}|^k |EY_1\Delta_{1,\lambda}|^{m-k} \\
&\leq \sum_{k=0}^m c_{m,k} E|Y_1\Delta_{1,\lambda}|^k |E\left(E\left(Y_1\Delta_{1,\lambda}|X = x\right)\right)|^{m-k} \\
&\leq \sum_{k=0}^m c_{m,k} E|Y_1\Delta_{1,\lambda}|^k |r(x)|^{m-k} \\
&\leq C \max_{k=0,1,2\ldots,m} E|Y_1\Delta_{1,\lambda}|^k \\
&\leq C \max_{k=0,1,2\ldots,m} \varphi_{x,\lambda}(h)^{1-k}.
\end{aligned}$$

La dernière inégalité est obtenue en utilisant 3.42 pour $k \geq 2$. Pour $k = 1$ on a :

$$\begin{aligned}
E\left(|Y_1\Delta_{1,\lambda}|\right) &= E\left(|E\left(Y_1\Delta_{1,\lambda}|X = x\right)|\right) \\
&= E\left(|\Delta_{1,\lambda} r(x)|\right) \\
&= |r(x)|,
\end{aligned}$$

alors : $E\left(|Y_1\Delta_{1,\lambda}|\right) = O(1)$. Puisque $\varphi_{x,\lambda}(h) \to 0$ lorsque $n \to \infty$ alors :

$$E|Y_1\Delta_{1,\lambda} - EY_1\Delta_{1,\lambda}| = O\left(\varphi_{x,\lambda}(h)\right)^{1-m}.$$

On prend :

$$a^2 = \varphi_{x,\lambda}(h)^{-1},$$

on a :

$$u_n = a^2\frac{\log(n)}{n} = \frac{\log(n)}{n\varphi_{x,\lambda}(h)}.$$

Le terme $u_n \to 0$ *lorsque* $n \to \infty$, alors :

$$\frac{1}{n}\sum_{k=1}^{n} T_i = O\left(\sqrt{u_n}\right) p.co = O\left(\sqrt{\frac{\log(n)}{n\varphi_{x,\lambda}(h)}}\right), \; p.co.$$

Enfin, nous obtenons :

$$\hat{r}_{2,\lambda} - E\hat{r}_{2,\lambda} = O\left(\sqrt{\frac{\log(n)}{n\varphi_{x,\lambda}(h)}}\right), \; p.co.$$

On prend $Y_i = 1$, en utilisant le même raisonnement ci-dessus on obtient :

$$\hat{r}_1(x) - 1 = O\left(\sqrt{\frac{\log(n)}{n\varphi_{x,\lambda}(h)}}\right), \; p.co.$$

(Voir [4]). □

Démonstration du lemme 23. On a :

$$\mathbb{E}r_{2,\lambda}(x) - r_\lambda(x) = \mathbb{E}\left(\left(r_\lambda(X_1) - r_\lambda(x)\right)\Delta_{1,\lambda}\right).$$

La propriété de Lipshutz de r_λ nous permet d'écrire :

$$|\mathbb{E}r_{2,\lambda}(x) - r_\lambda(x)| \leq C\ \mathbb{E}\left(d_\lambda^\beta(x,X)\,\Delta_{1,\lambda}\right).$$

Finalement, puisque $\mathbb{E}\Delta_1 = 1$ et que le support de la fonction noyau K est $[0,1]$, on a :

$$|\mathbb{E}r_{2,\lambda}(x) - r_\lambda(x)| \leq Ch^\beta.$$

□

Chapitre 4

Estimation de la fonction de distribution conditionnelle sur le modèle à indice fonctionnelle

4.1 Introduction

La fonction de distribution cumulative conditionnelle (FDC) est un outil statistique précieux pour examiner les relations entre les variables dépendantes et les variables potentiellement indépendantes. L'estimation de cette fonction est un défi majeur dans les statistiques non paramétriques, et elle a été largement étudiée par Ait Saïdi et al. [49] et Attaoui et al. [5] entre autres. En règle générale, la variable explicative X est supposée se trouver dans un espace de dimension infinie, tandis que la variable de réponse Y se trouve dans $\mathbb{R}$. L'échantillon de paires (X, Y) est souvent supposé être distribué de manière indépendante et identique (i.i.d) ou présenter une certaine forme de dépendance (voir Selk et al. [50], Ling et al. [40], Ferraty et al. [24]). De plus, le modèle à indice unique est une approche utile pour résoudre les problèmes de grande dimension dans la régression non paramétrique, et il a été largement exploré en économétrie et dans d'autres domaines. Ce modèle combine des variables explicatives multidimensionnelles en un seul indice, atténuant le problème de la « réduction de dimension » ou celui de la « malédiction de la dimensionnalité » dans la régression non paramétrique fonctionnelle. Des discussions détaillées sur le modèle à indice unique sont disponibles dans les travaux de Ait-Saïdi et al. [4], Hardle et al. [31], Bouzebda et al. [14] et Delecroix et al. [21] entre autres. Les progrès des outils informatiques ont permis de traiter des ensembles de données de plus en plus volumineux, en particulier ceux représentant des observations en temps continu ou des données fonctionnelles dans diverses disciplines, telles que la médecine, l'économie et les sciences de l'environnement. Ait-Saïdi et al. [49] ont exploré les estimateurs à noyau pour la fonction de distribution conditionnelle cumulative d'une variable de réponse scalaire Y étant donnée une variable aléatoire hilbertienne X dans le cadre d'un modèle fonctionnel à indice unique. Cette étude se concentre sur les aspects théoriques des résultats de Ait-Saïdi et al. [49] dans le cas de l'indépendance des variables explicatives pour l'estimation de la fonction de distribution conditionnelle cumulative en effectuant une simulation cohérente des résultats asymptotiques, tels que la cohérence et la convergence ponctuelle presque complète de l'estimateur en utilisant la méthode du noyau. La motivation découle de la nécessité, dans les domaines appliqués, d'étudier les processus stochastiques en temps continu pour la prévision des données de séries chronologiques, avec des applications dans des domaines tels que l'économie et les sciences de l'environnement. Dans cette étude, nous avons développé une simulation de la fonction de

distribution cumulative conditionnelle (FDC) en reformulant l'estimateur introduit par Ferraty et al [23] pour l'adapter à un modèle de régression fonctionnelle. Cette reformulation permet d'intégrer l'approche pratique aux méthodes non paramétriques existantes, offrant ainsi une solution plus flexible et applicable aux données fonctionnelles. Notre travail contribue à améliorer l'estimation de fonction de distribution conditionnelle dans le cadre de modèles semi-paramétriques et non paramétriques, élargissant les possibilités d'analyse de données complexes dans divers contextes. Notamment, l'aspect pratique de l'estimation de la fonction de distribution cumulative conditionnelle dans un modèle de régression d'indice fonctionnel est rare. Il s'agit d'un problème critique qui n'a pas reçu une attention considérable dans la littérature existante.

4.2 Construction de l'estimateur et sa reformulation

Soit (X, Y) un couple de variables aléatoires prenant leurs valeurs dans $\mathcal{H} \times \mathbb{R}$, où $\mathcal{H}$ est un espace d'Hilbert séparé muni de la norme $||.||$ induite par le produit scalaire $\langle ., . \rangle$. On considère l'échantillon $(X_i, Y_i)_{i=1,...,n}$ de couples indépendantes et identiquement distribués que le couple (X, Y). Supposant que la fonction de distribution conditionnelle de Y sachant X a la structure de modèle à indice fonctionnelle. Une telle structure suppose que l'explication de Y à partir de X se fait à travers un indice fonctionnel fixe λ dans $\mathcal{H}$. Plus précisément, la fonction de distribution conditionnelle de Y sachant $X = x$ qu'on note $R_\lambda(.|x)$, est donnée par :

$$\forall y \in \mathbb{R} : R_\lambda(y|x) = R(y|\langle \lambda, x \rangle).$$

On considère la semi-métrique d_λ, associée à $\lambda \in \Lambda_\mathcal{H} \subset \mathcal{H}$ définie par :

$$\forall x_1, x_2 \in \mathcal{H} : d_\lambda(x_1, x_2) = |\langle x_1 - x_2, \lambda \rangle|.$$

On note par $\hat{R}_\lambda(y, x)$ l'estimateur à noyau de la fonction de distribution $R_\lambda(y, x)$ tel que :

$$\forall y \in \mathbb{R} : \hat{R}_\lambda(y, x) = \frac{\sum_{i=1}^n G(\frac{1}{g}(y - Y_i)) K(\frac{1}{h} d_\lambda(X_i, x))}{\sum_{i=1}^n K(\frac{1}{h} d_\lambda(X_i, x))}, \tag{4.1}$$

la fonction K est le noyau de type I ou de type II et G est défini par $\forall t \in \mathbb{R}$, $G(t) = \int_{-\infty}^{t} K_0(v)dv$ où K_0 est un noyau de type 0 et $h = h(n)$ (resp. $g = g(n)$) est une suite de nombres réels positifs qui tend vers zéro lorsque n tend vers $+\infty$. Cette estimation étend, de manière différente, les travaux de Roussas [48] dans le cas réel et de Ferraty et al. [24] dans le cas fonctionnel. Notre contribution théorique consiste en la reformulation de l'estimateur de la fonction de distribution conditionnelle de Y étant donné $X = x$. introduit par [23], afin de le rendre applicable à la simulation. L'estimateur que nous avons considéré, tel que reformulé, est celui proposé par Ferraty et Vieu [24] pour cette fonction, tel que présenté dans l'équation (4.1). Dans cette étude, nous simplifions le processus d'estimation en conservant le noyau K et en remplaçant le noyau G par une fonction de distribution de Y, notée F_Y. Cette modification élimine la complexité associée à l'introduction d'un noyau supplémentaire k_0, ce qui nous permet d'utiliser uniquement le noyau K. Cette contribution est cruciale pour notre approche pratique. Par conséquent, l'estimateur reformulé est simplifié comme suit :

$$\hat{R}_\lambda(y|x) = \sum_{i=1}^{n} W_i F_Y\left(\frac{Y_i - y}{g}\right), \tag{4.2}$$

où

$$W_i = \sum_{i=1}^{n} \frac{K(\frac{1}{h} d_\lambda(X_i, x))}{\sum_{i=1}^{n} K(\frac{1}{h} d_\lambda(X_i, x))}, \tag{4.3}$$

et F_Y est la fonction de distribution de Y.

En utilisant le seul noyau K, notre approche simplifie la procédure d'estimation et réduit la complexité de calcul. L'utilisation de la fonction de distribution de Y au lieu d'un noyau supplémentaire k_0 rend la mise en œuvre plus simple et plus facile. Malgré cette simplification, notre estimateur conserve les propriétés statistiques souhaitées, garantissant une estimation robuste de la fonction de distribution conditionnelle. En d'autres termes, notre approche simplifie considérablement l'estimation de la fonction de distribution conditionnelle tout en maintenant la rigueur théorique et la cohérence statistique du modèle.

4.3 La convergence presque complète de la fonction de répartition conditionnelle

Nous ferons référence à plusieurs résultats théoriques discutés en détail par Ferraty et al. [24] dans leur ouvrage "Nonparametric Functional Data Analysis" ainsi qu'à leurs preuves, sur lesquelles nous nous appuierons pour mener à bien les aspects théoriques de cette thèse.

Soit $\mathcal{N}_x \subset \mathcal{H}$ un voisinage de x et soit $\mathcal{S}$ un sous-ensemble compact fixé de $\mathbb{R}$. Afin d'établir la convergence presque complète (p.co.) de notre estimateur, nous donnons quelques conditions de régularité nécessaires pour cette estimation introduite par Ferraty et al. [24].

(H_1) le modèle de régression est :

$$Y = E(Y|X) + \epsilon, \text{ où } E(\epsilon|X) = 0.$$

(H_2) La probabilité de la variable fonctionnelle sur les petites boules est non nulle.

$$\forall \epsilon > 0, \ P(X \in B_\lambda(x, \epsilon)) = \varphi_{x,\lambda}(\epsilon) > 0.$$

(H_3)

$$\begin{cases} h \text{ est une suite de nombre positifs tel que} \\ \lim\limits_{n\to+\infty} h = 0 \ \text{ et } \lim\limits_{n\to+\infty} \dfrac{1}{\varphi_{x,\lambda}(h)} . \dfrac{\log n}{n} = 0. \\ K \text{ est un noyau de type } I \text{ ou de type } II \text{ et} \\ \exists C > 0, \exists \epsilon > 0, \forall \epsilon < \epsilon_0, \int_0^1 \varepsilon \varphi_{x,\lambda}(u)\, du > C \varepsilon \varphi_{x,\lambda}(\varepsilon). \end{cases}$$

(H_4) On suppose que R_λ vérifie la condition de type Hölder suivante :

$$\begin{cases} \exists C_{\lambda,x} > 0 \text{ such that } \forall (y_1, y_2) \in \mathcal{S}^2, \forall (x_1, x_2) \in \mathcal{N}_x \times \mathcal{N}_x, \\ |R_\lambda(y_1, x_1) - R_\lambda(y_2, x_2)| \leq C_{\lambda,x} \left(d_\lambda^{\alpha_1}(x_1, x_2) + |y_1 - y_2|^{\alpha_2}\right), \alpha_1 > 0 \text{ et } \alpha_2 > 0. \end{cases}$$

Théorème 27. *Sous les hypothèses* $(H_1), (H_2), (H_3)$ *et* (H_4) *on a pour tout nombre réel* y

$$\hat{R}_\lambda(y, x) - R_\lambda(y, x) = O(h^{\alpha_1}) + O(g^{\alpha_2}) + O_{p.co.}\left(\frac{1}{\varphi_{x,\lambda}(h)} . \frac{\log(n)}{n}\right)^{\frac{1}{2}}.$$

Démonstration du théorème 27.
Rappelons que nous avons défini dans le chapitre 3 :

$$\Omega_{i,\lambda}(x) = \frac{K\left(\frac{d_\lambda(X_i,x)}{h}\right)}{EK\left(\frac{d_\lambda(X_i,x)}{h}\right)}.$$

et que :

$$c_1\phi_{x,\lambda}(h) \leq \mathbb{E}K\left(\frac{d_\lambda(X,x)}{h}\right) \leq c_2\phi_{x,\lambda}(h), \tag{4.4}$$

ce qui garantit l'existence de $\Omega_{i,\lambda}$ sous certaines hypothèses sur le noyau K. (voir Ferraty et Vieu, 2006, p. 43, Lemme 4.3, [24]).
Maintenant, pour traiter la preuve du théorème 27, nous avons besoin de la décomposition 4.5 et des lemmes 28 et 29 suivants :

$$\begin{aligned}\hat{R}_\lambda(y,x) - R_\lambda(y,x) = &\frac{1}{\hat{r}_{1,\lambda}(x)}\left[(\hat{r}_{2,\lambda}(x,y) - E\hat{r}_{2,\lambda}(x,y) - (R_\lambda(y,x) - E\hat{r}_{2,\lambda}(x,y))\right] \\ &- \frac{R_\lambda(y,x)}{\hat{r}_{1,\lambda}(x)}(\hat{r}_{1,\lambda}(x) - 1),\end{aligned} \tag{4.5}$$

où

$$\hat{r}_{1,\lambda}(x) = \frac{1}{n}\sum_{i=1}^{n}\Omega_{i,\lambda}(x) \text{ et } \hat{r}_{2,\lambda}(x,y) = \hat{r}_{1,\lambda}(x)\,\hat{R}_\lambda(y,x) = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i(y)\Omega_{i,\lambda}(x),$$

avec $\Omega_{i,\lambda}(x) = \dfrac{K\left(\frac{d_\lambda(X_i,x)}{h}\right)}{EK\left(\frac{d_\lambda(X_i,x)}{h}\right)}$ et $\Gamma_i(y) = G(\frac{y-Y_i}{g})$.

Ainsi, la preuve du théorème est une conséquence directe des lemmes suivants : □

Lemme 28. *i) Sous les hypothèses du théorème 27, lorsque n tend vers l'infini, on a :*

$$\hat{r}_{1,\lambda}(x) - 1 = O_{p.co.}\left(\sqrt{\frac{1}{\varphi_{x,\lambda}(h)}.\frac{\log n}{n}}\right).$$

ii) Sous les hypothèses du théorème 27, lorsque n tend vers l'infini, on a :

$$E\hat{r}_{2,\lambda}(x,y) - \hat{r}_{2,\lambda}(x,y) = O_{p.co.}\left(\sqrt{\frac{1}{\varphi_{x,\lambda}(h)}.\frac{\log n}{n}}\right).$$

Lemme 29. *Sous les hypothèses (H_1), (H_2) et (H_4), lorsque n tend vers l'infini, on a :*

$$R_\lambda(y,x) - E\hat{r}_{2,\lambda}(x,y) = O\left(h^{\alpha_1}\right) + O_{p.co.}\left(g^{\alpha_2}\right).$$

Les numérateurs de la décomposition sont traités en faisant les deux lemmes 28 et 29 ci-dessus, tandis que les dénominateurs sont traités en utilisant la première partie du lemme 28 et la première partie de la proposition $A.6$.(voir [24]).

4.4 La convergence presque complète uniforme

Dans cette section, nous présentons les principaux résultats relatifs à la fonction de distribution conditionnelle estimée, en nous appuyant notamment sur les travaux de Ait Saïdi (2016) [49]. Nous commençons par énoncer les théorèmes fondamentaux et les lemmes qui en découlent, avant de proposer des remarques complémentaires permettant d'éclairer leur interprétation et leur portée pratique. Dans un premier temps, nous limitons notre analyse à un noyau de type I. Par la suite, nous supposons que $S_{\mathbb{R}}$ est un sous-ensemble compact de $\mathbb{R}$, alors que $S_{\mathcal{H}}$ et $\Lambda_{\mathcal{H}}$ représentent les espaces de paramètres tels que

$$S_{\mathcal{H}} \subset \bigcup_{k=1}^{d_n^{S_{\mathcal{H}}}} B(x_k, r_n) \text{ et } \Lambda_{\mathcal{H}} \subset \bigcup_{j=1}^{d_n^{\Lambda_{\mathcal{H}}}} B(t_j, r_n),$$

avec x_k (resp. t_j)$\in$Hetd$_n^{S_{\mathcal{H}}}, d_n^{\Lambda_{\mathcal{H}}}$ sont des suites de nombres réels positifs qui tendent vers l'infini lorsque $n \to \infty$.

De plus, nous avons besoin des hypothèses suivantes :

(A_1) Il existe une fonction différentiable $\varphi(.)$ telle que : $\forall x \in S_{\mathcal{H}}$ et $\forall \lambda \in \Lambda_{\mathcal{H}}$,

$$0 < C\varphi(h) \leq \varphi_{x,\lambda}(h) \leq C'\varphi(h) < \infty \text{ et } \exists \eta_0 > 0, \forall \eta < \eta_0, \varphi'(h) < C.$$

(A_2) La fonction de répartition conditionnelle vérifie :
$\forall (y_1, y_2) \in S_{\mathbb{R}}^2, \forall (x_1, x_2) \in S_{\mathcal{H}}^2, \forall \lambda \in \Lambda_{\mathcal{H}}$,

$$|R_\lambda(y_1, x_1) - R_\lambda(y_2, x_2)| \leq C\left(d_\lambda^{\alpha_1}(x_1 - x_2) + |y_1 - y_2|^{\alpha_2}\right), \alpha_1 > 0, \alpha_2 > 0.$$

(A_3) Le noyau K satisfait la condition de type Lipschitz suivante :

$$\left|K\left(\frac{d_\lambda(X_i, x_1)}{h}\right) - K\left(\frac{d_\lambda(X_i, x_2)}{h}\right)\right| \leq \frac{C}{h} d_\lambda(x_1 - x_2).$$

(A_4) Pour tout $\gamma \in (0,1), gn^\gamma \to \infty$ quand $n \to \infty$, et pour $r_n = O(\frac{\log(n)}{n})$, les suites $d_n^{S_{\mathcal{H}}}$ et $d_n^{\Lambda_{\mathcal{H}}}$ verifient :

$$\frac{(\log(n))^2}{n\varphi(h)} < \log d_n^{S_H} + \log d_n^{\Lambda_{\mathcal{H}}} < \frac{n\varphi(h)}{\log(n)},$$

$$\text{et pour tout } \alpha > 1, \sum_{n=1}^{\infty} n^{\gamma + \frac{1}{2}} (d_n^{S_H} d_n^{\Lambda_{\mathcal{H}}})^{1-\alpha} < \infty.$$

Remarque 30. *([49]) Notons que les hypothèses (A_1) et (A_2) sont respectivement la version uniforme de (H_2) et (H_4). Ici, l'hypothèse sur K est reformulée par l'hypothèse (A_3). L'hypothèse (A_4) est une condition technique et est également similaire à celles de Ferraty et al. [24]*

Théorème 31. *Sous les hypothèses $(A_1) - (A_4)$ nous avons :*

$$\sup_{\lambda \in \Lambda_{\mathcal{H}}} \sup_{x \in S_H} \sup_{y \in S_{\mathbb{R}}} \left|\hat{R}_\lambda(y, x) - R_\lambda(y, x)\right| = O(h^{\alpha_1}) + O(g^{\alpha_2}) + O_{p.co.}\left(\sqrt{\frac{\log d_n^{S_H} + \log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right).$$

Démonstration du théorème 31. Tout d'abord, d'après (A_1) et puisque K est un noyau de type I, on a en utilisant le résultat du théorème 27,

$$\forall x \in S_{\mathcal{H}}, \forall \lambda \in \Lambda_{\mathcal{H}}, \exists\ 0 < C < C' < \infty, C\varphi(h) < EK\left(\frac{d_\lambda(X_i, x)}{h}\right) < C'\varphi(h). \tag{4.6}$$

□

A partir de ce point, le théorème 31 peut être dérivé des résultats intermédiaires suivants qui sont une version uniforme des lemmes 28-29.

Lemme 32. *Sous les hypothèses* $(A_1), (A_3)$ *et* (A_4)*, lorsque* n *tend vers l'infini, on a*

$$\sup_{\lambda\in\Lambda_{\mathcal{H}}}\sup_{x\in S_H}\left|\hat{r}_{1,\lambda}(x)-1\right| = O_{p.co}\left(\sqrt{\frac{\log d_n^{S_H}+\log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right).$$

Lemme 33. *Sous les hypothèses* (A_1)*,* (A_2)*, quand* n *tend vers l'infini, on a :*

$$\sup_{\lambda\in\Lambda_{\mathcal{H}}}\sup_{x\in S_H}\sup_{y\in S_{\mathbb{R}}}\left|E\hat{r}_{2,\lambda}(x,y)-R_\lambda(y,x)\right| = O\left(h^{\alpha_1}\right)+O\left(g^{\alpha_2}\right).$$

Lemme 34. *Sous les hypothèses du théorème 31 et lorsque* n *tend vers l'infini, on a :*

$$\sup_{\lambda\in\Lambda_{\mathcal{H}}}\sup_{x\in S_H}\sup_{y\in S_{\mathbb{R}}}\left|E\hat{r}_{2,\lambda}(x,y)-\hat{r}_{2,\lambda}(x,y)\right| = O_{p.co}\left(\sqrt{\frac{\log d_n^{S_H}+\log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right).$$

Corollaire 35. *Sous les hypothèses du Lemme 34, on a :*

$$\sum_{i=1}^{n} P\left(\inf_{\lambda\in\Lambda_{\mathcal{H}}}\inf_{x\in S_{\mathcal{H}}}\hat{r}_{1,\lambda}(x)<\frac{1}{2}\right)<\infty. \tag{4.7}$$

4.5 Simulation : approche semi paramétrique

Cette section met en évidence le comportement de convergence forte de l'estimateur vers sa fonction théorique à travers une comparaison directe. Nous nous concentrons spécifiquement sur le cas où la fonction théorique est dérivée de $E(Y|<\lambda,X>)$, où λ est un indice fonctionnel comme discuté dans la Section 3. Dans le contexte de l'estimation de la fonction de distribution conditionnelle sur le modèle de régression d'indice fonctionnel $Y=r_\lambda(X)+\epsilon$ où ϵ est une variable aléatoire $\rightsquigarrow\mathcal{N}(0,0.05)$ en utilisant l'estimateur nouvellement reformulé, nous considérons un exemple de simulation pour illustrer la convergence de l'estimateur vers la fonction théorique. La fonction de lien $r_\lambda(x)$ est définie par $r_\lambda(x)=2<\lambda,x>$, nous conduit à la fonction théorique R_λ, qui représente la fonction de distribution cumulative conditionnelle donnée $X=x$ définie par $R_\lambda(y,x)=\Phi\left(\frac{y-2\langle\lambda,x\rangle}{\sigma}\right)$, où ϕ désigne la fonction de distribution cumulative de la distribution normale standard, et $\sigma=0.05$

4.5.1 Génération de données

Dans le but d'évaluer la performance de l'approche semi-paramétrique proposée, nous générons un échantillon synthétique constitué de $n=100$ courbes fonctionnelles. Ces courbes simulent des trajectoires continues sur un intervalle $[0,1]$, et sont définies par la relation suivante :

$$X_i(t_j)=\left|\sin\left(\frac{1+\sqrt{2}}{2}\left(w_i+\pi\left(\frac{1+\sqrt{2}}{2}\right)t_j-1\right)\right)\right|,$$

où $0=t_1<t_2<\cdots<t_{99}<t_{100}=1$ représentent des points équidistants sur l'intervalle d'observation. Les w_i sont des variables aléatoires indépendantes et identiquement distribuées suivant une loi uniforme sur $[0,1]$. Cette forme fonctionnelle, bien que simple,

permet de générer une grande variété de formes de courbes tout en conservant une certaine régularité. La figure 4.1 illustre ces courbes générées.

Une fois les courbes construites, nous simulons un modèle d'indice fonctionnel selon les étapes suivantes :

1. **Paramètres de lissage :** Pour l'estimation par noyau, deux paramètres de lissage sont requis : $h = 0.1$ et $g = 0.001$. Ces paramètres sont sélectionnés par balayage sur une grille de valeurs afin d'assurer un bon compromis biais-variance.

2. **Choix du vecteur d'indice :** Nous fixons un vecteur de projection $\lambda(t_j)$ selon la formule suivante :
$$\lambda(t_j) = t_j \sin\left(\frac{1}{1+t_j^4}\right),$$
ce qui garantit une certaine variabilité tout en maintenant la continuité de la direction de projection.

3. **Fonction de lien :** Nous considérons une fonction de lien linéaire définie par $r_\lambda(x) = 2\langle\lambda, x\rangle$, où $\langle\lambda, x\rangle$ représente le produit scalaire entre la courbe x et le vecteur λ.

4. **Calcul des indices :** Pour chaque courbe X_i, nous calculons le produit scalaire $\langle\lambda, X_i\rangle$, ce qui permet de projeter les courbes dans une direction informative unique.

5. **Génération du bruit :** Nous générons, pour chaque individu $i = 1, \ldots, n$, un bruit additif ϵ_i issu d'une loi normale centrée de variance $0.05 \cdot \mathrm{Var}_{\mathrm{emp}}(r(\langle\lambda, x_i\rangle))$. Ce choix correspond à un rapport signal sur bruit fixé à 5%.

6. **Construction de la variable réponse :** Enfin, la variable dépendante Y_i est obtenue par le modèle suivant :
$$Y_i = r(\langle\lambda, X_i\rangle) + \epsilon_i.$$
Ce modèle correspond à une régression sur un indice fonctionnel unique, tout en intégrant une perturbation stochastique réaliste.

Cette procédure de simulation permet de générer un jeu de données complet pour tester l'efficacité de notre méthode d'estimation. Elle permet également de contrôler de manière précise la structure du signal, la difficulté du problème, ainsi que le niveau de bruit.

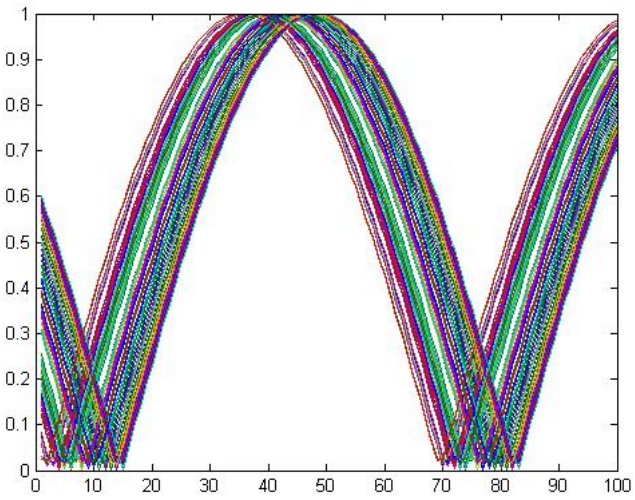

Fig. 4.1 : Courbes simulées (n=100).

4.5.2 Calcul de l'estimateur

La fonction noyau K est choisie comme gaussienne et la fonction théorique est prise comme R_λ. Pour calculer l'estimateur, commencez par calculer les poids W_i en faisant le noyau gaussien K. Ce noyau est appliqué aux produits scalaires entre la différence de chaque observation X_i et l'observation spécifiée x, projetée sur la direction définie par l'indice fonctionnel λ. Plus précisément, pour chaque i, on calcule $\langle X_i - x, \lambda \rangle$, qui correspond à la projection de X_i et de x sur λ. Cette projection est ensuite normalisée par la bande passante h. Ainsi, les poids sont donnés par la formule $K(\frac{\langle X_i - x, \lambda \rangle}{h})$, où le paramètre h contrôle l'influence relative des observations X_i sur l'estimation. Ensuite, estimez la fonction de distribution F_Y en utilisant la fonction de distribution empirique ou l'estimation de la densité du noyau (comme ksdensity dans MATLAB) appliquée à $\frac{(Y-y_i)}{g}$ où g est un paramètre de bande passante supplémentaire. De plus, nous avons utilisé les valeurs de l'erreur quadratique moyenne MSE pour évaluer la qualité de toute estimation de courbe de cette fonction.

$$MSE = \frac{1}{100} \sum_{i=1}^{100} \Big(R_\lambda(y, x) - \hat{R}_\lambda(y, x) \Big)^2 . \tag{4.8}$$

4.5.3 Visualisation des résultats

Les résultats de simulation sont résumés dans la figure 4.2 et le tableau 4.1 pour le cas où la distribution F_Y est estimée à l'aide de la fonction de distribution empirique. Dans le cas où la fonction F_Y est estimée à l'aide de la 'ksdensity' dans Matlab, les résultats sont présentés dans la figure 4.3 et le tableau 4.2.

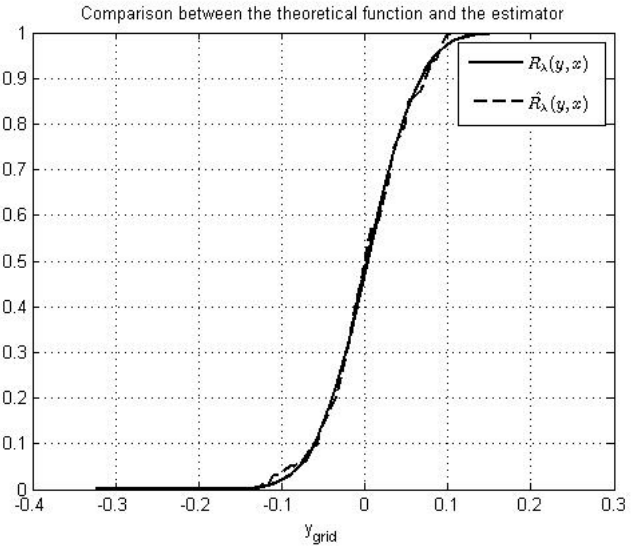

Fig. 4.2 : Estimation empirique.

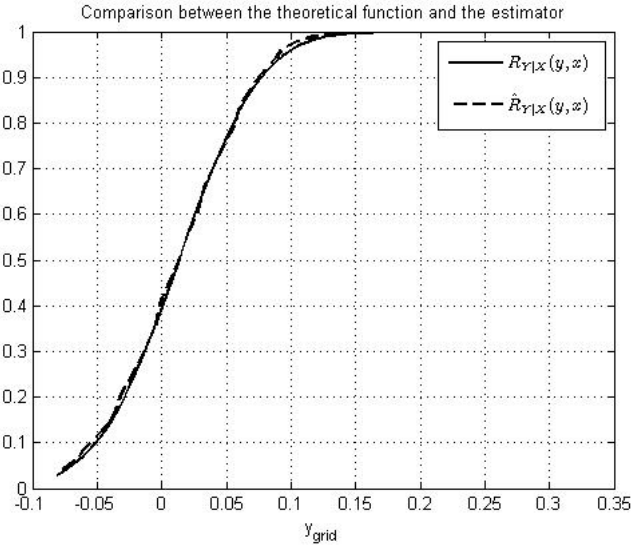

Fig. 4.3 : Estimation par 'ksdensity'

n	$R_\lambda(y;x)$	$\hat{R}_\lambda(y;x)$	MSE
	0.3096	0.3000	
	0.3553	0.3500	
	0.4032	0.4100	
	0.4526	0.4600	
	0.5028	0.5300	
	0.5529	0.5800	
	0.6021	0.5900	
	0.6498	0.6300	
	0.6953	0.6700	
	0.7379	0.7500	
21	0.7772	0.7700	10^{-4}
	0.8129	0.7900	
	0.8449	0.8600	
	0.8730	0.8600	
	0.8974	0.8700	
	0.9181	0.9000	
	0.9356	0.9300	
	0.9500	0.9400	
	0.9617	0.9700	
	0.9710	0.9900	
	0.9784	1.0000	

Tableau 4.1. Estimation empirique.

n	$R_\lambda(y;x)$	$\hat{R}_\lambda(y;x)$	MSE
	0.9920	0.9966	
	0.9936	0.9974	
	0.9968	0.9990	
	0.9949	0.9981	
	0.9975	0.9995	
	0.9981	0.9996	
	0.9985	0.9998	
	0.9989	0.9998	
	0.9991	0.9999	
	0.9993	0.9999	
21	0.9996	1.0000	5.10^{-5}
	0.9997	1.0000	
	0.9998	1.0000	
	0.9999	1.0000	
	0.9999	1.0000	
	0.9999	1.0000	
	0.9999	1.0000	
	1.0000	1.0000	
	1.0000	1.0000	
	1.0000	1.0000	
	1.0000	1.0000	

Tableau 4.2. Estimation par « ksdensity ».

En conclusion, la section ci-dessus illustre la construction d'un échantillon de courbes, le calcul de l'estimateur, ainsi que la comparaison avec la fonction théorique. Les courbes présentées dans les figures 4.2, 4.3 mettent en lumière la qualité de l'estimation, en montrant la convergence progressive de l'estimateur vers la fonction théorique. Cette convergence est confirmée par la proximité des deux courbes, témoignant de l'efficacité de l'approche. De plus, les valeurs numériques issues des tableaux 4.1 et 4.2 renforcent cette observation, avec des erreurs quadratiques moyennes (MSE) très faibles, notamment $MSE = 0.0001$ pour l'empirique et $MSE = 5 \times 10^{-5}$ pour "ksdensity" , respectivement. Ces résultats illustrent clairement la précision croissante de l'estimation, en particulier à mesure que n augmente, confirmant ainsi la robustesse de l'estimateur proposé.

4.6 Annexe B

Démonstration du lemme 29. Le fait que K soit supporté sur $[0,1]$ et le fait que $E\Omega_{1,\lambda} = 1$ conduisent directement à écrire :

$$\begin{aligned} E\hat{r}_{2,\lambda}(x,y) - R_\lambda(y,x) &= E\Gamma_1(y)\,\Omega_{1,\lambda}(x) - E\Omega_{1,\lambda}(x)R_\lambda(y,x) \\ &= E(E(\Gamma_1(y)\,\Omega_{1,\lambda}(x)|X)) - E\Omega_{1,\lambda}(x)R_\lambda(y,x) \\ &= E\left(\Omega_{1,\lambda}(x)I_{B(x,h)}(X)\left(E(\Gamma_{,}(y)\,|X) - R_\lambda(y,x)\right)\right). \end{aligned} \tag{4.9}$$

Puisque cette espérance finale peut être facilement calculée à l'aide du théorème de Fubini et du fait $H' = K_0$ on a :

$$\begin{aligned} E(\Gamma_1(y) \mid <\lambda, X>) &= \int_{\mathbb{R}} H\left(\frac{y-u}{g}\right) dP(u \mid <\lambda, X>) \\ &= \int_{\mathbb{R}} \int_{-\infty}^{\frac{y-u}{g}} K_0(v)\, dv dP(u \mid <\lambda, X>) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} K_0(v)\, I_{[v;\infty[}\left(\frac{y-u}{g}\right) dv dP(u \mid <\lambda, X>) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} K_0(v)\, I_{[v;\infty[}\left(\frac{y-u}{g}\right) dP(u \mid <\lambda, X>)\, dv \\ &= \int_{\mathbb{R}} K_0(v)\, R_\lambda(y - vg, X)\, dv. \end{aligned} \tag{4.10}$$

Comme K_0 s'intègre à 1, on obtient :

$$E(\Gamma_1(y) \mid <\lambda, X>) - R_\lambda(y, x) = \int_{\mathbb{R}} K_0(v)\left(R_\lambda(y - vg, X) - R_\lambda(y, x)\right). \tag{4.11}$$

Puisque h et g tendent vers zéro et que K_0 est de support $[-1, 1]$, la propriété de Hölder (H_4) permet d'écrire :

$$\sup_{v\in[-1,1]} 1_{B(x,h)}(X)|R_\lambda(y - vg, X) - R_\lambda(y, x)| \leq \sup_{v\in[-1,1]} 1_{B(x,h)}(X) C\left(d_\lambda(X, x)^{\alpha_1} + |vg|^{\alpha_2}\right). \tag{4.12}$$

Nous arrivons enfin à :

$$\begin{aligned} &\lim_{n\to\infty} \sup_{v\in[-1,1]} 1_{B(x,h)}(X)|R_\lambda(y - vg, X) - R_\lambda(y, x)| \\ &\leq \lim_{n\to\infty} \sup_{v\in[-1,1]} 1_{B(x,h)}(X) C\left(d_\lambda(X, x)^{\alpha_1} + |vg|^{\alpha_2}\right), \end{aligned} \tag{4.13}$$

qui peut être combiné avec 4.11 conduit à :

$$\lim_{n\to\infty} \sup_{v\in[-1,1]} 1_{B(x,h)} \left|E(\Gamma_1(y) \mid <\lambda, X>) - R_\lambda(y, x)\right| = O\left(h^{\alpha_1}\right) + O\left(g^{\alpha_2}\right). \tag{4.14}$$

Ce dernier résultat, ainsi que $E\Omega_1 = 1$ avec le résultat 4.9 suffisent à prouver le lemme 29. □

Démonstration du Lemme 28 deuxième partie [49]. Pour cela, nous utilisons la décomposition suivante

$$\hat{r}_{2,\lambda}(x, y) - E\hat{r}_{2,\lambda}(x, y) = \frac{1}{n}\sum_{i=1}^{n} T_i, \tag{4.15}$$

où

$$T_i = (Z_i - EZ_i) \text{ et } Z_i = \Omega_{i,\lambda}\Gamma_i(y).$$

Pour appliquer l'inégalité de type Bernstein, nous devons d'abord montrer que :

$$|T_i| \leq \frac{C}{\varphi_{x,\lambda}(h)} \text{ et } ET_i^2 \leq \frac{C}{\varphi_{x,\lambda}(h)}. \tag{4.16}$$

En utilisant le résultat 4.4 et en tenant compte du fait que K est de type I ou II ainsi que de l'hypothèse (H_3), on obtient :

$$C\varphi_{x,\lambda}(h) \leq EK\left(\frac{d_\lambda(X_i,x)}{h}\right) \leq C'\varphi_{x,\lambda}(h).$$

Et en utilisant le résultat 4.6 et puisque $\Gamma_i(y)$ est borné, nous obtenons :

$$|T_i| \leq \frac{C}{\varphi_{x,\lambda}(h)}.$$

D'un autre côté, on a :

$$EZ_i^2 \leq CE\Omega_{i,\lambda}^2$$

Puisque $0 < \int K^2 < \infty$, si K est de type I (resp. II) alors $\frac{K^2}{\int K^2}$ est aussi de type I (resp. II). Donc, en appliquant le résultat (**??**) on obtient

$$C\varphi_{x,\lambda}(h) \leq EK^2\left(\frac{d_\lambda(X_i,x)}{h}\right) \leq C'\varphi_{x,\lambda}(h), \tag{4.17}$$

et en utilisant le résultat 4.17, nous écrivons :

$$\frac{C}{\varphi_{x,\lambda}(h)} \leq E\Omega_{i,\lambda}^2 \leq \frac{C'}{\varphi_{x,\lambda}(h)},$$

ce qui implique :

$$ET_i^2 \leq EZ_i^2 \leq CE\Omega_{i,\lambda}^2 \leq \frac{C'}{\varphi_{x,\lambda}(h)}.$$

Grâce au résultat (4.16), nous sommes en mesure d'appliquer l'inégalité de type Bernstein donnée par le corollaire A.9-i (voir Ferraty et al. [24]), et nous obtenons :

$$\forall \varepsilon > 0, P\left(|\hat{r}_{2,\lambda}(x,y) - E\hat{r}_{2,\lambda}(x,y)| > \varepsilon\right) \leq 2\exp\left(\frac{\varepsilon^2 n\varphi_{x,\lambda}(h)}{2C'(1+\varepsilon)}\right). \tag{4.18}$$

Comme la suite $\dfrac{\log(n)}{n\varphi_{x,\lambda}(h)}$ tend vers zéro, en choisissant $\varepsilon = \varepsilon_0\sqrt{\dfrac{\log(n)}{n\varphi_{x,\lambda}(h)}}$ dans le résultat (4.18), on obtient directement :

$$P\left(|\hat{r}_{2,\lambda}(x,y) - E\hat{r}_{2,\lambda}(x,y)| > \varepsilon_0\sqrt{\frac{\log(n)}{n\varphi_{x,\lambda}(h)}}\right) \leq 2\exp\left(\frac{\varepsilon_0^2\log(n)}{2C'\left(1+\varepsilon_0\sqrt{\dfrac{\log(n)}{n\varphi_{x,\lambda}(h)}}\right)}\right)$$
$$\leq 2n^{-C\varepsilon_0^2},$$

et il s'ensuit que pour ε_0 suffisamment grand $\left(\varepsilon_0 > \sqrt{\frac{1}{C}}\right)$,

$$\sum_{i=1}^{\infty} P\left(|\hat{r}_{2,\lambda}(x,y) - E\hat{r}_{2,\lambda}(x,y)| > \epsilon_0\sqrt{\frac{\log(n)}{n\varphi_{x,\lambda}(h)}}\right) < \infty.$$

Donc, la preuve de la deuxième partie du lemme 28 est maintenant achevée. □

Démonstration du lemme 32 (Voir [5]). Pour tout $x \in S_H$, on pose :
$k(x) = \arg \min_{k\in\{1,\dots r_n\}} \|x - x_k\|$, et $j(\lambda) = \arg \min_{j\in\{1,\dots l_n\}} \|\lambda - t_j\|$.

Nous considérons la décomposition suivante :

$$\sup_{x\in S_H}\sup_{\lambda\in\Lambda_{\mathcal{H}}} |\hat{r}_{1,\lambda}(x) - E\hat{r}_{1,\lambda}(x)| \leq \underbrace{\sup_{x\in S_H}\sup_{\lambda\in\Lambda_{\mathcal{H}}} |\hat{r}_{1,\lambda}(x) - \hat{r}_{1,\lambda}(x_k)|}_{L_1}$$
$$+ \underbrace{\sup_{x\in S_H}\sup_{\lambda\in\Lambda_{\mathcal{H}}} \left|\hat{r}_{1,\lambda}(x_k) - \hat{r}_{1,t_j}(x_k)\right|}_{L_2} + \underbrace{\sup_{x\in S_H}\sup_{\lambda\in\Lambda_{\mathcal{H}}} \left|\hat{r}_{1,t_j}(x_k) - E\hat{r}_{1,t_j}(x_k)\right|}_{L_3}$$
$$+ \underbrace{\sup_{x\in S_H}\sup_{\lambda\in\Lambda_{\mathcal{H}}} \left|E\hat{r}_{1,t_j}(x_k) - E\hat{r}_{1,\lambda}(x_k)\right|}_{L_4} + \underbrace{\sup_{x\in S_H}\sup_{\lambda\in\Lambda_{\mathcal{H}}} |E\hat{r}_{1,\lambda}(x_k) - E\hat{r}_{1,\lambda}(x)|}_{L_5}.$$

Tout d'abord, nous étudions L_1, L_2, L_4 et L_5. Pour ces termes nous utilisons la condition de continuité de Hölder sur K et l'inégalité de Cauchy-Schwartz. Avec ces arguments nous obtenons :

$$\begin{aligned} L_1 &\leq \frac{C}{\varphi(h)} \sup_{x\in S_H}\sup_{\lambda\in\Lambda_{\mathcal{H}}} \frac{1}{n}\sum_{i=1}^{n} |K_{i,\lambda}(x) - K_{i,\lambda}(x_k)| \\ &\leq \frac{Cr_n}{h\varphi(h)}, \end{aligned}$$

et de manière analogue, pour L_2 on obtient :

$$L_2 \leq \frac{Cr_n}{h\varphi(h)} 1_{B(x,h)\cup B(x_k,h)}(X_i).$$

Ce dernier peut être traité par l'inégalité de Bernstein, avec

$$W_i = \frac{\varepsilon}{h\varphi(h)} 1_{B(x,h)\cup B(x_k,h)}(X_i),$$

ce qui donne, pour n tendant vers l'infini :

$$L_1 = O\left(\sqrt{\frac{\log d_n^{S_H} + \log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right) \text{ et } L_2 = O\left(\sqrt{\frac{\log d_n^{S_H} + \log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right). \tag{4.19}$$

De plus, en utilisant le fait que $L_4 \leq L_1$ et $L_5 \leq L_2$ pour obtenir, pour n tendant vers l'infini :

$$L_4 = O\left(\sqrt{\frac{\log d_n^{S_H} + \log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right) \text{ et } L_5 = O\left(\sqrt{\frac{\log d_n^{S_H} + \log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right). \tag{4.20}$$

Maintenant, nous traitons T_3. Pour tout $\eta > 0$, nous avons :

$$P\left(T_3 > \eta\sqrt{\frac{\log d_n^{S_H} + \log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right)$$
$$\leq d_n^{S_H} d_n^{\Lambda_{\mathcal{H}}} \max_{k\in\left\{1,\dots,d_n^{S_H}\right\}} \max_{j\in\left\{1,\dots,d_n^{\Lambda_{\mathcal{H}}}\right\}} P\left(\left|\hat{r}_{1,t_j}(x_k) - E\hat{r}_{1,t_j}(x_k)\right| > \eta\sqrt{\frac{\log d_n^{S_H} + \log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right)$$
$$= F,$$

nous posons :

$$\Phi_i = \frac{1}{\varphi(h)} \left(K_{i,\lambda}(x_k) - EK_{i,\lambda}(x_k)\right).$$

En appliquant l'inégalité exponentielle de Bernstein à Φ_i, et en tenant compte que K est supporté et borné sur$[-1,1]$, nous obtenons : $\forall j \leq d_n^{S_H}$ et $\forall k \leq d_n^{\Lambda_{\mathcal{H}}}$,

$$P\left(\left|\hat{r}_{1,t_j}(x_k) - E\hat{r}_{1,t_j}(x_k)\right| > \eta\sqrt{\frac{\log d_n^{S_H} + \log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right) = P\left(\frac{1}{n}\left|\sum_{i=1}^{n}\Phi_i\right| > \eta\sqrt{\frac{\log d_n^{S_H} + \log d_n^{\Lambda_{\mathcal{H}}}}{n\varphi(h)}}\right)$$
$$\leq 2\left(d_n^{S_H} d_n^{\Lambda_{\mathcal{H}}}\right)^{-C\eta^2}.$$

Par conséquent, en choisissant $C\eta^2 = \beta$ sous la condition (A_3) nous avons,

$$F = O\left(\left(d_n^{S_H} d_n^{\Lambda_{\mathcal{H}}}\right)^{1-C\eta^2}\right).$$

Finalement, le résultat peut être facilement déduit de ce dernier avec (4.19) and (4.20). □

Démonstration du lemme 33. Il suffit de combiner la preuve du lemme 32 avec l'hypothèse (A_2) où la condition de Lipschitz est supposée uniformément sur x, y et λ. □

Démonstration du lemme 34. On garde les mêmes notations que dans le lemme 32 et on utilise la compacité de S_H. On peut écrire :

$$S_H \subset \bigcup_{k=1}^{z_n}(y_j - l_n, y_j + l_n),$$

avec $l_n = n^{-\frac{(3\gamma+1)}{2}}$ et $z_n \leq Cn^{\frac{(3\gamma+1)}{2}}$. En prenant

$$j_y = \arg\min_{j\in\{1,\ldots,z_n\}}|y - t_j|.$$

On obtient la décomposition suivante :

$$\begin{aligned}\left|\hat{R}_\lambda(y,x) - E\hat{R}_\lambda(y,x)\right| \leq{}& \underbrace{\left|\hat{R}_\lambda(y,x) - \hat{R}_\lambda(y,x_k)\right|}_{F_1} + \underbrace{\left|\hat{R}_\lambda(y,x_k) - \hat{R}_{\lambda_j}(y,x_k)\right|}_{F_2}\\ &+\underbrace{\left|\hat{R}_{\lambda_j}(y_j,x_k) - \hat{R}_{\lambda_j}(y_j,x_k)\right|}_{F_3} + \underbrace{\left|\hat{R}_{\lambda_j}(y_j,x_k) - E\hat{R}_{\lambda_j}(y,x_k)\right|}_{F_4}\\ &+\underbrace{\left|E\hat{R}_{\lambda_i}(y_j,x_k) - E\hat{R}_{\lambda_j}(y,x_k)\right|}_{F_5} + \underbrace{\left|E\hat{R}_{\lambda_j}(y,x_k) - E\hat{R}_\lambda(y,x_k)\right|}_{F_6}\\ &+\underbrace{\left|E\hat{R}_\lambda(y,x_k) - E\hat{R}_\lambda(y,x)\right|}_{F_7}.\end{aligned}$$

En utilisant les mêmes idées que pour L_1, L_2, L_4 et L_5, nous obtenons, pour n tendant vers l'infini

$$F_1 = F_7 = O\left(\frac{\log(n)}{nh\varphi_{x,\lambda}(h)}\right) \text{ et } F_2 = F_5 = O\left(\frac{\log(n)}{ng\varphi_{x,\lambda}(h)}\right). \tag{4.21}$$

Concernant les termes F_3 et F_5, l'utilisation de la condition de Lipschitz sur le noyau H, nous permet d'écrire :

$$\left|\hat{R}_{\lambda_j}(y,x_k) - \hat{R}_{\lambda_j}(y_j,x_k)\right| \leq \frac{l_n}{g^2\varphi_{x,\lambda}(h)}.$$

Or, le fait que $\lim_{n\to\infty} n^{\gamma}h = 0$ et le choix de $l_n = n^{-\left(\frac{3\gamma+1}{2}\right)}$ implique que :

$$\frac{l_n}{g^2\varphi_{x,\lambda}(h)} = o\left(\sqrt{\frac{\log d_n^{S_H}}{ng\varphi_{x,\lambda}(h)}}\right).$$

Ainsi, pour n suffisamment grand, nous avons

$$F_3 = F_5 = O_{p.co}\left(\frac{\log d_n^{S_H}}{n^{1-\gamma}\varphi_{x,\lambda}(h)}\right). \tag{4.22}$$

Enfin, l'évaluation du terme (F_4) est très proche de (L_3) dans le lemme 32, en appliquant l'inégalité exponentielle de Bernstein à

$$\Omega_i = \frac{1}{h\varphi_{x,\lambda}(h)}\left(K_i(x_k)H_i\left(t_j\right) - E\left(K_i(x_k)H_i\left(t_j\right)\right)\right).$$

Tout d'abord, il résulte du fait que les noyaux K et H sont bornés, on obtient : $\forall j \leq z_n$

$$P\left(\left|\hat{R}_{\lambda_j}\left(y_j, x_k\right) - E\hat{R}_{\lambda_j}\left(y_j, x_k\right)\right| > \eta\sqrt{\frac{\log d_n^{S_H}}{ng\varphi_{x,\lambda}(h)}}\right) \leq 2\exp\left(-C\eta^2\log d_n^{S_H}\right).$$

En choisissant : $z_n = O\left(l_n^{-1}\right) = O\left(n^{\frac{3\gamma+1}{2}}\right)$, nous obtenons :

$$\begin{aligned} P\left(F_4 > \eta\sqrt{\frac{\log d_n^{S_H}}{ng\varphi_{x,\lambda}(h)}}\right) &\leq z_n d_n^{S_H} P\left(\left|\hat{R}_{\lambda_j}\left(y_j, x_k\right) - E\hat{R}_{\lambda_j}\left(y_j, x_k\right)\right| > \eta\sqrt{\frac{\log d_n^{S_H}}{ng\varphi_{x,\lambda}(h)}}\right) \\ &\leq C' z_n\left(d_n^{S_H}\right)^{1-C\eta^2}, \end{aligned}$$

en mettant $C\eta^2 = \beta$ et en utilisant (A_4), nous obtenons

$$F_4 = O\left(\sqrt{\frac{\log d_n^{S_H}}{ng\varphi_{x,\lambda}(h)}}\right). \tag{4.23}$$

Ainsi, le lemme 34 peut être facilement déduit de (4.21)-(4.23). □

4.7 Conclusion

Dans ce chapitre, nous avons exploré les aspects pratiques de l'estimation de la fonction de distribution conditionnelle pour des variables fonctionnelles dans un modèle de régression à indice fonctionnel en utilisant une méthode à noyau. En nous appuyant sur les travaux théoriques existants, nous avons introduit un estimateur reformulé qui permet au noyau de jouer le rôle d'une fonction de répartition (FDC) pour Y. Cette reformulation améliore la flexibilité et l'adaptabilité de l'estimateur, permettant une estimation plus précise dans une variété de scénarios pratiques. Nos résultats de simulation démontrent l'efficacité de cette approche, montrant de bonnes propriétés de convergence et validant les fondements théoriques de notre méthode. Cependant, la mise en œuvre de cet estimateur a révélé des défis importants, notamment dans la gestion des deux paramètres de lissage requis par la méthode du double noyau. La sélection des valeurs optimales pour ces paramètres est complexe et a un impact considérable sur la performance de l'estimateur.

Pour relever ce défi, nous proposons d'utiliser la méthode de validation croisée comme stratégie d'optimisation des paramètres de lissage. Cette approche rendrait l'estimateur plus robuste et améliorerait sa précision prédictive dans des applications pratiques. Dans l'ensemble, notre contribution fait progresser la compréhension et l'application pratique de l'estimation de la fonction de distribution conditionnelle pour des données fonctionnelles. En reformulant l'estimateur pour qu'il utilise le noyau en tant que fonction de répartition pour Y, nous offrons un outil plus polyvalent qui peut être efficacement appliqué dans divers contextes, ouvrant la voie à de nouvelles recherches et développements dans ce domaine.

Conclusion générale

Cette thèse a permis de développer des approches nouvelles pour l'estimation dans des modèles de régression fonctionnelle, en se concentrant particulièrement sur l'utilisation des méthodes à noyau pour les données fonctionnelles. Les contributions de ce travail combinent des bases théoriques solides avec des techniques de validation efficaces, permettant ainsi d'améliorer la précision et la fiabilité des estimations, tout en proposant des solutions pratiques aux défis rencontrés dans ce domaine. Nous avons d'abord étudié l'estimation de l'opérateur de régression dans un modèle à indice fonctionnel unique, en utilisant la méthode du noyau. La sélection du paramètre de lissage, grâce à des techniques comme la validation croisée Leave-One-Out (LOO), a joué un rôle clé pour optimiser les résultats et garantir la convergence théorique de l'estimateur. Cette méthode a démontré son efficacité en offrant un bon équilibre entre simplicité d'application et précision. Ces avancées contribuent de manière importante à l'analyse des données fonctionnelles en fournissant des outils performants pour l'estimation des opérateurs de régression. Nous nous sommes ensuite concentrés sur l'estimation de la fonction de répartition conditionnelle dans des modèles de régression fonctionnelle, en nous appuyant également sur la méthode du noyau. Nos résultats, validés par des simulations, ont confirmé la convergence presque complète des estimateurs et souligné leur pertinence dans des applications concrètes. Bien que l'approche par balayage pour la sélection des paramètres ait donné de bons résultats, nous avons également suggéré que des techniques plus avancées, comme la validation croisée, pourraient encore améliorer la précision des estimations, ouvrant ainsi des perspectives pour des recherches futures. L'un des apports majeurs de cette thèse réside dans la reformulation d'un estimateur pour la fonction de répartition conditionnelle, en intégrant un noyau qui joue le rôle de fonction de répartition pour la variable réponse Y. Cette reformulation a rendu l'estimation plus flexible et mieux adaptée à différents contextes pratiques. Les simulations ont montré l'efficacité de cette approche, non seulement par ses bonnes propriétés de convergence, mais aussi par sa capacité à fournir des estimations précises dans des cas réels. Cependant, la gestion des paramètres de lissage dans cette méthode a posé des défis importants, en particulier pour leur sélection optimale. Nous avons proposé la validation croisée comme méthode d'optimisation des paramètres, afin d'améliorer la robustesse et la précision des estimations dans des applications pratiques. Dans l'ensemble, ce travail a apporté des contributions importantes à l'étude des techniques d'estimation en régression fonctionnelle, en particulier dans les modèles semi-paramétriques et non-paramétriques. Les méthodes proposées, comme la reformulation de l'estimateur de la fonction de répartition conditionnelle, fournissent des outils fiables pour l'analyse des données fonctionnelles tout en ouvrant la voie à de futures améliorations méthodologiques. Les résultats obtenus offrent également des perspectives pour de nouvelles recherches et applications dans des domaines variés, comme la finance, la biostatistique, ou les sciences de l'environnement, où les modèles fonctionnels sont particulièrement utiles.

Bibliographie

[1] Khadidja Abdelhak, Abderrahmane Belguerna et Zeyneb Laala. « (R1483) Local Linear Estimator of the Conditional Hazard Function for Index Model in Case of Missing at Random Data ». In : *Applications and Applied Mathematics : An International Journal (AAM)* 17.1 (2022), p. 3.

[2] Rachid Agoune et Fateh Merahi. « ESTIMATE A KERNEL REGRESSION FOR FUNCTIONAL INDEPENDENT DATA IN SEMI-PARAMETRIC MODELS THROUGH THE APPLICATION OF THE LEAVE-ONE-OUT TECHNIQUE FOR BANDWIDTH SELECTION. » In : *Journal of Applied Probability & Statistics* 19.2 (2024).

[3] Rachid Agoune et Fateh Merahi. « Estimation in nonparametric functional-on regression models with real responses ». In : *Utilitas Mathematica* 120 (2023), p. 1954-1962.

[4] Ahmed Ait-Saïdi et al. « Cross-validated estimations in the single-functional index model ». In : *Statistics* 42.6 (2008), p. 475-494.

[5] Said Attaoui, Ali Laksaci et Elias Ould Said. « A note on the conditional density estimate in the single functional index model ». In : *Statistics & probability letters* 81.1 (2011), p. 45-53.

[6] Philippe C Besse, Hervé Cardot et David B Stephenson. « Autoregressive forecasting of some functional climatic variations ». In : *Scandinavian Journal of Statistics* 27.4 (2000), p. 673-687.

[7] U Bissinger, F Schimek et G Lenz. « Postoperative residual paralysis and respiratory status : a comparative study of pancuronium and vecuronium ». In : *Physiological research* 49.4 (2000), p. 455-462.

[8] Claus Borggaard et Hans Henrik Thodberg. « Optimal minimal neural interpretation of spectra ». In : *Analytical chemistry* 64.5 (1992), p. 545-551.

[9] Denis Bosq. *Linear processes in function spaces : theory and applications*. T. 149. Springer Science & Business Media, 2000.

[10] Denis Bosq. « Modelization, nonparametric estimation and prediction for continuous time processes ». In : *Nonparametric functional estimation and related topics* (1991), p. 509-529.

[11] Denis Bosq. « Parametric rates of nonparametric estimators and predictors for continuous time processes ». In : *The Annals of Statistics* 25.3 (1997), p. 982-1000.

[12] Denis Bosq et Michel Delecroix. « Nonparametric prediction of a Hilbert space valued random variable ». In : *Stochastic processes and their applications* 19.2 (1985), p. 271-280.

[13] Denis Bosq et Jean-Pierre Lecoutre. « Théorie de l'estimation fonctionnelle. Collection Economie et Statistiques Avancées ». In : *Economica, Paris* (1987).

[14] Salim Bouzebda, Ali Laksaci et Mustapha Mohammedi. « Single index regression model for functional quasi-associated times series data ». In : *REVSTAT-Statistical Journal* 20.5 (2022), p. 605-631.

[15] Hervé Cardot, Frédéric Ferraty et Pascal Sarda. « Functional linear model ». In : *Statistics & Probability Letters* 45.1 (1999), p. 11-22.

[16] Hervé Cardot et al. « Testing hypotheses in the functional linear model ». In : *Scandinavian Journal of Statistics* 30.1 (2003), p. 241-255.

[17] Gérard Collomb. « From non parametric regression to non parametric prediction : Survey of the mean square error and original results on the predictogram ». In : *Specifying Statistical Models : From Parametric to Non-Parametric, Using Bayesian or Non-Bayesian Approaches*. Springer. 1983, p. 182-204.

[18] Gérard Collomb. « Propriétés de convergence presque complète du prédicteur à noyau ». In : *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 66 (1984), p. 441-460.

[19] Bruce D Cuevas et al. « Tyrosine phosphorylation of p85 relieves its inhibitory activity on phosphatidylinositol 3-kinase ». In : *Journal of Biological Chemistry* 276.29 (2001), p. 27455-27461.

[20] Julien Damon et Serge Guillas. « The inclusion of exogenous variables in functional autoregressive ozone forecasting ». In : *Environmetrics* 13.7 (2002), p. 759-774.

[21] Michel Delecroix, Wolfgang Härdle et Marian Hristache. « Efficient estimation in conditional single-index regression ». In : *Journal of Multivariate Analysis* 86.2 (2003), p. 213-226.

[22] Laurent Delsol. « Advances on asymptotic normality in non-parametric functional time series analysis ». In : *Statistics* 43.1 (2009), p. 13-33.

[23] Frédéric Ferraty, Aldo Goia et Philippe Vieu. « Functional nonparametric model for time series : a fractal approach for dimension reduction ». In : *Test* 11 (2002), p. 317-344.

[24] Frédéric Ferraty et Philippe Vieu. *Nonparametric functional data analysis : theory and practice*. Springer Science & Business Media, 2006.

[25] Frédéric Ferraty et al. « Rate of uniform consistency for nonparametric estimates with functional variables ». In : *Journal of Statistical planning and inference* 140.2 (2010), p. 335-352.

[26] LLdiko E Frank et Jerome H Friedman. « A statistical view of some chemometrics regression tools ». In : *Technometrics* 35.2 (1993), p. 109-135.

[27] D Kh Fuk et Sergey V Nagaev. « Probability inequalities for sums of independent random variables ». In : *Theory of Probability & Its Applications* 16.4 (1971), p. 643-660.

[28] Ali Gannoun, Jérôme Saracco et Keming Yu. « Nonparametric prediction by conditional median and quantiles ». In : *Journal of statistical Planning and inference* 117.2 (2003), p. 207-223.

[29] Jiti Gao et Irene Gijbels. « Bandwidth selection in nonparametric kernel testing ». In : *Journal of the American Statistical Association* 103.484 (2008), p. 1584-1594.

[30] Shiferaw Gurmu, Paul Rilstone et Steven Stern. « Semiparametric estimation of count regression models ». In : *Journal of Econometrics* 88.1 (1999), p. 123-150.

[31] Wolfgang Hardle, Peter Hall et Hidehiko Ichimura. « Optimal smoothing in single-index models ». In : *The annals of Statistics* 21.1 (1993), p. 157-178.

[32] Wolfgang Härdle. *Applied nonparametric regression*. 19. Cambridge university press, 1990.

[33] Trevor Hastie et Colin Mallows. « [A statistical view of some chemometrics regression tools] : Discussion ». In : *Technometrics* 35.2 (1993), p. 140-143.

[34] Pao-Lu Hsu et Herbert Robbins. « Complete convergence and the law of large numbers ». In : *Proceedings of the national academy of sciences* 33.2 (1947), p. 25-31.

[35] T-C Hu, D Szynal et AI Volodin. « A note on complete convergence for arrays ». In : *Statistics & Probability Letters* 38.1 (1998), p. 27-31.

[36] Tien-Chung Hu et Andrei Volodin. « Addendum to "A note on complete convergence for arrays", Statist. Probab. Lett. 38 (1)(1998) 27–31 ». In : *Statistics & Probability Letters* 47.2 (2000), p. 209-211.

[37] R Leira et al. « Headache as a surrogate marker of the molecular mechanisms implicated in progressing stroke ». In : *Cephalalgia* 22.4 (2002), p. 303-308.

[38] Sue E Leurgans, Rana A Moyeed et Bernard W Silverman. « Canonical correlation analysis when the data are curves ». In : *Journal of the Royal Statistical Society Series B : Statistical Methodology* 55.3 (1993), p. 725-740.

[39] Yehua Li et Tailen Hsing. « Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data ». In : (2010).

[40] Nengxiang Ling, Zhihuan Li et Wenzhi Yang. « Conditional density estimation in the single functional index model for α-mixing functional data ». In : *Communications in Statistics-Theory and Methods* 43.3 (2014), p. 441-454.

[41] Wenceslao González Manteiga et Philippe Vieu. « Statistics for functional data ». In : *Computational Statistics & Data Analysis* 51.10 (2007), p. 4788-4792.

[42] Harald Martens et Tormod Naes. *Multivariate calibration*. John Wiley & Sons, 1992.

[43] William F Massy. « Principal components regression in exploratory statistical research ». In : *Journal of the American Statistical Association* 60.309 (1965), p. 234-256.

[44] Sergey Victorovich Nagaev. « Some refinements of probabilistic andmoment inequalities ». In : *Theory of Probability & Its Applications* 42.4 (1998), p. 707-713.

[45] Alex Preda. « The turn to things : Arguments for a sociological theory of things ». In : *The sociological quarterly* 40.2 (1999), p. 347-366.

[46] James O Ramsay et Bernard W Silverman. *Applied functional data analysis : methods and case studies*. T. 77. Springer, 2002.

[47] JO Ramsay et al. « Differential operators in functional data analysis ». In : *Functional Data Analysis* (1997), p. 217-238.

[48] George G Roussas. « Exact rates of almost sure convergence of a recursive kernel estimate of a probability densiy function : Application to regression and hazard rate estimation ». In : *Journal of Nonparametric Statistics* 1.3 (1992), p. 171-195.

[49] Ahmed Ait Saidi et Mecheri Kheira. « The conditional cumulative distribution function in single functional index model ». In : *Communications in Statistics-Theory and Methods* 45.16 (2016), p. 4896-4911.

[50] Leonie Selk et Jan Gertheiss. « Nonparametric regression and classification with functional, categorical, and mixed covariates ». In : *Advances in Data Analysis and Classification* (2022), p. 1-25.

[51] María Serón-Ferré et al. « The development of circadian rhythms in the fetus and neonate ». In : *Seminars in perinatology*. T. 25. 6. Elsevier. 2001, p. 363-370.

[52] James Victor Uspensky. « Introduction to mathematical probability ». In : (1937).

[53] Herman Wold. « Soft modelling by latent variables : the non-linear iterative partial least squares (NIPALS) approach ». In : *Journal of Applied Probability* 12.S1 (1975), p. 117-142.