Modèles linéaires fonctionnels avec des données partiellement observées

Chayma Daayeb

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE L'UNIVERSITÉ DE MONTPELLIER

En Biostatistique
École doctorale Information Structures Systèmes (I2S)
Unité de recherche Institut Montpelliérain Alexander Grothendieck (IMAG)
En Partenariat international
En Mathématiques Appliquées
avec École doctorale Sciences et Techniques de l'Ingénieur (STI)
Université de Tunis el Manar, École Nationale d'Ingénieurs de Tunis, Tunisie

MODÈLES LINÉAIRES FONCTIONNELS AVEC DES DONNÉES PARTIELLEMENT OBSERVÉES

Présentée par Chayma DAAYEB
Le 01 décembre 2022

Sous la direction de Ali GANNOUN, Mohamed MNIF,
Christophe CRAMBES et Yousri HENCHIRI

Devant le jury composé de

Mme. Sophie DABO NIANG, Professeure, Université de Lille	Rapportrice
M. Ali GANNOUN, Professeur, Université de Montpellier	Directeur
Mme. Amel BEN ABDA, Professeure, École National d'Ingénieur de Tunis	Examinatrice
M. Christophe CRAMBES, Maître de conférences habilité, Université de Montpellier	Co-directeur
M. Afif MASMOUDI, Professeur, Université de Sfax	Rapporteur
M. André MAS, Professeur, Université de Montpellier	Président
M. Mohamed MNIF, Professeur, École National d'Ingénieur de Tunis	Directeur
M. Yousri HENCHIRI, Maître assistant, Institut Supérieur des Arts du Multimédia de la Manouba	Co-directeur

UNIVERSITÉ DE MONTPELLIER

UNIVERSITE DE TUNIS EL MANAR

MÉMOIRE DE THÈSE DE DOCTORAT

MODÈLES LINEAIRES FONCTIONNELS AVEC DES DONNÉES PARTIELLEMENT OBSERVÉES

RÉDIGÉ PAR

CHAYMA DAAYEB

2022

RÉSUMÉ

Le traitement des données manquantes est un problème important dans le processus d'observation ou d'enregistrement des données. L'objectif de ce travail est l'étude de modèles fonctionnels de régression linéaire lorsque les variables sont partiellement observées.

La première partie de la thèse concerne la prédiction dans un modèle linéaire fonctionnel avec une variable explicative fonctionnelle partiellement observée et une réponse à valeurs réelles contenant des valeurs manquantes. Nous utilisons un opérateur de reconstruction qui vise à récupérer les parties non observées des courbes explicatives, puis nous nous intéressons à la méthode d'imputation simple par régression (sous l'hypothèse MAR; Missing At Random) des données manquantes sur la variable réponse, en utilisant la régression fonctionnelle sur composantes principales pour estimer le coefficient fonctionnel du modèle. Nous étudions le comportement asymptotique de l'erreur quadratique moyenne de prédiction. Le comportement pratique de la méthode est également étudié sur des données simulées et un jeu de données réelles.

Dans la deuxième partie, nous proposons une autre méthode pour compléter les données manquantes de la variable réponse réelle, une fois que les courbes de la covariable fonctionnelle sont reconstruites. Il s'agit de la méthode d'imputation multiple qui consiste à ajouter un résidu aléatoire à la valeur imputée par imputation simple. Une étude asymptotique donne une vitesse de convergence de l'erreur quadratique moyenne de prédiction. Des résultats numériques sur données simulées et réelles sont proposés.

La dernière partie concerne la prédiction dans un modèle linéaire fonctionnel dont la variable explicative et la variable réponse sont toutes deux fonctionnelles et partiellement observées. Les parties manquantes de la covariable sont reconstruites et les parties de courbes non observées de la réponse sont complétées par deux méthodes: imputation (sous l'hypothèse MAR) et reconstruction. Une fois l'ensemble de données reconstruit, nous calculons l'erreur quadratique moyenne de prédiction pour une nouvelle observation de la covariable. Les deux méthodes sont comparées d'un point de vue théorique et pratique.

Nous montrons dans cette thèse que la vitesse de convergence de l'erreur de prédiction est subordonnée à la vitesse de convergence de l'erreur de reconstruction de la

courbe explicative, la vitesse de convergence de l'erreur d'imputation de la réponse est secondaire.

REMERCIEMENTS

Je voudrais tout d'abord remercier mes Directeurs de thèse Messieurs les Professeurs Mohamed Mnif et Ali Gannoun, ainsi que mes co-encadrants Dr Christophe Crambes et Dr Yousri Henchiri. J'ai appris à votre contact à persévérer et ne jamais baisser les bras. Messieurs, travailler avec vous fut un réel plaisir.

Professeur Ali Gannoun, vous avez beaucoup compté pour moi durant ces 4 dernières années. Vous étiez à la fois un guide et un modèle. Je vous remercie pour nos nombreuses discussions et pour vos multiple conseils toujours très avisés. Merci aussi d'avoir su trouver les mots pour me remonter le moral aux moments difficiles de ma thèse. J'espère que nous allons garder le contact, tant sur le plan professionnel que personnel.

Merci à vous Dr Christophe Crambes, pour avoir su diriger mes travaux avec beaucoup de compétences et d'efficacité. Merci aussi d'avoir toujours été à mon écoute et d'avoir passé beaucoup de temps avec moi. Votre confiance, votre rigueur et votre disponibilité ont fait que cette thèse aboutisse. Ce fût un immense honneur et un plaisir de mener cette thèse à vos côtés.

Merci à Vous Dr Yousri Henchiri, de m'avoir proposé mon sujet de thèse et de m'avoir très bien accompagnée dans ce travail aussi bien en Tunisie qu'en France. Merci aussi de m'avoir ouvert les portes de l'Université Montpellier. Merci aussi pour toute votre aide, vos encouragements, vos exigences et votre confiance même à distance.

Mes remerciements les plus chaleureux vont aux Professeurs Mme Sophie Dabo-Niang et Monsieur Afif Masmoudi mes deux rapporteurs. Vous avez pris de votre temps pour évaluer mon travail et l'expertiser et je vous suis très reconnaissante pour toutes les remarques et suggestions pour améliorer mon travail.

Qu'il me soit permis de remercier chaudement Mme la Professeure Amel Ben Abda et Monsieur le Professeur André Mas qui ont accepté de siéger dans mon Jury de thèse. Votre présence malgré vos multiples activités est un honneur pour moi. Je vous suis très reconnaissante.

Je remercie aussi toute l'équipe de l'Institut Montpelliérain Alexander Grothendieck (IMAG) , ainsi que celle du Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur (ENIT, Tunisie). Mon passage dans ces deux laboratoires a été d'une très grande richesse aussi bien humainement que scientifiquement.

Enfin, je voudrais remercier les membres de ma famille pour leur patience, leur soutien inconditionnel. Merci d'avoir cru en moi tout le temps.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER 1

INTRODUCTION

1.1) Introduction générale sur les données fonctionnelles.

1.1.1) Analyse de données fonctionnelles.

Une variable aléatoire est dite fonctionnelle si ses valeurs sont dans un espace de dimension infinie, par exemple, l'espace $\mathbb{L}^2(\mathcal{I})$ des fonctions de carré intégrable sur un intervalle $\mathcal{I} \subseteq \mathbb{R}$. L'observation d'une variable fonctionnelle sur un ensemble $\mathcal{I}$ est appelée une donnée fonctionnelle, c'est à dire, une courbe ($\mathcal{I} \subseteq \mathbb{R}$), une image ($\mathcal{I} \subseteq \mathbb{R}^2$), etc...

L'Analyse des Données Fonctionnelles (ADF) consiste à traiter ce type de données comme une séquence de réalisations $(X_i)_{1 \leqslant i \leqslant n}$ d'une variable aléatoire à valeurs dans un espace de fonctions et observées à certains instant, où n est la taille de l'échantillon.

Les chercheurs ont commencé à s'intéresser à ce type de données, probablement en physique ou météorologie, par le barographe, inventé par l'Anglais Moreland en 1670 qui enregistre la pression atmosphérique dans la durée (voir Robert (1949)). Pearson (1901) était parmi les premiers statisticiens qui ont abordé l'analyse de ces données, puis Hötelling (1933). Karhunen (1947) et Loève (1948) ont développé l'analyse en composantes principales (ACP). L'idée principale de cette méthode statistique est de réduire le nombre de variables du modèle en un plus petit nombre de composantes afin de faciliter la visualisation des données. D'autre part, elle réduit également le risque de sur-ajustement du modèle en éliminant les variables à forte corrélation. Pour cette raison, l'ACP peut être considérée comme un outil de compression et de réduction des données.

L'ACP est une transformation linéaire qui transforme les données dans un nouveau système de coordonnées de telle sorte que le nouvel ensemble de variables, les composantes principales qui sont des fonctions linéaires des variables d'origine, ne sont pas corrélées et que la plus grande variance par toute projection des données vient à se situer

sur la première coordonnée, la deuxième plus grande variance sur la deuxième coordonnée, et ainsi de suite. Elle est également appelée "transformation discrète de Karhunen-Loève" en traitement du signal, "décomposition orthogonale propre" en génie mécanique et "décomposition des valeurs propres" en algèbre linéaire.

Soit $\{X(t), t \in \mathcal{I}\}$, une donnée fonctionnelle de carré intégrable, c'est à dire, $\mathbb{E}[\|X\|^2] < \infty$ où $\|.\|$ désigne la norme usuelle de $\mathbb{L}^2(\mathcal{I})$ associée au produit scalaire $\langle ., . \rangle$, défini par $\langle f, g \rangle = \int_{\mathcal{I}} f(t)g(t)\mathrm{d}t$ pour toutes fonctions f et g de $\mathbb{L}^2(\mathcal{I})$. Il existe une séquence de variables aléatoires réelles indépendantes $\{\xi_i, i \in \mathbb{N}\}$ et une série de fonctions $\{\phi_i(t)\}_{i\in\mathbb{N}}$ orthogonales dans $\mathbb{L}^2(\mathcal{I})$ telles que le processus X peut être écrit comme

$$X(t) = \mathbb{E}(X(t)) + \sum_{i=1}^{\infty} \xi_i \phi_i(t) \quad \text{pour tout } t \in \mathcal{I}.$$

L'analyse en composantes principales permet d'analyser la variabilité des données en étudiant la structure de leur matrice de covariance. En pratique, cela est réalisé en calculant la matrice de covariance pour l'ensemble des données complètes. Ensuite, les vecteurs propres et les valeurs propres de la matrice de covariance sont calculés et triés en fonction des valeurs propres décroissantes.

Rao C (1958) et Tucker L (1958) ont présenté la première approche de l'ACP qui associait les méthodes d'analyse factorielle aux modèles de courbe de croissance. Deville (1974) a généralisé l'analyse en composantes principales aux processus stochastiques. Puis, Dauxois and Pousse (1976); Dauxois et al. (1982) ont proposé un cadre mathématique et étudié la cohérence et les propriétés asymptotiques de l'analyse en composantes principales d'une fonction vectorielle aléatoire.

En approchant les fonctions aléatoires de dimension infinie par un nombre fini de vecteurs de scores aléatoires, l'analyse en composantes principales fonctionnelles (ACPF) apparaît comme une technique de réduction de dimension tout comme dans le cas multivarié et réduit la complexité des données.

1.1.2) Applications.

L'analyse de ce type de données s'est développé rapidement et pris un grand intérêt dans de nombreuses applications, comme par exemple, en médecine, économie, commerce, finance ou encore en physique. En effet, les progrès techniques récents permettent d'observer, de stocker et de traiter de grandes quantités de telles données.

La figure 1.1 représente des données issues du CEA et simulant l'évolution de la température dans un réacteur nucléaire lors d'un accident de perte de réfrigérant primaire. Ce type d'accident résulte d'un défaut de refroidissement du réacteur de type piscine dont le cœur comporte des crayons d'oxyde d'uranium UO_2. Les courbes présentent le comportement thermique de 1816 crayons de 0 à 5000 secondes.

L'analyse de données fonctionnelles ne se limite pas à étudier des quantités évoluant au cours du temps, que l'on les appelle des données longitudinales. Par exemple, en

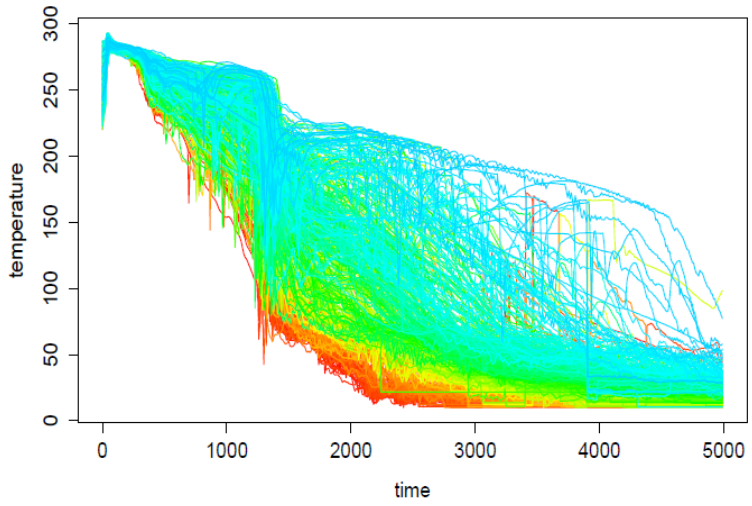

Figure 1.1: Simulations de l'évolution de la température (°C) en fonction du temps (s) dans un réacteur nucléaire lors d'un accident de perte de réfrigérant primaire.

biologie, en chimie et en agro-alimentaire, les données spectrométriques aident à étudier l'évolution de l'absorbance de la lumière en fonction de la longueur d'onde (voir Ferraty and Vieu (2002, 2006); Ferraty et al. (2007)).

Le champ de recherche d'analyse de données fonctionnelles a trouvé un réel écho auprès de la communauté des statisticiens, et a donc fait l'objet de nombreux travaux, tant théoriques que pratiques.
Ramsay and Silverman (2002) donnent une liste importante d'exemples qui montrent le large potentiel applicatif des différentes méthodes liées à l'ADF. Bosq (2000) a étudié la modélisation des variables aléatoires fonctionnelles dépendantes et Ferraty and Vieu (2006) ont développé des modèles non paramétriques pour les données fonctionnelles contenant une revue de contributions les récentes sur ce sujet.

L'étude de données fonctionnelles s'étend aussi à un cadre multivarié. Avec la croissance du marché des objets connectés, de plus en plus de données sont collectées pour un même individu. Par exemple dans le sport, les athlètes portent des appareils qui collectent des données pendant leur entraînement pour améliorer leurs performances et suivre leurs constantes physiques afin de prévenir les blessures, tels que le rythme cardiaque du joueur et l'altitude de son parcours. Ces données sont des données multivariées, présentés par plusieurs courbes qu'on peut écrire : $X = X(t)_{t\in\mathcal{I}}$ avec $X(t) = (X^1(t), ..., X^p(t))' \in \mathbb{R}^p$, $p \geqslant 2$. La figure 1.2 est un graphique de fréquence cardiaque d'un coureur au cours d'une course à pied à la mesure de deux appareils: la ceinture V800 avec OH1 (en bleu) et Fenix 5S avec ceinture Polar H10 (en rouge).

Figure 1.2: Un graphique de fréquence cardiaque (bpm) en fonction du trajet (m) réalisé avec deux appareils de mesure.

1.1.3) Modèles linéaires fonctionnels.

La statistique est l'étude d'un phénomène par la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. En général, les chercheurs s'intéressent souvent à la manière dont une variable d'intérêt Y peut être liée à une variable explicative X. Cette relation peut être écrite selon le modèle suivant

$$Y = r(X) + \varepsilon,$$

où ε est une variable aléatoire réelle centrée représentant *l'erreur du modèle*, de variance finie $\mathbb{E}(\varepsilon^2) = \sigma_\epsilon^2$, et indépendante de X, $\mathbb{E}(\varepsilon \mid X) = 0$. Nous considérons un échantillon $(X_i, Y_i)_{i=1}^n$ indépendant et identiquement distribué de même loi que (X, Y) pour estimer la fonction de régression $r(.)$ ou prédire une nouvelle valeur de réponse.

Un des modèles les plus populaires en analyse de données fonctionnelles est le modèle de régression linéaire fonctionnel, qui établit une relation linéaire de dépendance entre une variable réponse réelle Y et une variable aléatoire fonctionnelle $X = (X(t), t \in \mathcal{I})$. La variable explicative X est par exemple à valeurs dans l'espace de Hilbert $\mathrm{H} := \mathbb{L}^2(\mathcal{I})$ des fonctions de carré intégrable sur l'intervalle $\mathcal{I}$. Ce modèle est défini par

$$Y = \theta_0 + \int_{\mathcal{I}} \theta(t)\Big[X(t) - \mathbb{E}[X(t)]\Big]\mathrm{d}t + \varepsilon, \tag{1.1.1}$$

où $\theta_0 \in \mathbb{R}$, et la fonction de régression $\theta \in H$ est une fonction de carré intégrable définie sur $\mathcal{I}$ modélisant la relation entre la variable réponse Y et la variable explicative X. Le modèle (1.1.1) peut s'écrire sous la forme

$$Y = \theta_0 + \Theta X + \varepsilon,$$

où $\Theta : \mathrm{H} \to \mathbb{R}$ est l'opérateur linéaire continu défini par $\Theta u = \langle \theta, u \rangle$ pour toute fonction $u \in \mathrm{H}$.

Le modèle linéaire fonctionnel a été étudié par de nombreux auteurs. Une référence majeure est la monographie de Ramsay and Silverman (2005) qui donne un aperçu de la philosophie et des modèles de base impliquant des données fonctionnelles. Ce type de modèle est également étudié par Bosq (2000) dans un cadre plus général où la variable X est à valeurs dans un espace de Banach.

Le précèdent modèle peut se généraliser au cas où la réponse et la variable explicative sont des variables fonctionnelles, c'est-à-dire $X \in \mathbb{L}^2(\mathcal{I})$ et $Y \in \mathbb{L}^2(\mathcal{J})$, où $\mathcal{I}$ et $\mathcal{J}$ sont des intervalles de $\mathbb{R}$. Ce modèle est défini par

$$Y(t) = \theta_0(t) + \int_I \theta(s,t)\Big[X(s) - \mathbb{E}[X(s)]\Big]\mathrm{d}s + \varepsilon(t),$$

où la fonction de régression $(s,t) \mapsto \theta(s,t)$ appartient à $\mathbb{L}^2(\mathcal{I} \times \mathcal{J})$, et $\|\theta\|^2 = \int\int_{\mathcal{I}\times\mathcal{J}} \theta^2(s,t)\mathrm{d}s\mathrm{d}t < \infty$.

L'estimation de la fonction de régression fait partie des nombreuses questions étudiées dans les modèles fonctionnels de régression. Il existe deux approches principales, une approche paramétrique fonctionnelle, sous l'hypothèse que l'opérateur de régression r soit linéaire continu et une approche non-paramétrique fonctionnelle (voir Ferraty and Vieu (2006)). Cardot et al. (1999) ont étudié un modèle linéaire fonctionnel de régression dans lequel les variables explicatives sont des points d'échantillonnage d'un processus en temps continu et ils ont proposé un estimateur de régression au moyen d'une analyse en composantes principales fonctionnelle analogue à celle introduite par Bosq (1991). Müller H and Stadtmüller (2005) ont proposé un modèle de régression linéaire fonctionnel généralisé pour une situation de régression où la variable réponse est un scalaire et le prédicteur est une fonction aléatoire.

Des travaux théoriques sur le modèle linéaire fonctionnel Cai and Hall (2006); Hall and Horowitz (2007) ont permis d'obtenir des resultats sur des vitesses de convergence optimales.

La monographie Horváth and Kokoszka (2012) présente une introduction générale au cadre mathématique de l'ADF, puis il se concentre sur les méthodes inférentielles dans le cadre de données dans un espace de Hilbert, par exemple, les tests d'hypothèses. Une attention particulière est accordée aux méthodes basées sur les composantes principales fonctionnelles et les tests de spécification de modèle, y compris les tests de point de rupture. Par ailleurs, Crambes and Mas (2013) ont étudié les comportements asymptotiques de prédiction pour des réponses fonctionnelles en régression linéaire fonctionnelle.

Récemment, Kokoszka and Reimherr (2018) synthétisent les concepts les plus fondamentaux de l'ADF, notamment les projections sur une base, les fonctions de moyenne et

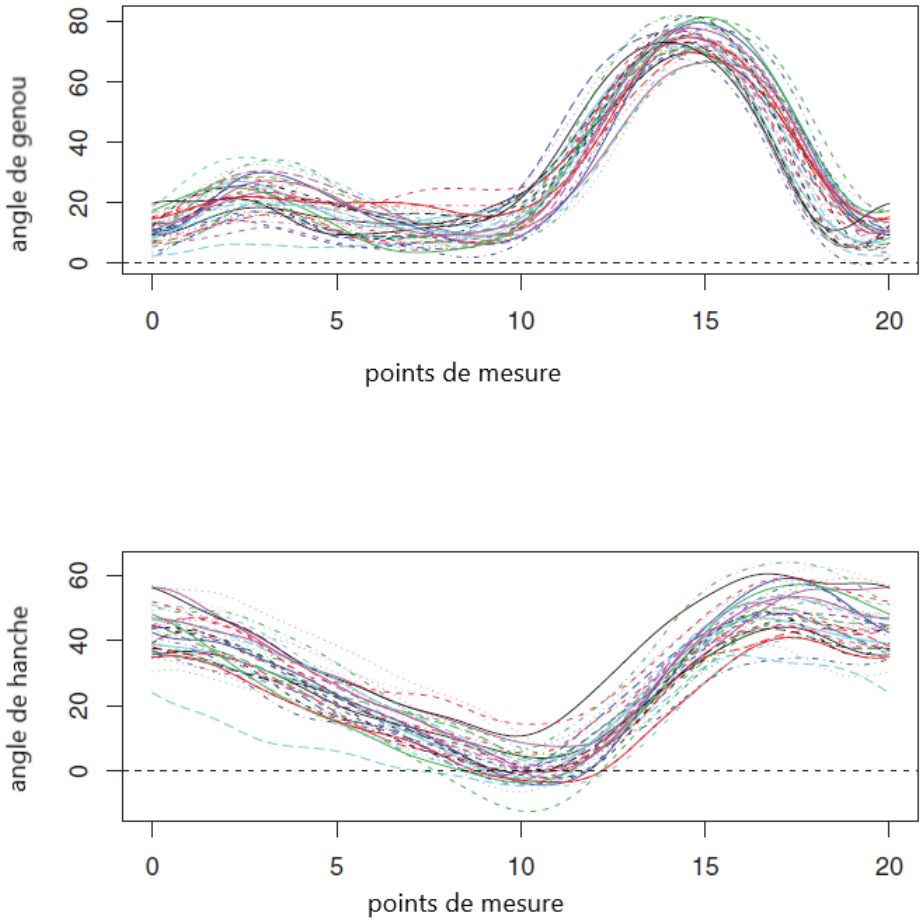

Figure 1.3: Courbes des angles de genou et de hanche en degrés sur un cycle de mouvement en 20 points pour 39 enfants.

de covariance, les composantes principales fonctionnelles et le lissage pénalisé, dans des modèles de régression fonctionnels. De plus, Aneiros et al. (2020) présentent les données de grande dimension et autres données complexes, abordant les aspects méthodologiques et informatiques, ainsi que les applications sur des données réelles.

La régression fonctionnelle trouve un intérêt dans de nombreuses applications. Un exemple classique dans le domaine de la pédiatrie vient de Ramsay et al. (2009) (figure 1.3) qui montre les courbes des angles (en degrés) des genoux et des hanches sur un cycle de mouvement en 20 points pour 39 enfants. Une question intéressante est de savoir dans quelle mesure l'angle de la hanche peut expliquer l'angle du genou.

Toute cette littérature se concentre sur des données fonctionnelles complètement observées. Dans la suite, nous allons nous intéresser à des situations où toutes les données ne sont pas disponibles.

1.2) Données manquantes.

1.2.1) Historique.

Toute collection de données conduit fréquemment à des données manquantes, à commencer par les plus anciennes opérations statistiques connues, les recensements dans les anciens empires. Le premier recensement connu a été effectué par les Babyloniens en 3800 avant J.C., il y a donc près de 6000 ans. Selon les archives, il était effectué tous les six ou sept ans et recensait le nombre de personnes et le bétail, ainsi que la quantité de beurre, de lait, de laine et de légumes, voir Kuhrt (1995). Mais la prise de conscience des problèmes soulevés par des données manquantes n'est que très récente.

Galton (1888) a été parmi les premiers à avoir étudié des situations avec des données manquantes. Il a rencontré des cas de mesures incomplètes dans ses travaux anthropométriques. Son étude était basée sur les données (longueur de coudée, longueur de main, palme, paume, etc ...) de 350 hommes, mais Galton disait que le nombre exact de 350 n'est pas conservé tout au long de l'étude car une blessure à un membre ou à un autre a réduit le nombre d'individus de 1, 2 ou 3 dans différents cas. Ensuite, Galton (1898) a considéré une distribution tronquée, en l'occurrence une loi normale tronquée à droite, en analysant des données extraites du *Wallace's Year Book*. Ces données consistaient en des temps de parcours, en vue d'une qualification, de coureurs devant parcourir un mile en 2 minutes et 30 secondes au maximum. Aucune trace n'était gardée des temps des coureurs les plus lents, d'où éliminés, dont le nombre est resté inconnu. Donc il a estimé la moyenne et il a repéré les quartiles pour estimer la dispersion à l'aide de l'intervalle interquartile. En 1931, le statisticien britannique Fisher, R. A a repris ce problème en utilisant la méthode du maximum de vraisemblance. Cette méthode a été utilisée dans l'article de Wilks (1932) pour estimer une matrice de covariance quand il y a des données manquantes. Il s'agissait d'un situation avec deux variables seulement. Après 20 ans, Lord (1955) l'a généralisé pour trois variables.

Dans les années de choléra, en étudiant la distribution du nombre de foyers où 0, 1, 2, 3, 4 cas de choléra ont été observés dans un village indien, McKendrick (1926), qui fut un des pionniers de l'épidémiologie mathématique, a constaté un nombre trop élevé de zéros pour une distribution de Poisson, alors que le véritable nombre d'observations de foyers touchés et le nombre des contaminés étaient inconnus.

Fisher (1934) a étudié le problème des albinos suivant : on ne peut pas distinguer les familles génétiquement capables d'avoir des enfants albinos mais qui n'en ont pas eu, et des familles incapables d'en avoir.

Pendant la deuxième guerre mondiale, en se basant sur les impacts de balle constatés au retour de mission, Abraham Wald a recommandé de blinder des bombardiers partout,

en particulier sur les moteurs. En effet, les avions qui étaient revenus étaient ceux qui n'avaient pas été touchés au moteur, ceux touchés au moteur étaient manquants. On peut trouver plus de détails dans les documents suivants : Mangel and Samaniego (1984), Wainer (2011) et Ellenberg (2014), dont Ellenberg (2018) est une traduction française.

Pour analyser les mesures des squelettes humains obtenus à partir du Jebel Moya au Soudan, Rao (1985) a utilisé les résultats d'une étude archéologique (Mukherjee and Rao (1955)) pourtant sur un échantillon de crânes. Certains de ces crânes étaient en bon état, c'est à dire, décrit par 4 variables (capacité, longueur, largeur et hauteur), alors que d'autres crânes étaient fracturés rendant certaines mesures impossibles.

"Missing data refers to a data value that should have been recorded but, for some reason, was not, Day (1999)."

C'est la première définition pour une valeur manquante, elle autorise des causes de données manquantes comme par exemple une panne d'un serveur lors d'enregistrement des données, qui entraine une perte partielle des données.

Maintenant, malgré la masse de données disponibles, qui augmente chaque jour et l'émergence du Big Data, les problématiques de données manquantes ou non observées restent très répandues dans les problèmes statistiques et nécessitent une approche particulière.

1.2.2) Mécanisme des données manquantes.

Le mécanisme engendrant les données manquantes est le processus qui produit les valeurs manquantes. En particulier, dans un modèle de régression, on peut par exemple considérer des mécanismes de données manquantes sur la variable réponse Y, liés d'une certaine façon à la variable explicative X et la réponse Y.

Supposons par exemple que la réponse Y contienne des données manquantes, que la covariable X soit complètement observée et considérons une variable aléatoire binaire $\delta^{[Y]}$ et un échantillon $(\delta_i^{[Y]})_{i=1,\dots,n}$ tel que $\delta_i^{[Y]} = 1$ si la valeur Y_i est observée et $\delta_i^{[Y]} = 0$ si la valeur Y_i est manquante, pour tout $i = 1, \dots, n$.

Le premier mécanisme est "Missing Completely At Random" (MCAR) où la probabilité qu'une donnée soit manquante est indépendante des données observées (voir Allison (2001) et Briggs et al. (2003)). Ainsi,

$$\mathbb{P}(\delta^{[Y]} = 1 \mid X, Y) = \mathbb{P}(\delta^{[Y]} = 1),$$

ce qui revient à dire que le fait qu'une donnée soit manquante ne dépend ni de la valeur de cette donnée ni des données observées. Dans un cadre non-paramétrique, Ferraty et al. (2013) ont considéré deux types d'estimation de la moyenne d'une réponse scalaire, basés sur un échantillon dans lequel une variable explicative est observée pour chaque sujet alors que les réponses sont manquantes.

	Y			
X	Complete	MCAR	MAR	MNAR
	Data for individual participants			
169	148	148	148	148
126	123	—	—	—
132	149	—	—	149
160	169	—	169	169
105	138	—	—	—
116	102	—	—	—
125	88	—	—	—
112	100	—	—	—
133	150	—	—	150
94	113	—	—	—
109	96	—	—	—
109	78	—	—	—
106	148	—	—	148
176	137	—	137	—
128	155	—	—	155
131	131	—	—	—
130	101	101	—	—
145	155	—	155	155
136	140	—	—	—
146	134	—	134	—
111	129	—	—	—
97	85	85	—	—
134	124	124	—	—
153	112	—	112	—
118	118	—	—	—
137	122	122	—	—
101	119	—	—	—
103	106	106	—	—
78	74	74	—	—
151	113	—	113	—

Figure 1.4: Mesures de la pression simulées (N = 30 participants) en janvier (X) et février (Y) avec des valeurs manquantes imposées par trois méthodes différentes (source Schafer and Graham (2002)).

Deuxièmement, on a le mécanisme "Missing At Random" (MAR) où les données non observées ne sont affectées que par les données complètes (voir Allison (2001)). Ce cas est une hypothèse plus faible que MCAR. Cette hypothèse est définie par :

$$\mathbb{P}(\delta^{[Y]} = 1 \mid X, Y) = \mathbb{P}(\delta^{[Y]} = 1 \mid X).$$

Dans le cadre MAR et MCAR, Chiou et al. (2014) ont étudié l'imputation des valeurs manquantes par une approche non paramétrique et la détection des valeurs aberrantes pour les données fonctionnelles de flux de trafic.

Enfin, le troisième mécanisme est "Missing Not At Random" (MNAR). C'est le mécanisme qui ne vérifie pas le mécanisme MAR. Du coup, la probabilité d'apparition des données manquantes dépend de la partie non-observée des données,

$$\mathbb{P}(\delta^{[Y]} = 1 \mid X, Y) = \mathbb{P}(\delta^{[Y]} = 1 \mid Y).$$

Bugni (2012) adapte un test de spécification pour les données fonctionnelles (des données économiques étudiées par Bugni et al. (2009)) avec la présence d'observations manquantes. Sa méthode est capable d'extraire les informations disponibles dans la partie observée des données tout en n'ayant pas de connaissance sur la nature de l'observation manquante.

Pour mieux expliquer la différence entre ces trois mécanismes, Schafer and Graham (2002) ont donné des exemples simples (voir aussi Little and Rubin (2020) et Graham (2012) pour les données multivariées). Prenons l'exemple d'analyse de sang de N participants enregistrées en janvier (X). Certains d'entre eux ont une deuxième lecture en février (Y), mais d'autres non. La figure 1.4 montre les mesures de la pression simulées pour 30 individus avec une moyenne est égale à 125 pour X et Y. Les deux premières colonnes du tableau montrent les données complètes pour X et Y. Les autres colonnes montrent les valeurs de Y qui restent après création des données manquantes par les trois critères. Dans la première méthode, les 7 mesures effectuées en février ont été aléatoirement parmi celles mesurées en janvier ; ce mécanisme est le MCAR. Dans la deuxième méthode, ceux qui sont revenus en février l'ont fait parce que leurs mesures de janvier dépassaient 140 ($X > 140$), un niveau utilisé pour le diagnostic de l'hypertension ; il s'agit d'un MAR mais pas d'un MCAR. Dans la troisième méthode, les personnes enregistrées en février étaient ceux dont les mesures de février dépassaient 140 ($Y > 140$). Cela pourrait se produire, par exemple, si tous les individus sont revenus en février, mais le personnel a décidé d'enregistrer la valeur de février uniquement si elle se situe dans le cas hypertension. Ce troisième mécanisme est un exemple de MNAR. D'autres mécanismes MNAR sont possibles, par exemple, la mesure de février peut être enregistrée seulement si elle est substantiellement différente de celle de janvier.

1.2.3) Imputation.

La plupart des travaux et des recherches en statistique considèrent que les données à analyser ne présentent pas de valeurs manquantes. Souvent, la solution naïve qui est adoptée consiste à supprimer les individus qui présentent des valeurs manquantes.

We are surrounded by missing data. Problems created by missing data in statistical analysis have long been swept under the carpet, Van Buuren (2018).

Cependant, cette méthode de suppression des données manquantes risque de fausser complètement les résultats de l'étude statistique.

Missing data are unobserved values that would be meaningful for analysis if observed ; in other words, a missing value hides a meaningful value, Little and Rubin (2020)

Pour pallier ce problème, des méthodes d'imputation ont été développées surtout dans le cas des enquêtes et recensements où les données manquantes sont nombreuses. Hansen and Madow (1953) ont utilisé cette méthode pour corriger une enquête américaine en 1948 sur *Survey of retail shares*, tandis que, Rancourt (2001a,b) a expliqué le mot imputation et son utilisation dans le domaine statistique.

La méthode d'imputation consiste à compléter le jeu de données (i.e., à prédire des valeurs estimées pour les données non observées). De nombreuses méthodes d'imputation des valeurs manquantes ont été développées. Elles peuvent être divisés en deux branches : imputation simple et imputation multiple. Les applications pratiques

de cette méthode sont de plus en plus nombreuses, parmi les plus récentes, nous pouvons citer par exemple Van Buuren (2007) et He et al. (2022).

Le principe d'imputation simple est de remplacer les valeurs manquantes des variables de l'enquête par des valeurs plausibles. Une fois l'imputation effectuée, les analyses se déroulent avec les valeurs imputées (voir Haziza and Rao (2006); Crambes and Henchiri (2019)).

Parfois, il est préférable de considérer plusieurs imputations du même jeu de données. On parle alors de l'imputation multiple (IM) développée par Rubin (1987). La création de l'IM a été facilitée par la technologie informatique et de nouvelles méthodes de simulation bayésienne découvertes à la fin des années 1980 (voir Schafer (1997)). Cette méthode consiste, comme son nom l'indique, à imputer q fois les valeurs manquantes avec $q > 1$. Cela permet de créer plusieurs ensembles de données complètes, afin de combiner les résultats pour réduire l'erreur due à l'imputation.
Plusieurs statisticiens ont exploré ce domaine, par exemple, Joseph et al. (1998) et Joseph and Schafer (2003) pour l'imputation multiple dans les problèmes multivariés. Haziza (2009) a étudié les différents estimateurs pour l'imputation déterministe (en appliquant plusieurs fois la méthode d'imputation, la valeur imputée reste toujours la même) et aléatoire (la valeur imputée diffère à chaque application de la méthode). Par ailleurs, He et al. (2011) ont développé une approche d'imputation multiple fonctionnelle modélisant la réponse longitudinale manquante sous un modèle fonctionnel à effets mixtes. Ils ont étudié un algorithme d'échantillonnage de Gibbs (une méthode de Monte-Carlo par chaînes de Markov) pour estimer les paramètres du modèle et les imputations des valeurs manquantes.

De plus, Preda et al. (2010) ont adapté une méthodologie basée sur l'algorithme NIPALS (Nonlinear Iterative Partial Least Squares), qui fournit une méthode d'imputation des données manquantes qui ont affecté les covariables fonctionnelles.

1.2.4) Données fonctionnelles partiellement observées.

La situation de données fonctionnelles partiellement observées est souvent qualifiée de données fonctionnelles fragmentées, tronquées, incomplètes et sa pertinence pratique a déclenché une série de travaux de recherche portant sur différents aspects de ce problème. Cette situation a d'abord été considérée dans des travaux appliqués Liebl (2013, 2019); Liebl and Rameseder (2019) pour modéliser et prévoir des fonctions du prix et la demande d'électricité en Allemagne. En ce qui concerne les travaux théoriques, Delaigle and Hall (2013) proposent une procédure de « shift-and-connect » pour reconstruire et classer des données fonctionnelles fragmentées et Delaigle and Hall (2016) intègrent un modèle de chaîne de Markov.

L'étude de Gromenko et al. (2017) est également liée, proposant une méthode d'inférence pour des courbes incomplètes spatialement et temporellement corrélées. Goldberg et al. (2014) considèrent le cas des données longitudinales dans lesquels chaque observation consiste en une série de mesures dans le temps qui sont échantillonnées à partir d'une courbe sous-jacente, éventuellement avec du bruit. Par exemple, les courbes de

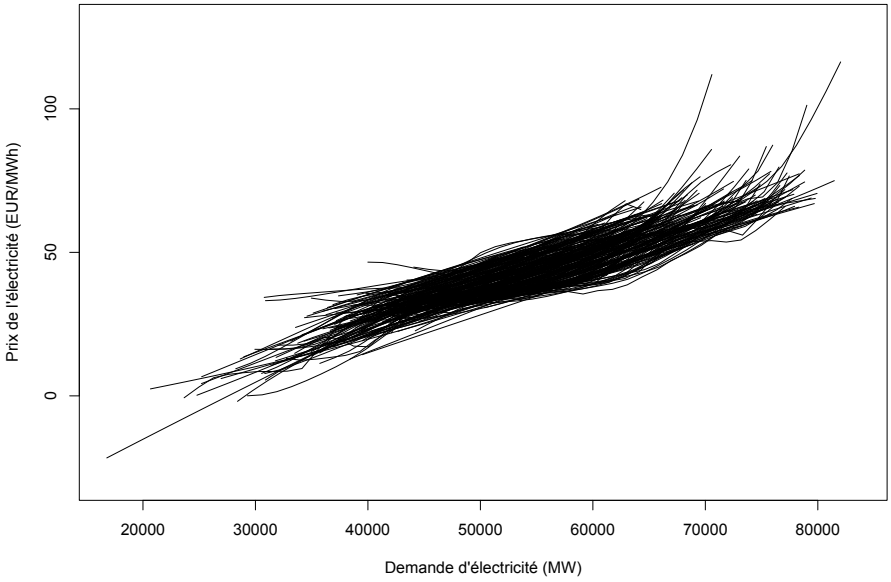

Figure 1.5: Courbes journalières des prix de l'électricité en fonction de la demande résiduelle.

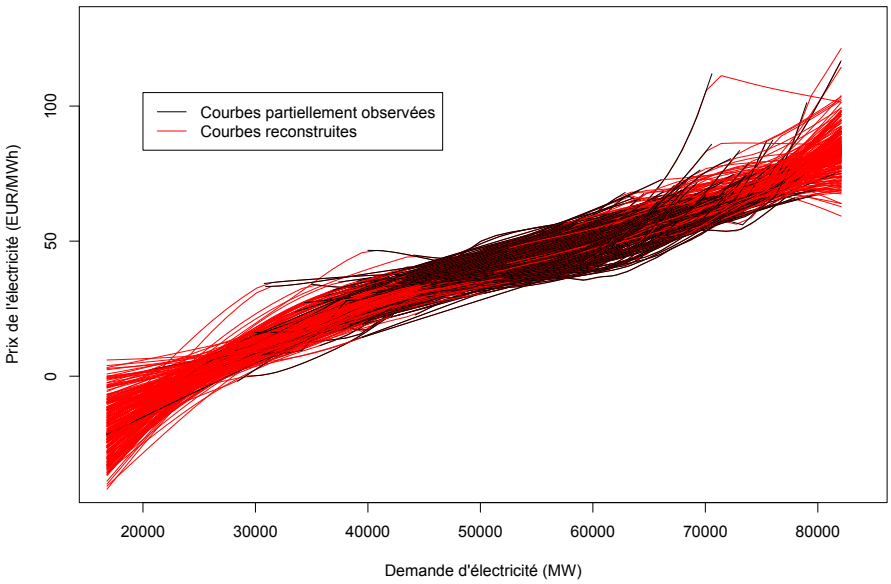

Figure 1.6: Courbes quotidiennes des prix de l'électricité reconstituées en fonction de la demande résiduelle.

croissance de différents individus et les taux d'arrivée d'appels vers un centre d'appels ou de patients vers une salle d'urgence au cours de différents jours. Ils supposent que les courbes ou les séries de mesures qui approximent ces courbes, ont été recueillies précédemment. Alors, ils cherchent à estimer la continuité d'une nouvelle courbe donnée au début, en utilisant le comportement des courbes précédemment collectées.

Kraus (2015) propose une méthode de reconstruction linéaire basée sur une régularisation de type ridge. Il a proposé une procédure de complétion fonctionnelle qui récupère la partie manquante en utilisant la partie observée de la courbe. Il a construit des intervalles de prédiction pour les scores principaux et des bandes pour les parties manquantes des trajectoires. Les problèmes de prédiction étant des problèmes inverses mal posés, il a utilisé une technique de régularisation pour obtenir une solution stable. Dans le même cadre, Kneip and Liebl (2020) étudient une méthode de la reconstruction par composantes principales qui s'est avérée asymptotiquement optimale.

Récemment, Kraus and Stefanucci (2020) donnent un résultat théorique plus fort que celui de Kraus (2015) avec la même méthode de reconstruction basée sur une régularisation ridge. Ce nouveau théorème montre que même si la solution optimale n'est pas un opérateur de régression (c'est-à-dire intégrable au sens Hilbert-Schmidt), on peut agir comme si c'était le cas, en utilisant la régularisation ridge, et atteindre de manière asymptotique l'optimalité.

La figure 1.5 montre des courbes de prix de l'électricité en Allemagne (en EUR/MWh) du 15 mars 2012 au 14 mars 2013 sur 241 jours de travail (mesurées toutes les heures), en fonction de la demande résiduelle (en MWh). La figure 1.6 présente les courbes reconstruites avec la méthode de Kneip and Liebl (2020).

D'autre part, une méthode de complétion de matrice de covariance pour l'analyse des données fonctionnelles a été développée récemment. Elle consiste à estimer les coefficients de matrice de covariance en présence de données manquantes. Descary and Panaretos (2019) ont construit un estimateur non paramétrique de la covariance basé sur des fragments discrètement observés, telle que chaque partie observée de courbe est un intervalle unique de longueur égale à $0 < \delta < 1$ avec absence d'information sur la covariance en dehors de la bande $\mathcal{B}_\delta = \{(s,t) \in [0,1]^2, |s-t| \leqslant \delta\}$. Plusieurs statisticiens ont étudié ce problème d'estimation de matrice de covariance en présence de données incomplètes, par exemple, Lin and Wang (Lin and Wang); Lin et al. (2021), pour les données longitudinales. Delaigle et al. (2020) utilisent uniquement la bande diagonale en minimisant la distance entre la matrice de covariance et son estimateur calculé sur la bande diagonale (voir figure 1.7).

1.3) Résumé.

Dans cette thèse, nous commençons par étudier le modèle linéaire fonctionnel dont la covariable $X = (X(t), t \in [a,b])$ est une donnée fonctionnelle centrée dans un espace Hilbert $L^2([a,b])$ avec $[a,b]$ un intervalle de $\mathbb{R}$ et la réponse Y est une variable à valeurs réelles.

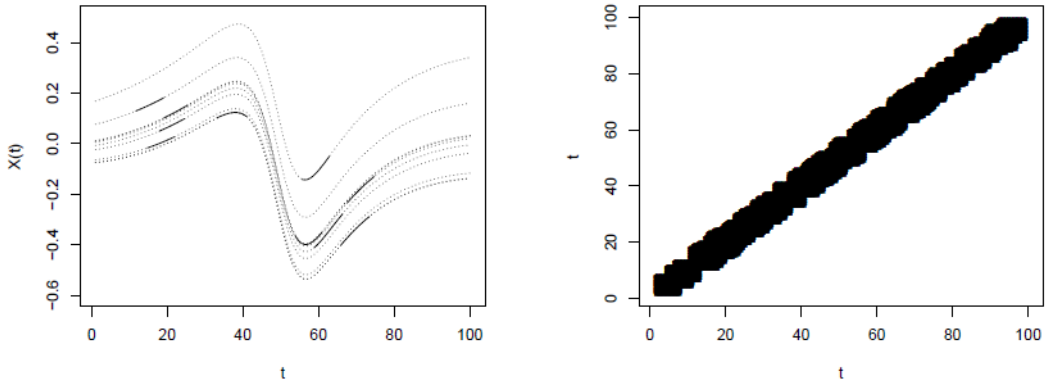

Figure 1.7: À gauche: sous-échantillon de 10 fragments de courbe $(-)$, pris à partir d'un échantillon $n = 100$ de fragments de courbes $X_i(t)$, pour $t \in I = [1, 100]$ et $i = 1, ..., n$ (les longues lignes pointillées montrent les 10 courbes non observées). À droite : nuage de points $(s, t) \in I \times I$ où au moins un couple $(X_i(t), X_i(s))$ est observé, pour $i = 1, ..., n$.

Ce modèle a été notamment étudié par Cardot et al. (1999); Cai and Hall (2006); Hall and Horowitz (2007); Crambes et al. (2009). Il est défini par

$$Y = \theta_0 + \int_a^b \theta(t)X(t)\mathrm{d}t + \varepsilon,$$

avec $\theta_0 \in \mathbb{R}$ et θ est une fonction carré intégrable sur $[a, b]$ qui sert à modéliser la relation entre la variable réelle Y et la fonction aléatoire intégrable X. L'erreur du modèle ε est une variable aléatoire réelle centrée indépendante de X avec une variance finie $\mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$. Dans cette partie, nous étudions le cadre où la variable fonctionnelle est partiellement observée et la réponse réelle contient des données manquantes. Pour cela, nous commençons par reconstruire la partie non observée de la courbe X, en adoptant la méthodologie de Kneip and Liebl (2020) pour la reconstruction fonctionnelle de la variable explicative. A partir de certaines hypothèses de régularité, des vitesses de convergence sont disponibles pour les courbes reconstruites.

Une fois les données fonctionnelles complétées, nous passons à la deuxième étape, celle d'imputation des données réelles manquantes de la réponse Y. Nous utilisons une méthode d'imputation simple, puis une imputation multiple. Nous estimons la fonction θ en utilisant la régression fonctionnelle sur composantes principales (RCP) qui sert à régresser la réponse sur les composantes principales liées aux plus grandes valeurs propres de l'opérateur de covariance du prédicteur fonctionnel. Cette procédure a été étudiée dans plusieurs travaux (voir Cardot et al. (1999, 2003); Hall and Hosseini-Nasab (2006), Hall and Horowitz (2007); Ferraty et al. (2013)). Comme dernière étape, nous prédisons une nouvelle observation et donnons des vitesses de convergence de l'erreur de prédiction.

Un objectif de la thèse est d'étudier aussi le modèle linéaire fonctionnel dont la réponse et la variable explicative sont toutes les deux des données fonctionnelles, en particulier, en présence de données partiellement observées. Nous nous intéressons à deux façons de procéder sur la variable réponse après avoir reconstruit la variable explicative: reconstruire la variable réponse, ou imputer la variable réponse à l'aide du modèle.

La thèse, écrite en anglais, est divisée en cinq chapitres. Nous étudions dans le chapitre 2, le modèle linéaire fonctionnel où la covariable est une variable fonctionnelle et la réponse est une variable réelle. Nous étudions ici le modèle en présence des données partiellement observées. Nous reconstruisons les courbes et nous utilisons la méthode d'imputation simple pour compléter les données réelles. Dans le chapitre 3, nous étudions l'imputation multiple sur les données réelles. Pour le chapitre 4, nous considérons la réponse comme variable fonctionnelle. Enfin nous présentons au chapitre 5, les conclusions et les perspectives de cette thèse. Voici un résumé des chapitres 2, 3 et 4.

Chapitre 2

Dans ce chapitre, nous proposons le modèle linéaire fonctionnel en présence de données partiellement observées avec la covariable fonctionnelle et la réponse réelle.

D'abord, nous construisons la variable explicative en adaptant la méthodologie de Kneip and Liebl (2020). Nous notons la partie observée par O_i et M_i la partie manquante de la courbe X_i, avec $O_i = \cup_{j=1}^{S} O_i^S$ où $O_i^1, \ldots, O_i^S$ sont S intervalles disjoints où la courbe X_i est observée. Pour simplifier, nous prenons $S = 1$ et nous utilisons la décomposition de Karhunen-Loève sur la courbe observée X_i^O dans $\mathbb{L}^2(O)$

$$X_i^O(t) = \sum_{k=1}^{+\infty} \xi_{ik}^O \phi_k^O(t),$$

où $t \in O$, ϕ_k^O sont les fonctions propres de la matrice de covariance des courbes observées et ξ_{ik}^O les scores des composantes principales définis par $\xi_{ik}^O = \langle \phi_k^O, X_i^O \rangle$ pour tout $i = 1, \ldots, n$ et $k \geqslant 1$. Nous avons $\mathbb{E}(\xi_{ik}^O) = 0$ et $\mathbb{E}(\xi_{ik}^O \xi_{i\ell}^O) = \lambda_k^O$ pour tout $k = \ell$ et zero pour $k \neq \ell$. Nous considérons que la partie manquante des courbes s'écrit à partir de la partie observée, sous la forme

$$X_i^M(s) = L(X_i^O(t)) + Z_i(s),$$

pour tout $t \in O$ et $s \in M$, où $L : \mathbb{L}^2(O) \to \mathbb{L}^2(M)$ est un opérateur linéaire de reconstruction et $Z_i \in \mathrm{L}^2(M)$ est l'erreur de reconstruction. L'opérateur optimal de reconstruction, qui minimise l'erreur $\mathbb{E}\left(\left(X_i^M(s) - L(X_i^O)(s)\right)^2\right)$, pour tout $s \in M$, est donné dans Kneip and Liebl (2020) par

$$\mathcal{L}(X_i^O)(s) = \sum_{k=1}^{+\infty} \xi_{ik}^O \widetilde{\phi}_k^O(s), \quad \text{pour tout } s \in M,$$

où $\widetilde{\phi}_k^O$ sont les fonctions propres de l'opérateur de covariance de la courbe reconstruite. Nous nous basons sur un estimateur tronqué, en utilisant des estimateurs par polynômes

locaux pour les scores des composantes principales et les fonctions propres de l'opérateur de covariance de la courbe observée. L'estimateur tronqué de l'opérateur de reconstruction est défini par

$$\widehat{\mathcal{L}}_{k_n}(X_i^O)(s) = \sum_{k=1}^{k_n} \widehat{\xi}_{ik}^O \widehat{\widetilde{\phi}}_k^O(s) \quad \text{pour tout } \ s \in M,$$

où k_n est un entier positif. Nous définissons alors la courbe reconstruite par

$$X_i^\star(t) = \begin{cases} X_i^O(t) & \text{si } \ t \in O, \\ \widehat{\mathcal{L}}_{k_n}(X_i^O)(t) & \text{si } \ t \in M. \end{cases}$$

Ensuite, nous estimons la fonction de régression θ avec la nouvelle courbe reconstruite $X^\star$. Puis, nous adaptons la méthodologie présentée dans Crambes and Henchiri (2019) pour compléter les données manquantes présentés dans la réponse réelle par l'imputation simple sous l'hypothèse MAR. Soit $Y_{i,imp}$ la valeur imputée sur la i$^{\text{ème}}$ observation, nous obtenons alors la nouvelle valeur de réponse $Y_i^\star$, définie par

$$Y_i^\star = Y_i \delta_i^{[Y]} + Y_{i,imp}(1 - \delta_i^{[Y]}) \quad \text{pour tout } \ i = 1, \ldots, n.$$

Une fois l'échantillon initial complété, nous présentons l'estimation du paramètre fonctionnel et prédisons de nouvelles valeurs pour la réponse. Nous étudions ensuite l'erreur quadratique moyenne de la prédiction. Des simulations et des résultats numériques sont également présentés à la fin du chapitre 2. Nous donnons également une illustration sur un jeu de données réelles.

Ce chapitre est sous forme d'article Crambes et al. (2022) soumis à *Journal of Nonparametric Statistics*.

Chapitre 3

Une fois que les courbes partiellement observées X sont reconstruites, nous adaptons dans ce chapitre la méthode d'imputation multiple pour compléter les données manquantes dans la réponse réelle Y par la méthode aléatoire. Cette méthode aléatoire consiste à ajouter un résidu aléatoire $\varepsilon_i^\star$ tiré au hasard, généralement avec remise, à partir d'une méthode déterministe, c'est-à-dire celle de l'imputation simple. La valeur imputée est définie par

$$\widetilde{Y}_\ell \triangleq Y_{\ell,imp} + \varepsilon_\ell^\star,$$

où $Y_{\ell,imp} = \widetilde{\theta}_0 + \langle \widetilde{\theta}, X_\ell^\star \rangle$ est la valeur imputée par la méthode de l'imputation déterministe. Les résidus $\varepsilon_\ell^\star$ sont tirés aléatoirement parmi les erreurs de prédiction de la méthode d'imputation simple observées sur les répondants. Nous imputons q fois la valeur manquante Y_ℓ avec $\delta_\ell^{[Y]} = 0$ et q un entier fixé supérieur à 1. Nous passons par les quatre étapes suivantes :

Étape 1 : Estimer les paramètres $\widetilde{\theta}_0$ et $\widetilde{\theta}$ de modèle linaire fonctionnel en utilisant l'échantillon complet $(X_i^\star, Y_i, \delta_i^{[Y]} = 1)$.

Étape 2 : Tirer le résidu $\varepsilon_\ell^{\star(w)}$ parmi les erreurs centrées réduites de prédiction de la méthode d'imputation simple observées sur les répondants pour tout $w = 1, ..., q$.

Étape 3 : Tirer les valeurs imputées des données manquantes ($\delta_\ell^{[Y]} = 0$) de

$$\widetilde{Y}_\ell^{(w)} = \widetilde{\theta}_0^{(w)} + < \widetilde{\theta}^{(w)}, X_\ell^\star > + \varepsilon_\ell^{\star(w)}.$$

Étape 4 : Répéter l'étape 2 et 3 indépendamment q fois, pour créer plusieurs ensembles d'imputations ($w = 1, ..., q$).

Une fois que la base de données a été reconstruite avec q imputations aléatoires, nous estimons pour $w = 1, \ldots, q$, l'intercept et la fonction de régression du modèle fonctionnel et nous obtenons la nouvelle réponse prédite $\widehat{Y}_{new}^{\star(w)}$.

Pour une nouvelle courbe X_{new}, la réponse est prédite par

$$\widehat{Y}_{new} = \frac{1}{q} \sum_{s=1}^{q} \widehat{Y}_{new}^{\star(w)}.$$

Nous étudions la vitesse convergence de l'erreur quadratique moyenne de la prévision. Nous donnons aussi à la fin des simulations et des résultats numériques.

Ce chapitre est un article soumis à *Communications in statistics : Theory and Methods*.

Chapitre 4

Dans ce chapitre, nous proposons un modèle linéaire fonctionnel dont la variable explicative prend ses valeurs dans l'espace $\mathbb{L}^2(\mathcal{T})$ et la variable à expliquer Y centrée dans $\mathbb{L}^2(\mathcal{S})$ où $\mathcal{T}$ et $\mathcal{S}$ sont deux intervalles dans $\mathbb{R}$. Le modèle s'écrit comme

$$Y(s) = \int_{\mathcal{T}} \theta(s,t) X(t) \mathrm{d}t + \varepsilon(s), \quad \mathbb{E}(\varepsilon \mid X) = 0, \tag{1.3.1}$$

où la fonction $(s,t) \mapsto \theta(s,t)$ est dans l'espace $\mathbb{L}^2(\mathcal{T} \times \mathcal{S})$ et ε est l'erreur du modèle.

Nous étudions le modèle (5.2.2) quand la variable explicative fonctionnelle et la réponse fonctionnelle sont partiellement observées.

D'abord, nous adaptons la même stratégie que dans le chapitre 2 Crambes et al. (2022), pour reconstruire les courbes de la covariable. Ensuite, nous cherchons à compléter les données non observées dans les courbes de la variable Y. Nous proposons deux méthodes.

La première méthode consiste à reconstruire les courbes de réponse en adaptant la même méthodologie que les courbes de la covariable. Nous introduisons un opérateur de reconstruction $\mathcal{J}$ pour Y, similaire à celui de X, et nous obtenons la nouvelle courbe reconstruite de la réponse, définie par

$$Y_i^\star(s) = \begin{cases} Y_i^{O^{[Y]}}(s) & \text{si } \ s \in O^{[Y]}, \\ \widehat{\mathcal{J}}_{j_n}(Y_i^{O^{[Y]}})(s) & \text{si } \ s \in M^{[Y]}, \end{cases}$$

où $\widehat{\mathcal{J}}_{j_n}$ est l'estimateur tronqué de l'opérateur de reconstruction linéaire $\mathcal{J}$ avec j_n entier positif, $Y_i^{O^{[Y]}}$ est la partie observée de la courbe de réponse Y_i, $O^{[Y]}$ et $M^{[Y]}$ sont respectivement les intervalles observés et non observés de la courbe. Ensuite, nous étudions l'estimation de la fonction θ et la prédiction de la réponse en donnant l'erreur quadratique moyenne de la prédiction.

Dans la deuxième méthode, nous imputons, sous l'hypothèse missing at random (MAR), les courbes partiellement observées pour compléter les courbes de réponse. Pour le mécanisme des données manquantes dans les courbes Y, nous proposons une variable fonctionnelle $\delta^{[Y]}$ menant à l'échantillon $(\delta_i^{[Y]})_{i=1,\dots,n}$ tel que,

$$\delta_i^{[Y]} = \begin{cases} 0 & \text{si } \ M_i^{[Y]} \neq \varnothing, \\ 1 & \text{si } \ O_i^{[Y]} = \mathcal{S}. \end{cases}$$

Cette méthode d'imputation consiste en une généralisation de ce qui a été fait dans le chapitre 2 à ce cadre où la réponse est fonctionnelle.

Une fois les données fonctionnelles complètes, nous déterminons la vitesse de convergence de la prédiction. Nous donnons à la fin des simulations et des résultats numériques, ainsi qu'une comparaison entre les deux méthodes.

CHAPTER 2

FUNCTIONAL LINEAR MODEL WITH PARTIALLY OBSERVED COVARIATE AND MISSING VALUES IN THE RESPONSE

Abstract.

Dealing with missing values is an important issue in data observation or data recording process. In this paper, we consider a functional linear regression model with partially observed covariate and missing values in the response. We use a reconstruction operator that aims at recovering the missing parts of the explanatory curves, then we are interested in regression imputation method of missing data on the response variable, using functional principal component regression to estimate the functional coefficient of the model. We study the asymptotic behavior of the prediction error when missing data are replaced by the imputed values in the original dataset. The practical behavior of the method is also studied on simulated data and a real dataset.

Keywords.
Functional linear model; Functional Principal Components; Missing data; Missing At Random; Regression imputation.

2.1) Introduction

The analysis of functional data has grown very significantly in recent years, as evidenced by the numerous literatures on the subject: Ramsay and Silverman (2005), Ferraty and

Vieu (2006), Hsing and Eubank (2015), Horváth and Kokoszka (2012) provide a non-exhaustive list of monographs giving an overview of this topic. One of the most popular model in functional data analysis is the functional linear model, when one is interested in considering a relationship between a real-valued variable Y and a covariate $X = (X(t), t \in [a,b])$ valued in a real separable Hilbert space H of functions defined on a compact interval $[a,b]$ of $\mathbb{R}$. We assume that X is centered, that is $\mathbb{E}(X(t)) = 0$ for all $t \in [a,b]$. In the following, we consider the space $H = L^2([a,b])$ of square integrable functions defined on $[a,b]$, endowed with its usual inner product defined by $\langle u, v\rangle = \int_a^b u(t)v(t)\mathrm{d}t$ for all functions $u, v \in H$, and its associated norm $\|.\|$. This model, studied by many authors as for instance Cardot et al. (1999), Cai and Hall (2006), Hall and Horowitz (2007), Crambes et al. (2009), is defined by

$$Y = \theta_0 + \int_a^b \theta(t)X(t)\mathrm{d}t + \varepsilon, \tag{2.1.1}$$

where $\theta_0 \in \mathbb{R}$ and θ is a square integrable function defined on $[a,b]$ modeling the relationship between the real random variable Y and the square integrable random function X. The error of the model ε is a centered real random variable independent of X with finite variance $\mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$. We can also write the functional linear regression model (2.1.1) as

$$Y = \theta_0 + \Theta X + \varepsilon,$$

where $\Theta : H \to \mathbb{R}$ is a linear continuous operator defined by $\Theta u = \langle \theta, u\rangle$ for any function $u \in H$. The existence and unicity of this regression function θ is discussed in Cardot et al. (2003). A smooth version of the functional principal components regression (SPCR) is introduced. It consists in considering the empirical covariance operator of the predictor X and diagonalizing it to select the eigenfunctions associated to the highest eigenvalues. Then, a least squares regression is performed with the response Y and the coordinates of the functional covariate X projected on the space spanned by the selected eigenfunctions.

Considering a sample $(X_i, Y_i)_{i=1,\dots,n}$ of independent and identically distributed couples with the same distribution as (X,Y), we define the empirical cross covariance operator $\widehat{\Delta}_n$ given by $\widehat{\Delta}_n u = \frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle Y_i$ for all $u \in H$, the empirical covariance operator $\widehat{\Gamma}_n$ given by $\widehat{\Gamma}_n u = \frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle X_i$ for all $u \in H$. Denoting $(\widehat{\phi}_j)_{j=1,\dots,k_n}$ the eigenfunctions associated to $\widehat{\Gamma}_n$ corresponding to the k_n highest eigenvalues $\widehat{\lambda}_1 > \dots > \widehat{\lambda}_{k_n} > 0$ (where k_n is an integer depending on n), we define the orthogonal projection operator $\widehat{\Pi}_{k_n}$ onto the subspace $\mathrm{Span}(\widehat{\phi}_1, \dots, \widehat{\phi}_{k_n})$ by $\widehat{\Pi}_{k_n} u = \sum_{j=1}^{k_n} \langle \widehat{\phi}_j, u\rangle \widehat{\phi}_j$ for all $u \in H$. Then, the functional principal component regression estimator $\widehat{\Theta}$ of Θ is defined by

$$\widehat{\Theta} = \langle \widehat{\theta}, . \rangle = \widehat{\Pi}_{k_n}\widehat{\Delta}_n(\widehat{\Pi}_{k_n}\widehat{\Gamma}_n\widehat{\Pi}_{k_n})^{-1}.$$

The corresponding estimator of θ is given by

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i, \widehat{\phi}_j\rangle Y_i}{\widehat{\lambda}_j}\widehat{\phi}_j = \sum_{j=1}^{k_n}\widehat{s}_j\widehat{\phi}_j, \tag{2.1.2}$$

with $\widehat{s}_j = \frac{1}{n\widehat{\lambda}_j}\sum_{i=1}^{n}\langle X_i, \widehat{\phi}_j\rangle Y_i$. In addition, the estimator of θ_0 is $\widehat{\theta}_0 = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Now, given $\widehat{\theta}_0$ and $\widehat{\theta}$, it is easy to obtain the residuals of the fit, given by $\widehat{\varepsilon}_{i,k_n} = Y_i - \widehat{\theta}_0 - \langle X_i, \widehat{\theta}\rangle$, for $i = 1, \ldots, n$, that can be used to estimate the error variance, σ^2_ε, through

$$\widehat{\sigma}^2_{\varepsilon,k_n} = \frac{1}{n - k_n - 1}\sum_{i=1}^{n}\widehat{\varepsilon}^2_{i,k_n}.$$

In the previously cited works on the functional linear model, data is fully observed. This may not always be the case, and missing data appear in many situations, for example when the measuring device breaks down or when an observation interval is not available. This topic has to be studied a lot in the multivariate framework, for example we refer the reader to Little and Rubin (2002) and Graham (2012). For functional data, the literature only starts developing. In functional linear regression, the work of Crambes and Henchiri (2019) considers a missing data mechanism on the response Y while the functional covariate is completely observed. A regression imputation methodology for the missing data is proposed and the authors propose an estimation of the functional parameter θ with the reconstructed dataset, as well as the prediction of new values. The method consistency is studied both from a theoretical and a practical point of view. The same problematic is studied in another paper from Febrero-Bande et al. (2019), although not exploring theoretical results. Other works explore the context of missing data in the response while the response is missing at random in a nonparametric setting (see Ferraty et al. (2013), Ling et al. (2015)) or in a functional partial linear regression setting (see Ling et al. (2019), Zhou and Peng (2020)) or while the response is not missing at random (see Li et al. (2018)). In our work, we want to consider the functional linear model where some observations of the real response are affected with missing data and the covariate is partially observed, which is an unexplored topic as far as we know.

For the missing data mechanism in the response, we consider a dichotomous random variable $\delta^{[Y]}$ leading to the sample $(\delta_i^{[Y]})_{i=1,\ldots,n}$ such that $\delta_i^{[Y]} = 1$ if the value Y_i is available and $\delta_i^{[Y]} = 0$ if the value Y_i is missing, for all $i = 1, \ldots, n$. Here, we consider that the data in the response is missing at random (MAR): the fact that the value Y is missing does not depend on the response of the model, but can possibly depend on the covariate, that is,

$$\mathbb{P}(\delta^{[Y]} = 1 \mid X, Y) = \mathbb{P}(\delta^{[Y]} = 1 \mid X).$$

As a consequence of this MAR assumption, the variable $\delta^{[Y]}$ (the fact that an observation is missing) is independent of the error of the model ε. In the following, the number of

missing values among $Y_1, \dots, Y_n$ is denoted

$$m_n^{[Y]} = \sum_{i=1}^{n} \mathbb{1}_{\{\delta_i^{[Y]}=0\}}.$$

For the missing data mechanism of the functional covariate, we adopt the paradigm of partially observed functions as in Kneip and Liebl (2020) or Kraus (2015). We also refer the reader to Delaigle et al. (2020) or Kraus and Stefanucci (2020) for recent contributions on this topic. More precisely, for each curve X_i, $i = 1, \dots, n$, we consider the observed part $O_i \subseteq [a, b]$ of X_i and the missing part $M_i = [a, b] \setminus O_i$. The observed part O_i refers to an interval (or several intervals) where the curve X_i is observed at some measure points of O_i. Based on the punctual observations, the whole curve can be reconstructed on O_i with usual methods (e.g. smoothing splines, regression splines, local polynomial smoothing, ...). On the contrary, no information is available on the missing part M_i. An example of such partially observed functions is given in section 2.5 of the paper.

The objective of this paper is to predict a new value of the response Y given a new test observation on the explanatory variable X once the partially observed curves X have been reconstructed and the missing data Y have been imputed. More precisely, we want to obtain convergence rates for this prediction error, and we want to analyse how these convergence rates depend on the convergence rates of the reconstruction of the missing parts of the covariate and the convergence rates of the imputation error. Moreover, we want to explore the interest of the imputation methodology compared to other methods, for example the naive method which would consist in simply ignoring the missing data and only using the observations when both X and Y are observed, or other imputation methods.

In the following, we give in section 2.2 theoretical results when the covariate is partially observed. Then, in section 2.3, we extend these results when the covariate is partially observed and some observations of the real response are affected with missing data. In section 2.4, we present some simulation results to show the behaviour of the method in practice. Section 2.5 is devoted to a real dataset application. Finally, all the proofs are postponed to section 2.6.

2.2) Partially observed covariate

2.2.1) Curve reconstruction

We write "O" and "M" to denote a given production of O_i and M_i. In addition, we denote the observed and missing parts of X_i by X_i^O and X_i^M. As noticed in Kneip and Liebl (2020, p. 7) all the following remains valid if we consider the more general case of several observed subintervals, that is $O_i = \cup_{j=1}^{J} O_i^J$ where $O_i^1, \dots, O_i^J$ are J disjoint

intervals where the curve X_i is observed. For the sake of simplicity, we will take $J = 1$ and $O_i = O_i^1$. We write the Karhunen-Loève (KL) decomposition of X_i^O in $\mathbb{L}^2(O)$

$$X_i^O(t) = \sum_{k=1}^{+\infty} \xi_{ik}^O \phi_k^O(t),$$

where $t \in O$. In this decomposition, the principal component scores are defined for all $i = 1, \ldots, n$ and $k \geqslant 1$ by $\xi_{ik}^O = \langle \phi_k^O, X_i^O \rangle$, where $\mathbb{E}(\xi_{ik}^O) = 0$ and $\mathbb{E}(\xi_{ik}^O \xi_{i\ell}^O) = \lambda_k^O$ for all $k = \ell$ and zero for all $k \neq \ell$. Moreover, the eigenfunctions satisfy

$$\phi_k^O(t) = \frac{\langle \phi_k^O, \gamma_t^O \rangle}{\lambda_k^O}, \tag{2.2.1}$$

for all $t \in O$ and $k \geqslant 1$, where $\gamma_t^O(s) = \gamma^O(t, s) = \mathbb{E}\left(X_i^O(t) X_i^O(s)\right)$, and the decreasing eigenvalues $\lambda_1^O > \lambda_2^O > \ldots > 0$ are tending to zero.

We consider a reconstruction problem relating the missing part of the curves to the observed part, writing

$$X_i^M(s) = L(X_i^O(t)) + Z_i(s),$$

for all $t \in O$ and $s \in M$, where $L : \mathbb{L}^2(O) \to \mathbb{L}^2(M)$ is a linear reconstruction operator and $Z_i \in \mathrm{L}^2(M)$ is the reconstruction error. This reconstruction estimator is estimated in Kneip and Liebl (2020) by

$$\mathcal{L}(X_i^O)(s) = \sum_{k=1}^{+\infty} \xi_{ik}^O \widetilde{\phi}_k^O(s) = \sum_{k=1}^{+\infty} \xi_{ik}^O \frac{\langle \phi_k^O, \gamma_s \rangle}{\lambda_k^O}, \tag{2.2.2}$$

for all $s \in M$, where $\gamma_s(t) = \mathbb{E}\left(X_i^M(s) X_i^O(t)\right)$ for all $t \in O$ and $s \in M$. The definition of $\widetilde{\phi}_k^O$ is a way to extend the relation (2.2.1) to the missing parts of the curves. It is shown in Kneip and Liebl (2020) that $\mathcal{L}(X_i^O)$ has a continuous and finite variance function and is unbiaised.

2.2.2) Estimation of the reconstruction in practice

We consider a discretization without measurement errors, that is $((W_{i1}, t_{i1}), \ldots, (W_{ip}, t_{ip}))$ denote the observable data pairs of the function X_i^O, namely $W_{ij} = X_i^O(t_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, where $t_{ij} \in O_i$. In order to estimate the curve X_i^O and the covariance function γ_s, a nonparametric curve estimation by local polynomials smoothers is used. The latter is similar to the procedure in Yao et al. (2005) or Hall et al. (2006). For the curve X_i^O, we use a kernel κ_1 and a bandwidth h_X, the local linear smoother of the curve X_i^O being denoted $\widehat{X}_i^O(t; h_X)$. Similarly for the covariance function γ_s, we use a kernel κ_2 and a bandwidth h_γ, the local linear smoother of the covariance function γ being denoted $\widehat{\gamma}(t, s; h_\gamma)$.
For estimating the eigenvalues λ_k^O and the eigenfunctions ϕ_k^O, we use the Fredholm integral equation

$$\int_O \widehat{\gamma}(t,u;h_\gamma)\widehat{\phi}_k^O(u)\mathrm{d}u = \widehat{\lambda}_k^O\widehat{\phi}_k^O(t),$$

for all $t \in O$. For the functional principal component scores $\xi_{ik}^O = \int_O X_i^O(t)\phi_k(t)\mathrm{d}t$, the estimator is defined by

$$\widehat{\xi}_{ik}^O = \sum_{j=1}^{p} \widehat{\phi}_k^O(t_{ij})W_{ij}(t_{ij} - t_{i,j-1}), \quad \text{with } t_{i0} = a.$$

Finally, to estimate $\mathcal{L}(X_i^O)$ in (2.2.2), considering a positive integer k_n, we define

$$\widehat{\mathcal{L}}_{k_n}(X_i^O)(s) = \sum_{k=1}^{k_n} \widehat{\xi}_{ik}^O \frac{\langle \widehat{\phi}_k^O, \widehat{\gamma}_s\rangle}{\widehat{\lambda}_k^O},$$

where $\widehat{\gamma}_s = \widehat{\gamma}(.,s;h_\gamma)$. At this step we are able to find the estimator of the missing parts of X_i^O

$$\widehat{X}_i^M(s) = \widehat{\mathcal{L}}_{k_n}(X_i^O)(t),$$

for all $t \in O$ and $s \in M$. In the following, we denote

$$X_i^\star(t) = \begin{cases} X_i^O(t) \text{ if } t \in O, \\ \widehat{\mathcal{L}}_{k_n}(X_i^O)(t) \text{ if } t \in M. \end{cases}$$

2.2.3) *Estimation of θ and prediction*

For estimating θ, we set

$$\widehat{\widehat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}_{j,rec}\rangle Y_i}{\widehat{\lambda}_{j,rec}}\widehat{\phi}_{j,rec} = \sum_{j=1}^{k_n} \widehat{\widehat{s}}_j\widehat{\phi}_{j,rec},$$

with $\widehat{\widehat{s}}_j = \frac{1}{n\widehat{\lambda}_{j,rec}}\sum_{i=1}^{n}\langle X_i^\star, \widehat{\phi}_{j,rec}\rangle Y_i$. The estimation of the operator Θ is given by

$$\widehat{\widehat{\Theta}} = \langle\widehat{\widehat{\theta}}, .\rangle = \widehat{\Pi}_{k_n,rec}\widehat{\Delta}_{n,rec}(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec})^{-1},$$

where $\widehat{\Delta}_{n,rec}$ is the reconstructed cross covariance operator given by $\widehat{\Delta}_{n,rec} = \frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star, .\rangle Y_i$, $\widehat{\Gamma}_{n,rec}$ is the reconstructed covariance operator given by $\widehat{\Gamma}_{n,rec} = \frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star, .\rangle X_i^\star$, and $\widehat{\Pi}_{k_n,rec}$ is the projection operator onto the subspace $\mathrm{Span}(\widehat{\phi}_{1,rec},\ldots,\widehat{\phi}_{k_n,rec})$, that is the subspace spanned by the k_n first eigenfunctions of the covariance operator $\widehat{\Gamma}_{n,rec}$. The eigenvalues of the covariance operator $\widehat{\Gamma}_{n,rec}$ are denoted $\widehat{\lambda}_{1,rec},\ldots,\widehat{\lambda}_{k_n,rec}$. Moreover, the estimator of θ_0 is defined by $\widehat{\widehat{\theta}}_0 = \overline{Y}$. Given $\widehat{\widehat{\theta}}_0$ and $\widehat{\widehat{\theta}}$, the residuals of the fit, $\widehat{\widehat{\varepsilon}}_{i,k_n} = Y_i - \widehat{\widehat{\theta}}_0 - \langle X_i^\star, \widehat{\widehat{\theta}}\rangle$, for $i = 1,\ldots,n$, can be used to estimate the error variance as follows

$$\widehat{\widehat{\sigma}}^2_{\varepsilon,k_n} = \frac{1}{n-k_n-1}\sum_{i=1}^{n}\widehat{\widehat{\varepsilon}}^2_{i,k_n}.$$

Finally, given a new observation of the covariate X, denoted X_{new}, possibly partially observed, we predict the corresponding value of the response Y by

$$\widehat{Y}_{new} = \widehat{\widehat{\theta}}_0 + \langle \widehat{\widehat{\theta}}, X^{\star}_{new} \rangle.$$

2.2.4) Assumptions

We present in this part the assumptions needed for our results. These assumptions are used in Kneip and Liebl (2020) in order to control the curve reconstruction for the covariate.

(A.1) The variable X has a finite four moment order, that is $\mathbb{E}\left(\|X\|^4\right) < \infty$.

(A.2) Let $np \to \infty$ when $n \to \infty$ and $p = p(n)$. We assume $p = n^{\eta_1}$ with $0 < \eta_1 < \infty$ in the following.

(A.3) The bandwidth h_X satisfies $h_X \to 0$ and $(ph_X) \to \infty$ as $p \to \infty$. For instance, we assume that $h_X = \frac{1}{n^{\eta_2}}$ with $0 < \eta_2 < \eta_1$. The bandwidth h_γ satisfies $h_\gamma \to 0$ and $(n(p^2-p)h_\gamma) \to \infty$ as $n(p^2-p) \to \infty$. For example, we can take $h_\gamma = \frac{1}{n^{\eta_3}}$ with $0 < \eta_3 < 2\eta_1 + 1$.

(A.4) Let κ_1 and κ_2 be nonnegative, second order univariate and bivariate kernel functions with support $[-1,1]$. For example, we can use univariate and bivariate Epanechnikov kernel functions with compact support $[-1,1]$, namely $\kappa_1(x) = \frac{3}{4}(1-x^2)\mathbb{1}_{[-1,1]}(x)$ and $\kappa_2(x,y) = \frac{9}{16}(1-x^2)(1-y^2)\mathbb{1}_{[-1,1]}(x)\mathbb{1}_{[-1,1]}(y)$.

(A.5) • For any subinterval $O \subseteq [a,b]$, we assume that the eigenvalues $\lambda_1 > \lambda_2 > \ldots > 0$ have multiplicity one. Moreover, we assume that there exist $a_O > 1$ and $0 < c_O < \infty$ such that `(i)` $\lambda^O_k - \lambda^O_{k+1} \geqslant c_O k^{-a_O-1}$, `(ii)` $\lambda^O_k = \mathcal{O}(k^{-a_O})$, `(iii)` $1/\lambda^O_k = \mathcal{O}(k^{a_O})$ as $k \to \infty$.
• $\mathbb{E}\left(\xi^4_k\right) = \mathcal{O}(\lambda^2_k)$,

(A.6) For any subinterval $O \subseteq [a,b]$, we assume that there exists $0 < D_O < \infty$ such that the eigenfunctions satisfy $\sup_{t\in[a,b]}\sup_{k\geqslant 1}\left|\widetilde{\phi}^O_k(t)\right| \leqslant D_O$.

Assumption ***(A.1)*** holds for many processes X (Gaussian processes, bounded processes). Assumption ***(A.2)*** is mild and can be satisfied even if the number of observation points p does not go fast to infinity. As in Kneip and Liebl (2020), we assume that $p = n^{\eta_1}$ with $0 < \eta_1 < \infty$. Assumptions ***(A.3)*** and ***(A.4)*** are classic in the context of local polynomials smoothers. Assumptions ***(A.5)*** and ***(A.6)***, related to eigenvalues and

eigenfunctions of the covariance operator of X, are given in Kneip and Liebl (2020). In particular, a polynomial decrease of the eigenvalues is required, allowing a large class of eigenvalues for the covariance operator of X.

2.2.5) Asymptotic results

Under assumptions ***(A.1)-(A.6)***, it is proved in Kneip and Liebl (2020) that, in the case where $p \sim n^{\eta_1}$ with $\eta_1 \leqslant 1/2$, we have for any $t \in [a, b]$

$$|X_i^\star(t) - X_i(t)| = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big) . \tag{2.2.3}$$

The previous result allows to obtain some bounds between quantities related to functional principal components analysis with the constructed curves and with the original curves. These bounds are given in the following proposition. For any linear continuous operator $T : H \to H$ or any linear continuous operator $S : H \to \mathbb{R}$, we define the operator norm of T as $\|T\|_\infty = \sup_{\|x\|=1} \|Tx\|$, and the operator norm of S as $\|S\|_\infty = \sup_{\|x\|=1} |Sx|$.

Proposition 2.2.1. *Under assumptions **(A.1)-(A.6)**, we have*

$$\begin{aligned}
&(i) \quad \left\|\widehat{\Gamma}_{n,rec} - \widehat{\Gamma}_n\right\|_\infty = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big) , \\
&(ii) \quad \left\|\widehat{\Delta}_{n,rec} - \widehat{\Delta}_n\right\|_\infty = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big) , \\
&(iii) \quad \forall k \geqslant 1, \quad \left\|\widehat{\phi}_{k,rec} - \widehat{\phi}_k\right\| = \mathcal{O}_p\big(k^{a_O+1} p^{-(a_O-1)/(2(a_O+2))}\big) , \\
&(iv) \quad \forall k \geqslant 1, \quad \left|\widehat{\lambda}_{k,rec} - \widehat{\lambda}_k\right| = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big) .
\end{aligned}$$

We finish this section with the main result giving a bound for the prediction error of Y_{new} with a new value of the covariate X_{new}.

Theorem 2.2.2. *Under assumptions **(A.1)-(A.6)**, if we take $k_n \sim p^{1/(a_O+2)}$ and $p \sim n^{\eta_1}$ with $\eta_1 \leqslant 1/2$, the prediction error is*

$$\mathbb{E}\left(\widehat{\widehat{\theta}}_0 + \langle \widehat{\widehat{\theta}}, X_{new}^\star \rangle - \theta_0 - \langle \theta, X_{new}^\star \rangle\right)^2 = \mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big) .$$

This prediction error rate $\mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)$ is related to the rate given in Corollary 4.1 in Kneip and Liebl (2020) (in the particular case where $\eta_1 = 1/2$). This means that, provided with some conditions on the number of observation points p and the number of principal components k_n are fulfilled, the prediction error rate has the same order as the curve reconstruction error rate. In other words, this means that, when reconstructing missing parts of the explanatory curves in a functional linear model and then predicting a new value of the response, the most important step is the curve reconstruction. This step is going to fix the convergence rate of the prediction.

Remark 2.2.3. *Due to the bound (2.2.3), the result of Theorem 2.2.2 remains valid if we replace $X^{\star}_{new}$ with X_{new}.*

Corollary 2.2.4. *Under the hypotheses of Theorem 2.2.2, in the favorable situation where $\eta_1 = 1/2$, the prediction error is*

$$\mathbb{E}\left(\widehat{\widehat{\theta}}_0 + \langle \widehat{\widehat{\theta}}, X^{\star}_{new}\rangle - \theta_0 - \langle \theta, X^{\star}_{new}\rangle\right)^2 = \mathcal{O}_p\big(n^{-(a_O-1)/(4(a_O+2))}\big)\,.$$

2.3) Partially observed covariate and missing data on the response

In this section, we are interested in the most general case of missing data in functional linear regression: when the covariate is partially observed and when the response is affected by missing data. We have seen in the previous section the methodology for reconstructing the missing parts of the explanatory curves. Concerning missing data on the response, we are going to apply the methodology presented in Crambes and Henchiri (2019), imputing missing values on the response using a regression imputation. Next, once the initial sample is completed, we will present the estimation of the functional parameter θ and predict new values for the response.

2.3.1) Regression imputation on the response

In this subsection, we use the methodology to impute a missing value of Y as in Crambes and Henchiri (2019). We consider the whole data, possibly with reconstructed explanatory curves, except the ones for which the value of Y is not available. We define the covariance operator with the reconstructed curves

$$\widehat{\Gamma}^{obs}_{n,rec} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \langle X_i^{\star}, .\rangle \delta_i^{[Y]} X_i^{\star}.$$

Let $\widehat{\Pi}^{obs}_{k_n,rec}$ be the projection operator onto the subspace Span$(\widehat{\phi}^{obs}_{1,rec}, \ldots, \widehat{\phi}^{obs}_{k_n,rec})$ where $\widehat{\phi}^{obs}_{1,rec}, \ldots, \widehat{\phi}^{obs}_{k_n,rec}$ are the k_n first eigenfunctions of the covariance operator $\widehat{\Gamma}^{obs}_{n,rec}$. With analogous notations, $\widehat{\lambda}^{obs}_{1,rec}, \ldots, \widehat{\lambda}^{obs}_{k_n,rec}$ represent the k_n first eigenvalues of $\widehat{\Gamma}^{obs}_{n,rec}$. We first estimate θ with the observed responses and the observed or reconstructed covariates

$$\widetilde{\theta} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n-m_n^{[Y]}} \sum_{j=1}^{k_n} \frac{\langle X_i^{\star}, \widehat{\phi}^{obs}_{j,rec}\rangle \delta_i^{[Y]} Y_i}{\widehat{\lambda}^{obs}_{j,rec}} \widehat{\phi}^{obs}_{j,rec} = \sum_{j=1}^{k_n} \widetilde{s}_j \widehat{\phi}^{obs}_{j,rec},$$

with $\widetilde{s}_j = \frac{1}{(n-m_n^{[Y]})\widehat{\lambda}_{j,rec}^{obs}} \sum_{i=1}^{n-m_n^{[Y]}} \langle X_i^\star, \widehat{\phi}_{j,rec}^{obs}\rangle \delta_i^{[Y]} Y_i$. We also estimate the intercept θ_0 with $\widetilde{\theta}_0 = \overline{Y}_{obs} = \frac{1}{n-m_n^{[Y]}} \sum_{i=1}^{n} \delta_i^{[Y]} Y_i$. Now, the residuals of the fit, $\widetilde{\varepsilon}_{i,k_n} = Y_i - \widetilde{\theta}_0 - \langle X_i^\star, \widetilde{\theta}\rangle$ for $i = 1, \ldots, n$, can be used to estimate the error variance as follows

$$\widetilde{\sigma}^2_{\varepsilon,k_n} = \frac{1}{n - m_n^{[Y]} - k_n - 1} \sum_{i=1}^{n} \delta_i^{[Y]} \widetilde{\varepsilon}^2_{i,k_n}.$$

Then, considering a missing value on the response, say Y_ℓ such that $\delta_\ell^{[Y]} = 0$, we define the imputed value $Y_{\ell,imp}$ by

$$Y_{\ell,imp} = \widetilde{\theta}_0 + \langle \widetilde{\theta}, X_\ell^\star \rangle = \widetilde{\theta}_0 + \sum_{j=1}^{k_n} \widetilde{\widetilde{s}}_j \langle X_\ell^\star, \widehat{\phi}_{j,rec}^{obs} \rangle,$$

with $\widetilde{\widetilde{s}}_j = \frac{1}{(n-m_n^{[Y]})\widehat{\lambda}_{j,rec}^{obs}} \sum_{\substack{i=1 \\ i\neq\ell}}^{n} \langle X_i^\star, \widehat{\phi}_{j,rec}^{obs}\rangle \delta_i^{[Y]} Y_i$. Let us remark that the imputation $Y_{\ell,imp}$ can also be written

$$Y_{\ell,imp} = \widehat{\Pi}_{k_n,rec}^{obs} \widehat{\Delta}_{n,rec}^{obs} \left(\widehat{\Pi}_{k_n,rec}^{obs} \widehat{\Gamma}_{k_n,rec}^{obs} \widehat{\Pi}_{k_n,rec}^{obs} \right)^{-1} X_\ell^\star,$$

where $\widehat{\Delta}_{n,rec}^{obs} = \frac{1}{n-m_n^{[Y]}} \sum_{i=1}^{n} \langle X_i^\star, . \rangle \delta_i^{[Y]} Y_i$.

2.3.2) Estimation of θ and prediction

Once the whole database has been reconstructed, we estimate the functional coefficient θ with

$$\widehat{\theta}^\star = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}_{j,rec}^\star \rangle Y_i^\star}{\widehat{\lambda}_{j,rec}^\star} \widehat{\phi}_{j,rec}^\star = \sum_{j=1}^{k_n} \widehat{s}_j^\star \widehat{\phi}_{j,rec}^\star,$$

where $\widehat{s}_j^\star = \frac{1}{n\widehat{\lambda}_{j,rec}^\star} \sum_{i=1}^{n} \langle X_i^\star, \widehat{\phi}_{j,rec}^\star \rangle Y_i^\star$ and $Y_i^\star = Y_i \delta_i^{[Y]} + Y_{i,imp}(1 - \delta_i^{[Y]})$ for all $i = 1, \ldots, n$. The estimation of the operator Θ is similarly given by

$$\widehat{\Theta}^\star = \langle \widehat{\theta}^\star, . \rangle = \widehat{\Pi}_{k_n,rec}^\star \widehat{\Delta}_{n,rec}^\star \left(\widehat{\Pi}_{k_n,rec}^\star \widehat{\Gamma}_{n,rec}^\star \widehat{\Pi}_{k_n,rec}^\star \right)^{-1},$$

where the cross covariance operator is $\widehat{\Delta}_{n,rec}^\star = \frac{1}{n} \sum_{i=1}^{n} \langle X_i^\star, . \rangle Y_i^\star$, the covariance operator is $\widehat{\Gamma}_{n,rec}^\star = \frac{1}{n} \sum_{i=1}^{n} \langle X_i^\star, . \rangle X_i^\star$, and $\widehat{\phi}_{1,rec}^\star, \ldots, \widehat{\phi}_{k_n,rec}^\star$ and $\widehat{\lambda}_{1,rec}^\star, \ldots, \widehat{\lambda}_{k_n,rec}^\star$ represent respectively the k_n first eigenfunctions and eigenvalues of the operator $\widehat{\Gamma}_{n,rec}^\star$. We use this estimation to predict a new value of the response Y when a new explanatory curve X_{new} is given

Table 2.1: Single and aggregate imputation mean square error convergence rates.

		single error	aggregate error
(i) $m_n^{[Y]} = \lfloor a_n n \rfloor$		$\mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)$	$\mathcal{O}_p\big(a_n n^{1-\eta_1(a_O-1)/(2(a_O+2))}\big)$
(ii) $m_n^{[Y]} \sim \lfloor \rho n \rfloor$		$\mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)$	$\mathcal{O}_p\big(n^{1-\eta_1(a_O-1)/(2(a_O+2))}\big)$
(iii) $n - m_n^{[Y]} = \lfloor n^\gamma \rfloor$	$\gamma \geqslant \frac{\eta_1(a_O+1)}{2(a_O+2)}$	$\mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)$	$\mathcal{O}_p\big(n^{1-\eta_1(a_O-1)/(2(a_O+2))}\big)$
	$\gamma < \frac{\eta_1(a_O+1)}{2(a_O+2)}$	$\mathcal{O}_p\big(n^{\eta_1/(a_O+2)-\gamma}\big)$	$\mathcal{O}_p\big(n^{1+\eta_1/(a_O+2)-\gamma}\big)$

$$\widehat{Y}_{new} = \widehat{\theta}_0^\star + \langle \widehat{\theta}^\star, X_{new}^\star \rangle = \widehat{\theta}_0^\star + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}_{j,rec}^\star \rangle \langle X_{new}^\star, \widehat{\phi}_{j,rec}^\star \rangle Y_i^\star}{\widehat{\lambda}_{j,rec}^\star} = \widehat{\theta}_0^\star + \sum_{j=1}^{k_n} \widehat{s}_j^\star \langle X_{new}^\star, \widehat{\phi}_{j,rec}^\star \rangle,$$

where $\widehat{\theta}_0^\star = \overline{Y}^\star = \frac{1}{n}\sum_{i=1}^n Y_i^\star$. Then, the residuals of the fit, $\widehat{\varepsilon}_{i,k_n}^\star = Y_i^\star - \widehat{\theta}_0^\star - \langle X_i^\star, \widehat{\theta}^\star \rangle$ for $i = 1, \ldots, n$, allow to estimate the error variance writing

$$(\widehat{\sigma}_{\varepsilon,k_n}^\star)^2 = \frac{1}{n - k_n - 1}\sum_{i=1}^{n} (\widehat{\varepsilon}_{i,k_n}^\star)^2.$$

2.3.3) Asymptotic results

The first result gives an error rate of the imputed values.

Theorem 2.3.1. *Under assumptions **(A.1)-(A.6)**, if we take $k_n \sim p^{1/(a_O+2)}$ and $p \sim n^{\eta_1}$ with $\eta_1 \leqslant 1/2$, we have*

$$\mathbb{E}\left(Y_{\ell,imp} - \theta_0 - \langle \theta, X_\ell^\star \rangle\right)^2 = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}}\right).$$

Moreover, the aggregate error for all the imputed values is given by

$$\sum_{\ell=1}^{n}(1-\delta_\ell^{[Y]})\mathbb{E}\left(Y_{\ell,imp} - \theta_0 - \langle \theta, X_\ell^\star \rangle\right)^2 = \mathcal{O}_p\left(m_n^{[Y]} n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{m_n^{[Y]} n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}}\right).$$

The following corollary explores some specific cases of the above error rates. The given results simply come from a comparison between the convergence rates of the above result, hence the proof is ommited.

Corollary 2.3.2. *We consider cases where the number of missing values on the response are (i) negligeable with respect to the sample size, (ii) proportional to the sample size, (iii) of the same order than the sample size. More precisely*

(i) $m_n^{[Y]} = \lfloor a_n n \rfloor$ *where* a_n *goes to zero when* n *goes to infinity,*
(ii) $m_n^{[Y]} \sim \lfloor \rho n \rfloor$ *with* $0 < \rho < 1$,
(iii) $n - m_n^{[Y]} = \lfloor n^\gamma \rfloor$ *with* $0 < \gamma < 1$.
We summarize the error rate for a single imputed value and the aggregate error in Table 2.1.

We finish the theoretical results with the prediction error of Y_{new} with a new value of the covariate X_{new}. The proof of this result is ommited as it uses previous results of Theorems 2.2.2 and 2.3.1 and follows exactly the same lines as the proof of Theorem 2.2.2.

Theorem 2.3.3. *Under assumptions* ***(A.1)-(A.6)****, and* $k_n \sim p^{1/(a_O+2)}$ *and* $p \sim n^{\eta_1}$ *with* $\eta_1 \leqslant 1/2$*, the prediction error is*

$$\mathbb{E}\left(\widehat{\theta}_0^\star + \langle \widehat{\theta}^\star, X_{new}^\star \rangle - \theta_0 - \langle \theta, X_{new}^\star \rangle\right)^2 = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}}\right).$$

In the particular case where $\eta_1 = 1/2$*, the first term in the convergence rate is* $\mathcal{O}_p\big(n^{-(a_O-1)/(4(a_O+2))}\big)$.

All our convergence rates depend in particular on the parameter $a_O > 1$, which is directly linked to the smoothness of the stochastic process X. The larger a_O is, the smoother X is. When a_O tends to 1 (non-smooth processes, for example a standard Brownian motion corresponds to $a_O = 2$), the convergence rate deteriorates. When a_O tends to infinity (very smooth processes), the convergence rate $n^{-\eta_1(a_O-1)/(2(a_O+2))}$ is equivalent to $n^{-\eta_1/2}$.

As before, we consider cases in the corollary below where the number of missing values on the response are (i) negligeable with respect to the sample size, (ii) proportional to the sample size, (iii) of the same order than the sample size.

Corollary 2.3.4. *In the cases* (i)*,* (ii) *and* (iii) *with* $\gamma \geqslant \frac{\eta_1(a_O+1)}{2(a_O+2)}$*, the prediction error of a new value of the response is*

$$\mathbb{E}\left(\widehat{\theta}_0^\star + \langle \widehat{\theta}^\star, X_{new}^\star \rangle - \theta_0 - \langle \theta, X_{new}^\star \rangle\right)^2 = \mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big).$$

In the case (iii) *with* $\gamma < \frac{\eta_1(a_O+1)}{2(a_O+2)}$*, the prediction error of a new value of the response is*

$$\mathbb{E}\left(\widehat{\theta}_0^\star + \langle \widehat{\theta}^\star, X_{new}^\star \rangle - \theta_0 - \langle \theta, X_{new}^\star \rangle\right)^2 = \mathcal{O}_p\big(n^{\eta_1/(a_O+2)-\gamma}\big).$$

In other words, in situations where the number of missing values on the response is negligeable or moderate with respect to the sample size, the convergence rate of the prediction error is given by the convergence rate obtained in Kneip and Liebl (2020) for the

curve reconstruction. As a conclusion, when dealing with a functional linear model with a partially observed covariate and missing values in the response, the convergence rate of the prediction error strongly depends on the curve reconstruction error, with respect to the response imputation error.

Remark 2.3.5. *As noticed at the end of the previous section, all the results obtained in this section remain valid if we replace $X^\star$ with X.*

2.4) Simulations

2.4.1) Model and samples

All the procedures described below were implemented with the R software. In the simulations, we deal with functions defined on the interval $[0, 1]$. We consider the model

$$Y = \theta_0 + \langle \theta, X \rangle + \varepsilon, \tag{2.4.1}$$

where the error ε is either a Gaussian noise $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.2$ and $\sigma_\varepsilon = 1.5$, or drawn from a centered Beta(2,2) law. We derived different models from (2.4.1), simulating more or less smooth processes X. For the sake of concision, we only give the results for the model presented below. Results for other models are available on demand to the authors.

In this model, as in Hall and Horowitz (2007), the functional covariate X is generated by a set of cosine basis functions $\phi_1 \equiv 1$ and $\phi_{j+1} = \sqrt{2}\cos(j\pi t)$ for $j > 1$, such that $X(t) = \sum_{j=1}^{150} \varrho_j \zeta_j \phi_j(t)$ for all $t \in [0, 1]$, where the ζ_j's are independently sampled from the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$ and the ϱ_j's are defined by $\varrho_j = (-1)^{j+1}(j)^{-\beta/2}$ with $\beta = 4$. The covariance function writes

$$cov(X(t), X(s)) = \sum_{j=1}^{150} \frac{2}{j^\beta} \cos(j\pi t)\cos(j\pi s).$$

The true parameters of the model are $\theta_0 = 3$ and θ defined for all $t \in [0, 1]$ by

$$\theta(t) = \sum_{j=1}^{50} b_j \phi_j(t),$$

with $b_1 = 0.3$ and $b_j = 4(-1)^{j+1} j^{-2}$ for all $j > 1$.

The trajectories of X_i for $i = 1, \ldots, N$ are discretized in $p = 100$ equidistant points. We consider $n = \frac{4}{5}N$ for the training sets $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $n_1 = \frac{1}{5}N$ for the test sets $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+n_1}, Y_{n+n_1})$, where $N = 1400$. In each simulation, we replicated $\boldsymbol{S} = 400$ samples.

2.4.2) Criteria

We used the following criteria, related to the prediction step with the test samples.

- Criterion 1: the mean square prediction errors ($MSPE$) averaged over $\boldsymbol{S}$ samples

$$\overline{MSPE} = \frac{1}{\boldsymbol{S}} \sum_{j=1}^{\boldsymbol{S}} MSPE(j),$$

  where $MSPE(j) = \frac{1}{n_1} \sum_{k=n+1}^{n+n_1} \left(\widehat{\theta}_0 + \langle \widehat{\theta}, X_k^{\star,j} \rangle - \theta_0 - \langle \theta, X_k^{\star,j} \rangle \right)^2$ is the mean square prediction error computed on the j^{th} simulated sample, $j \in \{1, \ldots, \boldsymbol{S}\}$.

- Criterion 2: the ratio respect to truth between the mean square prediction error and the mean square prediction error when the true mean is known averaged over $\boldsymbol{S}$ samples

$$\overline{RT} = \frac{1}{\boldsymbol{S}} \sum_{j=1}^{\boldsymbol{S}} RT(j),$$

  where $RT(j) = \frac{\sum_{k=n+1}^{n+n_1} \left(Y_k^j - \widehat{\theta}_0 - \langle \widehat{\theta}, X_k^{\star,j} \rangle \right)^2}{\sum_{k=n+1}^{n+n_1} \left(\varepsilon_k^j \right)^2}$ is the ratio between the mean square prediction error and the mean square prediction error when the true mean is known, computed on the j^{th} simulated sample.

Notice that all the criteria tend to zero when the sample size tends to infinity. Criterion RT is a rescaled version of $MSPE$ if we substitute the denominator by its limit (specifically, $MSPE(j) = RT(j)\sigma_\epsilon^2$).

2.4.3) Methodology

As in Crambes and Henchiri (2019), we use a smoothed version of the estimator (2.1.2) based on the Smooth Principal Components Regression (SPCR). We use a regression spline basis with parameters: the number κ of knots of the spline functions, the degree q of spline functions and the order m of derivative. Let us remark that, with appropriate conditions, all the theoretical results obtained in our work will also apply when using the SPCR estimation. We take $\kappa = 20$, $q = 3$ and $m = 2$. The choice of these parameters is not crucial in our study, especially in comparison with the choice of the number of principal components (see Crambes and Henchiri (2019) for more details). In this subsection, we firstly present the missing data simulation scenarios for the response and functional covariate. Secondly, we give a procedure to choose the optimal tuning parameter on a growing sequence of dimension $k_n = 2, \ldots, 22$.

Missing data simulation scenario

In our simulations, we have adopted the following scenario to determine the number of missing data on the response Y as in Crambes and Henchiri (2019): we simulate $\delta^{[Y]}$ according to the logistic functional regression. The variable δ follows the Bernoulli law with parameter $p(X)$ such that

$$\log\big(\frac{p(X)}{1-p(X)}\big) = \big\langle \alpha_0, X\big\rangle + c,$$

where $\alpha_0 = \sin(2\pi t)$ for all $t \in [0,1]$ and c is a constant allowing to take different levels of missing data. For exemple $c = 1$ for around 26.97% of missing data, $c = 0.2$ for around 44.99% of missing data and $c = -0.2$ for around 45.087% of missing data.

To deal with partially observed curves for the covariate, we adopted the missing data simulation scenario from Kneip and Liebl (2020) such that

- 70% (respectively 55%) of the curves are fully observed on $[0,1]$,
- for the 30% (respectively 45%) of partially observed curves, the curve X_i is fully observed on $[A_i, B_i] \subset [0,1]$ with A_i drawn with uniform law on the interval $[0, A]$ and $B_i = A_i + B$, with $A = 3/8$ and $B = 6/8$.

Choice of the optimal parameter

Theoretical results are generally obtained under assumptions concerning the rate of convergence of the integer k_n. In practice, this integer is selected by minimizing a certain empirical criterion, for example the Generalized Cross Validation (GCV) criterion, the Cross Validation (CV) criterion and the K-fold Cross Validation (K-fold CV) criterion (see Crambes and Henchiri (2019)). In our simulations, we chose the GCV procedure, known to be computationally fast. The GCV criterion is given below for imputation

$$\text{GCV}(k_n) = \frac{(n - m_n^{[Y]})\sum_{i=1}^{n}(\widehat{Y}_i - \theta_0 - \langle\theta, X_i\rangle)^2\delta_i}{((n - m_n^{[Y]}) - k_n)^2},$$

and the analogous criterion for prediction

$$\text{GCV}(k_n) = \frac{n\sum_{i=1}^{n}(\widehat{Y}_i - \theta_0 - \langle\theta, X_i\rangle)^2}{(n - k_n)^2}.$$

2.4.4) Analysis of results

The criteria were computed according to the different cases listed below.

- Case 1: **FULL**: X and Y are fully observed, this corresponds to the complete reference dataset,

- Case 2: **REC_X_IMP_Y**: X is partially observed and Y is affected with missing values, the missing parts of X are reconstructed and the missing values of Y are imputed, according to the method presented in this paper,

- Case 3: **REC_X_MEAN_IMP_Y**: X is partially observed and Y is affected with missing values, the missing parts of X are reconstructed and the missing values of Y are imputed by the mean of the response observed values,

- Case 4: **REC_X_RAND_IMP_Y**: X is partially observed and Y is affected with missing values, the missing parts of X are reconstructed and the missing values of Y are imputed by a random response observed value,

- Case 5: **REC_X_ZERO_IMP_Y**: X is partially observed and Y is affected with missing values, the missing parts of X are reconstructed and the missing values of Y are imputed by a value equal to zero,

- Case 6: **REC_X_DEL_Y**: X is partially observed and Y is affected with missing values, the missing parts of X are reconstructed and the missing values of Y are removed from the sample,

- Case 7: **DEL_X_DEL_Y**: X is partially observed and Y is affected with missing values, the individuals presenting either a partially observed curve or a missing response are removed from the sample.

Our results are presented in Tables 2.2, 2.3 and 2.4. Other intermediate cases have been examinated (when X is fully observed and Y is affected by missing values, or when X is partially observed and Y is not affected by missing values). Complete results are available on demand to the authors.

As it can be expected, the errors increase as the model error increases. The main point we want to discuss is related to the level of missing data in the sample. Our method (**REC_X_IMP_Y**) always behaves better than the other methods, specially with respect to the imputation with the value zero (**REC_X_ZERO_IMP_Y**) or the more naive methods where we delete missing data on the response (**REC_X_DEL_Y**) or where we delete all missing data (**DEL_X_DEL_Y**). The other imputation methods with the mean (**REC_X_MEAN_IMP_Y**) or with a random value drawn in the observed values (**REC_X_RAND_IMP_Y**) behave better than (**REC_X_DEL_Y**) and (**DEL_X_DEL_Y**). There is a more clear-cut difference between our method and the other ones when the percentage of missing data increases. We can empirically see the advantage of reconstructing the missing parts of the covariate. This echoes to our theoretical results where we remark that the prediction error rate is subordinate to the reconstruction error of the covariate.

Table 2.2: Mean and standard deviation errors for the predicted values based on 400 simulation replications with different levels of missing data and a sample size $N = 1400$. Partially observed curves are fully observed on $[3/8, 6/8]$ and the error $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.2$.

Rate of missing	26.97	26.99	44.99	45.01	55.07	54.91
data in Y in %	(1.41)	(1.22)	(1.47)	(1.58)	(1.50)	(1.43)
Rate of missing	30.03	45.01	30.07	44.81	29.89	44.94
data in X in %	(1.16)	(1.36)	(1.19)	(1.32)	(1.22)	(1.36)
(**FULL**) $\overline{MSPE} \times 10^3$	18.65	17.07	18.45	18.29	18.41	18.61
	(16.48)	(14.23)	(15.87)	(16.30)	(15.96)	(16.94)
$\overline{RT}$	1.47	1.44	1.47	1.46	1.47	1.47
	(0.41)	(0.39)	(0.42)	(0.42)	(0.40)	(0.44)
(**REC_X_IMP_Y**) $\overline{MSPE} \times 10^3$	31.46	30.92	49.33	52.54	68.51	67.95
	(28.39)	(27.65)	(39.08)	(48.53)	(59.54)	(59.67)
$\overline{RT}$	1.79	1.79	2.24	2.31	2.72	2.72
	(0.72)	(0.74)	(0.97)	(1.21)	(1.46)	(1.53)
(**REC_X_MEAN_IMP_Y**) $\overline{MSPE} \times 10^3$	31.58	31.44	52.82	56.15	72.59	70.48
	(27.40)	(25.46)	(36.49)	(40.56)	(42.02)	(44.07)
$\overline{RT}$	1.79	1.80	2.33	2.40	2.83	2.78
	(0.70)	(0.68)	(0.92)	(1.01)	(1.08)	(1.12)
(**REC_X_RAND_IMP_Y**) $\overline{MSPE} \times 10^3$	31.81	31.26	52.31	56.00	72.26	70.49
	(27.68)	(25.19)	(36.01)	(40.90)	(41.86)	(44.39)
$\overline{RT}$	1.80	1.79	2.31	2.40	2.83	2.78
	(0.71)	(0.68)	(0.91)	(1.02)	(1.07)	(1.12)
(**REC_X_ZERO_IMP_Y**) $\overline{MSPE} \times 10^2$	72.31	72.96	194.18	194.74	287.29	286.35
	(8.53)	(8.23)	(15.01)	(14.51)	(16.93)	(17.04)
$\overline{RT}$	19.27	19.38	49.63	49.96	72.94	73.22
	(2.62)	(2.67)	(5.57)	(5.23)	(7.47)	(7.68)
(**REC_X_DEL_Y**) $\overline{MSPE} \times 10^3$	39.55	42.69	72.04	78.83	96.71	96.87
	(32.79)	(36.91)	(53.85)	(59.65)	(70.89)	(75.14)
$\overline{RT}$	1.99	2.08	2.81	2.98	3.42	3.45
	(0.84)	(0.95)	(1.35)	(1.50)	(1.77)	(1.95)
(**DEL_X_DEL_Y**) $\overline{MSPE} \times 10^3$	48.66	57.22	84.33	103.34	112.44	123.62
	(46.92)	(53.12)	(69.73)	(91.15)	(89.72)	(107.01)
$\overline{RT}$	2.21	2.46	3.14	3.61	3.81	4.14
	(1.17)	(1.43)	(1.78)	(2.32)	(2.28)	(2.68)

Table 2.3: Mean and standard deviation errors for the predicted values based on 400 simulation replications with different levels of missing data and a sample size $N = 1400$. Partially observed curves are fully observed on $[3/8, 6/8]$ and the error $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 1.5$.

Rate of missing	27.12	27.14	45.16	45.23	54.91	54.84
data in Y in %	(1.35)	(1.28)	(1.57)	(1.43)	(1.49)	(1.46)
Rate of missing	29.92	45.16	30.00	45.08	30.06	44.83
data in X in %	(1.20)	(1.26)	(1.21)	(1.29)	(1.26)	(1.29)
(**FULL**) $\overline{MSPE} \times 10^3$	23.52	22.89	27.12	22.68	23.44	24.35
	(18.16)	(19.36)	(22.32)	(19.22)	(18.06)	(21.13)
$\overline{RT}$	1.01	1.01	1.01	1.01	1.01	1.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
(**REC_X_IMP_Y**) $\overline{MSPE} \times 10^3$	37.45	34.22	62.68	57.56	76.55	76.61
	(29.46)	(25.32)	(42.74)	(41.20)	(44.91)	(42.59)
$\overline{RT}$	1.02	1.02	1.03	1.03	1.03	1.04
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
(**REC_X_MEAN_IMP_Y**) $\overline{MSPE} \times 10^3$	38.84	36.41	63.90	57.45	79.56	79.80
	(33.00)	(31.36)	(50.87)	(52.04)	(64.31)	(64.38)
$\overline{RT}$	1.02	1.02	1.03	1.03	1.04	1.04
	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)	(0.04)
(**REC_X_RAND_IMP_Y**) $\overline{MSPE} \times 10^3$	39.55	35.75	63.88	60.15	77.23	76.95
	(31.29)	(27.00)	(44.83)	(44.00)	(47.60)	(44.50)
$\overline{RT}$	1.02	1.02	1.03	1.03	1.04	1.04
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
(**REC_X_ZERO_IMP_Y**) $\overline{MSPE} \times 10^2$	73.55	73.31	195.88	195.79	286.40	285.60
	(11.01)	(9.76)	(17.75)	(17.13)	(20.29)	(18.75)
$\overline{RT}$	1.33	1.34	1.88	1.88	2.30	2.29
	(0.09)	(0.09)	(0.16)	(0.15)	(0.20)	(0.20)
(**REC_X_DEL_Y**) $\overline{MSPE} \times 10^3$	44.07	48.93	85.43	81.84	110.13	113.27
	(36.69)	(41.30)	(67.26)	(64.54)	(85.67)	(82.88)
$\overline{RT}$	1.02	1.02	1.04	1.04	1.05	1.05
	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.05)
(**DEL_X_DEL_Y**) $\overline{MSPE} \times 10^3$	63.17	68.44	102.19	115.53	133.12	154.27
	(55.99)	(61.75)	(81.25)	(99.83)	(121.22)	(125.97)
$\overline{RT}$	1.03	1.03	1.05	1.05	1.06	1.07
	(0.04)	(0.04)	(0.05)	(0.06)	(0.07)	(0.08)

Table 2.4: Mean and standard deviation errors for the predicted values based on 400 simulation replications with different levels of missing data and a sample size $N = 1400$. Partially observed curves are fully observed on $[3/8, 6/8]$ and the error ε equals $\eta - 0.5$ with $\eta \sim$ Beta$(2, 2)$.

Rate of missing	26.98	26.90	45.04	45.06	54.96	54.95
data in Y in %	(1.38)	(1.27)	(1.50)	(1.37)	(1.52)	(1.43)
Rate of missing	29.92	45.14	29.89	45.01	30.08	44.92
data in X in %	(1.22)	(1.34)	(1.26)	(1.31)	(1.15)	(1.23)
(**FULL**) $\overline{MSPE} \times 10^3$	19.31	18.89	18.33	18.69	19.26	18.35
	(18.28)	(15.85)	(16.81)	(16.77)	(17.99)	(16.03)
$\overline{RT}$	1.38	1.38	1.37	1.37	1.39	1.37
	(0.38)	(0.33)	(0.35)	(0.34)	(0.37)	(0.34)
(**REC_X_IMP_Y**) $\overline{MSPE} \times 10^3$	32.16	33.62	48.88	52.85	69.84	68.75
	(29.96)	(29.51)	(44.84)	(51.28)	(44.40)	(59.57)
$\overline{RT}$	1.64	1.67	1.98	2.05	2.40	2.39
	(0.62)	(0.60)	(0.92)	(1.03)	(0.90)	(1.21)
(**REC_X_MEAN_IMP_Y**) $\overline{MSPE} \times 10^3$	32.24	34.30	54.38	56.56	70.74	70.03
	(26.73)	(27.56)	(39.92)	(38.24)	(44.25)	(42.84)
$\overline{RT}$	1.64	1.69	2.09	2.15	2.42	2.42
	(0.55)	(0.57)	(0.82)	(0.77)	(0.90)	(0.87)
(**REC_X_RAND_IMP_Y**) $\overline{MSPE} \times 10^3$	32.25	34.49	53.79	56.45	70.20	69.49
	(26.75)	(27.73)	(39.75)	(38.38)	(66.94)	(43.25)
$\overline{RT}$	1.64	1.69	2.08	2.14	2.41	2.41
	(0.55)	(0.57)	(0.82)	(0.78)	(1.38)	(0.88)
(**REC_X_ZERO_IMP_Y**) $\overline{MSPE} \times 10^2$	72.39	71.81	194.32	194.74	286.33	287.001
	(8.34)	(8.32)	(14.90)	(14.08)	(17.00)	(16.60)
$\overline{RT}$	15.57	15.47	40.15	39.98	58.39	58.61
	(1.96)	(1.98)	(4.11)	(3.78)	(5.24)	(4.83)
(**REC_X_DEL_Y**) $\overline{MSPE} \times 10^3$	40.28	41.28	69.12	74.00	98.81	98.99
	(33.20)	(35.94)	(53.94)	(62.68)	(77.89)	(76.62)
$\overline{RT}$	1.81	1.83	2.39	2.48	2.99	3.00
	(0.69)	(0.73)	(1.10)	(1.26)	(1.60)	(1.56)
(**DEL_X_DEL_Y**) $\overline{MSPE} \times 10^3$	49.27	53.20	79.46	95.38	110.30	126.28
	(43.41)	(49.08)	(66.09)	(95.39)	(90.27)	(113.34)
$\overline{RT}$	1.99	2.07	2.60	2.92	3.24	3.58
	(0.91)	(1.00)	(1.37)	(1.92)	(1.89)	(2.39)

2.5) Real dataset study

In this section, we are interested in a model involving electricity production, demand and prices of the German power market. Kneip and Liebl (2020) were already interested in the curve reconstruction problem of electricity prices curves (function of the demand). These data are provided from three different publicly available sources: The European Power Exange (www.epexspot.com), the European Network of Transmission System Operators for Electricity (www.entsoe.eu) and the European Energy Exchange (www.eex-transparency.com). The observation period corresponds to $n = 241$ working days from March 15, 2012 to March 14, 2013. The dataset consists in $n = 241$ daily electricity prices curves in Germany (measured every hour) in function of the residual electricity demand, which is the relevant value for considering electricity demand. It corresponds to germany's gross electricity demand minus infeeds from renewable energy sources plus net-imports from foreign countries. Some prices greater than 120 EUR/MWh have to be treated as outliers since they cannot be explained by the model and were set to the value 120. Negative prices are not impossible in this situation: electricity producers prefer to sell electricity at negative prices (meaning that they are paying for delivering electricity), it is sometimes more profitable than shutting off and restarting a central plant. Figure 2.1 shows the prices curves (in EUR/MWh) in function of the residual demand (in MWh), and Figure 2.2 shows the reconstructed curves with the method from Kneip and Liebl (2020). Price curves can be seen as partially observed curves, as some prices cannot be observed with respect to some residual demand values.

Here, the price-demand functions are observed on different domains. This distinguishes our functional data set from classical functional data sets, where all functions are observed on a common domain. We consider a standardized domain where the standardization can be achieved as follows: for $i = 1, \dots, n$, we consider a sequence from $\min_{1\leqslant j\leqslant p} t_{ij}$ to $\max_{1\leqslant j\leqslant p} t_{ij}$ with a regular step $(b-a)/p$, where $a := \min_{1\leqslant i\leqslant n} \min_{1\leqslant j\leqslant p} t_{ij}$ and $b := \max_{1\leqslant i\leqslant n} \max_{1\leqslant j\leqslant p} t_{ij}$.

Our experimental study is based on two steps. In the first treatment step, we do not observe the price-demand functions directly but we have to estimate each price-demand function by a local polynomial smoother estimator. Here, we choose the Gaussian kernel and we consider a cross validation criterion to select the optimal tuning bandwidth parameter from a grid of parameter values in the interval [1070,35000]. In the second step, we reconstructed the missing parts of the differents curves.

We introduce now the model

$$Y_i = \theta_0 + \langle \theta, X_i \rangle + \varepsilon_i,$$

for $i = 1, \dots, 241$, where X_i is the daily electricity price curve on day i (function of the residual demand), and Y_i is the value of electricity production (in MWh) on day i. The

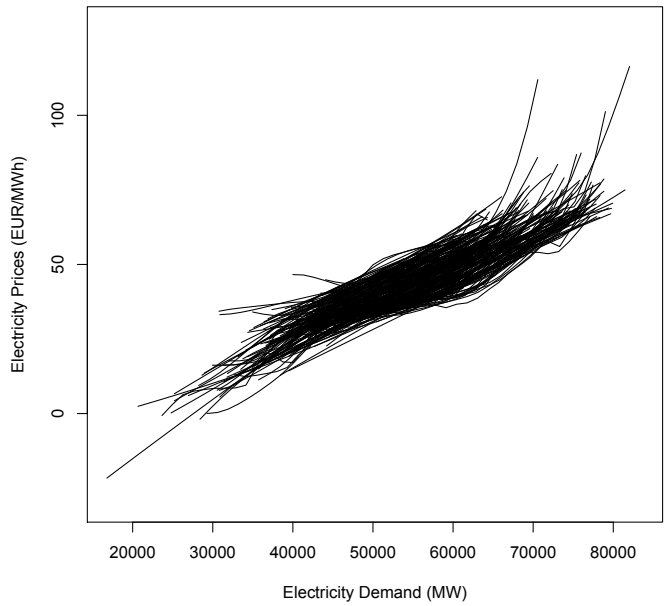

Figure 2.1: Daily electricity price curves in function of the residual demand.

production data come from https://www.agora-energiewende.de[1]. Only a graphic (with numerical values marked at the observation points) was available on this website to collect a data (neither a table nor an Excel file). It can be possible to use a software to get numerical values from a graphic (see https://automeris.io[2]). However, this software is not completely reliable and some numerical values, being not possible, can be considered as missing data for the response variable. In our case, the percentage of missing data is 13.26%.

We split the initial sample into a learning sample (the index set is denoted I_L) with size 181 and a test sample with size 60 (the index set is denoted I_T). Firstly, we reconstructed the missing parts of the differents curves (see Figure 2.2) and, on the learning sample, we imputed the missing values on the response. Then, on the test sample, we computed the prediction values for the response. In order to evaluate the quality of the prediction with our method (**REC_X_IMP_Y**), we calculated the mean square prediction error $MSPE = \frac{1}{60}\sum_{i\in I_T}(Y_i - \widehat{Y}_i)^2 = 40.44$ and the mean absolute error $MAE = \frac{1}{60}\sum_{i\in I_T}|Y_i - \widehat{Y}_i| = 5.35$. As a point of comparison, the MSPE is 40.50 for the method (**REC_X_MEAN_IMP_Y**), 41.40 for the method (**REC_X_RAND_IMP_Y**), 107.95 for the method (**REC_X_ZERO_IMP_Y**) and 40.54 for the method (**REC_X_DEL_Y**).

[1] agora-energiewende.de/en/service/recent-electricity-data/chart/power_generation/15.03.2012/14.03.2013/
[2] automeris.io/WebPlotDigitizer/

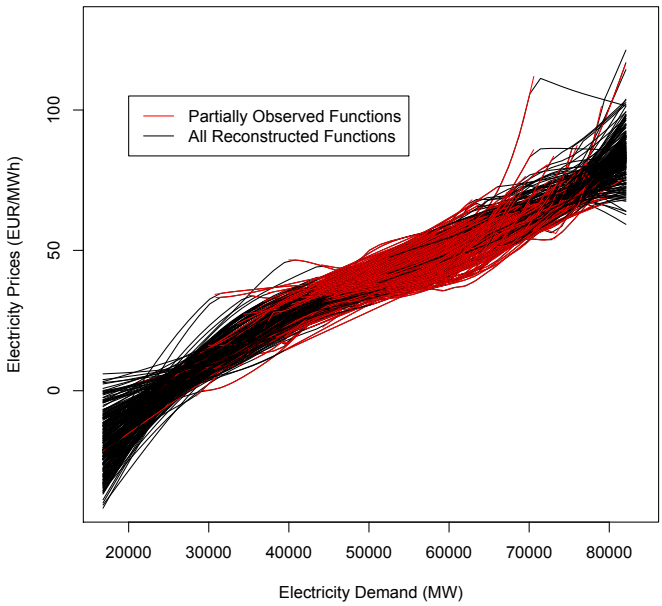

Figure 2.2: Reconstructed daily electricity price curves in function of the residual demand.

The MAE is 5.35 for the method (**REC_X_MEAN_IMP_Y**), 5.37 for the method (**REC_X_RAND_IMP_Y**), 8.89 for the method (**REC_X_ZERO_IMP_Y**) and 5.35 for the method (**REC_X_DEL_Y**). Again, our method performs better than the other ones, even if the differences are sometimes slight. Notice finally that, in this situation, the method (**DEL_X_DEL_Y**) would not be possible since all the curves are partially observed and this would cause removing all individuals in the sample.

2.6) Proofs

Proof of Proposition 2.2.1

For any $x \in H$ such that $\|x\| = 1$, we have

$$\widehat{\Gamma}_{n,rec}x - \widehat{\Gamma}_n x = \frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star - X_i, x\rangle X_i + \frac{1}{n}\sum_{i=1}^{n}\langle X_i, x\rangle (X_i^\star - X_i) + \frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star - X_i, x\rangle (X_i^\star - X_i)\,.$$

Using the Cauchy-Schwarz inequality, we get

$$\|\langle X_i^\star - X_i, x\rangle X_i\| \leqslant \|X_i^\star - X_i\|\,\|x\|\,\|X_i\|\,,$$

from which we deduce with (2.2.3) that

$$\|\langle X_i^\star - X_i, x\rangle X_i\| = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)\,.$$

We prove in the same way that

$$\|\langle X_i, x\rangle (X_i^\star - X_i)\| = \mathcal{O}_p\big(p^{-(a_O-1)/(2(a_O+2))}\big)\,,$$

and

$$\|\langle X_i^\star - X_i, x\rangle (X_i^\star - X_i)\| = \mathcal{O}_p\big(p^{-(a_O-1)/(a_O+2)}\big)\,,$$

which gives the first result (i). The result (ii) can be shown exactly the same way. Finally, we notice that $\widehat{\alpha}_k = \mathcal{O}_p(k^{-a_O-1})$ where we set $\widehat{\alpha}_1 = \widehat{\lambda}_1 - \widehat{\lambda}_2$ and $\widehat{\alpha}_k = \min\left(\widehat{\lambda}_{k-1} - \widehat{\lambda}_k; \widehat{\lambda}_k - \widehat{\lambda}_{k+1}\right)$ for all $k \geqslant 2$. This allows to show results (iii) and (iv) from (i) and respectively Lemma 2.3 and Lemma 2.2 in Horváth and Kokoszka (2012).

Proof of Theorem 2.2.2

We start with the decomposition

$$
\begin{aligned}
&\mathbb{E}\left(\langle\widehat{\widehat{\theta}}, X^{\star}_{new}\rangle - \langle\theta, X^{\star}_{new}\rangle\right)^2 \\
&= \mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Delta}_{n,rec}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X^{\star}_{new} - \Theta X^{\star}_{new}\right)^2 \\
&\leqslant 2\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\Theta\widehat{\Gamma}_{n,rec}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X^{\star}_{new}\right)^2 \\
&+ 2\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^{\star}, .\rangle\varepsilon_i\right)\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X^{\star}_{new} - \Theta X^{\star}_{new}\right)^2.
\end{aligned}
$$

Applying several times the identity $(a+b)^2 \leqslant 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, we get

$$
\begin{aligned}
\mathbb{E}\left(\langle\widehat{\widehat{\theta}}, X^{\star}_{new}\rangle - \langle\theta, X^{\star}_{new}\rangle\right)^2 \leqslant\ & 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,rec}X^{\star}_{new} - \Theta\widehat{\Pi}_{k_n}X^{\star}_{new}\right)^2 \\
&+ 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}X^{\star}_{new} - \Theta\widehat{\Pi}_{k_n}X_{new}\right)^2 \\
&+ 16\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}X_{new} - \Theta\Pi_{k_n}X_{new}\right)^2 \\
&+ 8\mathbb{E}\left(\Theta\Pi_{k_n}X_{new} - \Theta X_{new}\right)^2 \\
&+ 4\mathbb{E}\left(\Theta X_{new} - \Theta X^{\star}_{new}\right)^2 \\
&+ 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^{\star}, \left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X^{\star}_{new}\rangle\varepsilon_i\right)^2.
\end{aligned}
$$

We start with the first term in the above decomposition $A_1 := 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,rec}X^{\star}_{new} - \Theta\widehat{\Pi}_{k_n}X^{\star}_{new}\right)^2$. Applying Lemma 5.1 in Crambes and Henchiri (2019), we obtain

$$
A_1 = o\left(\frac{\widehat{\lambda}_{k_n}k_n^2}{n} + \frac{k_n}{n}\right).
$$

With Lemma 2.2 in Horváth and Kokoszka (2012), we get

$$
A_1 = o\left(\frac{\lambda_{k_n}k_n^2}{n} + \frac{k_n}{n}\right).
$$

Now, we use (2.2.3) to obtain

$$
A_2 := 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}X^{\star}_{new} - \Theta\widehat{\Pi}_{k_n}X_{new}\right)^2 = \mathcal{O}_p\left(p^{-(a_O-1)/(2(a_O+2))}\right).
$$

Moreover, again with Lemma 5.1 in Crambes and Henchiri (2019), we obtain

$$A_3 := 16\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}X_{new} - \Theta\Pi_{k_n}X_{new}\right)^2 = o\left(\frac{\lambda_{k_n}k_n^2}{n} + \frac{k_n}{n}\right).$$

We go on with $A_4 := 8\mathbb{E}\left(\Theta\Pi_{k_n}X_{new} - \Theta X_{new}\right)^2$. With Lemma 5.3 in Crambes and Henchiri (2019), we get

$$A_4 = 8\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2.$$

Next, using again (2.2.3), we can write

$$A_5 := 4\mathbb{E}\left(\Theta X_{new} - \Theta X^{\star}_{new}\right)^2 = \mathcal{O}_p\left(p^{-(a_O-1)/(2(a_O+2))}\right).$$

Finally, the last term of the decomposition comes from Lemma 5.2 in Crambes and Henchiri (2019) and gives

$$A_6 := 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^{\star}, \left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X^{\star}_{new}\rangle\varepsilon_i\right)^2 = \frac{2\sigma_\varepsilon^2 k_n}{n} + o\left(\frac{k_n}{n}\right).$$

We can now conclude the proof of Theorem 2.2.2. The decomposition from the beginning of the proof gives

$$\mathbb{E}\left(\langle\widehat{\widehat{\theta}}, X^{\star}_{new}\rangle - \langle\theta, X^{\star}_{new}\rangle\right)^2 = \mathcal{O}_p\left(\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2 + p^{-(a_O-1)/(2(a_O+2))} + \frac{\sigma_\varepsilon^2 k_n}{n}\right) + o\left(\frac{\lambda_{k_n}k_n^2}{n} + \frac{k_n}{n}\right).$$

The first term in the convergence rate is

$$\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2 = \sum_{j=k_n+1}^{+\infty}\lambda_j\left(\Theta\phi_j\right)^2 \leqslant \sum_{j=k_n+1}^{+\infty} j^{-a_O}.$$

Comparing the latter sum to an integral, we get

$$\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2 = \mathcal{O}\left(k_n^{-(a_O+1)}\right) = \mathcal{O}\left(p^{-(a_O+1)/(a_O+2)}\right) = \mathcal{O}\left(n^{-\eta_1(a_O+1)/(a_O+2)}\right).$$

The second term in the convergence rate is

$$p^{-(a_O-1)/(2(a_O+2))} \sim n^{-\eta_1(a_O-1)/(2(a_O+2))},$$

and the third term in the convergence rate is

$$\frac{\sigma_\varepsilon^2 k_n}{n} \sim \frac{\sigma_\varepsilon^2 n^{\eta_1/(a_O+2)}}{n} = \sigma_\varepsilon^2 n^{\eta_1/(a_O+2)-1}.$$

If we compare the different rates, with the condition $\eta_1 \leqslant 1/2$, we get

$$\mathbb{E}\left(\langle\widehat{\widehat{\theta}}, X_{new}^\star\rangle - \langle\theta, X_{new}^\star\rangle\right)^2 = \mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big).$$

Finally, we can write

$$\begin{aligned}\mathbb{E}\left(\widehat{\widehat{\theta}}_0 + \langle\widehat{\widehat{\theta}}, X_{new}^\star\rangle - \theta_0 - \langle\theta, X_{new}^\star\rangle\right)^2 &= \mathbb{E}\left(\overline{Y} - \theta_0 + \langle\widehat{\widehat{\theta}}, X_{new}^\star\rangle - \langle\theta, X_{new}^\star\rangle\right)^2 \\ &\leqslant 2\mathbb{E}\left(\overline{Y} - \mathbb{E}(Y)\right)^2 + 2\mathbb{E}\left(\langle\widehat{\widehat{\theta}}, X_{new}^\star\rangle - \langle\theta, X_{new}^\star\rangle\right)^2.\end{aligned}$$

The first term of the right-hand side is given by $\mathbb{E}\left(\overline{Y} - \mathbb{E}(Y)\right)^2 = \mathcal{O}_p(n^{-1})$ (with Bienaymé-Tchebychev inequality), and the second term of the right-hand side gives a convergence rate in probability of $n^{-\eta_1(a_O-1)/(2(a_O+2))}$, which gives the desired result

$$\mathbb{E}\left(\widehat{\widehat{\theta}}_0 + \langle\widehat{\widehat{\theta}}, X_{new}^\star\rangle - \theta_0 - \langle\theta, X_{new}^\star\rangle\right)^2 = \mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big).$$

Proof of Theorem 2.3.1

This proof follows the same lines as the proof of Theorem 2.2.2. We write the decomposition

$$\begin{aligned}\mathbb{E}\left(\langle\widetilde{\theta}, X_\ell^\star\rangle - \langle\theta, X_\ell^\star\rangle\right)^2 \leqslant\ & 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,rec}^{obs} X_\ell^\star - \Theta\widehat{\Pi}_{k_n} X_\ell^\star\right)^2 \\ &+ 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}^{obs} X_\ell^\star - \Theta\widehat{\Pi}_{k_n} X_\ell\right)^2 \\ &+ 16\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n} X_\ell - \Theta\Pi_{k_n} X_{new}\right)^2 \\ &+ 8\mathbb{E}\left(\Theta\Pi_{k_n} X_\ell - \Theta X_\ell\right)^2 \\ &+ 4\mathbb{E}\left(\Theta X_\ell - \Theta X_\ell^\star\right)^2 \\ &+ 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \langle X_i^\star, \left(\widehat{\Pi}_{k_n,rec}^{obs}\widehat{\Gamma}_{n,rec}^{obs}\widehat{\Pi}_{k_n,rec}^{obs}\right)^{-1} X_\ell^\star\rangle \delta_i^{[Y]}\varepsilon_i\right)^2.\end{aligned}$$

The first term in the above decomposition $B_1 := 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n,rec}^{obs} X_\ell^\star - \Theta\widehat{\Pi}_{k_n} X_\ell^\star\right)^2$. Applying Lemma 5.1 in Crambes and Henchiri (2019) and Lemma 2.2 in Horváth and Kokoszka (2012), we get

$$B_1 = o\left(\frac{\lambda_{k_n} k_n^2}{n - m_n^{[Y]}} + \frac{k_n}{n - m_n^{[Y]}}\right).$$

Now, we use (2.2.3) to obtain

$$B_2 := 32\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n}^{obs} X_\ell^\star - \Theta\widehat{\Pi}_{k_n} X_\ell\right)^2 = \mathcal{O}_p\left(p^{-(a_O-1)/(2(a_O+2))}\right).$$

Again with Lemma 5.1 in Crambes and Henchiri (2019), we obtain

$$B_3 := 16\mathbb{E}\left(\Theta\widehat{\Pi}_{k_n} X_\ell - \Theta\Pi_{k_n} X_{new}\right)^2 = o\left(\frac{\lambda_{k_n} k_n^2}{n} + \frac{k_n}{n}\right).$$

The next term is $B_4 := 8\mathbb{E}\left(\Theta\Pi_{k_n} X_\ell - \Theta X_\ell\right)^2$. With Lemma 5.3 in Crambes and Henchiri (2019), we get

$$B_4 = 8\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2.$$

Then, using again (2.2.3), we can write

$$B_5 := 4\mathbb{E}\left(\Theta X_\ell - \Theta X_\ell^\star\right)^2 = \mathcal{O}_p\left(p^{-(a_O-1)/(2(a_O+2))}\right).$$

Finally, the last term of the decomposition comes from Lemma 5.2 in Crambes and Henchiri (2019) and gives

$$B_6 := 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star, \left(\widehat{\Pi}_{k_n,rec}^{obs}\widehat{\Gamma}_{n,rec}^{obs}\widehat{\Pi}_{k_n,rec}^{obs}\right)^{-1} X_\ell^\star\rangle \delta_i^{[Y]}\varepsilon_i\right)^2 = \frac{2\sigma_\varepsilon^2 k_n}{n - m_n^{[Y]}} + o\left(\frac{k_n}{n - m_n^{[Y]}}\right).$$

We can now conclude the proof of Theorem 2.3.1. Coming back to the decomposition from the beginning, we get

$$\mathbb{E}\left(\langle\widetilde{\theta}, X_\ell^\star\rangle - \langle\theta, X_\ell^\star\rangle\right)^2 = \mathcal{O}_p\left(\sum_{j=k_n+1}^{+\infty}\left(\Theta\Gamma^{1/2}\phi_j\right)^2 + p^{-(a_O-1)/(2(a_O+2))} + \frac{\sigma_\varepsilon^2 k_n}{n - m_n^{[Y]}}\right) + o\left(\frac{\lambda_{k_n} k_n^2}{n - m_n^{[Y]}} + \frac{k_n}{n - m_n^{[Y]}}\right).$$

Comparing the convergence rates, we obtain

$$\mathbb{E}\left(\langle\widetilde{\theta}, X_\ell^\star\rangle - \langle\theta, X_\ell^\star\rangle\right)^2 = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}}\right).$$

Finally, we can get the desired result including the intercept. We follow the end of the proof of Theorem 2.2.2 to write

$$\mathbb{E}\left(\widetilde{\theta}_0+\langle\widetilde{\theta},X_\ell^\star\rangle-\theta_0-\langle\theta,X_\ell^\star\rangle\right)^2=\mathbb{E}\left(\overline{Y}_{obs}-\theta_0+\langle\widehat{\widetilde{\theta}},X_\ell^\star\rangle-\langle\theta,X_\ell^\star\rangle\right)^2$$
$$\leqslant 2\mathbb{E}\left(\overline{Y}_{obs}-\mathbb{E}(Y)\right)^2+2\mathbb{E}\left(\langle\widetilde{\theta},X_\ell^\star\rangle-\langle\theta,X_\ell^\star\rangle\right)^2.$$

The first term of the right-hand side is given by $\mathbb{E}\left(\overline{Y}_{obs}-\mathbb{E}(Y)\right)^2=\mathcal{O}_p\Big((n-m_n^{[Y]})^{-1}\Big)$ (with Bienaymé-Tchebychev inequality), and the second term of the right-hand side gives a convergence rate in probability of $n^{-\eta_1(a_O-1)/(2(a_O+2))}+\frac{n^{\eta_1/(a_O+2)}}{n-m_n^{[Y]}}$, which gives

$$\mathbb{E}\left(\widetilde{\theta}_0+\langle\widetilde{\theta},X_\ell^\star\rangle-\theta_0-\langle\theta,X_\ell^\star\rangle\right)^2=\mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))}+\frac{n^{\eta_1/(a_O+2)}}{n-m_n^{[Y]}}\right).$$

Proof of Theorem 2.3.3

Following the same lines of previous proofs but first we write the cross covariance operator as

$$\begin{aligned}\widehat{\Delta}_{n,rec}^\star=&\frac{1}{n}\sum_{i=1}^n\langle X_i^\star,.\rangle Y_i^\star\\=&\frac{1}{n}\sum_{i=1}^n\langle X_i^\star,.\rangle\Big(Y_i\delta_i^{[Y]}+Y_{i,imp}(1-\delta_i^{[Y]})\Big)\\=&\frac{1}{n}\sum_{i=1}^n\langle X_i^\star,.\rangle\delta_i^{[Y]}Y_i+\frac{1}{n}\sum_{i=1}^n\langle X_i^\star,.\rangle(1-\delta_i^{[Y]})Y_{i,imp}.\end{aligned}$$

Next, we observe that

$$\begin{aligned}&\mathbb{E}\left(\langle\widehat{\theta}^\star,X_{new}^\star\rangle-\langle\theta,X_{new}^\star\rangle\right)^2\\&=\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Delta}_{n,rec}^\star\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X_{new}^\star-\Theta X_{new}^\star\right)^2\\&\leqslant 2\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\frac{1}{n}\sum_{i=1}^n\langle X_i^\star,.\rangle\delta_i^{[Y]}Y_i\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X_{new}^\star\right)^2\\&+2\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\left(\frac{1}{n}\sum_{i=1}^n\langle X_i^\star,.\rangle Y_{i,imp}(1-\delta_i^{[Y]})\right)\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X_{new}^\star-\Theta X_{new}^\star\right)^2.\end{aligned}$$

The first term is given by the result of Theorem 2.2.2. For the second term

$$\begin{aligned}
&\mathbb{E}\left(\widehat{\Pi}_{k_n,rec}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star,.\rangle Y_{i,imp}(1-\delta_i^{[Y]})\right)\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X_{new}^\star-\Theta X_{new}^\star\right)^2\\
&\leqslant 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star,\left(\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X_{new}^\star(Y_{i,imp}-Y_i)(1-\delta_i^{[Y]})\right)\\
&+2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star,\left(\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X_{new}^\star\rangle Y_i(1-\delta_i^{[Y]})-\Theta X_{new}^\star\right)^2.
\end{aligned}$$

We notice that the first term above is exactly the same as in Theorem 2.3.1 and the second term is directly the result of the Theorem 2.2.2. So, comparing the convergence rates, we get

$$\mathbb{E}\left(\langle\widehat{\theta}^\star,X_{new}^\star\rangle-\langle\theta,X_{new}^\star\rangle\right)^2=\mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))}+\frac{n^{\eta_1/(a_O+2)}}{n-m_n^{[Y]}}\right),$$

which gives the desired result.

CHAPTER 3

MULTIPLE IMPUTATION IN THE FUNCTIONAL LINEAR MODEL WITH PARTIALLY OBSERVED COVARIATE AND MISSING VALUES IN THE RESPONSE

Abstract.

Missing data problems are common and difficult to handle in data analysis. Ad hoc methods such as simply removing cases with missing values can lead to invalid analysis results. In this paper, we consider a functional linear regression model with partially observed covariate and missing values in the response. We use a reconstruction operator that aims at recovering the missing parts of the explanatory curves, then we are interested in regression imputation method of missing data on the response variable, using functional principal component regression to estimate the functional coefficient of the model. We study the asymptotic behavior of the prediction error when missing values in an original dataset are imputed by multiple sets of plausible values.

Keywords.

Functional linear model, Missing data, Functional Principal Components, Missing At Random, Multiple imputation.

3.1) Introduction

Functional data analysis (FDA) can be seen as a important field of statistics that has reached a certain maturity. FDA methods have been applied quite broadly in medicine, science, business, engineering, ..., while new theoretical and methodological developments regularly appear. For a more comprehensive treatment of FDA theory and methods, readers are referred to the classic monographs (Ramsay and Silverman, 2002, 2005; Ramsay et al., 2009), recent monographs (Hsing and Eubank, 2015; Srivastava and Klassen, 2016; Kokoszka and Reimherr, 2018) and review papers (Morris, 2015; Wang et al., 2016).

The functional linear model with scalar response in which a functional random variable is used to predict a real random variable has been the object of considerable attention in the literature. Several procedures have been proposed to the prediction and estimation problems under this model including, for example, functional principal component regression (Febrero-Bande et al., 2017).

This procedure has been considered by many authors Cardot et al. (2003); Hall and Hosseini-Nasab (2006); Cai and Hall (2006); Hall and Horowitz (2007) and Wang et al. (2016). Considering the functional linear regression methodology described in Ramsay and Silverman (2005, Chapter 10), we observe the sample $\mathcal{D}_n \triangleq \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where the X_i's are centered independent and identically distributed with the same law as a random function X taking values in the space $\mathbb{L}_2(\mathcal{I})$ of square integrable functions defined on an interval $\mathcal{I} \subset \mathbb{R}$, and the real responses Y_i's are generated by the regression model

$$Y_i = \alpha + \int_{\mathcal{I}} \theta(t) X_i(t) \mathrm{d}t + \varepsilon_i, \tag{3.1.1}$$

for all $i = 1, \ldots, n$. Here, α is a constant corresponding to the intercept of the model, and θ is a square integrable function belonging to $\mathbb{L}_2(\mathcal{I})$, representing the slope function. It is supposed that the errors ε_i's are independent and identically distributed with finite variance and zero mean and independent from the explanatory variables X_i's.

The functional principal component regression methodology is based on spectral expansions of both the covariance operator of X and its estimator. We define the empirical cross covariance operator $\widehat{\Delta}_n$ given by $\widehat{\Delta}_n u = \frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle Y_i$ for all $u \in \mathbb{L}_2(\mathcal{I})$, the empirical covariance operator $\widehat{\Gamma}_n$ given by $\widehat{\Gamma}_n u = \frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle X_i$ for all $u \in \mathbb{L}_2(\mathcal{I})$. Denoting $(\widehat{\phi}_j)_{j=1,\ldots,k_n}$ the eigenfunctions associated to $\widehat{\Gamma}_n$ corresponding to the k_n highest eigenvalues $\widehat{\lambda}_1 > \ldots > \widehat{\lambda}_{k_n} > 0$ (where k_n is an integer depending on n), we define the orthogonal projection operator $\widehat{\Pi}_{k_n}$ onto the subspace $\mathrm{Span}(\widehat{\phi}_1, \ldots, \widehat{\phi}_{k_n})$ by $\widehat{\Pi}_{k_n} u = \sum_{j=1}^{k_n} \langle \widehat{\phi}_j, u\rangle \widehat{\phi}_j$ for all $u \in \mathbb{L}_2(\mathcal{I})$. Considering

$$\eta(X) \triangleq \alpha + \int_{\mathcal{I}} \theta(t) X(t) \mathrm{d}t,$$

we first estimate η based on a training sample $\mathcal{D}_n$. Let ℓ_n be a functional data fit that measures how well η fits the data. Then, the functional principal component regression estimator $\widehat{\eta}_n$ of η is given by

$$\widehat{\eta}_n \triangleq \operatorname{argmin}_{\eta_0} \left(\ell_n \left(\eta_0 \mid \mathcal{D}_n \right) \right),$$

where the minimization is taken over

$$\left\{ \eta_0 \mid \eta_0(X) = \alpha_0 + \int_{\mathcal{I}} \theta_0(t) X(t) \mathrm{d}t : \alpha_0 \in \mathbb{R}, \theta_0 \in \mathsf{Span}\left(\widehat{\phi}_1, \ldots, \widehat{\phi}_{k_n} \right) \right\}.$$

The most common choice of the functional data fit is the mean square error

$$\ell_n(\eta_0 \mid \mathcal{D}_n) \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \eta_0(X_i) \right)^2.$$

In general, ℓ_n is chosen such to be convex in η_0 and $\mathbb{E}(\ell_n(\eta_0))$ in uniquely minimized by η. Equivalently, the minimization can be taken over (α_0, θ_0) to obtain estimates for both the intercept and slope, denoted by $\widehat{\theta}$ and $\widehat{\alpha}$, as follows

$$\widehat{\theta} = \sum_{j=1}^{k_n} \widehat{\mathsf{s}}_j \widehat{\phi}_j, \quad \text{with} \quad \widehat{\mathsf{s}}_j = \frac{1}{n\widehat{\lambda}_j} \sum_{i=1}^{n} \langle X_i, \widehat{\phi}_j \rangle Y_i, \tag{3.1.2}$$

and $\widehat{\alpha} = \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.
In this work, we focus on the prediction problem. Let $\widehat{\eta}_n$ be a prediction rule given by

$$\widehat{\eta}_n(X_{new}) \triangleq \widehat{\alpha} + \int_{\mathcal{I}} \widehat{\theta}(t) X_{new}(t) \mathrm{d}t,$$

where X_{new} is a copy of X independent of $X_1, \ldots, X_n$. The prediction accuracy can be naturally measured by the excess risk

$$\begin{aligned} \mathcal{E}(\widehat{\eta}_n)(X_{new}) &\triangleq \mathbb{E}^\star \left(\widehat{\eta}_n(X_{new}) - \eta(X_{new}) \right)^2 \\ &= \mathbb{E}^\star \left(\widehat{\alpha} + \langle \widehat{\theta}, X_{new} \rangle - \alpha - \langle \theta, X_{new} \rangle \right)^2, \end{aligned}$$

where $\mathbb{E}^\star$ stands for the expectation with respect to X_{new}.

Earlier works on functional data focused in large part on regular functional data where data are fully observed. This may not always be the case, and missing data appear in many situations, for example when the measuring device breaks down. Many methods for the imputation of missing values have been developed. They can be divided into two branches, *single imputation* and *multiple imputation*. Single imputation consists in creating a single imputed value to replace a missing value. This procedure does not reflect the uncertainty about the prediction of the missing values during the imputation process. Multiple imputation is a statistical technique designed to take advantage in imputing a missing data several times. Each missing value is replaced by two or more

imputed values in order to represent the uncertainty of the value to be imputed. For a comprehensive review of missing data mechanism and imputation methods, we refer the readers to a non-exhaustive list of monographs giving an overview of this topic: Rubin (1987); Graham (2012); Little and Rubin (2020); He et al. (2022).

In recent years, applications producing partially observed functional data have emerged. Sometimes each individual trajectory is collected only over individual-specific subintervals, densely or sparsely, within the whole domain of interest. Several recent works have begun addressing the estimation of covariance functions for short functional segments observed at sparse and irregular grid points, called *functional snippets* (Lin and Wang, Lin and Wang; Lin et al., 2021) or for *fragmented functional data* observed on small subintervals (Delaigle et al., 2020). For densely observed partial data, existing studies have focused on estimating the unobserved part of curves (Kneip and Liebl, 2020; Kraus and Stefanucci, 2020), prediction (Goldberg et al., 2014), classification (Kraus and Stefanucci, 2018; Park, 2019), functional regression (Gellar et al., 2014), and inferences (Kraus, 2019; Park et al., 2021).

To go further, we describe two types of missing data mechanisms that will be the subject of our paper. The first one is related to the real response and the second one is related to the functional covariate. Concerning the missing data mechanism on the real response, we consider a dichotomous random variable $\delta^{[Y]}$ leading to the sample $(\delta_i^{[Y]})_{i=1,\dots,n}$ such that $\delta_i^{[Y]} = 1$ if the value Y_i is available and $\delta_i^{[Y]} = 0$ if the value Y_i is missing, for all $i = 1, \dots, n$. We consider that the data in the response is missing at random (MAR): the fact that the value Y is missing does not depend on the response of the model, but can possibly depend on the covariate, that is,

$$\mathbb{P}(\delta^{[Y]} = 1 \mid X, Y) = \mathbb{P}(\delta^{[Y]} = 1 \mid X).$$

MAR assumption implies that the distribution of Y is the same for units such that $\delta_i^{[Y]} = 1$ (observed units) as for those such that $\delta_i^{[Y]} = 0$ (non-observed units), conditionally on X. As a consequence, the variable $\delta^{[Y]}$ (the fact that an observation is missing or not) is independent of the error of the model ε. In the following, the number of missing values among $Y_1, \dots, Y_n$ is denoted

$$m_n^{[Y]} = \sum_{i=1}^{n} \mathbb{1}_{\left\{\delta_i^{[Y]}=0\right\}}.$$

Concerning the missing data mechanism on the functional covariate, we adopt the paradigm of partially observed functions as in Kneip and Liebl (2020) or Kraus (2015). More precisely, for each curve X_i, $i = 1, \dots, n$, we consider the observed part $O_i \subseteq \mathcal{I}$ of X_i and the missing part $M_i = \mathcal{I} \setminus O_i$. The observed part O_i refers to an interval (or several intervals) where the curve X_i is observed at some measure points of O_i. Based on the punctual observations, the whole curve can be reconstructed on O_i with usual methods (e.g. smoothing splines, regression splines, local polynomial smoothing, ...). On the contrary, no information is available on the missing part M_i. For the rest of

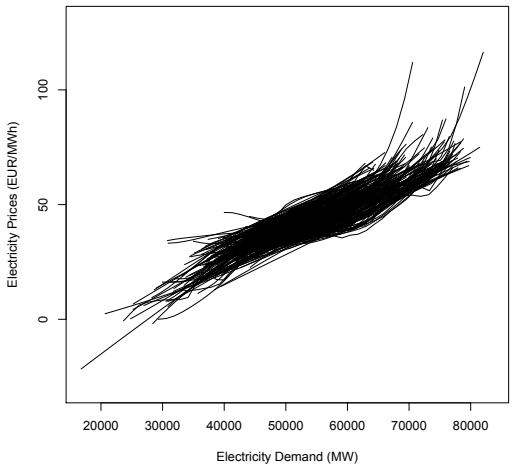

Figure 3.1: Daily electricity price curves in function of the residual demand.

paper, we write "O" and "M" to denote a given production of O_i and M_i. In addition, we denote the observed and missing parts of X_i by X_i^O and X_i^M. As an example, we consider a data set from energy economics presenting demand and prices of the German power market which is shown in Figure 3.1. The data consist of partially observed price functions. The observation period corresponds to 241 working days from March 15, 2012 to March 14, 2013. Price curves can be seen as partially observed curves, as some prices cannot be observed with respect to some residual demand values. Here, the price-demand functions are observed on different domains. This distinguishes our functional data set from classical functional data sets, where all functions are observed on a common domain. We consider a standardized domain where the standardization can be achieved as follows: for $i = 1, \ldots, 241$, we consider a sequence from $\min_{1\leqslant j\leqslant p} t_{ij}$ to $\max_{1\leqslant j\leqslant p} t_{ij}$ with a regular step $(b-a)/p$, where $a := \min\limits_{1\leqslant i\leqslant 241} \min\limits_{1\leqslant j\leqslant p} t_{ij}$ and $b := \max\limits_{1\leqslant i\leqslant 241} \max\limits_{1\leqslant j\leqslant p} t_{ij}$.

The objective of this paper is to predict a new value of the response Y given a new test observation on the explanatory variable X once the partially observed curves X have been reconstructed and the missing data Y have been imputed with the multiple imputation method. More precisely, we want to obtain convergence rates for this prediction error, and we want to analyze how these convergence rates depend on the convergence rates of the reconstruction of the missing parts of the covariate and the convergence rates of the imputation error. We show the difference between the deterministic regression imputation, the random regression imputation and the multiple regression imputation, and its effect on the mean square error of prediction.

In the following, we give in section 3.2 theoretical results of the partially observed covariate. Then, in section 3.3, we study different methods of imputation and the prediction error when the covariate is partially observed and some observations of the real response are affected with missing data. Next, in section 3.4 we give theoretical results related

to the prediction error. In section 3.5, we present some simulation results to show the behavior of the methods in practice. Section 3.6 is devoted to a real dataset application. Finally, all the proofs are postponed to section 3.7.

3.2) Reconstruction of partially observed covariate

In this work, we have to deal with the situation in which some of the real responses of a data set generated from the functional linear model with scalar response are missing at random. This situation has been only considered in Crambes and Henchiri (2019); Febrero-Bande et al. (2019). Other recent works explore this context but in a nonparametric setting (Wang et al., 2019; Rachdi et al., 2020) or in a functional partial linear regression setting (Ling et al., 2019; Zhou and Peng, 2020) or while the response is not missing at random (Li et al., 2018). More recently, Crambes et al. (2022) are interested in a more general case of missing data in functional linear regression: when the covariate is partially observed and when the response is affected by missing data. Following this latter paper Crambes et al. (2022, Subsection 2.1 and Subsection 2.2), $\widehat{\eta}_n$ can be calculated using the curve reconstruction method of Kneip and Liebl (2020, Section 2). We give here some essential elements for our work: we consider a reconstruction problem relating the missing part of the curves to the observed part, writing

$$X_i^M(s) = L(X_i^O(t)) + \mathcal{Z}_i(s),$$

for all $t \in O$ and $s \in M$, where $L : \mathbb{L}_2(O) \to \mathbb{L}_2(M)$ is a linear reconstruction operator and $\mathcal{Z}_i \in \mathbb{L}_2(M)$ is the reconstruction error. Then, the optimal linear reconstruction operator, minimizing the following expected risk

$$\mathbb{E}\left(\left(X_i^M(u) - L(X_i^O)(u)\right)^2\right), \quad \text{for all} \quad u \in M,$$

is given by $\mathcal{L}(X_i^O)(u)$. This operator is estimated in (Kneip and Liebl, 2020, Section 2) by $\widehat{\mathcal{L}}_{k_n}(X_i^O)$, where the truncation parameter k_n is a positive integer that can be fixed automatically with a grid search. Note that the data structure implies that we are faced with two simultaneous estimation problems. One is efficient estimation of $\mathcal{L}(X_i^O)(u)$ for $u \in M$, the other one is the best possible estimation of the function $X_i^O(t)$ for $t \in O$ observed at p discretization points $((W_{i1}, t_{i1}), \ldots, (W_{ip}, t_{ip}))$ with $W_{ij} = X_i^O(t_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, where $t_{ij} \in O$. In order to estimate the curve X_i^O and the covariance function $\gamma_s(t) = Cov(X_i^M(s), X_i^O(t))$ a nonparametric curve estimation by local polynomials smoothers is used. Let κ_1 be a kernel and h_X be a bandwidth of the local linear smoothers of the curve X_i^O. Moreover, let κ_2 be a bivariate kernel and h_γ be a bandwidth of the local linear smoothers of the covariance function γ_s.

The goal is to rebuild a reconstruction function that allows us to recover the full functions from their partial observations. Coming back to the introducing example,

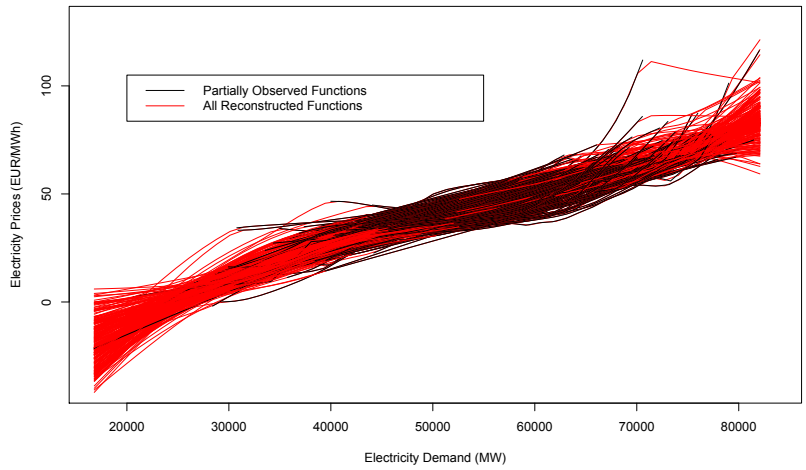

Figure 3.2: Reconstructed daily electricity price curves in function of the residual demand.

Figure 3.2 shows the reconstructed curves with the method from Kneip and Liebl (2020).

In the following, we consider the whole sample $\widetilde{\mathcal{D}}_n \triangleq \left\{(X_1^\star, \delta_1^{[Y]}, Y_1), \ldots, (X_n^\star, \delta_n^{[Y]}, Y_n)\right\}$, with possibly reconstructed explanatory curves

$$X_i^\star(t) = \begin{cases} X_i^O(t) & \text{if } t \in O, \\ \widehat{\mathcal{L}}_{k_n}(X_i^O)(t) & \text{if } t \in M. \end{cases} \tag{3.2.1}$$

Once the curves are reconstructed, we complete missing values in the response with deterministic and random imputation.

3.3) Multiple regression imputation

We may classify regression imputation methods into two classes : *deterministic* (or simple) and *random*. Deterministic regression method yields to a fixed imputed value given the observed sample if the imputation process were repeated as opposed to random methods that do not necessarily yield to the same imputed value. The deterministic method strengthens the relationships in the data and may lead to imputations which seem to be perfect for the model generated from the observed data. However, once the imputation is done, analyses then typically proceed as if the imputed values were the truth. This leads to overly optimistic measures of uncertainty and the potential for substantial bias Van Buuren (2018). To deal with this problem, we consider the random regression imputation that can be seen as a deterministic regression imputation with a random noise $\varepsilon^\star$ (Haziza, 2009, Subsection 2.2). This is a powerful concept, which also builds the basis

of many modern missing values imputation approaches, as it takes into account the inherent uncertainty about missing values. The random noise, $\varepsilon^\star$, is drawn from the observed standardized residuals observed of the prediction errors.

In the following, we are interested in multiple imputation. This method consists in repeating q times the random regression imputation with $q \geqslant 2$. Multiple imputation creates multiple predictions for each missing value, the corresponding statistical analysis takes into account the uncertainty in the imputations and hence, yields to a more reliable standard error. In simple terms, if there is less information in the observed data regarding the missing values, the imputations will be more variable, leading to higher standard errors in the analysis. However, if the observed data allow to predict the missing values, the imputations will be more consistent across the multiple imputed data sets, resulting in smaller and more reliable standard errors Greenland and Finkle (1995). Finally, we will predict a new value under the functional linear model as the mean of all the predictive values.

3.3.1) Deterministic regression imputation

In this section, we follow the same steps as in Crambes et al. (2022). Using the exponent notation "obs" to make reference to the units for which the response is observed, we define the covariance operator with the reconstructed curves (3.2.1) as follows

$$\widehat{\Gamma}_{n,rec}^{obs} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \langle X_i^\star, . \rangle \delta_i^{[Y]} X_i^\star .$$

Let $\widehat{\Pi}_{k_n,rec}^{obs}$ be the projection operator onto the subspace Span$(\widehat{\phi}_{1,rec}^{obs}, \ldots, \widehat{\phi}_{k_n,rec}^{obs})$ where $\widehat{\phi}_{1,rec}^{obs}, \ldots, \widehat{\phi}_{k_n,rec}^{obs}$ are the k_n first eigenfunctions of the covariance operator $\widehat{\Gamma}_{n,rec}^{obs}$. With analogous notations, $\widehat{\lambda}_{1,rec}^{obs}, \ldots, \widehat{\lambda}_{k_n,rec}^{obs}$ represent the k_n first eigenvalues of $\widehat{\Gamma}_{n,rec}^{obs}$.

The functional principal component regression estimator $\widetilde{\eta}_n$ of η is given by

$$\widetilde{\eta}_n \triangleq \text{argmin}_{\widetilde{\eta}_0} \left(\widetilde{\ell}_n \left(\eta_0 \mid \widetilde{\mathcal{D}}_n \right) \right),$$

where the minimization is taken over

$$\left\{ \eta_0 \mid \eta_0(X^\star) = \alpha_0 + \int_{\mathcal{I}} \theta_0(t) X^\star(t) \mathrm{d}t : \alpha_0 \in \mathbb{R}, \theta_0 \in \text{Span} \left(\widehat{\phi}_{1,rec}^{obs}, \ldots, \widehat{\phi}_{k_n,rec}^{obs} \right) \right\},$$

and

$$\widetilde{\ell}_n(\eta_0 \mid \widetilde{\mathcal{D}}_n) \triangleq \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \delta_i^{[Y]} \left(Y_i - \eta_0(X_i^\star) \right)^2 .$$

Equivalently, the minimization can be taken over (α_0, θ_0) to obtain estimates for both the intercept and slope, for imputation, denoted by $\widetilde{\alpha}$ and $\widetilde{\theta}$ such that

$$\widetilde{\alpha} = \overline{Y}_{obs} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \delta_i^{[Y]} Y_i, \tag{3.3.1}$$

and

$$\widetilde{\theta} = \sum_{j=1}^{k_n} \widetilde{\mathsf{s}}_j \widehat{\phi}_{j,rec}^{obs}, \quad \text{with} \quad \widetilde{\mathsf{s}}_j = \frac{1}{(n - m_n^{[Y]})\widehat{\lambda}_{j,rec}^{obs}} \sum_{i=1}^{n} \langle X_i^\star, \widehat{\phi}_{j,rec}^{obs} \rangle \delta_i^{[Y]} Y_i. \tag{3.3.2}$$

For $i = 1, \ldots, n$ such that $\delta_i^{[Y]} = 1$, let $\widehat{Y}_i$ be the predicted value of Y_i given by

$$\widehat{Y}_i \triangleq \widetilde{\alpha} + \int_{\mathcal{I}} \widetilde{\theta}(t) X_i^\star(t) \mathrm{d}t. \tag{3.3.3}$$

Considering a missing value on the response, say Y_ℓ, such that $\delta_\ell^{[Y]} = 0$, we define the imputed value $Y_{\ell,imp}$ by

$$Y_{\ell,imp} = \widetilde{\eta}_n(X_\ell^\star) \triangleq \widetilde{\alpha} + \sum_{j=1}^{k_n} \widetilde{\mathsf{s}}_j \langle X_\ell^\star, \widehat{\phi}_{j,rec}^{obs} \rangle.$$

Finally, we obtain the complete sample $(X_i^\star, Y_i^\star)$ for $i = 1, \ldots, n$, with

$$Y_i^\star = \delta_i^{[Y]} Y_i + \left(1 - \delta_i^{[Y]}\right) Y_{i,imp}. \tag{3.3.4}$$

The imputation accuracy is measured by the excess risk

$$\mathcal{E}(\widetilde{\eta}_n)(X_\ell) = \mathbb{E}^\star \left(\widetilde{\alpha} + \langle \widetilde{\theta}, X_\ell^\star \rangle - \alpha - \langle \theta, X_\ell^\star \rangle \right)^2,$$

where $\mathbb{E}^\star$ stands for the expectation with respect to X_ℓ.

3.3.2) Random regression imputation

We define the missing value Y_ℓ

$$\widetilde{Y}_\ell = \breve{\eta}_n(X_\ell^\star) \triangleq Y_{\ell,imp} + \varepsilon_\ell^\star, \tag{3.3.5}$$

where $\varepsilon_\ell^\star$ is drawn in the set

$$\left\{ e_i \mid e_i = \widetilde{e}_i - \overline{e}, i = 1, \ldots, n, \delta_i^{[Y]} = 1 \right\}, \tag{3.3.6}$$

using (3.3.3) and (3.3.4), we have

$$\widetilde{e}_i = \widetilde{\sigma}^{-1} \left(Y_i^\star - \widehat{Y}_i \right),$$

$$\widetilde{\sigma} = \sqrt{\frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \delta_i^{[Y]} \left(Y_i^\star - \widehat{Y}_i\right)^2},$$

and

$$\overline{e} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \delta_i^{[Y]} \widetilde{e}_i.$$

This method is nonparametric as no distribution is assumed for the distribution of the standardized residuals observed e_i's.
Finally, we obtain the complete sample $(X_i^\star, \breve{Y}_i^\star)$ for $i = 1, \ldots, n$, with

$$\breve{Y}_i^\star = \delta_i^{[Y]} Y_i + \left(1 - \delta_i^{[Y]}\right) \widetilde{Y}_i.$$

Here, the imputation accuracy is measured by the excess risk

$$\mathcal{E}(\breve{\eta}_n)(X_\ell) = \mathbb{E}^\star \left(\widetilde{\alpha} + \langle \widetilde{\theta}, X_\ell^\star \rangle + \varepsilon_\ell^\star - \alpha - \langle \theta, X_\ell^\star \rangle\right)^2.$$

3.3.3) *Multiple regression imputation*

Let i be an index for the observed cases and ℓ be an index for the incomplete cases. The multiple imputation algorithm is sketched as follows:

Algorithm 1 The multiple imputation algorithm

step 1. Estimating parameters $\widetilde{\alpha}$ and $\widetilde{\theta}$ from the functional linear model using complete sample $(X_i^\star, Y_i, \delta_i^{[Y]} = 1)$, for $i = 1, \ldots, n$, as in (3.3.1) and (3.3.2).
step 2.. Drawing $\varepsilon_\ell^{\star(w)}$ from the set of $\left\{e_i \mid e_i = \widetilde{e}_i - \overline{e}, i = 1, \ldots, n, \delta_i^{[Y]} = 1\right\}$, as in (3.3.6), for $\ell \in \widetilde{\mathcal{D}}_m$, where $\widetilde{\mathcal{D}}_m$ is the set of missing responses of size $m_n^{[Y]}$.
step 3. Drawing the imputed values of missing data, as in (3.3.5), from

$$\widetilde{Y}_\ell^{(w)} = \widetilde{\alpha} + < \widetilde{\theta}, X_\ell^\star > + \varepsilon_\ell^{\star(w)},$$

for $\ell \in \widetilde{\mathcal{D}}_m$.
step 4. Repeat Steps 2 to 3, q times independently to create multiple sets of imputations $(w = 1, ..., q)$.

Finally, we obtain the multiple sets of complete data $(X_i^\star, Y_i^{\star(w)})$, for $w = 1, \ldots, q$, with

$$Y_i^{\star(w)} = \delta_i^{[Y]} Y_i + \left(1 - \delta_i^{[Y]}\right) \widetilde{Y}_i^{(w)}.$$

Here, the imputation accuracy is measured by the excess risk

$$\mathcal{E}(\breve{\eta}_n)(X_\ell) = \mathbb{E}^\star \left(\frac{1}{q} \sum_{w=1}^{q} \left(\widetilde{\alpha} + < \widetilde{\theta}, X_\ell^\star > + \varepsilon_\ell^{\star(w)}\right) - \alpha - \langle \theta, X_\ell^\star \rangle\right)^2.$$

3.3.4) Prediction

Once the whole database has been reconstructed, we obtain estimates for both the intercept and slope, denoted by $(\widehat{\alpha}^\star, \widehat{\theta}^\star)$ and $(\breve{\alpha}^\star, \breve{\theta}^\star)$ respectively after deterministic regression imputation and after random regression imputation such that

$$\widehat{\alpha}^\star = \frac{1}{n}\sum_{i=1}^{n} Y_i^\star,$$

$$\widehat{\theta}^\star = \sum_{j=1}^{k_n} \widehat{s}_j^\star \widehat{\phi}_{j,rec}^\star, \quad \text{with} \quad \widehat{s}_j^\star = \frac{1}{n\widehat{\lambda}_{j,rec}^\star}\sum_{i=1}^{n}\langle X_i^\star, \widehat{\phi}_{j,rec}^\star\rangle Y_i^\star, \tag{3.3.7}$$

$$\breve{\alpha}^\star = \frac{1}{n}\sum_{i=1}^{n} \breve{Y}_i^\star,$$

$$\breve{\theta}^\star = \sum_{j=1}^{k_n} \breve{s}_j^\star \widehat{\phi}_{j,rec}^\star, \quad \text{with} \quad \breve{s}_j^\star = \frac{1}{n\widehat{\lambda}_{j,rec}^\star}\sum_{i=1}^{n}\langle X_i^\star, \widehat{\phi}_{j,rec}^\star\rangle \breve{Y}_i^\star, \tag{3.3.8}$$

where $\widehat{\phi}_{1,rec}^\star, \ldots, \widehat{\phi}_{k_n,rec}^\star$ and $\widehat{\lambda}_{1,rec}^\star, \ldots, \widehat{\lambda}_{k_n,rec}^\star$ represent respectively the k_n first eigenfunctions and eigenvalues of the covariance operator $\widehat{\Gamma}_{n,rec}^\star = \frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star, .\rangle X_i^\star$.

In multiple regression imputation setting, for $w = 1, \ldots, q$, given either the observed values or the random imputations $Y_1^{\star(w)}, \ldots, Y_n^{\star(w)}$, we estimate the parameters α and θ in model (3.1.1) with

$$\widehat{\alpha}^{(w)} = \frac{1}{n}\sum_{i=1}^{n} Y_i^{\star(w)}$$

and

$$\widehat{\theta}^{(w)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}_{j,rec}^\star\rangle Y_i^{\star(w)}}{\widehat{\lambda}_{j,rec}^\star}\widehat{\phi}_{j,rec}^\star = \sum_{j=1}^{k_n} \widehat{\mathsf{s}}_j^{(w)}\widehat{\phi}_{j,rec}^\star, \tag{3.3.9}$$

with

$$\widehat{\mathsf{s}}_j^{(w)} = \frac{1}{n\widehat{\lambda}_{j,rec}^\star}\sum_{i=1}^{n}\langle X_i^\star, \widehat{\phi}_{j,rec}^\star\rangle Y_i^{\star(w)}.$$

For a new curve X_{new}, we predict the response value as follows

$$\widehat{Y}_{new} = \frac{1}{q}\sum_{w=1}^{q} \widehat{Y}_{new}^{\star(w)},$$

where

$$\widehat{Y}_{new}^{\star(w)} = \widehat{\alpha}^{(w)} + \langle \widehat{\theta}^{(w)}, X_{new}^\star\rangle.$$

An asymptotic behavior of the prediction error is given in Crambes et al. (2022) when the missing parts of the covariate are reconstructed and the missing values on the response are imputed by deterministic regression imputation. In the next section, we will study the convergence rate of this prediction error with multiple regression imputation.

3.4) Theoretical results

3.4.1) Assumptions

In this subsection, we give the assumptions needed for our theoretical results. Some assumptions are used in Kneip and Liebl (2020) and Crambes et al. (2022) in order to control the curve reconstruction of the covariate.

(A.1) Let $np \to \infty$ when $n \to \infty$ and $p = p(n)$. We assume $p = n^{\eta_1}$ with $0 < \eta_1 < \infty$ in the following.

(A.2) • For any subinterval $O \subseteq \mathcal{I}$, we assume that the eigenvalues $\lambda_1 > \lambda_2 > \ldots > 0$ have multiplicity one. Moreover, we assume that there exist $a_O > 1$ and $0 < c_O < \infty$ such that (i) $\lambda_k^O - \lambda_{k+1}^O \geqslant c_O k^{-a_O-1}$, (ii) $\lambda_k^O = \mathcal{O}(k^{-a_O})$, (iii) $1/\lambda_k^O = \mathcal{O}(k^{a_O})$ as $k \to \infty$.
• $\mathbb{E}\left(\xi_k^4\right) = \mathcal{O}(\lambda_k^2)$.

(A.3) For any subinterval $O \subseteq \mathcal{I}$, we assume that there exists $0 < D_O < \infty$ such that the eigenfunctions satisfy $\sup_{t\in\mathcal{I}} \sup_{k\geqslant 1} |\widetilde{\phi}_k^O(t)| \leqslant D_O$, where $\widetilde{\phi}_k^O(s) = \langle \phi_k^O, \gamma_s \rangle / \lambda_k^O$.

(A.4) The bandwidth h_X satisfies $h_X \to 0$ and $(ph_X) \to \infty$ as $p \to \infty$. For instance, we assume that $h_X = \frac{1}{n^{\eta_2}}$ with $0 < \eta_2 < \eta_1$. The bandwidth h_γ satisfies $h_\gamma \to 0$ and $(n(p^2-p)h_\gamma) \to \infty$ as $n(p^2-p) \to \infty$. For example, we can take $h_\gamma = \frac{1}{n^{\eta_3}}$ with $0 < \eta_3 < 2\eta_1 + 1$.

(A.5) Let κ_1 and κ_2 be nonnegative, second order univariate and bivariate kernel functions with support $[-1, 1]$. For example, we can use univariate and bivariate Epanechnikov kernel functions with compact support $[-1, 1]$, namely $\kappa_1(x) = \frac{3}{4}(1-x^2)\mathbb{1}_{[-1,1]}(x)$ and $\kappa_2(x, y) = \frac{9}{16}(1-x^2)(1-y^2)\mathbb{1}_{[-1,1]}(x)\mathbb{1}_{[-1,1]}(y)$.

(A.6) The random variables X and Y are almost surely bounded, respectively in $\mathbb{L}_2(\mathcal{I})$ and $\mathbb{R}$.

Assumption **(A.1)** is mild and can be satisfied even if the number of observation points p does not go fast to infinity. Assumptions **(A.2)** and **(A.3)**, related to eigenvalues and eigenfunctions of the covariance operator of X, are given in Kneip and Liebl (2020) in order to control the curve reconstruction for the covariate. In particular, a polynomial decrease of the eigenvalues is required, allowing a large class of eigenvalues for the covariance operator of X. Assumptions **(A.4)** and **(A.5)** are classic in the context of local polynomials smoothers. For Assumption **(A.6)**, we can find in practice a large enough interval such that it is satisfied.

3.4.2) Asymptotic result

To start this subsection, we give the main result from Crambes et al. (2022) for the prediction error when the missing parts of the covariate are reconstructed and the completion of the missing data in the response is done by deterministic imputation. Let Y_{new} be the predicted value of the response given a new observation X_{new} of the covariate.

Proposition 3.4.1. *Under assumptions* **(A.1)-(A.6)**, *and* $k_n \sim p^{1/(a_O+2)}$ *and* $p \sim n^{\eta_1}$ *with* $\eta_1 \leqslant 1/2$, *the prediction error, based on the deterministic regression imputation, is*

$$\mathbb{E}\left(\widehat{\alpha} + \langle \widehat{\theta}, X^{\star}_{new}\rangle - \alpha - \langle \theta, X^{\star}_{new}\rangle\right)^2 = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}}\right).$$

In the particular case where $\eta_1 = 1/2$, *the first term in the convergence rate is* $\mathcal{O}_p\big(n^{-(a_O-1)/(4(a_O+2))}\big)$.

This result shows that the prediction error rate with the deterministic regression imputation in the response is subordinate to the reconstruction error of the covariate. We now give our main result.

Theorem 3.4.2. *Under assumptions* **(A.1)-(A.6)**, *if we additionally take* $k_n \sim p^{1/(a_O+2)}$ *and* $p \sim n^{\eta_1}$ *with* $\eta_1 \leqslant 1/2$, *as well as* $m_n^{[Y]} = o\big(n^{1-\eta_1(a_O+3)/4(a_O+2)}\big)$, *the prediction error, based on the multiple regression imputation, is*

$$\mathbb{E}\left(\widehat{Y}_{new} - \alpha - \langle \theta, X^{\star}_{new}\rangle\right)^2 = \mathcal{O}_p\left(\frac{n^{-\eta_1(a_O-1)/(2(a_O+2))}}{q} + \frac{n^{\eta_1/(a_O+2)}}{q(n - m_n^{[Y]})}\right).$$

This result, giving the convergence rate of the prediction error after q random imputations, is asymptotically comparable to the convergence rate obtained in Proposition 3.4.1 in the case of a deterministic regression imputation. We let the value of q appear in the convergence rate to highlight the fact that the constant when the convergence rate should be better in the case of several random imputations instead of a single deterministic one.

Remark 3.4.3. *Theoretical results are generally obtained under assumptions concerning the rate of convergence of the integer* k_n. *In practice, this integer is selected by minimizing a certain empirical criterion. We chose the Generalized Cross Validation (GCV) procedure, known to be computationally fast. The GCV criterion is given below for imputation*

$$GCV(k_n) = \frac{(n - m_n^{[Y]}) \sum_{i=1}^{n} \left(\widetilde{\alpha} + \langle \widetilde{\theta}, X_i^{\star}\rangle - \alpha - \langle \theta, X_i^{\star}\rangle\right)^2 \delta_i^{[Y]}}{\left((n - m_n^{[Y]}) - k_n\right)^2},$$

and the analogous criterion for prediction

$$GCV(k_n) = \frac{n\sum_{i=1}^{n}\left(\widehat{\alpha}^{(w)} + \langle\widehat{\theta}^{(w)}, X_i^{\star}\rangle - \alpha - \langle\theta, X_i\rangle\right)^2}{\left(n - k_n\right)^2}, \quad \textit{for } w = 1, \ldots, q.$$

3.5) Simulations

3.5.1) Methodology

We generated the functional covariate in a similar way to that adopted in Hall and Horowitz (2007). More specifically, the functional covariates were identically and independently generated as:

$$X_i(t) = \sum_{j=1}^{150} \zeta_{ij}\varrho_j\phi_j(t), \quad i = 1, \ldots, N,$$

where $\phi_1 \equiv 1$, $\phi_{j+1} = \sqrt{2}\cos(j\pi t)$, for $j \geqslant 2$, the ϱ_j's are defined by $\varrho_j = (-1)^{j+1}(j)^{-2}$ and the ζ_j's are independently sampled from the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$. The covariance function writes

$$cov(X(t), X(s)) = \sum_{j=1}^{150} \frac{2}{j^4}\cos(j\pi t)\cos(j\pi s).$$

These covariates are sampled at $p = 100$ equally spaced points between 0 and 1. The responses are generated from (3.1.1), where $\alpha = 3$ and θ defined, for all $t \in [0, 1]$, by

$$\theta(t) = \sum_{j=1}^{50} b_j\phi_j(t),$$

where $b_1 = 0.3$ and $b_j = 4(-1)^{j+1}j^{-2}$ for all $j > 1$. The random errors, ε_i's, are generated as $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon^2 = 0.2$. In each simulation replicate we randomly generate $n = \frac{4}{5}N$ independent copies of (X_i, Y_i) for training and $n_1 = \frac{1}{5}N$ copies for testing, with $N = 1400$. To better assess prediction performance of model, we repeat the simulation procedure $\boldsymbol{S} = 250$ times.

To deal with partially observed curves for the covariate, we adopted the missing data simulation scenario from Crambes et al. (2022) such that

- 70% (respectively 55%) of the curves are fully observed on $[0, 1]$,
- for the 30% (respectively 45%) of partially observed curves, the curve X_i is fully observed on $[A_i, B_i] \subset [0, 1]$ with A_i drawn with uniform law on the interval $[0, A]$ and $B_i = A_i + B$, with $A = 3/8$ and $B = 6/8$.

We simulate the number of missing data on the response Y and the indicator $\delta^{[Y]}$ by the logistic functional regression. The variable δ follows the Bernoulli law with parameter $p(X)$ such that

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \langle c, X\rangle + ct,$$

where $c = \sin(2\pi t)$ for all $t \in [0,1]$ and ct is a constant allowing to take different levels of missing data. For exemple $ct = 1$ for around 26.903% of missing data, $ct = 0.2$ for around 44.941% of missing data and $ct = -0.2$ for around 54.793% of missing data.

We estimate the parameters of the model and we obtain the predicted values of the response with imputation methods. Notice that, we use a smoothed version of the different estimators (3.1.2), (3.3.2), (3.3.7), (3.3.8) and (3.3.9) based on the Smooth Principal Components Regression (SPCR). Let us remark that, with appropriate conditions, all the theoretical results obtained in our work will also apply when using the SPCR estimation. We use a regression spline basis with 20 knots, a degree 3 and the order of derivation 2. The choice of these parameters is not crucial in our study, especially in comparison with the choice of the number of principal components. The choice of this optimal tuning parameter is made on a growing sequence of dimension $k_n = 2, \ldots, 22$.

3.5.2) Criteria

Our objective is to predict the response in the test samples. We use two criteria.

- Criterion 1: the average mean square prediction error

$$\overline{MSPE} = \frac{1}{\boldsymbol{S}}\sum_{j=1}^{\boldsymbol{S}} MSPE(j),$$

where $MSPE(j) = \frac{1}{n_1}\sum_{k=n+1}^{n+n_1}\left(\widehat{\alpha} + \langle\widehat{\theta}, X_k^{\star,j}\rangle - \alpha - \langle\theta, X_k^{\star,j}\rangle\right)^2$ is the mean square prediction error computed on the j^{th} simulated sample, $j \in \{1, \ldots, \boldsymbol{S}\}$. The criterion $\overline{MSPE}$ tends to zero when the sample size tends to infinity.

- Criterion 2: the average ratio respect to truth, based on a deterministic regression imputation,

$$\overline{RT} = \frac{1}{\boldsymbol{S}}\sum_{j=1}^{\boldsymbol{S}} RT(j),$$

where $RT(j) = \frac{\sum_{k=n+1}^{n+n_1}\left(\widehat{\alpha} + \langle\widehat{\theta}, X_\ell^{\star,j}\rangle - Y_k^j\right)^2}{\sum_{k=n+1}^{n+n_1}\left(\varepsilon_k^j\right)^2}$ is the ratio between the mean square prediction error and the mean square prediction error when the true parameters are known, computed on the j^{th} simulated sample. The criterion $\overline{RT}$ tends to one when the sample size tends to infinity.

Table 3.1: Mean and standard deviation errors for the predicted values based on 250 simulation replications with different levels of missing data and a sample size $N = 1400$. Partially observed curves are fully observed on $[3/8, 6/8]$ and the error ε is a Gaussian noise: $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.2$.

Rate of missing	26.903	26.877	44.941	45.218	54.793	55.109
data in Y in %	(1.298)	(1.409)	(1.563)	(1.515)	(1.337)	(1.460)
Rate of missing	30.047	44.952	29.995	45.030	30.086	45.164
data in X in %	(1.112)	(1.230)	(1.238)	(1.280)	(1.216)	(1.317)
(**FULL**) $\overline{MSPE} \times 10^3$	17.602	16.785	18.145	16.960	19.150	18.055
	(16.058)	(15.640)	(15.990)	(13.681)	(16.149)	(15.709)
$\overline{RT} \times 10$	14.421	14.231	14.639	14.144	14.733	14.580
	(4.337)	(3.887)	(4.134)	(3.405)	(4.042)	(4.096)
(**DETER_IM**) $\overline{MSPE} \times 10^3$	30.786	29.748	51.942	48.223	66.907	70.525
	(28.722)	(27.327)	(47.172)	(44.261)	(57.530)	(67.268)
$\overline{RT} \times 10$	17.751	17.540	23.320	21.925	26.695	27.758
	(7.624)	(6.914)	(12.195)	(10.902)	(14.482)	(16.921)
(**RAND_IM**) $\overline{MSPE} \times 10^3$	45.463	45.833	67.350	65.721	85.999	90.581
	(36.723)	(39.229)	(50.395)	(49.144)	(66.256)	(73.653)
$\overline{RT} \times 10$	2.138	2.160	2.716	2.623	3.139	3.286
	(0.959)	(1.018)	(1.304)	(1.219)	(1.664)	(1.878)
(**RAND_NORM_IM**) $\overline{MSPE} \times 10^3$	30.732	29.927	52.298	48.412	67.055	70.693
	(28.284)	(27.798)	(47.330)	(44.427)	(58.103)	(67.278)
$\overline{RT} \times 10$	17.735	17.589	23.411	21.981	26.721	27.799
	(7.505)	(7.024)	(12.237)	(10.972)	(14.605)	(16.941)
(**MUL_IM** (q=5)) $\overline{MSPE} \times 10^3$	34.405	31.449	55.165	52.568	69.329	74.675
	(29.976)	(28.133)	(48.511)	(44.218)	(56.368)	(69.281)
$\overline{RT} \times 10$	18.663	17.978	24.166	23.058	27.310	28.763
	(7.875)	(7.147)	(12.561)	(10.911)	(14.194)	(17.467)
(**MUL_NORM_IM** (q=5)) $\overline{MSPE} \times 10^3$	30.819	29.698	51.988	48.054	66.978	70.689
	(28.606)	(27.233)	(47.249)	(44.095)	(57.747)	(67.771)
$\overline{RT} \times 10$	17.756	17.526	23.332	21.885	26.713	27.797
	(7.603)	(6.895)	(12.206)	(10.829)	(14.550)	(17.032)
(**MUL_IM** (q=10)) $\overline{MSPE} \times 10^3$	30.998	30.255	53.437	49.395	68.639	73.111
	(28.554)	(27.931)	(47.390)	(44.601)	(56.621)	(67.923)
$\overline{RT} \times 10$	17.807	17.667	23.692	22.224	27.125	28.386
	(7.640)	(7.053)	(12.223)	(10.934)	(14.196)	(17.100)
(**MUL_NORM_IM** (q=10)) $\overline{MSPE} \times 10^3$	30.680	29.627	51.890	48.178	66.629	70.699
	(28.664)	(27.221)	(47.210)	(44.347)	(57.206)	(67.557)
$\overline{RT} \times 10$	17.721	17.510	23.304	21.915	26.620	27.801
	(7.601)	(6.891)	(12.200)	(10.926)	(14.392)	(16.982)

Table 3.2: Mean and standard deviation errors for the predicted values based on 250 simulation replications with different levels of missing data and a sample size $N = 1400$. Partially observed curves are fully observed on $[3/8, 6/8]$ and the error ε is a Gaussian noise: $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.2$.

(**MUL_IM** (q=30)) $\overline{MSPE} \times 10^3$	30.326	29.207	**51.259**	48.253	**66.384**	71.054
	(28.978)	(26.953)	**(47.133)**	(44.117)	**(56.415)**	(67.691)
$\overline{RT} \times 10$	17.627	17.419	**23.149**	21.919	**26.547**	27.884
	(7.713)	(6.837)	**(12.183)**	(10.857)	**(14.189)**	(17.048)
(**MUL_NORM_IM** (q=30)) $\overline{MSPE} \times 10^3$	30.695	29.731	51.951	48.029	66.812	70.553
	(28.701)	(27.490)	(47.299)	(44.102)	(57.459)	(67.418)
$\overline{RT} \times 10$	17.726	17.538	23.324	21.876	26.669	27.765
	(7.619)	(6.954)	(12.226)	(10.867)	(14.467)	(16.957)
(**MUL_IM** (q=100)) $\overline{MSPE} \times 10^3$	**30.114**	**29.130**	51.605	**47.395**	66.611	**70.374**
	(28.662)	**(27.545)**	(47.355)	**(43.987)**	(57.918)	**(67.553)**
$\overline{RT} \times 10$	**17.57**4	**17.392**	23.225	**21.719**	26.620	**27.730**
	(7.614)	**(6.953)**	(12.234)	**(10.835)**	(14.573)	**(16.988)**
(**MUL_NORM_IM**(q=100)) $\overline{MSPE} \times 10^3$	30.700	29.693	51.913	48.110	66.742	70.507
	(28.663)	(27.370)	(47.154)	(44.196)	(57.487)	(67.309)
$\overline{RT} \times 10$	17.727	17.527	23.314	21.897	26.652	27.755
	(7.608)	(6.924)	(12.192)	(10.884)	(14.474)	(16.929)
(**MEAN_IM**) $\overline{MSPE} \times 10^3$	30.913	30.746	54.404	53.923	74.127	74.266
	(24.018)	(23.151)	(37.310)	(37.364)	(44.807)	(45.275)
$\overline{RT} \times 10$	17.755	17.820	23.961	23.374	28.672	28.768
	(6.199)	(6.153)	(9.777)	(9.314)	(11.481)	(11.631)
(**RANDO_IM**) $\overline{MSPE} \times 10^3$	30.870	30.909	54.121	54.127	73.924	73.478
	(24.108)	(23.192)	(37.878)	(37.973)	(45.867)	(44.904)
$\overline{RT} \times 10$	17.740	17.852	23.885	23.433	28.622	28.568
	(6.214)	(6.312)	(9.897)	(9.480)	(11.771)	(11.563)
(**ZERO_IM**) $\overline{MSPE} \times 10^2$	72.025	71.648	194.283	195.935	283.638	287.324
	(8.039)	(7.951)	(14.874)	(14.811)	(15.420)	(17.570)
$\overline{RT} \times 10$	191.134	190.669	501.892	501.050	728.003	736.894
	(24.954)	(28.526)	(55.713)	(58.105)	(71.678)	(70.364)
(**REM_Y**) $\overline{MSPE} \times 10^3$	40.047	37.844	78.278	72.577	94.632	100.989
	(34.437)	(32.908)	(58.280)	(61.814)	(71.559)	(81.000)
$\overline{RT} \times 10$	20.052	19.568	29.985	28.085	33.687	35.381
	(8.923)	(8.399)	(15.204)	(15.412)	(18.126)	(20.334)
(**REM_X,Y**) $\overline{MSPE} \times 10^3$	48.448	60.808	91.500	90.137	117.749	135.675
	(47.901)	(61.016)	(74.047)	(81.080)	(94.352)	(126.053)
$\overline{RT} \times 10$	22.284	25.280	33.257	32.779	39.728	44.123
	(13.086)	(15.150)	(18.983)	(20.142)	(24.442)	(31.855)

3.5.3) Results

Tables (3.1) and (3.2) presents the criteria for the complete dataset (**FULL**) and the imputation methods presented in this paper, with reconstructed curves :

- **DETER_IM** : Deterministic regression imputation, as described in subsection 3.3.1.
- **RAND_IM** : Random regression imputation, as described in subsection 3.3.2.
- **RAND_NORM_IM** : Parametric approach of random regression imputation, where the error term $\varepsilon^\star$ is drawn from the distribution of the residuals, here assuming the residuals are normally distributed, thus $\varepsilon^\star \sim N(0, \widehat{\sigma}^2_{\varepsilon^\star})$, with $\widehat{\sigma}^2_{\varepsilon^\star}$ being estimated from the residuals of the formerly fitted functional linear model. This parametric method is easy to implement. It seems natural to test the performance of this method on simulations.
- **MUL_IM** : Multiple regression imputation with different values of q ($q = 5, 10, 30, 100$), as described in subsection 3.3.3.
- **MUL_NORM_IM** : Parametric approach of multiple regression imputation with different values of q ($q = 5, 10, 30, 100$). Here, the error term $\varepsilon^\star$ is drawn as described above, thus $\varepsilon^\star \sim N(0, \widehat{\sigma}^2_{\varepsilon^\star})$.
- **MEAN_IM** : Mean imputation,
- **RANDO_IM** : Random imputation (imputation by a random response drawn in the set of observed values),
- **ZERO_IM** : Zero imputation (imputation by zero).

Moreover, we propose two other cases :

- **REM_Y** : Reconstruct X and remove all the missing values in Y from the sample,
- **REM_X,Y** : Either a partially observed curve or a missing response are removed from the sample.

As it can be expected, the errors increase as the percentage of missing values in X and Y increase. Moreover, when the number of iterations q increases, we recover the $\overline{MSPE}$ and $\overline{RT}$ of the deterministic imputation (**DETER_IM**). Furthermore, when q is large enough ($q = 30$ and $q = 100$), our method (**MUL_IM**) behaves better than the other imputation methods, specially where we delete the missing values (**REM_Y** and **REM_X,Y**). Comparing (**MUL_IM**) and (**MUL_NORM_IM**), we notice that (**MUL_NORM_IM**), behaves better for small values of q while (**MUL_IM**) behaves better for larger values of q.

3.6) Real dataset study

Our experimental study is based on two steps. In the first treatment step, we do not observe the price-demand functions directly but we have to estimate each price-demand function by a local polynomial smoother estimator. Here, we choose the Gaussian kernel and we consider a cross validation criterion to select the optimal tuning bandwidth parameter from a grid of parameter values in the interval [1070,35000]. In the second step, we reconstructed the missing parts of the differents curves. Now, X_i, $i = 1, \ldots, 241$, is the daily electricity price curve on day i (function of the residual demand), and Y_i is the value of electricity production (in MWh) on day i. The production data come from https://www.agora-energiewende.de[1]. Only a graphic (with numerical values marked at the observation points) was available on this website to collect a data (neither a table nor an Excel file). It can be possible to use a software to get numerical values from a graphic (see https://automeris.io[2]). However, this software is not completely reliable and some numerical values, being not possible, can be considered as missing data for the response variable. In our case, the percentage of missing data is 13.26%.

We split the initial sample into a learning sample (the index set is denoted I_L) with size 181 and a test sample with size 60 (the index set is denoted I_T). Firstly, we reconstructed the missing parts of the differents curves and, on the learning sample, we imputed the missing values on the response. We tested the residuals normality, the shapiro test gives a p-value equal to 0.905, hence the normality of the residuals cannot be rejected. Then, on the test sample, we computed the prediction values for the response. In order to evaluate the quality of the prediction, we calculated, for $q = 100$, the mean squared prediction error $MSPE = \frac{1}{60}\sum_{i\in I_T}(Y_i - \widehat{Y}_i)^2 = 40.440$ and the mean absolute prediction error $MAPE = \frac{1}{60}\sum_{i\in I_T}|Y_i - \widehat{Y}_i| = 5.349$. Table 3.3 gives the $MSPE$ and the $MAPE$ for different imputation methods.

Comparing (**MUL_IM**) and (**MUL_NORM_IM**), we notice that (**MUL_NORM_IM**) behaves better for larger values of q, even if the differences are sometimes slight, because the normality of the residuals. Notice finally that, in this situation, the method (**REM_X,Y**) would not be possible since all the curves are partially observed and this would cause removing all individuals in the sample.

Missing values are imputed directly from the regression model, reducing the prediction error with respect to the missing rate but not taking into account the uncertainty of missing values or unseen data. Multiple regression imputation takes this into account by adding a random error term from the regression model residual distribution. This does not reduce the mean square prediction error but when the number of iteration increases, we can recover that of the deterministic regression imputation. Furthermore, multiple imputations are more realistic depending on the quality of the training data set the regression

[1] agora-energiewende.de/en/service/recent-electricity-data/chart/power_generation/15.03.2012/14.03.2013/
[2] automeris.io/WebPlotDigitizer/

Table 3.3: The mean square prediction error and the mean absolute prediction error with standard deviation errors for deterministic, random and multiple imputation methods.

Imputation methods	$MSPE$	$MAPE$
DETER_IM	40.443 (45.615)	5.354 (3.461)
RAND_IM	40.468 (45.662)	5.356 (3.462)
RAND_NORM_IM	40.533 (46.097)	5.363 (3.463)
MUL_IM ($q = 5$)	40.452 (45.613)	5.355 (3.461)
MUL_NORM_IM ($q = 5$)	40.479 (45.577)	5.357 (3.460)
MUL_IM ($q = 50$)	40.448 (45.624)	5.354 (3.461)
MUL_NORM_IM ($q = 50$)	40.269 (45.474)	5.345 (3.450)
MUL_IM ($q = 100$)	40.440 (45.625)	5.349 (3.461)
MUL_NORM_IM ($q = 100$)	40.211 (45.363)	5.343 (3.443)
REM_Y	40.543 (45.947)	5.354 (3.475)

model was trained under.

3.7) Proof of Theorem 3.4.2

Considering the decomposition of $\widehat{\theta}^{(w)}$, we write

$$\begin{aligned}
\widehat{\theta}^{(w)} &= \frac{1}{n}\sum_{\substack{i=1\\ \delta_i^{[Y]}=1}}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i^\star, \widehat{\phi}_{j,rec}^\star\rangle Y_i}{\widehat{\lambda}_{j,rec}^\star}\widehat{\phi}_{j,rec}^\star \\
&\quad + \frac{1}{n}\sum_{\substack{i=1\\ \delta_i^{[Y]}=0}}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i^\star, \widehat{\phi}_{j,rec}^\star\rangle \left(Y_{i,imp} + \varepsilon_i^{\star(w)}\right)}{\widehat{\lambda}_{j,rec}^\star}\widehat{\phi}_{j,rec}^\star \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i^\star, \widehat{\phi}_{j,rec}^\star\rangle Y_i^\star}{\widehat{\lambda}_{j,rec}^\star}\widehat{\phi}_{j,rec}^\star \\
&\quad + \frac{1}{n}\sum_{\substack{i=1\\ \delta_i^{[Y]}=0}}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i^\star, \widehat{\phi}_{j,rec}^\star\rangle \left(Y_{i,imp} + \varepsilon_i^{\star(w)}\right)}{\widehat{\lambda}_{j,rec}^\star}\widehat{\phi}_{j,rec}^\star,
\end{aligned}$$

hence

$$\widehat{Y}_{new}^{\star(w)} - \alpha - \langle \theta, X_{new}^{\star} \rangle = \widehat{\alpha}^{(w)} + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i^{\star}, \widehat{\phi}_{j,rec}^{\star} \rangle Y_i^{\star}}{\widehat{\lambda}_{j,rec}^{\star}} \langle \widehat{\phi}_{j,rec}^{\star}, X_{new}^{\star} \rangle - \alpha - \langle \theta, X_{new}^{\star} \rangle$$
$$+ \frac{1}{n} \sum_{\substack{i=1 \\ \delta_i^{[Y]}=0}}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i^{\star}, \widehat{\phi}_{j,rec}^{\star} \rangle \left(Y_{i,imp} + \varepsilon_i^{\star(w)} \right)}{\widehat{\lambda}_{j,rec}^{\star}} \langle \widehat{\phi}_{j,rec}^{\star}, X_{new}^{\star} \rangle.$$

We obtain from Crambes et al. (2022) the convergence rate for the first term of the decomposition

$$\mathbb{E}\left(\widehat{\alpha}^{(w)} + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i^{\star}, \widehat{\phi}^{\star}{}_{j,rec} \rangle Y_i^{\star}}{\widehat{\lambda}_{j,rec}^{\star}} \langle \widehat{\phi}_{j,rec}^{\star}, X_{new}^{\star} \rangle - \alpha - \langle \theta, X_{new}^{\star} \rangle \right)^2$$
$$= \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}} \right).$$

For the second term, we first use the boundedness of X and Y, which allows to bound $\varepsilon_i^{\star(w)}$, hence

$$\mathbb{E}\left(\frac{1}{n} \sum_{\substack{i=1 \\ \delta_i^{[Y]}=0}}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i^{\star}, \widehat{\phi}_{j,rec}^{\star} \rangle \left(Y_{i,imp} + \varepsilon_i^{\star(w)} \right)}{\widehat{\lambda}_{j,rec}^{\star}} \langle \widehat{\phi}_{j,rec}^{\star}, X_{new}^{\star} \rangle \right)^2$$
$$= \mathcal{O}_p\left(\frac{(m_n^{[Y]})^2 k_n^2}{n^2} \right).$$

As a consequence, with the assumptions

$$k_n \sim n^{\eta_1/(a_O+2)} \text{ and } m_n^{[Y]} = o\left(n^{1-\eta_1(a_O+3)/4(a_O+2)} \right),$$

we get

$$\mathbb{E}\left(\frac{1}{n} \sum_{\substack{i=1 \\ \delta_i^{[Y]}=0}}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i^{\star}, \widehat{\phi}_{j,rec}^{\star} \rangle \left(Y_{i,imp} + \varepsilon_i^{\star(w)} \right)}{\widehat{\lambda}_{j,rec}^{\star}} \langle \widehat{\phi}_{j,rec}^{\star}, X_{new}^{\star} \rangle \right)^2$$
$$= o_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} \right),$$

and the second term in the decomposition of $\widehat{Y}_{new}^{\star(w)} - \alpha - \langle \theta, X_{new}^{\star} \rangle$ is negligeable with respect to the first one. As a result, we obtain

$$\mathbb{E}\left(\widehat{Y}_{new}^{\star(w)} - \alpha - \langle \theta, X_{new}^{\star} \rangle\right)^2 = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}}\right).$$

Finally, the mean over q iterations of the random imputation gives

$$\begin{aligned}\mathbb{E}\left(\widehat{Y}_{new} - \alpha - \langle \theta, X_{new}^{\star} \rangle\right)^2 &= \frac{1}{q^2}\sum_{w=1}^{q}\mathbb{E}\left(\widehat{Y}_{new}^{\star(w)} - \alpha - \langle \theta, X_{new}^{\star} \rangle\right)^2 \\ &= \mathcal{O}_p\left(\frac{n^{-\eta_1(a_O-1)/(2(a_O+2))}}{q} + \frac{n^{\eta_1/(a_O+2)}}{q(n - m_n^{[Y]})}\right).\end{aligned}$$

CHAPTER 4

PREDICTION IN FUNCTION-ON-FUNCTION LINEAR MODEL WITH PARTIALLY OBSERVED FUNCTIONAL COVARIATE AND RESPONSE

Abstract.

In this work, we are interested in a function-on-function linear model in which the response and the covariate are partially observed curves. First, we reconstruct the missing part of the covariate using the observed parts. Then, we consider two strategies for dealing with the missing part of the response. The first one consists in a reconstruction in the same way as for the covariate. The second one uses regression imputation. Once the dataset is reconstructed, we estimate the slope function and give the mean square prediction error for a new observation of the covariate. Both methods are compared from a theoretical and a practical point of view.

Keywords.

Functional linear model, Missing parts, Imputation, Partially observed functional data, Functional Principal Components, Reconstruction operator.

Functional data analysis (FDA) is becoming progressively more and more important in Statistic (Bosq, 2000; Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Hsing and Eubank, 2015; Kokoszka and Reimherr, 2018). The functional linear model is specially a very popular model in both theoretical and applied research such as climatology, meteorology, economy, image analysis and many other fields. There is a large amount of work

done on the functional linear model. In the case where the response is a real variable, the model has been widely studied (e.g. Cardot et al., 2003; Cai and Hall, 2006; Li and Hsing, 2007; Hall and Horowitz, 2007; Crambes et al., 2009; Comte and Johannes, 2012; Cai and Yuan, 2012; Brunel et al., 2016). In contrast, fewer researchers have tackled the problem of function-on-function linear models where the covariate and the response are both functional (e.g. Ramsay and Silverman, 2005; Yao et al., 2005; Prchal and Sarda., 2007; Aguilera et al., 2008; Lian, 2011; Ferraty et al., 2012; Crambes and Mas, 2013; Crambes et al., 2016; Luo and Qi, 2017; Benatia et al., 2017; Imaizumi and Kato, 2016; Sun et al., 2018).

In recent years, applications producing partially observed functional data have emerged. Sometimes each individual trajectory is collected only over individual specific subintervals, densely or sparsely, within the whole domain of interest. Several recent works have begun addressing the estimation of covariance functions for short functional segments observed at sparse and irregular grid points, called *functional snippets* (Lin and Wang, 2022; Lin et al., 2021) or for *fragmented functional data* observed on small subintervals (Delaigle et al., 2020). For densely observed partial data, existing studies have focused on estimating the unobserved part of curves (Kneip and Liebl, 2020; Kraus and Stefanucci, 2020), prediction (Goldberg et al., 2014), classification (Kraus and Stefanucci, 2018; Park, 2019), functional regression (Gellar et al., 2014), and inferences (Kraus, 2019; Park et al., 2021).

To our knowledge, few articles investigate the theoretical properties of the slope operator or kernel estimators in the framework of partially observed data. In the framework of a real response, we can notice the works from Crambes and Henchiri (2019) and Crambes et al. (2022). The objective of this paper is to study the prediction problem when the covariate and the response are partially observed in the function-on-function linear regression setting, for which we propose two methods. The paper is organized as follows. We present the model with fully observed functional data in section 4.1. In section 4.2, we study the case of partially observed covariate and response. We reconstruct the missing parts of the functional explanatory variable and the functional response using the work Kneip and Liebl (2020), and we get theoretical results for the prediction error rate. In section 4.3, we propose an alternative method which consists in imputing the functional response after having reconstructed the missing parts of the functional covariate as in the previous section. We also obtain a convergence rate for the mean square prediction error. In section 4.4, we conduct a numerical study over simulated data in order to compare the methods in practice. All the proofs are postponed to section 4.5.

4.1) The centered function-on-function linear model

4.1.1) Functional principal components regression

Let $f : \mathcal{T} \to \mathbb{R}$ and $g : \mathcal{T} \times \mathcal{S} \to \mathbb{R}$ be two square integrable functions defined in the Hilbert space $\mathbb{L}^2(\mathcal{T})$ $\big($resp. $\mathbb{L}^2(\mathcal{T}\times\mathcal{S})\big)$ i.e. the space of square integrable functions on the

interval $\mathcal{T} \subseteq \mathbb{R}$ $\big($resp. $\mathcal{S} \subseteq \mathbb{R}\big)$. For all $t \in \mathcal{T}$ $\big($resp. $s \in \mathcal{S}\big)$, we set $\|f\| = \Big(\int_{\mathcal{T}} f^2(t)\mathrm{d}t\Big)^{1/2}$ and $\|g\| = \Big(\int_{\mathcal{S}}\int_{\mathcal{T}} g^2(t,s)\mathrm{d}t\mathrm{d}s\Big)^{1/2}$.

We use the following notation for the tensor product, defining $u \otimes v : \mathbb{L}^2(\mathcal{T}) \to \mathbb{L}^2(\mathcal{T})$, by $u \otimes v = \langle u, .\rangle v$, for any functions $u, v \in \mathbb{L}^2(\mathcal{T})$. We define the Hilbert-Schmidt integral operator $G : \mathbb{L}^2(\mathcal{T} \times \mathcal{S}) \to \mathbb{L}^2(\mathcal{T})$ by $(G \cdot u)(s) = \int_{\mathcal{T}} u(t)g(t,s)\mathrm{d}t$ for all g belonging to $\mathbb{L}^2(\mathcal{S} \times \mathcal{T})$ and $\int_{\mathcal{S}}\int_{\mathcal{T}} |g(t,s)|^2\mathrm{d}t\mathrm{d}s < \infty$. The function g is the Hilbert-Schmidt kernel corresponding to the operator G. Moreover, the operator G is continuous (i.e bounded) and compact.

We assume here that $\mathcal{X}$ takes values in the space $\mathbb{L}^2(\mathcal{T})$ and $\mathcal{Y}$ in $\mathbb{L}^2(\mathcal{S})$, $\mathcal{X} \triangleq \{\mathcal{X}(t) \mid t \in \mathcal{T}\}$ is the predictor variable and $\mathcal{Y} \triangleq \{\mathcal{Y}(s) \mid s \in \mathcal{S}\}$ the response variable. We observe a sample $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^n$, of identically distributed and independent copies $(\mathcal{X}, \mathcal{Y})$, with $\mathbb{E}(\|\mathcal{X}\|^2) < \infty$ and $\mathbb{E}(\|\mathcal{Y}\|^2) < \infty$. In practice, functional responses $\mathcal{Y}_i(s)$ and functional covariates $\mathcal{X}_i(t)$ are observed on grid points $\boldsymbol{s}_i = (s_{i1}, \ldots, s_{iq_i})$ and $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{ip_i})$. For the sake of simplicity, we assume those to be identical vectors $\boldsymbol{s}, \boldsymbol{t}$ with lengths q and p, respectively, for each observation i.

We consider a centered function-on-function linear regression model (FFLRM) characterizing a linear relationship between the functional response and the functional predictor

$$\big[\mathcal{Y} - \mathbb{E}\big[\mathcal{Y}\big]\big](s) = \int_{\mathcal{T}} \big[\mathcal{X} - \mathbb{E}\big[\mathcal{X}\big]\big](t)\theta(t,s)\mathrm{d}t + \epsilon(s),$$

where the bivariate functional coefficient $(t,s) \mapsto \theta(t,s)$ is assumed to be in $\mathbb{L}^2(\mathcal{T} \times \mathcal{S})$, that is, $\int_{\mathcal{S}}\int_{\mathcal{T}} |\theta(t,s)|^2\mathrm{d}t\mathrm{d}s < \infty$. Natural estimators of $\mathbb{E}\big[\mathcal{X}\big]$ and $\mathbb{E}\big[\mathcal{Y}\big]$ are the empirical means, $\overline{\mathcal{X}} \triangleq 1/n\sum_{i=1}^n \mathcal{X}_i$ and $\overline{\mathcal{Y}} \triangleq 1/n\sum_{i=1}^n \mathcal{Y}_i$. The centered random error function ϵ is assumed to be independent of $\mathcal{X}$, and $\mathbb{E}\|\epsilon\|^2 < \infty$.
We define the elements of the centered model as follows, $X \triangleq \mathcal{X} - \mathbb{E}\big[\mathcal{X}\big]$ and $Y \triangleq \mathcal{Y} - \mathbb{E}\big[\mathcal{Y}\big]$, and the centered FFLRM writes

$$Y(s) = \int_{\mathcal{T}} X(t)\theta(t,s)\mathrm{d}t + \epsilon(s). \tag{4.1.1}$$

Denote

$$\mathrm{C}_X(t_1,t_2) = \mathbb{E}\Big(X(t_1)X(t_2)\Big), \quad t_1, t_2 \in \mathcal{T},$$

and

$$\mathrm{C}_Y(s_1,s_2) = \mathbb{E}\Big(Y(s_1)Y(s_2)\Big), \quad s_1, s_2 \in \mathcal{S},$$

as the covariance functions of X and Y, respectively. As $\mathrm{C}_X \in \mathbb{L}^2(\mathcal{T}^2)$ and $\mathrm{C}_Y \in \mathbb{L}^2(\mathcal{S}^2)$ are a positive self-adjoint operators, then, according to Mercer's theorem (Hsing and Eubank, 2015, Theorem 4.6.5), C_X and C_Y admit spectral expansions in $\mathbb{L}^2(\mathcal{T}^2)$ and

$\mathbb{L}^2(\mathcal{S}^2)$ respectively

$$\mathrm{C}_X(t_1,t_2)=\sum_{k=1}^{+\infty}\lambda_k\phi_k(t_1)\phi_k(t_2) \quad \text{and} \quad \mathrm{C}_Y(s_1,s_2)=\sum_{j=1}^{+\infty}\mu_j\psi_j(s_1)\psi_j(s_2),$$

where $\lambda_1 > \lambda_2 > \ldots > 0$ and $\mu_1 > \mu_2 > \ldots > 0$ are the eigenvalue sequences of the covariance functions C_X and C_Y, respectively, while $\{\phi_k\}_{k\geqslant 1}$ and $\{\psi_j\}_{j\geqslant 1}$ are the corresponding orthonormal bases of eigenfunctions in $\mathbb{L}^2(\mathcal{T})$ and $\mathbb{L}^2(\mathcal{S})$.

The Karhunen–Loève (KL) expansions of the curves X and Y (Hsing and Eubank, 2015, Theorem 7.3.5) in $\mathbb{L}^2(\mathcal{T})$ and $\mathbb{L}^2(\mathcal{S})$ are respectively

$$X(t)=\sum_{k=1}^{+\infty}\xi_k\phi_k(t) \quad \text{and} \quad Y(s)=\sum_{j=1}^{+\infty}\beta_j\psi_j(s), \tag{4.1.2}$$

for all $t \in \mathcal{T}$ and $s \in \mathcal{S}$, where $\xi_k = \int_{\mathcal{T}} X(t)\phi_k(t)\mathrm{d}t$ and $\beta_j = \int_{\mathcal{S}} Y(s)\psi_j(s)\mathrm{d}s$ are uncorrelated random variables with zero mean and variances $\mathbb{E}(\xi_k^2)=\lambda_k$ and $\mathbb{E}(\beta_j^2)=\mu_j$ for all $k,j \geqslant 1$. These coefficients ξ_k and β_j are called functional principal components scores. By Parseval's identity and Fubini's theorem, we have

$$\sum_{k=1}^{+\infty}\mathbb{E}(\xi_k^2)<\infty \quad \text{and} \quad \sum_{j=1}^{+\infty}\mathbb{E}(\beta_j^2)<\infty.$$

(Ramsay and Silverman, 2005, Chapter 16, Section 1.1), Park and Qian (2012), Crambes and Mas (2013) and Imaizumi and Kato (2016) suggest that $\theta(t,s)$ can be expressed in terms of a double basis expansion

$$\theta(t,s)=\sum_{k=1}^{+\infty}\sum_{j=1}^{+\infty}\theta_{k,j}\phi_k(t)\psi_j(s). \tag{4.1.3}$$

Using (5.2.1) and (4.1.3), we notice that

$$\begin{aligned}\int_{\mathcal{T}}\theta(t,s)X(t)\mathrm{d}t &= \int_{\mathcal{T}}\sum_{k=1}^{+\infty}\sum_{j=1}^{+\infty}\theta_{k,j}\phi_k(t)\psi_j(s)\sum_{r=1}^{+\infty}\xi_r\phi_r(t)\mathrm{d}t,\\ &\quad \Big[\int_{\mathcal{T}}\phi_r(t)\phi_k(t)\mathrm{d}t = 0 \; if \; r\neq k \; and \; 1 \; otherwise.\Big]\\ &= \sum_{k=1}^{+\infty}\sum_{j=1}^{+\infty}\theta_{k,j}\xi_k\psi_j(s),\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\Big(\xi_k Y(s)\Big) =& \mathbb{E}\Big(\xi_k \mathbb{E}\big(Y \mid X\big)(s)\Big),\\ =& \mathbb{E}\left(\xi_k \sum_{r=1}^{+\infty}\sum_{j=1}^{+\infty}\theta_{r,j}\xi_r\psi_j(s)\right),\\ =& \sum_{j=1}^{+\infty}\mathbb{E}\big(\xi_k\xi_k\big)\theta_{k,j}\psi_j(s),\\ =& \lambda_k\sum_{j=1}^{+\infty}\theta_{k,j}\psi_j(s),\end{aligned}$$

we obtain the following characterization of the bivariate functional coefficient $\theta(t,s)$ as

$$\theta(t,s) = \sum_{k=1}^{+\infty}\frac{\mathbb{E}\Big(\xi_k Y(s)\Big)}{\lambda_k}\phi_k(t).$$

This characterization leads to a method for estimating $\theta(t,s)$. For example, Park and Qian (2012) and Crambes and Mas (2013) use the following truncation estimator with $k_n \to \infty$ as $n \to \infty$,

$$\widehat{\theta}(t,s) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{k_n}\frac{\widehat{\xi}_{ik}Y_i(s)}{\widehat{\lambda}_k}\widehat{\phi}_k(t), \tag{4.1.4}$$

where $\widehat{\lambda}_k$ and $\widehat{\phi}_k$ are the estimators of the eigenvalues λ_k and the eigenfunctions ϕ_k and the estimates of the FPC scores are $\widehat{\xi}_{ik} = \int_{\mathcal{T}} X_i(t)\widehat{\phi}_k(t)\mathrm{d}t$. We will use this characterization in this paper.

Benatia et al. (2017) propose an estimator similar to (4.1.4). The sum over k is not truncated, but regularized with the term $\widehat{\lambda}_k + \kappa$

$$\widehat{\theta}^{\ddagger}(t,s) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\frac{\widehat{\xi}_{ik}Y_i(s)}{\widehat{\lambda}_k + \kappa}\widehat{\phi}_k(t).$$

Ramsay and Silverman (2005) obtain the following characterization of the bivariate functional coefficient, for some small or large numbers ι_1 and ι_2, as

$$\theta^{\dagger}(t,s) = \sum_{k=1}^{\iota_1}\sum_{j=1}^{\iota_2}\theta_{k,j}\phi_k(t)\psi_j(s), \quad \text{where } \theta_{k,j} = \frac{\mathbb{E}\big(\xi_k\beta_j\big)}{\lambda_k}.$$

Therefore Imaizumi and Kato (2016) obtain the following characterization

$$\theta^{\dagger\dagger}(t,s) = \sum_{k=1}^{+\infty}\frac{\mathbb{E}\big(\xi_k\beta_j\big)}{\lambda_k}\phi_k(t)\psi_j(s).$$

Notice that $\theta^{\dagger}(t,s)$ and $\theta^{\dagger\dagger}(t,s)$ based on truncating the double series, namely, the double truncation, which will not be discussed here.

4.1.2) Operatorial point of view

We notice in this subsection that the model (4.1.1) can be seen from an operatorial point of view. Indeed, we can write the model

$$Y(\boldsymbol{s}) = \Big(\Theta \cdot X\Big)(\boldsymbol{s}) + \epsilon(\boldsymbol{s}),$$

where the slope operator $\Theta : \mathbb{L}^2(\mathcal{T} \times \mathcal{S}) \to \mathbb{L}^2(\mathcal{S})$ is an integral continuous operator defined for all $u \in \mathbb{L}^2(\mathcal{T})$ by $(\Theta \cdot u)(s) = \int_{\mathcal{T}} u(t)\theta(t,s)\mathrm{d}t$ which kernel is $\theta \in \mathbb{L}^2(\mathcal{T} \times S)$.

To close this section, we introduce the operators that will be used through the following theorems in which we describe prediction convergence rates. The covariance operator of X, denoted Γ, is defined by

$$\Gamma(u) = \mathbb{E}\big(X \otimes X(u)\big), \quad \text{for all } u \in \mathbb{L}^2(\mathcal{T}).$$

Note that the covariance operator is a natural extension of the covariance matrix, in the infinite dimensional framework. We also introduce the cross-covariance operator Δ of (X, Y) given by

$$\Delta(u) = \mathbb{E}\big(Y \otimes X(u)\big), \quad \text{for all } u \in \mathbb{L}^2(\mathcal{T}).$$

For any integer k, we define Π_k the orthogonal projection operator on the subspace $\mathrm{Span}(\phi_1, \ldots, \phi_k)$, given by

$$\Pi_k = \sum_{j=1}^{k} \phi_j \otimes \phi_j.$$

Empirical counterparts of Γ, Δ and Π_k, respectively, are denoted $\widehat{\Gamma}_n$, $\widehat{\Delta}_n$ and $\widehat{\Pi}_{k_n}$. These operators are naturally defined by $\widehat{\Gamma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i$, $\widehat{\Delta}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \otimes X_i$ and $\widehat{\Pi}_{k_n} = \sum_{j=1}^{k_n} \widehat{\phi}_j \otimes \widehat{\phi}_j$. The functional principal component regression estimator $\widehat{\Theta}$ of Θ is defined by

$$\widehat{\Theta} = \widehat{\Pi}_{k_n}\widehat{\Delta}_n(\widehat{\Pi}_{k_n}\widehat{\Gamma}_n\widehat{\Pi}_{k_n})^{-1}.$$

4.2) The centered Function-on-function linear model with partially observed covariate and response: Reconstructing X and Y.

In this section, we are interested in the most general case of missing data in the centered function-on-function linear regression: when both the functional covariate and the functional response are partially observed. We follow the methodology studied in Kneip and Liebl (2020) for reconstructing the missing parts of the curves. Next, once the initial sample is completed, we will present the estimation of the slope operator Θ or its kernel θ and predict new values for the response.

In the following, we denote "$O^{[Y]}$", "$M^{[Y]}$", "$O^{[X]}$" and "$M^{[X]}$" a given production of $O_i^{[Y_i]}$, $M_i^{[Y_i]}$, $O_i^{[X_i]}$ and $M_i^{[X_i]}$, respectively corresponding to the observed domain and the missing domain of Y_i included in $\mathcal{S}$ and the observed domain and the missing domain of X_i included in $\mathcal{T}$. In addition, the corresponding parts of the curve X_i are denoted $X_i^{O^{[X]}}$ and $X_i^{M^{[X]}}$. Similarly, the corresponding parts of Y_i are denoted $Y_i^{O^{[Y]}}$ and $Y_i^{M^{[Y]}}$.

4.2.1) Curve reconstruction of the covariate and the response

The Karhunen–Loève expansions of the observed sample curves $X_i^{O^{[X]}}$ in $\mathbb{L}^2(O^{[X]})$ and $Y_i^{O^{[Y]}}$ in $\mathbb{L}^2(O^{[Y]})$ are written

$$X_i^{O^{[X]}}(t) = \sum_{k=1}^{+\infty} \xi_{ik}^{O^{[X]}} \phi_k^{O^{[X]}}(t) \quad \text{and} \quad Y_i^{O^{[Y]}}(s) = \sum_{j=1}^{+\infty} \beta_{ij}^{O^{[Y]}} \psi_j^{O^{[Y]}}(s),$$

where $\xi_{ik}^{O^{[X]}} = \int_{\mathcal{T}} X_i^{O^{[X]}}(t)\phi_k^{O^{[X]}}(t)\mathrm{d}t$ and $\beta_{ij}^{O^{[Y]}} = \int_{\mathcal{S}} Y_i^{O^{[X]}}(s)\psi_j^{O^{[X]}}(s)\mathrm{d}s$ are uncorrelated random variables with zero mean and variances $\mathbb{E}\left(\xi_k^{O^{[X]}}\right)^2 = \lambda_k^{O^{[X]}}$ and $\mathbb{E}\left(\beta_j^{O^{[Y]}}\right)^2 = \mu_j^{O^{[Y]}}$ for all $k, j \geqslant 1$.

We consider a reconstruction problem relating the missing part of the curves to the observed part, writing

$$X_i^{M^{[X]}}(t_2) = L^{[X]}(X_i^{O^{[X]}}(t_1)) + \mathcal{Z}_i^{[X]}(t_2), \quad \text{for all } t_1 \in O^{[X]} \text{ and } t_2 \in M^{[X]},$$

and

$$Y_i^{M^{[Y]}}(s_2) = L^{[Y]}(Y_i^{O^{[Y]}}(s_1)) + \mathcal{Z}_i^{[Y]}(s_2), \quad \text{for all } s_1 \in O^{[Y]} \text{ and } s_2 \in M^{[Y]},$$

where $L^{[X]} : \mathbb{L}_2(O^{[X]}) \to \mathbb{L}_2(M^{[X]})$ and $L^{[Y]} : \mathbb{L}_2(O^{[Y]}) \to \mathbb{L}_2(M^{[Y]})$ are linear reconstruction operators and $\mathcal{Z}_i^{[X]} \in \mathbb{L}_2(M^{[X]})$ and $\mathcal{Z}_i^{[Y]} \in \mathbb{L}_2(M^{[Y]})$ are reconstruction errors. We need to minimize the mean square error between the curves with fragmentary data and the linear reconstruction operators $L^{[X]}$ and $L^{[Y]}$ as follows

$$\mathbb{E}\left(\left\|X_i^{M^{[X]}} - \mathrm{L}^{[X]}(X_i^{O^{[X]}})\right\|^2\right) \quad \text{and} \quad \mathbb{E}\left(\left\|Y_i^{M^{[Y]}} - \mathrm{L}^{[Y]}(Y_i^{O^{[Y]}})\right\|^2\right). \tag{4.2.1}$$

Thes operator $\mathrm{L}^{[X]}$ $\left(\text{or } \mathrm{L}^{[Y]}\right)$ has been studied by Kraus (2015) and Kneip and Liebl (2020). Kraus (2015) proposed to use the ridge regularization method. The estimator is introduced by $\mathcal{L}^{(\alpha)} = \mathcal{R}_{MO}\mathcal{R}_{OO}^{(\alpha)-1}$ where $\mathcal{R}$ is the covariance operator defined on $\mathbb{L}^2(\mathcal{T})$ and $\mathcal{R}_{OO}^{(\alpha)-1} = \mathcal{R}_{OO} + \alpha\mathcal{J}_O$ where α is a positive parameter and $\mathcal{J}_O$ is the identity operator on $\mathbb{L}^2(O^{[X]})$. Besides, Kraus and Stefanucci (2020) prove that the reconstruction operator (considered in Kraus (2015)) can be seen as Hilbert-Schmidt integral operator from $\mathbb{L}^2(O^{[X]})$ to $\mathbb{L}^2(O^{[X]})$ writing

$$\mathrm{L}(X_i^{O^{[X]}}) = \int_{O^{[X]}} a_i(.,t)X_i^{O^{[X]}}(t)\mathrm{d}t,$$

for all $i = 1, \ldots, n$, where a is a square integrable function on $M^{[X]} \times O^{[X]}$.

The optimal reconstruction operators minimizing (4.2.1) are denoted $\mathcal{L}(X_i^{O^{[X]}})$ and $\mathcal{J}(Y_i^{O^{[Y]}})$ and defined by

$$\mathcal{L}(X_i^{O^{[X]}})(t_2) = \sum_{k=1}^{+\infty} \xi_{ik}^{O^{[X]}} \frac{\langle \phi_k^{O^{[X]}}, \gamma_{t_2} \rangle}{\lambda_k^{O^{[X]}}} \text{ and } \mathcal{J}(Y_i^{O^{[Y]}})(s_2) = \sum_{j=1}^{+\infty} \beta_{ij}^{O^{[Y]}} \frac{\langle \psi_j^{O^{[Y]}}, \gamma_{s_2} \rangle}{\mu_j^{O^{[Y]}}}, \quad (4.2.2)$$

where $\gamma_{t_2}(t_1) = \mathbb{E}\left(X_i^{M^{[X]}}(t_2) X_i^{O^{[X]}}(t_1)\right)$, for all $t_1 \in O^{[X]}$ and $t_2 \in M^{[X]}$, and $\gamma_{s_2}(s_1) = \mathbb{E}\left(Y_i^{M^{[Y]}}(s_2) Y_i^{O^{[Y]}}(s_1)\right)$, for all $s_1 \in O^{[Y]}$ and $s_2 \in M^{[Y]}$.

The operators in relation (4.2.2) are estimated as in (Kneip and Liebl, 2020, Section 2) and (Crambes et al., 2022, Subsection 2.1 and Subsection 2.2) by $\widehat{\mathcal{L}}_{k_n}(X_i^{O^{[X]}})$ and $\widehat{\mathcal{J}}_{j_n}(Y_i^{O^{[Y]}})$, where the truncation parameters k_n and j_n are positive integers that can be fixed automatically with a grid search. The solution of (4.2.2) uses local linear smoothers for unknown quantities in (4.2.2), considering the following notations. Let κ_1 and κ_1' be kernels and h_X and h_Y be bandwidths of the local linear smoothers of the curves $X_i^{O^{[X]}}$ and $Y_i^{O^{[Y]}}$, respectively. Moreover, let κ_2 and κ_2' be bivariate kernels and $h_{\gamma_{t_2}}$ and $h_{\gamma_{s_2}}$ be bandwidths of the local linear smoothers of the covariance functions γ_{t_2} and γ_{s_2}, respectively.

In the following, we consider the whole sample $\mathcal{D}_n \triangleq \{(X_1^\star, Y_1^\star), \ldots, (X_n^\star, Y_n^\star)\}$, with possibly reconstructed curves on the missing parts, that is

$$X_i^\star(t) = \begin{cases} X_i^{O^{[X]}}(t) & \text{if } t \in O^{[X]}, \\ \widehat{\mathcal{L}}_{k_n}(X_i^{O^{[X]}}(t)) & \text{if } t \in M^{[X]} \end{cases} \quad \text{and} \quad Y_i^\star(s) = \begin{cases} Y_i^{O^{[Y]}}(s) & \text{if } s \in O^{[Y]}, \\ \widehat{\mathcal{J}}_{j_n}(Y_i^{O^{[Y]}})(s) & \text{if } s \in M^{[Y]}. \end{cases}$$

4.2.2) Estimation of slope operator and its kernel and prediction

We estimate the kernel function θ with

$$\widehat{\theta}^\star(t,s) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{k_n} \frac{\widehat{\xi}_{ik,rec}^\star Y_i^\star(s)}{\widehat{\lambda}_{k,rec}^\star} \widehat{\phi}_{k,rec}^\star(t), \quad (4.2.3)$$

where $\widehat{\phi}_{1,rec}^\star, \ldots, \widehat{\phi}_{k_n,rec}^\star$ and $\widehat{\lambda}_{1,rec}^\star, \ldots, \widehat{\lambda}_{k_n,rec}^\star$ represent respectively the k_n first eigenfunctions and eigenvalues of the covariance operator $\widehat{\Gamma}_{n,rec}^\star$ and $\widehat{\xi}_{ik,rec}^\star = \int_{\mathcal{T}} X_i^\star(t) \widehat{\phi}_{k,rec}^\star(t) \mathrm{d}t$ are the estimates of the FPC scores.

From an operatorial point of view, the covariance operator $\Gamma_{rec}^\star$ of $X^\star$ and the cross-covariance operator $\Delta_{rec}^\star$ of $(X^\star, Y^\star)$ are given by $\Gamma_{rec}^\star = \mathbb{E}[X^\star \otimes X^\star]$ and $\Delta_{rec}^\star = \mathbb{E}[Y^\star \otimes X^\star]$, for all $u \in \mathbb{L}^2(\mathcal{T})$. The orthogonal projection operator $\Pi_{k,rec}^\star$ on the subspace $\mathrm{Span}(\phi_{1,rec}^\star, \ldots, \phi_{k,rec}^\star)$ is given by $\Pi_k^\star = \sum_{j=1}^k \phi_{j,rec}^\star \otimes \phi_{j,rec}^\star$. The empirical counterparts of $\Gamma_{rec}^\star$, $\Delta_{rec}^\star$ and $\Pi_{k,rec}^\star$, are denoted respectively $\widehat{\Gamma}_{n,rec}^\star$, $\widehat{\Delta}_{n,rec}^\star$ and $\widehat{\Pi}_{k_n,rec}^\star$. The estimator

of the slope operator Θ is given by

$$\widehat{\Theta}^{\star} = \langle \widehat{\theta}^{\star}(\cdot, s), \cdot \rangle = \widehat{\Pi}^{\star}_{k_n,rec} \widehat{\Delta}^{\star}_{n,rec} \left(\widehat{\Pi}^{\star}_{k_n,rec} \widehat{\Gamma}^{\star}_{n,rec} \widehat{\Pi}^{\star}_{k_n,rec} \right)^{-1}.$$

Finally, we obtain the prediction of the response when a new explanatory curve X_{new} is given, by

$$Y^{\star}_{new}(s) = \int_{\mathcal{T}} X^{\star}_{new}(t) \widehat{\theta}^{\star}(t, s) \mathrm{d}t = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{k_n} \frac{\widehat{\xi}^{\star}_{ik,rec} \widehat{\xi}^{\star}_{new;k,rec} Y^{\star}_i(s)}{\widehat{\lambda}^{\star}_{k,rec}},$$

for all $s \in \mathcal{S}$, where $\widehat{\xi}^{\star}_{new;k,rec} = \int_{\mathcal{T}} X^{\star}_{new}(t) \widehat{\phi}^{\star}_{k,rec}(t) \mathrm{d}t$. Alternatively, it is also given with the operatorial quantities

$$Y^{\star}_{new} = \widehat{\Pi}^{\star}_{k_n,rec} \widehat{\Delta}^{\star}_{n,rec} \left(\widehat{\Pi}^{\star}_{k_n,rec} \widehat{\Gamma}^{\star}_{n,rec} \widehat{\Pi}^{\star}_{k_n,rec} \right)^{-1} X^{\star}_{new}.$$

4.2.3) Assumptions

To achieve our theoretical results, we need first to adopt some classical assumptions which have been similarly used in Kneip and Liebl (2020) and Crambes et al. (2022) to control the curve reconstruction for the covariate and the response.

(A.1) X and Y have finite fourth moment order.

(A.2) Let $np \to \infty$ when $n \to \infty$, where $p = p(n)$ is the number of observation points of the covariate. Similarly, $nq \to \infty$ when $n \to \infty$, where $q = q(n)$ is the number of observation points of the response. We assume in the following that $p = n^{\eta_1}$ with $0 < \eta_1 < \infty$ and $q = n^{\zeta_1}$ with $0 < \zeta_1 < \infty$.

(A.3) • The bandwidth h_X satisfies $h_X \to 0$ and $(ph_X) \to \infty$ as $p \to \infty$. For instance, we assume that $h_X = \frac{1}{n^{\eta_2}}$ with $0 < \eta_2 < \eta_1$. The bandwidth $h_{\gamma_{t_2}}$ satisfies $h_{\gamma_{t_2}} \to 0$ and $(n(p^2 - p)h_{\gamma_{t_2}}) \to \infty$ as $n(p^2 - p) \to \infty$. For example, we can take $h_{\gamma_{t_2}} = \frac{1}{n^{\eta_3}}$ with $0 < \eta_3 < 2\eta_1 + 1$.
• Let κ_1 and κ_2 be nonnegative, second order univariate and bivariate kernel functions with support $[-1, 1]$. For example, we can use univariate and bivariate Epanechnikov kernel functions with compact support $[-1, 1]$, namely $\kappa_1(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{[-1,1]}(x)$ and $\kappa_2(x, y) = \frac{9}{16}(1 - x^2)(1 - y^2)\mathbb{1}_{[-1,1]}(x)\mathbb{1}_{[-1,1]}(y)$.

(A.4) • For any subinterval $O^{[X]} \subseteq \mathcal{T}$, we assume that the eigenvalues $\lambda_1 > \lambda_2 > \ldots > 0$ have multiplicity one. Moreover, we assume that there exist $a_O > 1$ and $0 < c_O < \infty$ such that, (a) $\lambda_k^{O^{[X]}} - \lambda_{k+1}^{O^{[X]}} \geqslant c_O k^{-a_O - 1}$, (b) $\lambda_k^{O^{[X]}} = \mathcal{O}(k^{-a_O})$, (c) $1/\lambda_k^{O^{[X]}} = \mathcal{O}(k^{a_O})$ as $k \to \infty$.
• $\mathbb{E}\left(\xi_k^4\right) = \mathcal{O}(\lambda_k^2)$.
• For any subinterval $O^{[X]} \subseteq \mathcal{T}$, we assume that there exists $0 < A_O < \infty$ such that the eigenfunctions satisfy $\sup_{t \in \mathcal{T}} \sup_{k \geqslant 1} \left| \widetilde{\phi}_k^{O^{[X]}}(t) \right| \leqslant A_O$, where $\widetilde{\phi}_k^{O^{[X]}}(t_2) = \langle \phi_k^{O^{[X]}}, \gamma_{t_2} \rangle / \lambda_k^{O^{[X]}}$.

(A.5) • The bandwidth h_Y satisfies $h_Y \to 0$ and $(qh_Y) \to \infty$ as $q \to \infty$. Moreover, we assume that $h_Y = \frac{1}{n^{\zeta_2}}$ with $0 < \zeta_2 < \zeta_1$. The bandwidth $h_{\gamma_{s_2}}$ satisfies $h_{\gamma_{s_2}} \to 0$ and $(n(q^2-q)h_{\gamma_{s_2}} \to \infty$ as $n(q^2-q) \to \infty$. For example, we can take $h_{\gamma_{s_2}} = \frac{1}{n^{\zeta_3}}$ with $0 < \zeta_3 < 2\zeta_1 + 1$.
• Let $\kappa_1^{'}$ and $\kappa_2^{'}$ be nonnegative, second order univariate and bivariate kernel functions with support $[-1, 1]$.

(A.6) • For any subinterval $O^{[Y]} \subseteq \mathcal{S}$, we assume that $\mu_1 > \mu_2 > \ldots > 0$ have multiplicity one and we assume that there exist $b_O > 1$ and $0 < d_O < \infty$ such that, (a) $\mu_j^{O^{[Y]}} - \mu_{j+1}^{O^{[Y]}} \geqslant d_O j^{-b_O-1}$, (b) $\mu_j^{O^{[Y]}} = \mathcal{O}\big(j^{-b_O}\big)$, (c) $1/\mu_j^{O^{[Y]}} = \mathcal{O}\big(j^{b_O}\big)$ as $j \to \infty$.
• $\mathbb{E}\left(\beta_k^4\right) = \mathcal{O}(\mu_k^2)$.
• For any subinterval $O^{[Y]} \subseteq \mathcal{S}$, we assume that there exists $0 < B_O < \infty$ such that $\sup_{s\in\mathcal{S}} \sup_{j\geqslant 1} \left|\widetilde{\psi}_j^{O^{[Y]}}(s)\right| \leqslant B_O$, where $\widetilde{\psi}_j^{O^{[Y]}}(s_2) = \langle \psi_j^{O^{[Y]}}, \gamma_{s_2}\rangle / \lambda_j^{O^{[Y]}}$.

Assumption ***(A.1)*** holds for many processes X and Y (Gaussian processes, bounded processes). Assumption ***(A.2)*** is mild and can be satisfied even if the number of observation points p and q do not go fast to infinity. ***(A.3)*** and ***(A.5)*** are classic assumptions in the context of local polynomials smoothers. Assumptions ***(A.4)*** and ***(A.6)*** are similar to the ones in Kneip and Liebl (2020).

4.2.4) Asymptotic results

Under assumptions ***(A.1)-(A.6)***, it is proved in Kneip and Liebl (2020) that, in the case where $p \sim n^{\eta_1}$ and $q \sim n^{\zeta_1}$ with $\eta_1, \zeta_1 \leqslant 1/2$ we have for any $t \in \mathcal{T}$ and $s \in \mathcal{S}$

$$|X_i^\star(t) - X_i(t)| = \mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big) \text{ and } |Y_i^\star(s) - Y_i(s)| = \mathcal{O}_p\big(n^{-\zeta_1(b_O-1)/(2(b_O+2))}\big) .$$

The previous result allows to obtain some bounds between quantities related to functional principal components analysis between the reconstructed curves and with the original curves.

We finish this subsection with the main result giving a bound for the prediction error of Y_{new} with a new value of the covariate X_{new}.

Theorem 4.2.1. *Under assumptions **(A.1)-(A.6)**, taking $k_n \sim p^{1/(a_O+2)}$, $p \sim n^{\eta_1}$, $j_n \sim q^{1/(b_O+2)}$ and $q \sim n^{\zeta_1}$, with $\eta_1, \zeta_1 \leqslant 1/2$, we get*

$$\mathbb{E}\left(\left\|\left(\widehat{\Theta}^\star \cdot X_{new}^\star - \Theta \cdot X_{new}^\star\right)\right\|^2\right) = \mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))} + n^{\eta_1/(a_O+2)-1-\zeta_1(b_O-1)/(b_O+2)}\big) .$$

Corollary 4.2.2. *We make a comparison between the parameters to find the best convergence error. We summarize the error rates in Table 4.1.*

Table 4.1: Convergence error rates depending on the observation points and the regularity of the curves X and Y.

(i) $\eta_1 = \zeta_1$	$a_O \leqslant b_O$	$\mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)$
	$a_O > b_O$	$\mathcal{O}_p\big(n^{\eta_1/(a_O+2)-1-\eta_1(b_O-1)/(b_O+2)}\big)$
(ii) $\eta_1 < \zeta_1$		$\mathcal{O}_p\big(n^{\eta_1/(a_O+2)-1-\zeta_1(b_O-1)/(b_O+2)}\big)$
(iii) $\eta_1 > \zeta_1$		$\mathcal{O}_p\big(n^{-\eta_1(a_O-1)/(2(a_O+2))}\big)$

Comparing the parameters, we remark all the convergence rates depend in particular on the parameter $a_O > 1$, which is directly linked to the smoothness of the stochastic process X. The larger a_O is, the smoother X is. In the case (i) when the number of observation points of the covariate is equivalent to that of the response for $a_O > b_O$ and in the case (ii), the convergence rates depend of the parameter a_O and also of the parameter $b_O > 1$, which is directly linked to the smoothness of the stochastic process Y. In these cases, the final rate of convergence will be linked to the parameter a_O or b_O corresponding to the less smooth process (either X or Y).

4.3) The centered function-on-function with partially observed covariate and response: Reconstructing X and imputing Y.

We have seen in the previous section the methodology for reconstructing the missing parts of the explanatory curves. In this section, we try another strategy to deal with missing data on the response. After reconstructing the missing parts of the covariate X, we apply the regression imputation methodology as presented in Crambes et al. (2022) for a real response. Next, we will present the estimation of the kernel function θ and predict the new response once all sample is completed.

4.3.1) Regression imputation on the functional response

We consider the following missing data mechanism for the response, through a variable $\delta^{[Y]}$ leading to the sample $(\delta_i^{[Y]})_{i=1,\dots,n}$ such that,

$$\delta_i^{[Y]} = \begin{cases} 0 \text{ if } M_i^{[Y]} \neq \varnothing, \\ 1 \text{ if } O_i^{[Y]} = \mathcal{S}. \end{cases}$$

We assume that the response is missing at random (MAR), which means that the fact that Y contains missing parts does not depend on the response of the model, but can

possibly depend on the reconstructed covariate,

$$\mathbb{P}(\delta^{[Y]} = 1 \mid X^\star, Y) = \mathbb{P}(\delta^{[Y]} = 1 \mid X^\star).$$

We denote the number of curves partially observed by

$$m_n^{[Y]} = \sum_{i=1}^{n} \mathbb{1}_{\{\delta_i^{[Y]}=0\}}.$$

Using the exponent notation "obs" to make reference to the units for which the response is completely observed, we define the covariance operator Γ_{rec}^{obs} and the cross-covariance operator Δ_{rec}^{obs} with the reconstructed curves by $\Gamma_{rec}^{obs} = \mathbb{E}\left(\delta^{[Y]} X^\star \otimes X^\star\right)$ and $\Delta_{rec}^{obs} = \mathbb{E}\left(\delta^{[Y]} Y \otimes X^\star\right)$. The empirical counterparts of Γ_{rec}^{obs} and Δ_{rec}^{obs} are denoted respectively $\widehat{\Gamma}_{n,rec}^{obs}$ and $\widehat{\Delta}_{n,rec}^{obs}$.

Let Y_ℓ be a response curve such that $\delta_\ell^{[Y]} = 0$, we define the imputed response $Y_{\ell,imp}$ by

$$Y_{\ell,imp}(s) = \int_{\mathcal{T}} X_\ell^\star(t)\widetilde{\theta}(t,s)\mathrm{d}t,$$

with

$$\widetilde{\theta}(t,s) = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n}\sum_{k=1}^{k_n} \frac{\widehat{\xi}_{ik,rec}^{obs}\delta_i^{[Y]}(s)Y_i(s)}{\widehat{\lambda}_{k,rec}^{obs}}\widehat{\phi}_{k,rec}^{obs}(t), \tag{4.3.1}$$

for all $(t,s) \in \mathcal{T} \times \mathcal{S}$, where $\widehat{\xi}_{ik,rec}^{obs} = \int_{\mathcal{T}} X_i^\star(t)\widehat{\phi}_{k,rec}^{obs}(t)\mathrm{d}t$ and $\widehat{\phi}_{1,rec}^{obs}, \ldots, \widehat{\phi}_{k_n,rec}^{obs}$ and $\widehat{\lambda}_{1,rec}^{obs} \ldots, \widehat{\lambda}_{k_n,rec}^{obs}$ represent respectively the k_n first eigenfunctions and eigenvalues of the covariance operator $\widehat{\Gamma}_{n,rec}^{obs}$.

Alternatively, if we denote $\widehat{\Pi}_{k_n,rec}^{obs}$ the projection on the space spanned by the k_n first eigenfunctions of $\widehat{\Gamma}_{n,rec}^{obs}$, the estimation of the slope operator Θ is given by

$$\widetilde{\Theta} = \widehat{\Pi}_{k_n,rec}^{obs}\widehat{\Delta}_{n,rec}^{obs}\left(\widehat{\Pi}_{k_n,rec}^{obs}\widehat{\Gamma}_{k_n,rec}^{obs}\widehat{\Pi}_{k_n,rec}^{obs}\right)^{-1}.$$

and the imputation $Y_{\ell,imp}$ can also be written

$$Y_{\ell,imp}(s) = \widehat{\Pi}_{k_n,rec}^{obs}\widehat{\Delta}_{n,rec}^{obs}\left(\widehat{\Pi}_{k_n,rec}^{obs}\widehat{\Gamma}_{k_n,rec}^{obs}\widehat{\Pi}_{k_n,rec}^{obs}\right)^{-1} X_\ell^\star(t),$$

4.3.2) Estimation of the slope operator and its kernel and prediction

Once the whole database has been reconstructed, we estimate the bivariate functional coefficient θ with

$$\widehat{\theta}^{\star\star}(t,s) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{k_n} \frac{\widehat{\xi}_{ik,rec}^{\star}Y_i^{\star\star}(s)}{\widehat{\lambda}_{k,rec}^{\star}}\widehat{\phi}_{k,rec}^{\star}(t), \tag{4.3.2}$$

for all $(t,s) \in \mathcal{T} \times \mathcal{S}$, where $Y_i^{\star\star} = Y_i \delta_i^{[Y]} + Y_{i,imp}(1-\delta_i^{[Y]})$ for all $i = 1, \ldots, n$. The estimation of the operator Θ is similarly given by

$$\widehat{\Theta}^{\star\star} = \widehat{\Pi}^{\star}_{k_n,rec} \widehat{\Delta}^{\star\star}_{n,rec} \left(\widehat{\Pi}^{\star}_{k_n,rec} \widehat{\Gamma}^{\star}_{n,rec} \widehat{\Pi}^{\star}_{k_n,rec} \right)^{-1},$$

where $\widehat{\Delta}^{\star\star}_{n,rec}$ is the empirical counterpart of the cross-covariance operator $\Delta^{\star\star}_{rec}$.

We use this estimation to predict a new value of the response Y when a new explanatory curve X_{new} is given by

$$\widehat{Y}_{new}(s) = \int_{\mathcal{T}} X^{\star}_{new}(t) \widehat{\theta}^{\star\star}(t,s) \mathrm{d}(t),$$

which can also be written

$$\widehat{Y}_{new} = \widehat{\Pi}^{\star}_{k_n,rec} \widehat{\Delta}^{\star\star}_{n,rec} \left(\widehat{\Pi}^{\star}_{k_n,rec} \widehat{\Gamma}^{\star}_{k_n,rec} \widehat{\Pi}^{\star}_{k_n,rec} \right)^{-1} X^{\star}_{new}.$$

4.3.3) Asymptotic results

The proof of these results follows the same lines as the proof of **Theorem (3.1)** and **Theorem (3.3)** in Crambes et al. (2022).

Theorem 4.3.1. *Under assumptions **(A.1)-(A.4)**, if we take $k_n \sim p^{1/(a_O+2)}$ and $p \sim n^{\eta_1}$ with $\eta_1 \leqslant 1/2$, we obtain*

$$\mathbb{E}\left(\left\| \left(\widetilde{\Theta} \cdot X^{\star}_{\ell} - \Theta \cdot X^{\star}_{\ell} \right) \right\|^2 \right) = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}} \right),$$

for $\ell \in \widetilde{\mathcal{D}}_m$, where $\widetilde{\mathcal{D}}_m$ is the set of missing responses of size $m_n^{[Y]}$.

Theorem 4.3.2. *Under assumptions **(A.1)-(A.4)**, and $k_n \sim p^{1/(a_O+2)}$ and $p \sim n^{\eta_1}$ with $\eta_1 \leqslant 1/2$, the prediction error is*

$$\mathbb{E}\left(\left\| \left(\widehat{\Theta}^{\star\star} \cdot X^{\star}_{new} - \Theta \cdot X^{\star}_{new} \right) \right\|^2 \right) = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + \frac{n^{\eta_1/(a_O+2)}}{n - m_n^{[Y]}} \right).$$

Remark 4.3.3. *Comparing parameters as $n^{\eta_1/(a_O+2)-1-\zeta_1(b_O-1)/(b_O+2)} \lesssim \frac{n^{\eta_1/(a_O+2)}}{n-m_n^{[Y]}}$, we find that the prediction error with reconstruction (obtained in Section 4.2) is asymptotically at least the same than the prediction error with imputation.*

Remark 4.3.4. *Theoretical results are generally obtained under assumptions concerning the rate of convergence of the integer k_n. In practice, this integer is selected by minimizing a certain empirical criterion such as Generalized Cross Validation (GCV) criterion, Cross Validation (CV) criterion, or K-fold Cross Validation (K-fold CV) criterion. We*

chose in the following simulation section the GCV procedure, known to be computationally fast. The GCV criteria is given as follows for imputation

$$GCV(k_n) = \frac{(n - m_n^{[Y]}) \sum_{i=1}^{n} \left\| \left(\widetilde{\Theta} \cdot X_i^{\star} - \Theta \cdot X_i^{\star} \right) \right\|^2 \delta_i^{[Y]}}{\left((n - m_n^{[Y]}) - k_n \right)^2},$$

and analogously for prediction.

4.4) Simulations

4.4.1) Methodology

In this section, we conduct Monte Carlo experiments to illustrate the finite-sample performance of the proposed methods presented in section 4.2 and section 4.3. We set $\mathcal{S} = \mathcal{T} = [0, 1]$. Each response and predictor curve is observed at $q = 90$ and $p = 100$ equally spaced points in their domains, respectively. For computational simplicity we consider equidistant points $s_j = j/(q-1)$, $j = 0, \ldots, q-1$, and $t_k = k/(p-1)$, $k = 0, \ldots, p-1$. Two simulated data mechanisms are generated. Each model is defined by

$$Y^{(w)}(s) = \int_0^1 X^{(w)}(t)\theta^{(w)}(t, s)\mathrm{d}t + \epsilon^{(w)}(s), \tag{4.4.1}$$

for $t \in \mathcal{T}$ and $s \in \mathcal{S}$, where $w = 1, 2$. We approximate the integral in (4.4.1) using a Riemann sum over the grid $\boldsymbol{t}$. The analytical expressions of the kernels and processes are given below.

SCENARIO 1 The kernel is given by

$$\theta^{(1)}(t, s) = \sum_{k=1}^{50} \sum_{j=1}^{50} \theta_{k,j} \phi_k(t) \psi_j(s),$$

where $\phi_k(t) = \sqrt{2}\cos(k\pi t)$, $\psi_j(s) = \sqrt{2}\cos(j\pi s)$ and $\theta_{k,j} = 4(-1)^{k+j} k^{-2.5} j^{-3}$. The input is the random function $X^{(1)}(t) = \sum_{k=1}^{50} k^{-1}\xi_k \phi_k(t)$, where the ξ_k's are independently sampled from the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$. Finally, the noise is given by $\epsilon^{(1)}(s) = \sum_{j=1}^{50} \beta_j' \psi_j(s)$ and the β_j''s are independent mean-zero Gaussian random variables with variances equal to $j^{-1.1}$. Then, the covariance function $\mathrm{C}_{X^{(1)}}$ of $X^{(1)}$ is given by

$$\mathrm{C}_{X^{(1)}}(t_1, t_2) = \sum_{k=1}^{50} \frac{2}{k^2} \cos(k\pi t_1) \cos(k\pi t_2), \quad \text{where } t_1, t_2 \in \mathcal{T}.$$

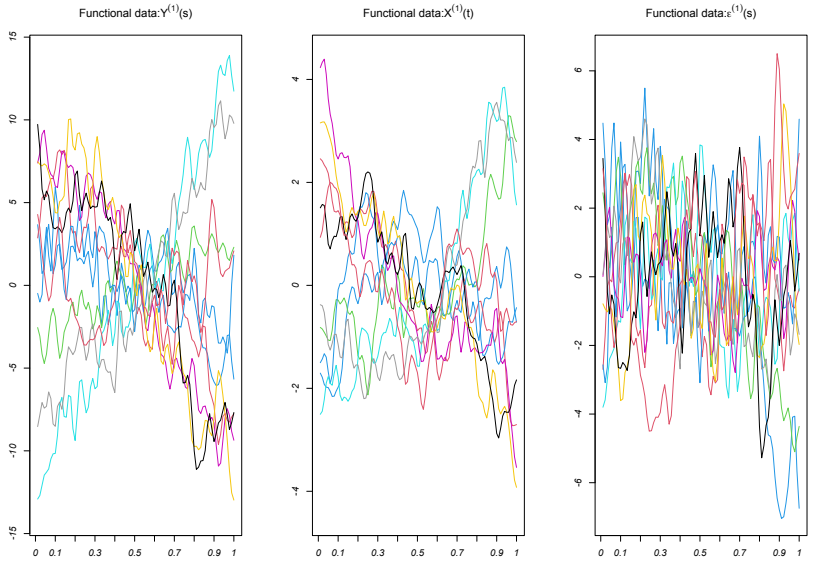

Figure 4.1: Examples of simulated functions with SCENARIO 1.

SCENARIO 2 The kernel is defined by $\theta^{(2)}(t,s) = s^3 + \sin(2\pi t)^3$ and the noise $\epsilon^{(2)}$ is generated according to a standard Brownian motion divided by 20. In addition, the functional covariate $X^{(2)}$ is generated through its covariance function, defined, for all $t_1, t_2 \in \mathcal{T}$, by

$$\mathrm{C}_{X^{(2)}}(t_1,t_2) = \frac{\sigma_1^2 \exp\left(-|t_1-t_2|^{\alpha}\right)}{\varsigma},$$

with $\sigma_1 = 1$, $\alpha = 2$ and $\varsigma = 0.2$. In this setting, even if a polynomial decrease of the eigenvalues of the covariance operator of $X^{(2)}$ is required in our theoretical results (see assumption ***(A.4)***), we want to see how the method works in practice if this assumption is no more satisfied, namely here in the case of an exponential decay.

Figure 4.1 shows 10 discretized predictor functions $X_i^{(1)}$, the error functions $\epsilon_i^{(1)}$ and the response functions $Y_i^{(1)}$. Figure 4.2 shows 10 discretized predictor functions $X_i^{(2)}$, the error functions $\epsilon_i^{(2)}$ and the response functions $Y_i^{(2)}$.

Figure 4.3 shows the covariance functions of the covariate for SCENARIO 1 and 2. Figure 4.4 shows kernel functions for SCENARIO 1 and 2.

To deal with partially observed curves for the covariate and response, we adopted the missing data simulation scenario from Kneip and Liebl (2020) and Crambes et al. (2022) such that

- 70% (respectively 55%) of the curves are fully observed on $[0,1]$,

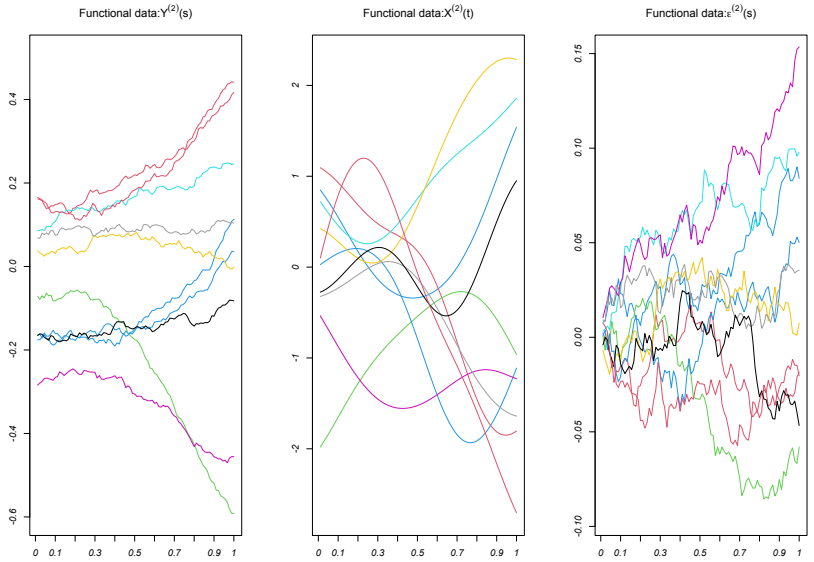

Figure 4.2: Examples of simulated functions with SCENARIO 2.

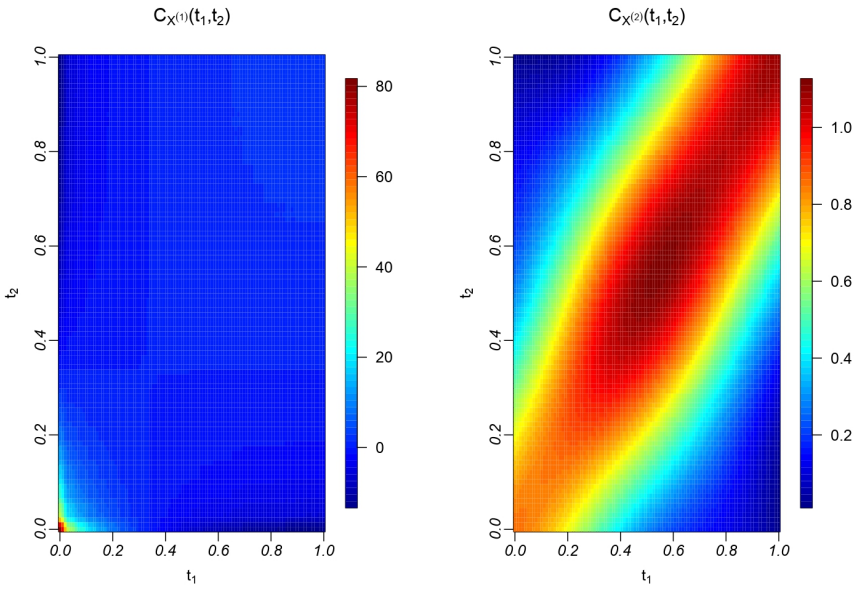

Figure 4.3: The covariance functions for SCENARIO 1 and 2.

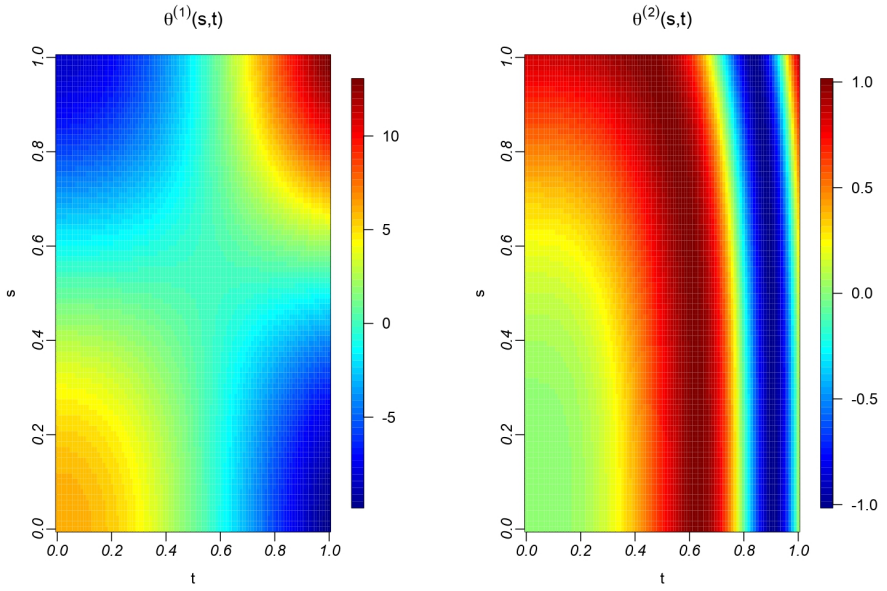

Figure 4.4: The kernel functions for SCENARIO 1 and 2.

- for the 30% (respectively 45%) of partially observed curves, the curves X_i and Y_i are fully observed on $[A_i, B_i] \subset [0, 1]$ with A_i drawn with uniform law on the interval $[0, A]$ and $B_i = A_i + B$, with either $A = 1/50$ and $B = 49/50$ or $A = 3/50$ and $B = 47/50$ for SCENARIO 1. We take either $A = 1/50$ and $B = 49/50$ or $A = 5/50$ and $B = 45/50$ for SCENARIO 2.

We simulate the number of missing data on the response Y and the indicator $\delta^{[Y]}$ by the logistic functional regression. The variable δ follows the Bernoulli law with parameter $p(X)$ such that

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \left[\int_S \left(\int_{\mathcal{T}} |s-t|X(t)dt\right)^2 ds\right]^{1/2} + ct,$$

where ct is a constant allowing to take different levels of missing data. For exemple, $ct = 2$ gives around 9.727% of missing data and $ct = 0.2$ gives around 38.801% of missing data.

Let us notice that we use a spline smoothed version of the different estimators (4.1.4), (4.2.3), (4.3.1) and (4.3.2), according to the so-called Smooth Principal Components Regression (SPCR) from Cardot et al. (2003). Let us remark that, with appropriate conditions on the spline parameters, all the theoretical results obtained in our work will also apply when using the SPCR estimation. We use a regression spline basis with 20 knots, a degree 3 and the order of derivation 2. The choice of these parameters is not crucial in our study, especially in comparison with the choice of the number of

principal components. The choice of this optimal tuning parameter is made on a growing sequence of dimension $k_n = 2, \ldots, 22$.

The dataset of size N is randomly splitted into a training set of size $n = \frac{4}{5}N$ and a test set of size $n_1 = \frac{1}{5}N$. We consider sample sizes N = 500, 1300. For each scenario, we use 200 Monte Carlo runs for the model assessment. In all numerical experiments, the proposed estimators have been carried out with the free software R.

4.4.2) Criteria

We use the following criteria to evaluate the performance of the methods.

- Criterion 1: $\overline{MSPE} = \frac{1}{\textbf{\textit{Sim}}} \sum_{j=1}^{\textbf{\textit{Sim}}} MSPE(j)$ is the average mean square prediction error. This criterion tends to zero when the sample size tends to infinity, where $MSPE(j) = \frac{1}{n_1} \sum_{\ell=n+1}^{n+n_1} \left\| \left(\widehat{\Theta} \cdot X_\ell^{\star,j} - \Theta \cdot X_\ell^{\star,j} \right) \right\|^2$ is the mean square prediction error computed on the j^{th} simulated sample, $j \in \{1, \ldots, \textbf{\textit{Sim}}\}$.
- Criterion 2: $\overline{RT} = \frac{1}{\textbf{\textit{Sim}}} \sum_{j=1}^{\textbf{\textit{Sim}}} RT(j)$ is the average ratio respect to truth. This criterion tends to one when the sample size tends to infinity, where $RT(j) = \frac{\sum_{\ell=n+1}^{n+n_1} \left\| \left(\widehat{\Theta} \cdot X_\ell^{\star,j} \right) - Y_\ell^j \right\|^2}{\sum_{\ell=n+1}^{n+n_1} \left\| \epsilon_\ell^j \right\|^2}$ is the ratio between the mean square prediction error and the mean square prediction error when the true parameters are known, computed on the j^{th} simulated sample.

We consider another criterion which is the determination coefficient R^2. In this context of functional regression setting, several definitions exist. Given the fitted values $\widehat{Y}_i(s)$, we used the definition as in Harezlak et al. (2007) given by

$$R^2 = \frac{1}{|\mathcal{S}|} \int_{\mathcal{S}} R^2(s) \mathrm{d}s = \frac{1}{|\mathcal{S}|} \int_{\mathcal{S}} \left(1 - \frac{\sum_{i=1}^{n} \left(Y_i(s) - \widehat{Y}_i(s) \right)^2}{\sum_{i=1}^{n} Y_i(s)^2} \right) \mathrm{d}s.$$

4.4.3) Simulation results

We denote the methods presented in this paper by :

- **Reconst_X_Y** : X and Y are partially observed , the missing parts of X and Y are reconstructed.
- **Reconst_X, Imp_Y** : X and Y are partially observed, the missing parts of X are reconstructed and Y imputed.

Moreover, we compare to other methods :

- **Full_X_Y** : X and Y are fully observed, this corresponds to the complete reference dataset.

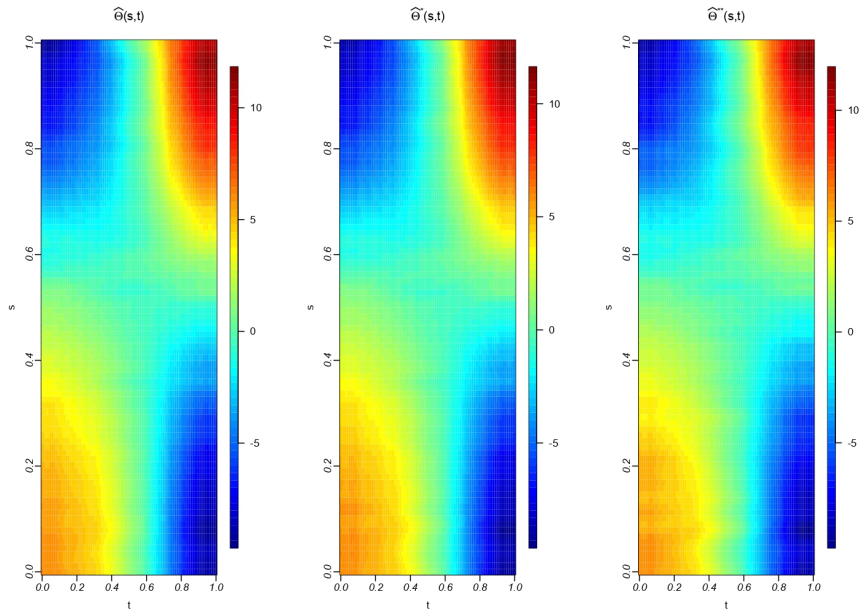

Figure 4.5: The estimated coefficient functions for SCENARIO 1.

- **Reconst_X, Remov_Y** : X and Y are partially observed, the missing parts of X are reconstructed and the missing part of Y are removed from the sample.
- **Remov _X_Y** : X and Y are partially observed, the individuals presenting partially observed curves are removed from the sample.

Even if our main goal is prediction, Figure 4.5 show estimates of the kernel function in SCENARIO 1 (with a sample size N = 400) for the dimension $k_n^\star$ chosen by the GCV criterion, respectively with full data ($\widehat{\Theta}$), reconstruction of the missing parts of X and Y ($\widehat{\Theta}^\star$) and reconstruction of the missing parts of X and imputation of Y ($\widehat{\Theta}^{\star\star}$). The missing part is 12% for both curves X and Y, the observed part being [3/50,47/50]. Moreover, 39.375% of curves Y (with ct = 0.1) are affected by missing data and 42.250% of curves X are affected by missing data. We remark that the estimators look graphically quite close, $\widehat{\Theta}^\star$ seems to be a little closer to $\widehat{\Theta}$ than $\widehat{\Theta}^{\star\star}$. A similar plot is obtained for SCENARIO 2 (Figure 4.6) with 37.812% of curves Y (with ct = 0.1) affected by missing data and 46.750% of curves X affected by missing data. In this situation, $\widehat{\Theta}^\star$ seems much closer to $\widehat{\Theta}$ than $\widehat{\Theta}^{\star\star}$.

We give in Table 4.2 the values of the determination coefficient R^2 and the value $k_n^\star$ chosen by the GCV criterion both scenarios 1 and 2. In scenario 1, we get a worse R^2 coefficient, maybe due to the fact that the curves X are not so smooth and do not seem easy to reconstruct.

Tables 4.3, 4.4, 4.5 and 4.6 give the values of the criteria $\overline{MSPE}$ and $\overline{RT}$ for scenarios 1 and 2 with different values of sample size, and different levels of missing data.

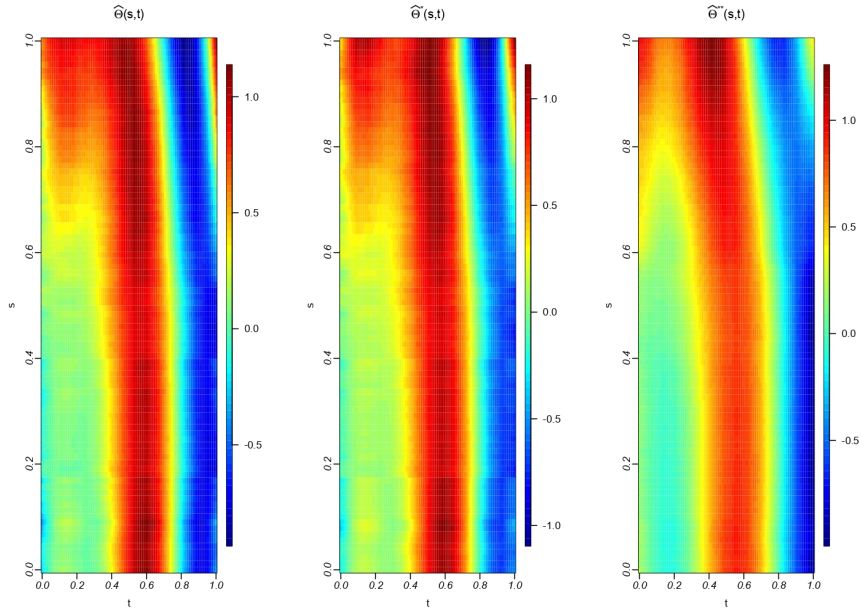

Figure 4.6: The estimated coefficient functions for SCENARIO 2.

Table 4.2: R^2 and $k_n^\star$ for scenarios 1 and 2.

Methods		SCENARIO 1	SCENARIO 2
Full_X_Y	R^2	68.996 %	98.652 %
	$k_n^\star$	2	6
Reconst_X_Y	R^2	68.859 %	98.641 %
	$k_n^\star$	2	6
Reconst_X, Imp_Y	R^2	68.853 %	98.630 %
	$k_n^\star$	2	5

The first conclusion is the fact that the errors decrease as the sample size increases. Secondly, these errors increase with the percentage of missing data on X or on Y. The rate of missing data on Y seems to have a more important impact on the errors, whatever the scenario we consider. In all cases, the method **Reconst_X_Y** reconstructing both curves X and Y has a better behaviour than the method **Reconst_X, Imp_Y** reconstructing X and imputing Y, which is quite in accordance to our theoretical results. The part of the observed curve is an important parameter in the curve reconstruction: as it can be expected, the results are better when the curve reconstruction is easier (for example when the observed part is $[1/50, 49/50]$, corresponding to 4% of missing information on the curves). Results tend to deteriorate when the curve reconstruction is harder (for example when the observed part is $[3/50, 47/50]$, corresponding to 12% of missing information on the curves). Finally, these two methods behave better than the other more naive methods (**Reconst_X, Remov_Y** and **Remov _X_Y**) that partially or completely ignore missing individuals affected by missing data.

Table 4.3: Mean and standard deviation errors for the predicted values based on 200 simulation replications with different levels of missing data and a sample size 500 (left panel) and a sample size 1300 (right panel). Partially observed curves are fully observed on $[1/50, 49/50]$ with SCENARIO 1.

Rate of missing	10.150	10.181	40.599	40.542
data in Y in %	(1.407)	(1.481)	(2.742)	(2.596)
Rate of missing	30.114	44.814	29.778	44.958
data in X in %	(1.998)	(2.498)	(2.354)	(2.180)
Full_X_Y : $\overline{MSPE} \times 10^3$	27.618	27.472	27.620	27.312
	(7.073)	(7.162)	(7.584)	(7.077)
$\overline{RT} \times 10$	10.073	10.067	10.071	10.081
	(0.066)	(0.058)	(0.057)	(0.066)
Reconst_X_Y : $\overline{MSPE} \times 10^3$	28.159	28.234	28.843	28.804
	(7.149)	(7.270)	(8.222)	(7.654)
$\overline{RT} \times 10$	10.075	10.069	10.072	10.084
	(0.068)	(0.059)	(0.059)	(0.070)
Reconst_X, Imp_Y : $\overline{MSPE} \times 10^3$	30.388	30.547	43.315	41.853
	(7.754)	(8.388)	(13.043)	(10.940)
$\overline{RT} \times 10$	10.080	10.074	10.113	10.120
	(0.070)	(0.061)	(0.073)	(0.082)
Reconst_X, Remov_Y : $\overline{MSPE} \times 10^3$	30.588	30.609	43.981	42.239
	(8.032)	(8.464)	(13.870)	(11.563)
$\overline{RT} \times 10$	10.080	10.074	10.114	10.120
	(0.071)	(0.061)	(0.075)	(0.084)
Remov_X_Y : $\overline{MSPE} \times 10^3$	40.575	48.537	58.408	70.458
	(11.757)	(14.961)	(19.486)	(21.764)
$\overline{RT} \times 10$	10.105	10.112	10.147	10.198
	(0.101)	(0.108)	(0.113)	(0.138)

Rate of missing	10.174	10.236	40.345	40.293
data in Y in %	(0.991)	(0.933)	(1.694)	(1.659)
Rate of missing	30.085	44.919	30.030	44.911
data in X in %	(1.238)	(1.352)	(1.326)	(1.391)
Full_X_Y : $\overline{MSPE} \times 10^3$	12.418	12.473	12.686	12.164
	(2.692)	(3.016)	(3.498)	(2.636)
$\overline{RT} \times 10$	10.029	10.031	10.030	10.033
	(0.026)	(0.027)	(0.026)	(0.025)
Reconst_X_Y : $\overline{MSPE} \times 10^3$	12.909	13.231	13.906	13.399
	(2.726)	(3.013)	(3.982)	(3.101)
$\overline{RT} \times 10$	10.030	10.034	10.033	10.036
	(0.027)	(0.028)	(0.027)	(0.028)
Reconst_X, Imp_Y : $\overline{MSPE} \times 10^3$	14.110	14.111	19.008	19.135
	(3.114)	(3.178)	(4.963)	(4.669)
$\overline{RT} \times 10$	10.032	10.036	10.047	10.052
	(0.026)	(0.029)	(0.031)	(0.035)
Reconst_X, Remov_Y : $\overline{MSPE} \times 10^3$	14.138	14.152	19.538	19.401
	(3.103)	(3.227)	(5.113)	(4.769)
$\overline{RT} \times 10$	10.032	10.036	10.049	10.052
	(0.026)	(0.029)	(0.032)	(0.035)
Remov_X_Y : $\overline{MSPE} \times 10^3$	18.444	22.823	25.462	30.748
	(4.846)	(5.951)	(6.564)	(8.185)
$\overline{RT} \times 10$	10.047	10.058	10.061	10.077
	(0.036)	(0.045)	(0.044)	(0.068)

Table 4.4: Mean and standard deviation errors for the predicted values based on 200 simulation replications with different levels of missing data and a sample size 500 (left panel) and a sample size 1300 (right panel). Partially observed curves are fully observed on $[3/50, 47/50]$ with SCENARIO 1.

Rate of missing	10.250	10.193	40.324	40.250
data in Y in %	(1.502)	(1.607)	(2.412)	(2.415)
Rate of missing	29.915	44.870	29.830	44.879
data in X in %	(2.010)	(2.301)	(2.067)	(2.156)
Full_X_Y : $\overline{MSPE} \times 10^3$	26.989	27.259	28.020	27.176
	(7.726)	(7.507)	(7.759)	(7.059)
$\overline{RT} \times 10$	10.072	10.069	10.078	10.068
	(0.063)	(0.060)	(0.063)	(0.062)
Reconst_X_Y : $\overline{MSPE} \times 10^3$	34.408	38.372	36.244	38.411
	(8.040)	(8.323)	(8.611)	(7.747)
$\overline{RT} \times 10$	10.089	10.098	10.098	10.095
	(0.078)	(0.081)	(0.077)	(0.083)
Reconst_X, Imp_Y : $\overline{MSPE} \times 10^3$	36.752	40.348	49.018	51.103
	(8.608)	(8.839)	(10.995)	(11.675)
$\overline{RT} \times 10$	10.094	10.105	10.123	10.127
	(0.081)	(0.082)	(0.088)	(0.094)
Reconst_X, Remov_Y : $\overline{MSPE} \times 10^3$	36.891	40.403	49.436	51.639
	(8.717)	(8.860)	(11.610)	(12.131)
$\overline{RT} \times 10$	10.095	10.105	10.124	10.127
	(0.082)	(0.083)	(0.087)	(0.094)
Remov_X_Y : $\overline{MSPE} \times 10^3$	39.619	47.940	56.390	69.935
	(10.117)	(14.026)	(16.332)	(21.684)
$\overline{RT} \times 10$	10.100	10.125	10.147	10.174
	(0.087)	(0.111)	(0.111)	(0.118)

Rate of missing	10.174	10.033	40.423	40.514
data in Y in %	(0.965)	(0.898)	(1.426)	(1.548)
Rate of missing	30.182	45.107	30.108	44.896
data in X in %	(1.232)	(1.459)	(1.285)	(1.290)
Full_X_Y : $\overline{MSPE} \times 10^3$	12.573	12.141	12.608	12.303
	(3.071)	(2.684)	(3.125)	(2.797)
$\overline{RT} \times 10$	10.034	10.031	10.033	10.034
	(0.026)	(0.027)	(0.027)	(0.029)
Reconst_X_Y : $\overline{MSPE} \times 10^3$	19.898	22.796	20.396	23.408
	(3.654)	(3.328)	(3.815)	(3.604)
$\overline{RT} \times 10$	10.052	10.060	10.053	10.057
	(0.038)	(0.044)	(0.041)	(0.045)
Reconst_X, Imp_Y : $\overline{MSPE} \times 10^3$	21.153	23.828	26.195	28.732
	(3.999)	(3.659)	(5.075)	(5.266)
$\overline{RT} \times 10$	10.053	10.063	10.067	10.068
	(0.040)	(0.044)	(0.045)	(0.049)
Reconst_X, Remov_Y : $\overline{MSPE} \times 10^3$	21.180	23.847	26.477	29.101
	(3.969)	(3.694)	(5.229)	(5.365)
$\overline{RT} \times 10$	10.054	10.063	10.068	10.069
	(0.040)	(0.044)	(0.046)	(0.049)
Remov_X_Y : $\overline{MSPE} \times 10^3$	18.892	22.105	25.893	31.192
	(4.792)	(5.681)	(6.873)	(8.474)
$\overline{RT} \times 10$	10.047	10.058	10.067	10.073
	(0.042)	(0.051)	(0.050)	(0.063)

4.5) Proofs

4.5.1) Proof of Theorem 4.2.1

Starting with the reconstruction cross covariance operator,

$$\begin{aligned}
\widehat{\Delta}^{\star}_{n,rec} =& \frac{1}{n}\sum_{i=1}^{n} Y_i^{\star} \otimes X_i^{\star} \\
=& \frac{1}{n}\sum_{i=1}^{n} \Big(Y_i + (Y_i^{\star} - Y_i)\Big) \otimes X_i^{\star}, \\
=& \frac{1}{n}\sum_{i=1}^{n} Y_i \otimes X_i^{\star} + \frac{1}{n}\sum_{i=1}^{n} (Y_i^{\star} - Y_i) \otimes X_i^{\star}, \\
=& \Theta\widehat{\Gamma}_{n,rec} + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \otimes X_i^{\star} + \frac{1}{n}\sum_{i=1}^{n} (Y_i^{\star} - Y_i) \otimes X_i^{\star}.
\end{aligned}$$

Table 4.5: Mean and standard deviation errors for the predicted values based on 200 simulation replications with different levels of missing data and a sample size 500 (left panel) and a sample size 1300 (right panel). Partially observed curves are fully observed on $[1/50, 49/50]$ with SCENARIO 2.

Rate of missing	9.727	9.757	38.801	38.865
data in Y in %	(1.575)	(1.456)	(2.297)	(2.383)
Rate of missing	29.826	44.816	30.032	45.152
data in X in %	(2.069)	(2.192)	(2.253)	(2.288)
Full_X_Y : $\overline{MSPE} \times 10^6$	20.654	19.593	20.815	19.787
	(10.200)	(8.722)	(9.723)	(9.110)
$\overline{RT} \times 10$	10.158	10.143	10.194	10.153
	(0.219)	(0.210)	(0.234)	(0.202)
Reconst_X_Y : $\overline{MSPE} \times 10^6$	20.706	19.614	22.114	20.838
	(10.140)	(8.743)	(10.572)	(9.366)
$\overline{RT} \times 10$	10.158	10.142	10.208	10.162
	(0.219)	(0.211)	(0.241)	(0.209)
Reconst_X, Imp_Y : $\overline{MSPE} \times 10^6$	22.473	21.336	32.038	31.248
	(11.216)	(9.245)	(16.620)	(15.914)
$\overline{RT} \times 10$	10.170	10.159	10.291	10.240
	(0.229)	(0.228)	(0.298)	(0.254)
Reconst_X, Remov_Y : $\overline{MSPE} \times 10^6$	22.473	21.340	32.052	31.312
	(11.213)	(9.239)	(16.651)	(15.948)
$\overline{RT} \times 10$	10.171	10.159	10.291	10.241
	(0.229)	(0.228)	(0.298)	(0.256)
Remov_X_Y : $\overline{MSPE} \times 10^6$	30.380	38.660	46.638	58.802
	(13.412)	(19.485)	(21.421)	(30.753)
$\overline{RT} \times 10$	10.206	10.293	10.403	10.518
	(0.322)	(0.438)	(0.451)	(0.577)

Rate of missing	9.726	9.678	39.232	39.113
data in Y in %	(0.947)	(0.974)	(1.551)	(1.417)
Rate of missing	29.847	45.090	30.013	45.147
data in X in %	(1.290)	(1.365)	(1.203)	(1.425)
Full_X_Y : $\overline{MSPE} \times 10^6$	7.646	7.919	8.067	7.931
	(3.042)	(3.148)	(3.528)	(3.736)
$\overline{RT} \times 10$	10.062	10.065	10.065	10.055
	(0.085)	(0.086)	(0.079)	(0.082)
Reconst_X_Y : $\overline{MSPE} \times 10^6$	7.740	8.009	8.917	8.860
	(3.070)	(3.224)	(3.640)	(3.916)
$\overline{RT} \times 10$	10.062	10.066	10.072	10.063
	(0.085)	(0.088)	(0.082)	(0.087)
Reconst_X, Imp_Y : $\overline{MSPE} \times 10^6$	8.441	8.809	12.872	12.691
	(3.599)	(3.633)	(5.396)	(6.048)
$\overline{RT} \times 10$	10.067	10.072	10.103	10.098
	(0.091)	(0.096)	(0.101)	(0.121)
Reconst_X, Remov_Y : $\overline{MSPE} \times 10^6$	8.442	8.809	12.938	12.702
	(3.599)	(3.632)	(5.390)	(6.052)
$\overline{RT} \times 10$	10.067	10.072	10.104	10.098
	(0.009)	(0.010)	(0.010)	(0.012)
Remov_X_Y : $\overline{MSPE} \times 10^6$	12.175	15.342	17.406	23.095
	(5.109)	(6.675)	(7.836)	(10.587)
$\overline{RT} \times 10$	10.106	10.107	10.146	10.175
	(0.138)	(0.154)	(0.144)	(0.189)

Next, we obtain

$$
\begin{aligned}
&\mathbb{E}\left(\left\|\widehat{\Theta}^{\star}\cdot X^{\star}_{new} \;-\; \Theta\cdot X^{\star}_{new}\right\|^2\right)\\
&= \mathbb{E}\left(\left\|\widehat{\Pi}_{k_n,rec}\widehat{\Delta}^{\star}_{n,rec}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X^{\star}_{new} - \Theta\cdot X^{\star}_{new}\right\|^2\right)\\
&\leqslant 4\mathbb{E}\left(\left\|\widehat{\Pi}_{k_n,rec}\Theta\widehat{\Gamma}_{n,rec}\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X^{\star}_{new} - \Theta\cdot X^{\star}_{new}\right\|^2\right)\\
&+ 4\mathbb{E}\left(\left\|\widehat{\Pi}_{k_n,rec}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\otimes X^{\star}_i\right)\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X^{\star}_{new}\right\|^2\right)\\
&+ 2\mathbb{E}\left(\left\|\widehat{\Pi}_{k_n,rec}\left(\frac{1}{n}\sum_{i=1}^{n}(Y^{\star}_i - Y_i)\otimes X^{\star}_i\right)\left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1}X^{\star}_{new}\right\|^2\right).
\end{aligned}
$$

Table 4.6: Mean and standard deviation errors for the predicted values based on 200 simulation replications with different levels of missing data and a sample size 500 (left panel) and a sample size 1300 (right panel). Partially observed curves are fully observed on $[5/50, 45/50]$ with SCENARIO 2.

Rate of missing	9.520	9.684	39.322	38.958
data in Y in %	(1.629)	(1.552)	(2.519)	(2.277)
Rate of missing	30.069	45.063	30.075	45.245
data in X in %	(2.096)	(2.232)	(2.106)	(2.213)
Full_X_Y : $\overline{MSPE} \times 10^6$	19.781	20.747	20.282	19.100
	(9.412)	(9.748)	(10.321)	(8.074)
$\overline{RT} \times 10$	10.169	10.170	10.176	10.160
	(0.217)	(0.229)	(0.222)	(0.217)
Reconst_X_Y : $\overline{MSPE} \times 10^6$	24.236	26.528	32.528	32.342
	(10.292)	(10.650)	(12.805)	(11.271)
$\overline{RT} \times 10$	10.198	10.204	10.274	10.271
	(0.222)	(0.240)	(0.261)	(0.258)
Reconst_X, Imp_Y : $\overline{MSPE} \times 10^6$	24.832	27.026	34.701	35.151
	(10.676)	(11.140)	(14.806)	(14.391)
$\overline{RT} \times 10$	10.209	10.203	10.288	10.281
	(0.235)	(0.249)	(0.295)	(0.281)
Reconst_X, Remov_Y : $\overline{MSPE} \times 10^6$	24.845	27.032	34.788	35.305
	(10.685)	(11.154)	(14.898)	(14.574)
$\overline{RT} \times 10$	10.209	10.205	10.289	10.281
	(0.235)	(0.249)	(0.295)	(0.282)
Remov_X_Y : $\overline{MSPE} \times 10^6$	29.948	39.318	44.115	56.488
	(13.803)	(18.900)	(20.894)	(31.158)
$\overline{RT} \times 10$	10.259	10.345	10.390	10.458
	(0.346)	(0.387)	(0.430)	(0.548)

Rate of missing	9.629	9.673	39.176	39.195
data in Y in %	(0.963)	(0.946)	(1.550)	(1.502)
Rate of missing	29.976	44.939	29.942	44.967
data in X in %	(1.284)	(1.488)	(1.325)	(1.392)
Full_X_Y : $\overline{MSPE} \times 10^6$	7.923	7.563	7.556	7.916
	(3.677)	(3.144)	(3.433)	(3.391)
$\overline{RT} \times 10$	10.056	10.065	10.063	10.067
	(0.082)	(0.079)	(0.087)	(0.078)
Reconst_X_Y : $\overline{MSPE} \times 10^6$	10.834	11.297	16.881	17.037
	(3.977)	(3.315)	(5.605)	(4.891)
$\overline{RT} \times 10$	10.082	10.096	10.132	10.139
	(0.088)	(0.086)	(0.117)	(0.109)
Reconst_X, Imp_Y : $\overline{MSPE} \times 10^6$	11.062	11.409	14.049	15.743
	(4.226)	(3.421)	(5.293)	(5.268)
$\overline{RT} \times 10$	10.082	10.097	10.118	10.127
	(0.094)	(0.091)	(0.107)	(0.117)
Reconst_X, Remov_Y : $\overline{MSPE} \times 10^6$	11.067	11.412	14.086	15.773
	(4.222)	(3.423)	(5.289)	(5.290)
$\overline{RT} \times 10$	10.082	10.098	10.119	10.128
	(0.094)	(0.091)	(0.108)	(0.118)
Remov_X_Y : $\overline{MSPE} \times 10^6$	12.702	15.481	16.638	22.266
	(5.567)	(7.721)	(7.140)	(11.556)
$\overline{RT} \times 10$	10.092	10.125	10.151	10.195
	(0.120)	(0.156)	(0.150)	(0.224)

Applying several times the identity $(a+b)^2 \leqslant 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, we get

$$\begin{aligned}
\mathbb{E}\left(\left\|\widehat{\Theta}^{\star} \cdot X^{\star}_{new} \;-\; \Theta \cdot X^{\star}_{new}\right\|^2\right) \leqslant\; & 64\mathbb{E}\left(\left\|\Theta\widehat{\Pi}_{k_n,rec}X^{\star}_{new} - \Theta\widehat{\Pi}_{k_n}X^{\star}_{new}\right\|^2\right) \\
& + 64\mathbb{E}\left(\left\|\Theta\widehat{\Pi}_{k_n}X^{\star}_{new} - \Theta\widehat{\Pi}_{k_n}X_{new}\right\|^2\right) \\
& + 32\mathbb{E}\left(\left\|\Theta\widehat{\Pi}_{k_n}X_{new} - \Theta\Pi_{k_n}X_{new}\right\|^2\right) \\
& + 16\mathbb{E}\left(\left\|\Theta\Pi_{k_n}X_{new} - \Theta X_{new}\right\|^2\right) \\
& + 8\mathbb{E}\left(\left\|\Theta X_{new} - \Theta X^{\star}_{new}\right\|^2\right) \\
& + 4\mathbb{E}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\langle X^{\star}_i, \left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X^{\star}_{new}\rangle\epsilon_i\right\|^2\right) \\
& + 2\mathbb{E}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\langle X^{\star}_i, \left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X^{\star}_{new}\rangle(Y^{\star}_i - Y_i)\right\|^2\right).
\end{aligned}$$

Results of terms in the above decomposition are in Crambes et al. (2022), exceptionally the last term, let be noted by $P_n = \frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star, \left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X_{new}^\star\rangle(Y_i^\star - Y_i)$. Hence, using the Cauchy-Schwarz inequality, we have

$$\mathbb{E}(\|P_n\|^2) \leqslant \sqrt{\mathbb{E}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\langle X_i^\star, \left(\widehat{\Pi}_{k_n,rec}\widehat{\Gamma}_{n,rec}\widehat{\Pi}_{k_n,rec}\right)^{-1} X_{new}^\star\rangle\right\|^4\right)\mathbb{E}\left(\|Y_i^\star - Y_i\|^4\right)}.$$

The result comes from Lemma 5.2 in Crambes and Henchiri (2019) and the result (4.2.4) that gives us

$$\begin{aligned}\mathbb{E}(\|P_n\|^2) &= o\left(\frac{k_n}{n}\right) + \mathcal{O}\left(n^{-\zeta_1(b_O-1)/(b_O+2)}\right)\\ &= \mathcal{O}\left(n^{\eta_1/(a_O+2)-1-\zeta_1(b_O-1)/(b_O+2)}\right).\end{aligned}$$

Summarizing, we get

$$\mathbb{E}\left(\left\|\widehat{\Theta}^\star \cdot X_{new}^\star - \Theta \cdot X_{new}^\star\right\|^2\right) = \mathcal{O}_p\left(n^{-\eta_1(a_O-1)/(2(a_O+2))} + n^{\eta_1/(a_O+2)-1-\zeta_1(b_O-1)/(b_O+2)}\right).$$

CHAPTER 5

CONCLUSIONS ET PERSPECTIVES

5.1) Conclusions Générales

Cette thèse a contribué à l'étude de modèles linéaires fonctionnels, en prenant en compte à la fois des covariables partiellement observées et des données manquantes sur la réponse. Nous avons étudié les questions théoriques et pratiques sur la prédiction dans les modèles fonctionnels suivants

1. La covariable fonctionnelle X est partiellement observée et la réponse réelle Y contient des données manquantes, les parties manquantes de X sont reconstruites et les valeurs manquantes de Y sont imputées par l'imputation simple (Chapitre 2).

2. La covariable fonctionnelle X est partiellement observée et la réponse réelle Y contient des données manquantes, les parties manquantes de X sont reconstruites et les valeurs manquantes de Y sont imputées par l'imputation multiple (Chapitre 3).

3. La covariable fonctionnelle X et la réponse fonctionnelle Y sont partiellement observées, les parties manquantes de X sont reconstruites et les parties de courbes non observées de Y sont complétées par deux méthodes : Imputation et reconstruction (Chapitre 4).

Dans les trois chapitres, nous concluons que la vitesse de convergence de la prédiction est subordonnée à la vitesse de convergence de la reconstruction de la courbe, la vitesse de convergence de l'imputation de la réponse est moins importante.

5.2) Perspectives

Il reste encore de nombreuses questions à étudier dans les futures recherches. Nous en décrivons ici quelques-uns.

- Trouver une application du chapitre 4, dont la variable explicative fonctionnelle et la variable à expliquer fonctionnelle sont partiellement observées, en utilisant la méthode d'imputation et la méthode de reconstruction.

- Reconstruire la partie non-observée de la réponse, Y^M à partir de toutes les parties observées, Y^O et X^O.

  On utilise l'expression de Karhunen–Loève (KL) pour les courbes X et Y dans $\mathbb{L}^2(\mathcal{T})$ et $\mathbb{L}^2(\mathcal{S})$ respectivement,

  $$X(t) = \sum_{k=1}^{+\infty} \xi_k \phi_k(t) \quad \text{et} \quad Y(s) = \sum_{j=1}^{+\infty} \beta_j \psi_j(s),$$

  pour tout $t \in \mathcal{T}$ et $s \in \mathcal{S}$, où $\xi_k = \int_{\mathcal{T}} X(t)\phi_k(t)\mathrm{d}t$ et $\beta_j = \int_{\mathcal{S}} Y(s)\psi_j(s)\mathrm{d}s$ sont les composantes principales fonctionnelles. Ce sont des variables aléatoires non corrélées avec une moyenne nulle et des variances $\mathbb{E}(\xi_k^2) = \lambda_k$ et $\mathbb{E}(\beta_j^2) = \mu_j$ pour tout $k, j \geqslant 1$.

  On écrit la courbe partiellement observée de la réponse sous forme d'un opérateur de reconstruction et une erreur de reconstruction,

  $$Y_i^M = \mathcal{J}(Y_i^O) + W_i, \quad \text{pour tout } i \in \{1, \ldots, n\},$$

  et on cherche à résoudre les deux problèmes de minimisation

  $$\begin{cases} \min\limits_{\Theta} \|Y - \Theta X\|^2, \\ \min\limits_{\mathcal{J}} \|Y^M - \mathcal{J} Y^O\|^2. \end{cases}$$

  L'opérateur de reconstruction de la courbe réponse est un opérateur de Hilbert-Schmidt sur les espaces $\mathbb{L}^2$ correspondent aux opérateurs de régression linéaire, qui s'écrit sous la forme suivante,

  $$\mathcal{J}(Y_i^O)(s) = \int_{\mathcal{T}} b(t,s) Y_i^O(t)\mathrm{d}t + \omega_i(s), \tag{5.2.1}$$

  où $b \in \mathbb{L}^2(\mathcal{T} \times \mathcal{S})$ est le noyau de l'opérateur et ω_i est l'erreur. En utilisant l'écriture du modèle, l'équation (5.2.1) devient

  $$\begin{aligned} \mathcal{J}(Y_i^O)(s) &= \int_{\mathcal{T}} b(t,s) \left(\int_{\mathcal{S}} \theta(t,s) X_i^O(s) + \epsilon_i(s)\mathrm{d}s \right) \mathrm{d}t + \omega_i(s) \\ &= \int_{\mathcal{T}} \int_{\mathcal{S}} b(t,s)\theta(t,s) X_i^O(s)\mathrm{d}s\mathrm{d}t + Z_i(s). \end{aligned}$$

On peut commencer à estimer la fonction de régression θ à partir du modèle, en utilisant la partie observée de la réponse. On reconstruit la courbe partiellement observée de la variable explicative à l'aide de l'opérateur de reconstruction étudié par Kneip and Liebl (2020) (suivant les mêmes démarches dans le chapitre 2).

Ensuite, on peut écrire le noyau b sous la forme suivante,

$$b(t,s) = \sum_{k=1}^{+\infty} \frac{\mathbb{E}\left(\beta_k Y(s)\right)}{\mu_k} \psi_k(t).$$

On estime cette fonction et on obtient l'estimateur de l'opérateur de reconstruction. Finalement, on prédit les nouvelles courbes.

- Estimation de quantiles conditionnels par projection sur un espace de Hilbert dans un modèle linéaire fonctionnel. Le modèle de régression sur quantiles s'écrit

$$Y = \langle \Psi_\tau, X \rangle + \varepsilon_\tau,$$

où Y et ε_τ sont à valeurs dans $\mathbb{R}$, X est à valeurs dans $\mathbb{L}^2(\mathcal{I})$ et Ψ_τ est un opérateur de $\mathbb{L}^2(\mathcal{I})$ dans $\mathbb{R}$. La réponse Y contient des données manquantes et la variable explicative X est partiellement observée. Le bruit ε_τ vérifie $\mathbb{P}\left(\varepsilon_\tau \leqslant 0 | X_i\right) = \tau$ où $\tau \in]0;1[$ est l'ordre du quantile.

L'estimateur du quantile conditionnel d'ordre τ est la solution du problème de minimisation suivant

$$\Psi_\tau(x) = \min_{a \in \mathbb{R}} \mathbb{E}\left(\rho_\tau(Y-a) | X = x\right),$$

pour tout $x \in \mathbb{L}^2(\mathcal{I})$, où la fonction ρ_τ est la "check function" définie par $\rho_\tau(u) = |u| + (2\tau - 1)u$, pour tout $u \in \mathbb{R}$. On se base sur les travaux de Cardot et al. (2005) (un estimateur est basé sur des fonctions splines) et Crambes et al. (2013) (un estimateur basé sur la méthode Support Vector Machine (SVM)).

L'utilisation de ce modèle pourrait être une alternative pour l'imputation de données manquantes sur la réponse Y, comme cela a été fait dans le chapitre 2.

BIBLIOGRAPHY

Aguilera, A., F. Ocaña, and M. Valderrama (2008). Estimation of functional regression models for functional responses by wavelet approximation. In S. Dabo-Niang and F. Ferraty (Eds.), *In Functional and Operatorial Statistics: Contributions to Statistics*, Chapter 3, pp. 15–21. Physica-Verlag/Springer, Heidelberg.

Allison, P. (2001). *Missing data - Quantitative applications in the social sciences*. Thousand Oaks, CA: Sage.

Aneiros, G., I. Horová, M. Hušková, and P. Vieu (2020). *Functional and High-Dimensional Statistics and Related Fields*. Springer International.

Benatia, D., M. Carrasco, and J.-P. Florens (2017). Functional linear regression with functional response. *Journal of Econometrics 201*(2), 269–291. THEORETICAL AND FINANCIAL ECONOMETRICS: ESSAYS IN HONOR OF C. GOURIEROUX.

Bosq, D. (1991). Nonparametric statistics for stochastic processes. *NATO, ASI Series*, 509–52.

Bosq, D. (2000). *Linear processes in function spaces: Theory and applications*. New York: Springer Verlag.

Briggs, A., T. Clark, J. Wolstenholme, and P. Clarke (2003). Missing.... presumed at random: cost-analysis of incomplete data. *Health Economics 12*.

Brunel, E., A. Mas, and A. Roche (2016). Non-asymptotic adaptive prediction in functional linear models. *Journal of Multivariate Analysis 143*, 208–232.

Bugni, F. (2012). Specification test for missing functional data. *Econ. Theory 28*, 959–1002.

Bugni, F., P. Hall, J. L. Horowitz, and G. R. Neumann (2009). Goodness-of-fit tests for functional data. *Econometrics Journal 12*(s1), S1–S18.

Cai, T. and P. Hall (2006). Prediction in functional linear regression. *Annals of Statistics 34*, 2159–2179.

Cai, T. T. and M. Yuan (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association 107*(499), 1201–1216.

Cardot, H., C. Crambes, and P. Sarda (2005). Quantile regression when the covariates are functions. *Journal of Nonparametric Statistics 17*, 841–856.

Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics and Probability Letters 45*, 11–22.

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Statistica Sinica 13*, 571–591.

Chiou, J.-M., Y.-C. Zhang, W.-H. Chen, and C.-W. Chang (2014). A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B: Transport Dynamics 2*, 106–129.

Comte, F. and J. Johannes (2012). Adaptive functional linear regression. *The Annals of Statistics 40*(6), 2765–2797.

Crambes, C., C. Daayeb, A. Gannoun, and Y. Henchiri (2022). Functional linear model with partially observed covariate and missing values in the response. *Journal of Nonparametric Statistics.*.

Crambes, C., A. Gannoun, and Y. Henchiri (2013). Support vector machine quantile regression approach for functional data : Simulation and application studies. *Journal of Multivariate Analysis 121*, 50–68.

Crambes, C. and Y. Henchiri (2019). Regression imputation in the functional linear model with missing values in the response. *Journal of Statistical Planning and Inference 201*, 103–119.

Crambes, C., N. Hilgert, and T. Manrique (2016). Estimation of the noise covariance operator in functional linear regression with functional outputs. *Statistics and Probability Letters 113*, 7–15.

Crambes, C., A. Kneip, and P. Sarda (2009). Smoothing splines estimators for functional linear regression. *The Annals of statistics 37*, 35–72.

Crambes, C. and A. Mas (2013). Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli 19*.

Dauxois, J. and A. Pousse (1976). Les analyses factorielles en calcul des probabilités et en statistique: Essai d'étude synthétique, university of toulouse.

Dauxois, J., A. Pousse, and Y. Romain (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis 12*, 136–154.

Day, S. (1999). *Dictionary for clinical trials*. Hoboken: John Wiley & Sons.

Delaigle, A. and P. Hall (2013). Classification using censored functional data. *the American Statistical Association 108*, 1269–1283.

Delaigle, A. and P. Hall (2016). Approximating fragmented functional data by segments of markov chains. *Biometrika 103*, 779–799.

Delaigle, A., P. Hall, W. Huang, and A. Kneip (2020). Estimating the covariance of fragmented and other related types of functional data. *Journal of the American Statistical Association 116*, 35–72.

Descary, M. H. and V. M. Panaretos (2019). Recovering covariance from functional fragments. *Biometrika 106*, 145–160.

Deville, J. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Ann. Insee 15*.

Ellenberg, J. (2014). *How Not to Be Wrong: The Power of Mathematical Thinking*. Penguin UK: Westminster.

Ellenberg, J. (2018). *L'art de ne pas dire n'importe quoi : ce que le bon sens doit aux mathématiques*. Gillingham: Cassini.

Febrero-Bande, M., P. Galeano, and W. Gonzalez-Manteiga (2019). Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random. *Computational Statistics and Data Analysis 131*, 91–103.

Febrero-Bande, M., P. Galeano, and W. González-Manteiga (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *International Statistical Review 85*, 61–83.

Ferraty, F., A. Mas, and P. Vieu (2007). Nonparametric regression on functional data : inference and practical aspects. *Aust. N. Z. J. Stat. 49*, 267–286.

Ferraty, F., M. Sued, and P. Vieu (2013). Mean estimation with data missing at random for functional covariables. *Statistics 47*, 688–706.

Ferraty, F., I. Van Keilegom, and P. Vieu (2012). Regression when both response and predictor are functions. *Journal of Multivariate Analysis 109*, 10–28.

Ferraty, F. and P. Vieu (2002). The functional nonparametric model and application to spectrometric data. *Comput. Statist. 17*, 545–564.

Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis: Theory and practice*. New York: Springer Verlag.

Fisher, R. A. (1934). Effect of methods of ascertainment upon the estimation of frequencies. *Annals of Human Genetics 6*, 13–25.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society 45*, 135–45.

Galton, F. (1898). An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London 62*, 310–315.

Gellar, J. E., E. Colantuoni, D. M. Needham, and C. M. Crainiceanu (2014). Variable-domain functional regression for modeling icu data. *Journal of the American Statistical Association 109*, 1425–1439.

Goldberg, Y., Y. Ritov, and A. Mandelbaum (2014). Predicting the continuation of a function with applications to call center data. *Journal of Statistical Planning and Inference 147*, 53–65.

Graham, J. W. (2012). *Missing data analysis and design*. New York: Springer Verlag.

Greenland, S. and W. Finkle (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology 142*, 1255–1264.

Gromenko, O., P. Kokoszka, and J. Sojka (2017). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *The Annals of Applied Statistics, 11*, 898–918.

Hall, P. and J. Horowitz (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics 35*, 70–91.

Hall, P. and M. Hosseini-Nasab (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society Series B (Statistical Methodology) 68*, 109–126.

Hall, P., H.-G. Müller, and J.-L. Wang (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics 34*, 1493–1517.

Hansen, M.H., H. W. and W. Madow (1953). *Sample Survey Methods and Theory*, Volume 1. New York: Wiley.

Harezlak, J., B. A. Coull, N. M. Laird, S. R. Magari, and D. C. . Christiani (2007). Penalized solutions to functional regression problems. *Computational Statistics and Data Analysis 51*, 4911–4925.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In D. Pfeffermann and C. Rao (Eds.), *Handbook of Statistics: Sample Surveys: Design, Methods and Applications*, Volume 29, pp. 215–256.

Haziza, D. and J. N. Rao (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology, Statistics Canada 32*.

He, Y., R. Yucel, and T. Raghunathan (2011). A functional multiple imputation approach to incomplete longitudinal data. *Stat. Med 30*, 1137–1156.

He, Y., G. Zhang, and C. Hsu (2022). *Multiple imputation of missing data in practice: Basic theory and analysis strategies*. New York: John Wiley and Sons.

Horváth, L. and P. Kokoszka (2012). *Inference for functional data with applications*. New York: Springer Verlag.

Hsing, T. and R. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley and Sons: Wiley series in probability and statistics.

Hötelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*, 417–441 and 498–520.

Imaizumi, M. and K. Kato (2016). Pca-based estimation for functional linear regression with functional responses.

Joseph, L. and J. L. Schafer (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica 57*(1), 19–35.

Joseph, L., J. L. Schafer, and M. K. Olsen (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research 33*(4), 545–571.

Karhunen, K. (1947). *Uber lineare methoden in der wahrscheinlichkeitsrechnung*. Annales Academiae Scientarum Fennicae.

Kneip, A. and D. Liebl (2020). On the optimal reconstruction of partially observed functional data. *The Annals of Statistics 48*, 1692–1717.

Kokoszka, P. and M. Reimherr (2018). *Introduction to functional data analysis*. New York: Chapman and Hall.

Kraus, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B 77*, 777–801.

Kraus, D. (2019). Inferential procedures for partially observed functional data. *Journal of Multivariate Analysis 173*, 583–603.

Kraus, D. and M. Stefanucci (2018). Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika 106*, 161–180.

Kraus, D. and M. Stefanucci (2020). Ridge reconstruction of partially observed functional data is asymptotically optimal. *Statistics and Probability Letters 165*.

Kuhrt, A. (1995). *The Ancient Near East (Routledge History of the Ancient World) c. 3000–330 B.C.E.* London: Routledge.

Li, T., F. Xie, X. Feng, J. Ibrahim, H. Zhu, and the Alzheimers Disease Neuroimaging Initiative (2018). Functional linear regression models for nonignorable missing scalar responses. *Statistica Sinica 28*, 1867–1886.

Li, Y. and T. Hsing (2007). On rates of convergence in functional linear regression. *Journal of Multivariate Analysis 98*, 1782–1804.

Lian, H. (2011). Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics 5*, 31–40.

Liebl, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *The Annals of Applied Statistics*, 1562–1592.

Liebl, D. (2019). Nonparametric testing for differences in electricity prices: The case of the fukushima nuclear accident. *The Annals of Applied Statistics 13*, 1128–1146.

Liebl, D. and S. Rameseder (2019). Partially observed functional data: The case of systematically missing parts. *Computational Statistics & Data Analysis 131*, 104–115.

Lin, Z. and J.-L. Wang. Mean and covariance estimation for functional snippets. *Journal of the American Statistical Association (Just-accepted)*. https://doi.org/10.1080/01621459.2020.1777138.

Lin, Z. and J.-L. Wang (2022). Mean and covariance estimation for functional snippets. *Journal of the American Statistical Association 117*, 348–360.

Lin, Z., J.-L. Wang, and Q. Zhong (2021). Basis expansions for functional snippets. *Biometrika 108*, 709–726.

Ling, N., R. Kan, P. Vieu, and S. Meng (2019). Semi-functional partially linear regression model with responses missing at random. *Metrika 82*, 39–70.

Ling, N., L. Liang, and P. Vieu (2015). Nonparametric regression estimation for functional stationary ergodic data with missing at random. *Journal of Statistical Planning and Inference, 162*, 75–87.

Little, R. and D. B. Rubin (2002). *Statistical analysis with missing data (Second edition)*. John Wiley, New York.

Little, R. J. A. and D. B. Rubin (2020). *Statistical analysis with missing data (Third edition)*. New York: John Wiley and Sons.

Lord, F. M. (1955). Estimation of parameters from incomplete data. *Journal of the American Statistical Association 271*, 870–876.

Loève, M. (1948). *Fonctions aleatoires du second ordre. Processus Stochastiques et Mouvement Brownien*. P. Levy (ed.).

Luo, R. and X. Qi (2017). Function-on-function linear regression by signal compression. *Journal of the American Statistical Association 112*(518), 690–705.

Mangel, M. and F. J. Samaniego (1984). Abraham wald's work on aircraft survivability. *Journal of the American Statistical Association 386*, 259–267.

McKendrick, A. (1926). Applications of mathematics to medical problems. *Proceeding of the Edinburgh Mathematical Society 44*, 98–130.

Müller H, G. and U. Stadtmüller (2005). Generalized functional linear models. *The Annals of Statistics 33*, 774–805.

Morris, J. (2015). Functional regression. *Annual Review of Statistics and Its Application 2*, 321–359.

Mukherjee, R. and C. R. Rao (1955). *The ancient inhabitants of Jebel Moya, Sudan*, Volume 123. England: Cambridge University Press.

Park, J. and J. Qian (2012). Functional regression of continuous state distribution. *Journal of Econometrics 167*, 397–412.

Park, Y., S. D. (2019). Robust probabilistic classi
cation applicable to irregularly sampled functional data. *Computational Statistics and Data Analysis 131*.

Park, Y., X. Chen, and D. S. Simpson (2021). Robust inference for partially observed functional response data. Preprint at http://www3.stat.sinica.edu.tw/preprint/SS-2020-0358_Preprint.pdf.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal in Science*, 559–572.

Prchal, L. and P. Sarda. (2007). Spline estimator for functional linear regression with functional response. Unpublished, Preprint at https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.583.1816&rep=rep1&type=pdf.

Preda, C., G. Saporta, and M. Hadj (2010). The nipals algorithm for functional data. *Revue Roumaine de Mathématique Pures et Appliquées 55*, 315–326.

Rachdi, M., A. Laksaci, Z. Kaid, A. Benchiha, and F. A. Al-Awadhi (2020). kNN local linear regression for functional and missing data at random. *Statistica Neerlandica 28*, 1867–1886.

Ramsay, J. O., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB*. New York: Springer Verlag.

Ramsay, J. O. and B. W. Silverman (2002). *Applied Functional data analysis: Methods and case studies*. New York: Springer Verlag.

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis (Second edition)*. New York: Springer Verlag.

Rancourt, E. (2001a). Edit and imputation : From suspicious to scientific techniques. *Proceeding of the international association of survey statisticians, Seoul 53rd session 2001*, 634–655.

Rancourt, E. (2001b). *Histoire du mot imputation*. Bulletin d'imputation.

Rao, C. R. (1985). *Weighted Distributions Arising Out of Methods of Ascertainment: What Population Does a Sample Represent?. In: Atkinson A.C., Fienberg S.E. (eds)*. New York: A Celebration of Statistics. Springer.

Rao C, R. (1958). Some statistical methods for the comparison of growth curves. *Biometrics 14*, 1–17.

Robert, G. (1949). *Leçons sur les instruments de bord et les équipements divers*.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley and Sons.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton: CRC Press.

Schafer, J. L. and J. W. Graham (2002). Missing data: Our view of the state of the art. *Psychological Methods 7*(2), 147–177.

Srivastava, A. and E. P. Klassen (2016). *Functional and Shape Data Analysis*. New York: Springer Verlag.

Sun, X., P. Du, X. Wang, and P. Ma (2018). Optimal penalized function-on-function regression under a Reproducing Kernel Hilbert Space framework. *Journal of the American Statistical Association 113*(524), 1601–1611.

Tucker L, R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika 23*, 19–23.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specication. *Statistical Methods in Medical Research 16*, 219–242.

Van Buuren, S. (2018). *Flexible imputation of missing data (Second edition)*. New York: Chapman and Hall.

Wainer, H. (2011). *Uneducated Guesses: Using Evidence to Uncover Misguided Education Policies*. Princeton: Princeton University Press.

Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016). Review of functional data analysis. *Annual Review of Statistics and Its Application 3*, 257–295.

Wang, L., R. Cao, J. Du, and Z. Zhang (2019). A nonparametric inverse probability weighted estimation for functional data with missing response data at random. *Journal of the Korean Statistical Society*.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics 3*, 163–195.

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics 33*(6), 2873–2903.

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association 100*, 577–590.

Zhou, J. and Q. Peng (2020). Estimation for functional partial linear models with missing responses. *Statistics and Probability Letters 156*.

Titre : Modèles linéaires fonctionnels avec des données partiellement observées.

Résumé : Le traitement des données manquantes est un problème important dans le processus d'observation ou d'enregistrement des données. L'objectif de ce travail est l'étude du modèle linéaire fonctionnel avec sortie réelle ou fonctionnelle lorsque les variables sont partiellement observées. Nous utilisons un opérateur de reconstruction des courbes explicatives avant d'imputer la variable réponse par différentes méthodes (imputation déterministe et imputation multiple dans le cas d'une réponse réelle, opérateur de reconstruction et imputation déterministe dans le cas d'une réponse fonctionnelle). Le comportement asymptotique de l'erreur quadratique moyenne de prédiction est étudié, ainsi que le comportement de la méthode en pratique sur des données simulées et réelles.

Mots-clès : Modèle linéaire fonctionnel, composantes principales fonctionnelles, données partiellement observées, données manquantes, imputation par régression, Prédiction.

Title : Functional linear models with partially observed data.

Abstract : The treatment of missing data is an important problem in the process of observation or data recording. The objective of this work is to study the functional linear model with real or functional output when the variables are partially observed. We use a reconstruction operator of the explanatory curves before imputing the response variable by different methods (deterministic imputation and multiple imputation in the case of a real response, reconstruction operator and deterministic imputation in the case of a functional response). The asymptotic behavior of the mean square error of prediction is studied, as well as the behavior of the method in practice on simulated and real data.

Keywords : Functional linear model, Functional Principal Components, partially observed data, Missing data, Regression imputation, Prediction.