Problèmes d'estimation dans des modèles de régression non paramétrique et de Poisson

Bilel Bousselmi

DOCTORAT BRETAGNE LOIRE MATHSTIC

INSA RENNES

THESE DE DOCTORAT DE

L'INSTITUT NATIONAL DES SCIENCES APPLIQUEES RENNES

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Mathématiques et leurs Interactions*

Par

Bilel BOUSSELMI

Problèmes d'estimation dans des modèles de régression non paramétrique et de Poisson

Thèse présentée et soutenue à l'INSA Rennes, le 06 / 12 / 2021.
Unité de recherche : IRMAR-UMR CNRS 6625
Thèse N° : 21 ISAR 30 / D21 - 30

Rapporteurs avant soutenance :

Fabienne COMTE	PR. Université Paris Descartes
Agathe GUILLOUX	PR. Université d'Évry Val d'Essonne

Composition du Jury :

Président :	Jean-Michel LOUBES	PR. Université de Toulouse
Examinateurs :	Fabienne COMTE	PR. Université Paris Descartes
	Valérie GARÉS	MCF. INSA Rennes
	Agathe GUILLOUX	PR. Université d'Évry Val d'Essonne
	Jean-Michel LOUBES	PR. Université de Toulouse
Dir. de thèse :	Jean-François DUPUY	PR. INSA de Rennes
Co-dir. de thèse :	Abderrazek KAROUI	PR. Faculté des sciences de Bizerte

Intitulé de la thèse :

Problèmes d'estimation dans des modèles de régression non paramétrique et de Poisson.

Bilel BOUSSELMI

En partenariat avec :

Remerciements

Comme beaucoup de doctorants, je me rends compte en faisant le bilan de ces trois années de thèse que celles-ci ont été intenses tant professionnellement que personnellement. Beaucoup de personnes qui ont croisé mon chemin et ont contribué ou non à ma thèse, avant ou pendant ont rendu cette aventure unique.

Je tiens en premier lieu à remercier Jean-François DUPUY et Abderrazek KAROUI, mes directeurs de thèse. Il est difficile de résumer ces trois années en quelques mots, mais je vous remercie sincèrement pour votre écoute, votre patience, votre disponibilité, votre soutien et la confiance que vous m'avez accordée durant ces trois années. Merci de m'avoir encouragé, laissé expérimenter, et guidé. J'ai énormément progressé et appris grâce à vous et j'espère être un jour aussi expérimentée et clairvoyant que vous. Ce fut un réel plaisir de travailler avec vous et j'espère que cela pourra continuer.

Un grand merci à Valérie GARES pour sa gentillesse et ses précieux conseils durant ces trois années.

Je souhaite remercier tout particulièrement Fabienne COMTE et Agathe GUILLOUX d'avoir eu la gentillesse de rapporter cette thèse. Je suis également très honoré de la présence de Valérie GARES et Jean-Michel LOUBES dans mon jury. Je vous en remercie.

Je remercie également mon comité de suivie de thèse : Aziz BELMILOUDI et Jean-Christophe BRETON d'avoir eu le temps de suivre ma thèse et de me donner des conseils précieux.

J'adresse également mes remerciements à l'ensemble des membres du département Génie Mathématiques de l'INSA Rennes: Aziz, Camar-Eddine, Léo, Martine, Mounir, Olivier, Patricia et en particulier Pierrette et Valérie pour leur écoute et leurs précieux conseils. Je remercie aussi plus largement l'équipe de statistique de l'IRMAR et le l'équipe du Laboratoire GAMA de la faculté des sciences de Bizerte pour leur gentillesse et leur écoute.

Merci à toutes les personnes que j'ai eu la chance de rencontrer dans le cadre de ma thèse, pour

les discussions et les conseils reçus. Merci aux doctorants que j'ai pu rencontrer pendant ces trois années, un grand à merci notamment à El Hassene, Essoham, Huan, Mériadec et Trinh pour toutes nos discussions (toujours optimistes !), qui continueront je l'espère. Merci à Audrey, Lorenza, Marie, Miriam, Othman. Merci à Mourad et Mustapha, j'ai passé avec vous de très bons moments.

Pour finir, je souhaite remercier de tout coeur ma (grande) famille, et en particulier mes parents pour leur soutien et leur confiance depuis toujours. Merci "Baba" et merci "Ommi" pour tout. Merci à mes frères Mahdi, Mohamed Ali et ma soeur Rym d'être toujours là et de me soutenir, et merci à Youssef, Meriem et Haroun. Je me sens extrêmement chanceuse d'être entourée d'une famille aussi merveilleuse, je vous aime infiniment.

À mes parents,

ceux qui m'ont fait naître dans ce monde,
et ceux qui m'ont aidé à y grandir.

Abstract

This thesis has two main contributions. The first contribution deals with study of random projection based estimators for solving nonparametric regression or linear functional regression problems. More precisely, the projection kernels we consider in the construction of our estimators are given by the Gegenbauer polynomials Christoffel-Darboux kernel and the convolution Sinc- kernel. In particular, we provide error and convergence analyses of the proposed estimators under some regularity assumptions on the class of regression functions. Also, we study an orthogonal projection based estimator for the fast and stable solution of a linear functional regression (LFR) problem. This problem is solved under the usual assumption that the unknown slope function to be estimated is well approximated by its projection on a finite dimensional subspace of a Hilbert space. Finally, we conduct a simulation study to assess finite-sample properties of these estimators.

The second contribution of this work deals with censored count data regression with missing censoring information. Note that Poisson regression is widely used to investigate the relationship between covariates and acountresponse. We consider the situation where the count of interest is randomly right-censored (for example,in a study of health-care utilization, patients reporting their number of visits to a doctor as“8 visits or more” provide a censored count, that is, a lower bound on the true unobserved count). The literature on censored count data analysis already contains several approaches for handling such data, we additionally suppose that the censoring indicator, which tells whether an observed count is censored or not, is missing for some subjects. For this second contribution, we propose several estimators based on the regression calibration, multiple imputation and augmented inverse probability weighting methods. Consistency, asymptotic normality and variance estimation of our estimators are rigorously established under appropriate regularity conditions. Simulation experiments are carried out to investigate the finite sample behaviour and relative performance of the proposed estimates.

Key words: nonparametric regression, linear functional regression, count data analysis, missing data, asymptotic properties, simulations.

Résumé

Cette thèse a deux contributions principales. La première contribution porte sur l'étude des estimateurs basés sur la projection aléatoire pour résoudre des problèmes de régression non paramétrique ou de régression fonctionnelle linéaire. Plus précisément, les noyaux de projection que nous considérons dans la construction de nos estimateurs sont donnés par le noyau des polynômes de Gegenbauer, Christoffel-Darboux et le noyau de convolution Sinc. En particulier, nous fournissons des analyses d'erreur et de convergence des estimateurs proposés sous certaines hypothèses de régularité sur la classe des fonctions de régression. Nous étudions également un estimateur basé sur la projection orthogonale pour une résolution stable d'un problème de régression fonctionnelle linéaire. Ce problème est résolu sous l'hypothèse habituelle que la fonction de pente inconnue à estimer est bien approximée par sa projection sur un sous-espace de dimension finie d'un espace de Hilbert. Enfin, nous menons une étude de simulation pour évaluer les propriétés de ces estimateurs.

La deuxième contribution de ce travail concerne la régression de poisson censurée en présence d'indicatrices de censure manquantes. Notons que la régression de Poisson est largement utilisée pour étudier la relation entre un ensemble des covariables et une variable de comptage. Nous considérons la situation où la variable de comptage observée peut être censurée aléatoirement à droite (par exemple, dans une étude sur l'utilisation des soins de santé, les patients déclarant leur nombre de visites chez un médecin comme « 8 visites ou plus » fournissent un nombre censuré, c'est-à-dire, une borne inférieure sur le vrai nombre non observé). La littérature sur l'analyse des données de comptage censuré contient déjà plusieurs approches pour traiter de telles données, nous supposons en outre que les indicatrices de censure, qui indiquent si un comptage observé est censuré ou non, sont manquantes pour certains individus de l'échantillon. Pour cette deuxième contribution, nous proposons plusieurs méthodes d'estimation : regression calibration, imputation multiple et estimation AIPW (pondération par l'inverse de la probabilité de sélection augmentée). La consistance, la normalité asymptotique et l'estimation de la variance de nos estimateurs sont rigoureusement établies sous des conditions de régularité appropriées. Une étude de simulation comparant ces différentes méthodes est décrite.

Mots clés: Régression non paramétrique, régression linéaire fonctionnelle, analyse des données de comptage, données manquantes, propriétés asymptotiques, simulations.

Notations

$\mathbb{P}(A)$	:	Probability of the event A
$\mathbb{E}(X)$	:	Mathematical expectation of a random variable X
$\text{var}(X)$	:	Variance of a random variable X
$\text{cov}(X, Y)$	:	Covariance of the two random variables X and Y
$X_n \xrightarrow{\mathbb{P}} X$	:	The sequence of random variables X_n converges in probability to X
$H^s_{\omega_\alpha, A}(I)$	:	The adapted weighted Sobolev space over I
$L^2(J)$	:	The Hilbert space of square integrable functions over J
$\mathbb{R}$	:	Set of real numbers
$X^\top$	:	Transpose of X
i.i.d.	:	independent and identically distributed
GLM	:	Generalized linear model
GLMs	:	Generalized linear models
TCL	:	Theorem central limit
IPW	:	Inverse probability weighting
AIPW	:	Augmented inverse probability weighting
MAR	:	Missing at random
MCAR	:	Missing Completely at Random
MNAR	:	Missing Not at Random
KRR	:	Kernel ridge regression
LFR	:	Linear functional regression
RKHS	:	Reproducing kernel Hilbert space
PSWFs	:	Prolate spheroidal wave functions
MISE	:	Mean integrated squared error

List of Tables

Contents

1 Introduction

La régression à pour but de modéliser la relation entre une variable réponse, et un ensemble de variables explicatives. Il existe plusieurs exemples de modèles de régression notamment: le modèle de régression linéaire et les modèles de comptage classiques (modèle de Poisson, ...). Ces modèles sont des cas particuliers des modèles linéaires généralisés, introduit par Nelder and Wedderburn (1972).
La régression non paramétrique permet de décrire la tendance entre une variable réponse et un ou plusieurs variables explicatives. Considérons l'ensemble de n observations $\{(X_i, Y_i),\ 1 \leqslant i \leqslant n\}$, le modèle de régression non paramétrique est défini par

$$Y_i = f(X_i) + \eta_i, \quad 1 \leqslant i \leqslant n,$$

où $(\eta_i)_{1\leqslant i\leqslant n}$ sont des variables aléatoires réelles i.i.d centrées et de variance σ^2. Ce modèle diffère des modèles de régression classiques en ce qu'il ne repose pas sur des hypothèses fortes concernant la forme de la relation entre les variables.
La régression non paramétrique est un outil d'exploration et de visualisation des données dont l'objectif de formaliser une idée préliminaire des données avant de procéder à un ajustement du modèle.
Les modèles de régression semi-paramétrique peuvent être obtenu en combinant les modèles de régression paramétrique et non paramétrique, voir par exemple (Ruppert et al., 2003).
Les approches de la régression non paramétrique comprennent l'estimation à noyau, la régression polynomiale locale, les modèles de régression basés sur les splines, les arbres de régression. Dans ce travail, nous avons développé un estimateur empirique basé sur un opérateur de projection orthogonale. Les noyaux de projection que nous avons considéré dans ce travail sont donnés par le noyau de Christoffel-Darboux Gegenbauer et le noyau de convolution Sinus-cardinal.
Nous avons également étudié un modèle de régression linéaire fonctionnelle donné par :

$$Y_i = \int_J X_i(s)\beta_0(s)\, ds + \varepsilon_i, \quad i = 1, ..., n$$

où Y_i sont des réponses scalaires, $X_i \in L_2(J)$ sont des prédicteurs fonctionnels et $(\varepsilon_i)_{1\leqslant i\leqslant n}$ sont des variables aléatoires réelles i.i.d centrées et variance σ^2 indépendantes de X_i et $\beta_0(\cdot)$ est la fonction inconnue à estimer.

La modélisation statistique des données de comptage est une question importante dans divers domaines: l'agriculture, l'économétrie, l'épidémiologie, les applications industrielles et l'assurance.

Le modèle de régression de Poisson est l'approche utilisée dans l'analyse des données de comptage qui a été étendue pour prendre en compte les données de comptage censurées. Bien que la censure soit généralement associée à l'analyse des données de survie, les données de comptage peuvent être également censurées.

Les modèles de comptage censurés peuvent être spécifiés de manière pratique en introduisant un indicateur de censure qui est égal à 1 si la variable réponse (le comptage) n'est pas censuré et à 0 sinon. Dans ce travail, nous considérons la situation où l'indicateur de censure est manquant pour certains individus de l'échantillon.
L'estimation dans le modèle de régression de Poisson censuré en présence d'indicatrices de censure manquantes reste un problème ouvert. Notre objectif dans ce travail est de fournir et de comparer plusieurs estimations adaptées à ce contexte.

Les problèmes de données manquantes ont donné lieu à une riche littérature et plusieurs méthodes d'estimation adaptées ont été proposées. Si, en amont, la maîtrise du processus de collecte des données est essentielle pour limiter au maximum l'impact des données manquantes, des variables collectées sont parfois incomplètes et leur mode de gestion le plus courant consiste à restreindre les analyses aux sujets pour lesquels l'ensemble des variables est entièrement renseigné (analyse dite cas-complet). Cette méthode généralement appliquée par défaut induit une perte systématique de puissance statistique et un risque potentiel de biais des estimations.

De nombreuses méthodes alternatives de traitement des données manquantes ont été proposées et appliquées dans divers domaines notamment en épidémiologie. Pour notre travail de recherche, nous avons fait le choix d'appliquer trois méthodes différentes : régression calibration, imputation multiple et estimation IPW (pondération par l'inverse de la probabilité de sélection) augmenté. Régression calibration est une méthode générale de traitement des variables manquantes ou mal mesurées. Elle consiste à remplacer les données manquantes par leur espérance conditionnelle compte tenu des données observées. Issue des travaux de Rubin et basée sur les statistiques bayésiennes, l'imputation multiple consiste à remplacer chaque donnée manquante par un jeu de données estimées à partir des données observées, ce qui permet de prendre en compte l'incertitude associée à chaque étape du processus d'imputation. Chacune des bases complètes ainsi générées fournit alors une estimation du paramètre d'intérêt, puis un estimateur unique est obtenu en calculant la moyenne de ces estimations. Le régression calibration et l'imputation multiple nécessitent tout les deux un modèle pour les données manquantes compte tenu des données observées. Comme pour l'analyse des cas complets, la pondération par probabilité inverse n'utilise que des cas complets, mais des poids sont utilisés pour rééquilibrer l'ensemble des cas complets. Le calcul de ces poids nécessite un modèle pour la probabilité

qu'un individu ait des données complètes. La pondération par probabilité inverse augmentée a ensuite été proposée pour assurer la robustesse contre une mauvaise spécification du modèle.

Nous résumons le contenu des différents chapitres ainsi que les résultats obtenus.

Dans le **chapitre 2**, nous introduisons les outils mathématiques qui seront nécessaires dans ce travail de thèse. D'abord, nous présentons quelques rappels essentiels sur les modèles de régression non paramétriques et régression linéaire fonctionnelle ainsi qu'une revue bibliographique de deux méthodes d'estimation dans ces deux modèles. Ensuite, nous énonçons des rappels sur les modèles linéaires généralisés, qui constituent le cadre de la modélisation des données de comptage. Puis, un rappel sur les données manquantes et les différentes approches qui ont été utilisées dans la littérature dans le cadre de la régression de Poisson censurée et on finira ce chapitre par un résumé sur les processus empiriques.

Dans le **chapitre 3**, qui est notre première contribution de cette thèse, nous nous sommes intéressés au problème de l'estimation de la fonction de régression dans un modèle de régression non-paramétrique en proposant un estimateur empirique basé sur un opérateur de projection orthogonale (les noyaux de projection considérés étant le noyau de polynômes de Gegenbauer Christoffel-Darboux et le noyau de convolution Sinc). Nous avons ensuite proposé un estimateur aléatoire pseudo-inverse, pour résoudre un problème de régression fonctionnelle linéaire, lorsque la fonction à estimer appartient à un sous-espace de dimension finie d'un espace de Hilbert. Les estimateurs proposés ont également été étudiés numériquement au moyen de simulations.

Contributions du chapitre :

- Régression non paramétrique
  - → Nous montrons que les opérateurs empiriques de projection orthogonale associés aux noyaux Gegenbauer et Sinc fournissent des approximations stables et assez précises de la vraie fonction de régression f et ils ont l'avantage d'être stables sans avoir besoin d'un processus de régularisation supplémentaire.
  - → Nous fournissons des analyses d'erreur et de convergence des estimateurs proposés sous l'hypothèse que la fonction de régression appartient à certains espaces fonctionnels appropriés et qu'ils ont un ordre de convergence similaire à celui des schémas connus dans la littérature à savoir de régression ridge à noyau (KRR).
- Régression linéaire fonctionnelle
  - → Sous certaines conditions, nous prouvons que notre estimateur aléatoire basé sur le pseudo-inverse fournit une approximation très proche de la vraie fonction de pente.
  - → Nous montrons que notre estimateur est stable et ne nécessite pas de processus de régularisation supplémentaire, contrairement aux schémas KRR classiques.

La deuxième contribution de cette thèse est décrite dans le **chapitre 4**, où nous intéressons au problème de l'inférence statistique dans le modèle de régression de Poisson censuré, en présence d'indicatrices de censure manquantes. Nous proposons plusieurs méthodes d'estimation des paramètres du modèle et nous établissons rigoureusement les propriétés asymptotiques des estimateurs. Nous comparons également l'ensemble de ces estimateurs au moyen de simulations.

Contributions du chapitre :

- Proposition de plusieurs méthodes d'estimation des paramètres du modèle (imputation par régression, imputation multiple, méthode robuste par pondération par l'inverse de la probabilité d'une donnée manquante).
- Etude théorique des propriétés asymptotiques des différents estimateurs (convergences en probabilité et en loi, estimation convergente de la variance asymptotique des estimateurs proposés).

Nos travaux de thèse ont donné lieu à des articles et conférences dont:

Articles

✍ B. Bousselmi, J.-F. Dupuy, A. Karoui. Censored count data regression with missing censoring information, *Electronic Journal of Statistics.* **15**(2): 4343-4383, **2021.** (DOI : 10.1214/21-EJS1897.)

✍ B. Bousselmi, J.-F. Dupuy, A. Karoui (2021). Random orthogonal projections based schemes for solving nonparametric and linear functional regressions problems, *article soumis.* Disponible sous arXiv.2001.11213.

Conférences

☞ B. Bousselmi, J.-F. Dupuy, A. Karoui. Régression fonctionnelle linéaire et non paramétrique basée sur des projections orthogonales aléatoires. *52èmes Journées de Statistique*. Nice, conférence virtuelle, **7-11 Juin 2021.**

☞ B. Bousselmi, J.-F. Dupuy, A. Karoui. Random orthogonal projections based schemes for solving nonparametric and linear functional regressions problems. *19th Conference of the Applied Stochastic Models and Data Analysis (ASMDA) International Society*. Athènes, conférence virtuelle, **1-4 juin 2021.**

☞ B. Bousselmi, J.-F. Dupuy, A. Karoui. Censored count data regression with missing data. *19th Conference of the Applied Stochastic Models and Data Analysis (ASMDA) International Society*. Athènes, conférence virtuelle, **1-4 juin 2021.**

☞ B. Bousselmi, J.-F. Dupuy, A. Karoui. Censored count data regression with missing censoring information. *13th International Conference of the ERCIM WG on Computational and Methodological Statistics(CMStatistics 2020)*. Londres, conférence virtuelle, **19-21 décembre 2020.**

2 Préliminaires mathématiques

Contents

2.1 Problèmes d'apprentissage et de régression linéaire fonctionnelle et non paramétrique

2.1.1 Problèmes d'apprentissage et de régression: définitions et propriétés

Dans la théorie de l'apprentissage, les observations sont tirées des deux ensembles: un espace métrique $\mathcal{X}$ ("input") et un espace de sortie $\mathcal{Y}$ ("output"). La relation entre les variables $x \in \mathcal{X}$ et $y \in \mathcal{Y}$ n'est pas déterministe et est décrite par une distribution de probabilité ρ, qui est connue seulement au travers de m observations i.i.d $Z = \{(X_i, Y_i),\ 1 \leqslant i \leqslant m\}$ tirées de $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ selon ρ.

Rappelons qu'en théorie de l'apprentissage, le problème de régression vise à obtenir de bonnes approximations de la fonction de régression construites par des algorithmes d'apprentissage à partir de l'ensemble d'observations aléatoires $Z = \{(X_i, Y_i),\ 1 \leqslant i \leqslant m\}$.

Lorsque $\mathcal{Y}$ est un sous-ensemble mesurable de $\mathbb{R}$, ce problème d'apprentissage est connu sous le nom d'un problème de régression non paramétrique.

Soit $\mathcal{Y} = \mathbb{R}$, l'erreur des moindres carrés pour une fonction $f : \mathcal{X} \to \mathcal{Y}$ est donnée par

$$\mathcal{E}(f) = \int_{\mathcal{Z}} (f(x) - y)^2 d\rho(x, y).$$

Soit $\rho(y|x)$ (resp. ρ_X) une mesure de probabilité conditionnelle (resp. marginale) sur $\mathcal{Y}$ (resp. $\mathcal{X}$). La relation entre $\rho, \rho(y|x)$ et ρ_X est donnée par l'égalité suivante

$$\int_{\mathcal{Z}} \varphi(x, y) d\rho(x, y) = \int_{\mathcal{X}} \Big(\int_{\mathcal{Y}} \varphi(x, y) d\rho(y|x) \Big) d\rho_X(x),$$

où $\varphi : \mathcal{Z} \to \mathbb{R}$ est une fonction intégrable.

La vraie fonction de régression associée au problème de régression est donnée par

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x), \quad x \in \mathcal{X}. \tag{2.1.1}$$

Proposition 2.1.1. *(Cucker and Smale, 2002)*

$$\forall f : \mathcal{X} \to \mathcal{Y}, \quad \mathcal{E}(f) = \int_{\mathcal{X}} (f(x) - f_\rho(x))^2 d\rho_X(x) + \sigma_\rho^2(x),$$

où $\sigma_\rho^2(x) = \int_{\mathcal{X}} \sigma^2(x) d\rho_X(x)$ *et* $\sigma^2(x) = \int_{\mathcal{Y}} (y - f_\rho(x))^2 d\rho(y|x)$.

Le premier terme du côté droit de la proposition 2.1.1 fournit une moyenne sur $\mathcal{X}$ de l'erreur déduite de l'utilisation de f comme modèle pour f_ρ.
D'après cette proposition, f_ρ minimise l'erreur de régression en norme L^2 c'est-à-dire qu'elle a la plus petite erreur possible parmi toutes les fonctions $f : \mathcal{X} \to \mathcal{Y}$ puisque σ_ρ^2 est indépendante de f.

Example 2.1.2. *(Cucker and Smale, 2002) L'apprentissage d'une loi physique par ajustement de la courbe aux données est un exemple classique d'apprentissage. Supposons que cette loi est modélisée par une fonction inconnue* $f : \mathbb{R} \to \mathbb{R}$ *qui a une forme spécifique et que l'espace de toutes les fonctions ayant cette forme puisse être paramétré par* N *nombres réels.*
Supposons par exemple que l'on dispose d'un ensemble d'observations $\{(X_i, Y_i), 1 \leqslant i \leqslant m\}$ *et que* f *est un polynôme de degré* d*, alors* $N = d + 1$ *et les paramètres sont les coefficients inconnus* $\omega_0, \omega_1, \ldots, \omega_d$ *de* f*. Dans ce cas, la recherche du meilleur ajustement par la méthode des moindres carrés permet d'estimer* f *à partir de l'ensemble d'observations .*
Si les mesures générant cet ensemble étaient exactes, alors $f(X_i)$ *serait égal à* Y_i*. Mais en général, les valeurs* Y_i *sont bruitées. Définissons pour* $m > N$*, la quantité suivante:*

$$\sum_{i=1}^{m} (f_\omega(x_i) - y_i)^2, f_\omega(x) = \sum_{j=1}^{m} \omega_j x^j, \tag{2.1.2}$$

On calcule donc le vecteur de coefficients ω *tel que 2.1.2 soit minimale. La technique des moindres carrés, remontant à Gauss et Legendre et qui est efficace en termes de calcul et repose sur l'algèbre linéaire numérique, résout ce problème de minimisation.*

Soit $Z \in \mathcal{Z}^m, Z = \{(X_i, Y_i), 1 \leqslant i \leqslant m\}$ un échantillon d'observations indépendantes de $\mathcal{Z}^m$ c'est-à-dire m observations tirés indépendamment selon la mesure de probabilité ρ. Ici, $\mathcal{Z}^m$ désigne le produit cartésien (m-fois) de Z.
L'erreur empirique de f est donnée par l'expression suivante

$$\mathcal{E}_Z(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$

Nous donnons maintenant quelques définitions et propriétés d'un espace de Hilbert à noyaux auto-reproduisant (RKHS), voir Aronszajn (1950) pour plus de détails. L'importance de ces espaces découle de la propriété (2.1.3) et du théorème du représentant 2.1.6.

2.1.2 Noyau défini positif et RKHS

On commence par noter que si $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un noyau défini positif (voir la définition suivante), alors pour tout ensemble fini de points distincts $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, la matrice $G_n = \Big[K(x_i, x_j)\Big]_{1 \leqslant i,j \leqslant n}$ est une matrice semi-définie positive.

Définition 2.1.3. *Un noyau défini positif sur un ensemble $\mathcal{X}$ est une fonction $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ qui est symétrique:*

$$\forall (x, x') \in \mathcal{X}^2, \quad K(x, x') = K(x', x),$$

et telle que pour tout $n \in \mathbb{N}$, $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ et $(a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$, on a

$$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \geqslant 0.$$

Définition 2.1.4. *Soit $\mathcal{X}$ un ensemble quelconque et $\mathcal{H}_K \subset R^{\mathcal{X}}$ une classe de fonctions formant un espace de Hilbert muni du produit interne $< ., . >_{\mathcal{H}_K}$. La fonction $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est appelée noyau auto-reproduisant de $\mathcal{H}_K$ si*

- *$\mathcal{H}_K$ contient toutes les fonctions de la forme*

$$\forall x \in \mathcal{X}, \quad K_x : t \mapsto K(x, t).$$

- *Soit $f \in \mathcal{H}_K$, alors*

$$f(x) =< f, K_x >_{\mathcal{H}_K}, \; x \in \mathcal{X}. \tag{2.1.3}$$

Si un tel noyau reproduisant existe, alors $\mathcal{H}_K$ est appelé un espace de Hilbert à noyau auto-reproduisant (RKHS).

Le théorème suivant combiné avec les définitions précédentes permet de trouver une relation entre un noyau défini positif et un RKHS.

Théorème 2.1.5. *(Aronszajn, 1950) K est un noyau défini positif sur l'ensemble $\mathcal{X}$ si et seulement s'il existe un RKHS $\mathcal{H}_K$ et une application:*

$$\phi : \mathcal{X} \mapsto \mathcal{H}_K,$$

telle que $\forall x, x' \in \mathcal{X}$:

$$K(x, x') =< \phi(x), \phi(x') > .$$

Le théorème du représentant suivant, voir par exemple (Argyriou et al., 2009), est fondamental pour l'utilisation des RKHS dans la résolution des problèmes de régression non paramétrique

Théorème 2.1.6. *Soient $\mathcal{X}$ un ensemble non vide, K un noyau défini positif sur $\mathcal{X} \times \mathcal{X}$, $\mathcal{H}_K$ le RKHS associé et $R : H \to \mathbb{R}$ une fonction différentiable. Étant donné n observations aléatoires $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ et une fonction d'erreur $E : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R}$, le problème suivant*

$$f^\star = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \{E((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))) + R(f)\},$$

admet une représentation de la forme

$$f^\star(\cdot) = \sum_{i=1}^{n} c_i K(x_i, x), \quad c_i \in \mathbb{R}$$

si et seulement s'il existe une fonction croissante $h : [0, \infty) \to \mathbb{R}$ pour laquelle $R(f) = h(\|f\|)$.

Nous illustrons cette section par quelques exemples de noyaux auto-reproduisants et leurs propriétés d'approximation spectrale.

Example 2.1.7. *(Noyau de Christoffel-Darboux Gegenbauer)*
Pour un paramètre réel $\alpha > -1/2$ et un entier positif N, le noyau de Christoffel-Darboux pour les polynômes de Gegenbauer est donné par:

$$K_N^\alpha(x, y) = \sum_{k=0}^{N} \left[\widetilde{C}_k^\alpha(x) \widetilde{C}_k^\alpha(y) \right] = A_N^\alpha \begin{cases} \dfrac{\widetilde{C}_{N+1}^\alpha(x)\widetilde{C}_N^\alpha(y) - \widetilde{C}_N^\alpha(x)\widetilde{C}_{N+1}^\alpha(y)}{x - y}, & x \neq y \\ \widetilde{C}_N^\alpha(x) \dfrac{d}{dx}\widetilde{C}_{N+1}^\alpha(x) - \widetilde{C}_{N+1}^\alpha(x) \dfrac{d}{dx}\widetilde{C}_N^\alpha(x), & x = y \end{cases} . \quad (2.1.4)$$

Ici $x, y \in I$, $A_N^\alpha = \frac{1}{2}\sqrt{\frac{(N+1)(N+2\alpha)}{(N+\alpha)(N+\alpha+1)}}$ et $\widetilde{C}_k^\alpha$ sont les polynômes de Gegenbauer orthonormaux de paramètre $\alpha > -\frac{1}{2}$ et de degré $k \geqslant 0$, voir par exemple [(Olver et al., 2010), pp.438-439].
Les polynômes de Gegenbauer non normalisés (également appelés polynômes ultra-sphériques) C_k^α avec un paramètre $\alpha > -\frac{1}{2}$ et de degré $k \geqslant 0$ sont donnés par la formule de récurrence suivante

$$C_{k+1}^\alpha(x) = \frac{2(k+\alpha)}{k+1}\, x\, C_k^\alpha(x) - \frac{k + 2\alpha - 1}{k+1} C_{k-1}^\alpha(x), \qquad k \geqslant 1,$$

et

$$C_0^\alpha(x) = 1, \quad C_1^\alpha(x) = 2\alpha x.$$

Les polynômes orthogonaux de Gegenbauer généralisent divers autres polynômes orthogonaux classiques. Par exemple, lorsque $\alpha = \frac{1}{2}$, les $C_k^\alpha(x)$ se réduisent aux polynômes de Legendre. Les polynômes

de Chebyshev sont également un cas particulier des $C_k^\alpha(x)$ avec $\alpha = 0$. Les polynômes de Gegenbauer sont aussi un cas particulier des polynômes de Jacobi orthonormaux notés par $\widetilde{P}_k^{\alpha,\beta}$ pour $\alpha, \beta > -1$ et $k \geqslant 0$. Cette relation est donnée par $\widetilde{C}_k^\alpha = \widetilde{P}_k^{\alpha-\frac{1}{2},\alpha-\frac{1}{2}}$. Les polynômes orthonormaux de Gegenbauer sont donnés par

$$\widetilde{C}_k^\alpha(x) = \frac{1}{\sqrt{h_k^\alpha}} C_k^\alpha(x), \qquad h_k^\alpha = \frac{2^{1-2\alpha}\pi\Gamma(k+2\alpha)}{k!(k+\alpha)\Gamma(\alpha)^2}.$$

On a la relation d'orthogonalité suivante

$$\int_{-1}^{1} \widetilde{C}_k^\alpha(x)\widetilde{C}_\ell^\alpha(x)\,\omega_\alpha(x)\,dx = \delta_{k\ell}, \qquad \omega_\alpha(x) = (1-x^2)^{\alpha-\frac{1}{2}}.$$

De plus, les $\widetilde{C}_k^\alpha$ forment une base orthonormale de $L^2_{\omega_\alpha}$: l'espace pondéré des fonctions réelles mesurables sur I avec

$$\|f\|^2_{\omega_\alpha} =< f, f >_{\omega_\alpha} = \int_I (f(x))^2 \omega_\alpha(x)\,dx < \infty.$$

À partir des bornes supérieures des polynômes de Jacobi orthonormaux, voir par exemple (Ben Saber and Karoui, 2021), on obtient l'inégalité suivante

$$\sup_{x\in I} |\widetilde{C}_k^\alpha(x)| \leqslant \frac{5}{4}\, k^{\alpha-\frac{1}{2}}\sqrt{k+\alpha}, \quad k \geqslant 2, \quad \alpha \geqslant 0, \tag{2.1.5}$$

et

$$\sup_{x\in I} |\widetilde{C}_0^\alpha(x)| = 2^{-\alpha}\Big(\beta\big(\alpha+\frac{1}{2},\alpha+\frac{1}{2}\big)\Big)^{-1/2}, \quad \sup_{x\in I} |\widetilde{C}_1^\alpha(x)| \leqslant \sqrt{2+2\alpha}\Big(2^{-\alpha}\Big(\beta\big(\alpha+\frac{1}{2},\alpha+\frac{1}{2}\big)\Big)^{-1/2}\Big). \tag{2.1.6}$$

Ici $\beta(x,y)$ est la fonction Béta.
La base de Gegenbauer est bien adaptée pour l' approximation spectrale des fonctions qui appartiennent à l'espace de Sobolev adapté à poids $H^s_{\omega_\alpha,A}(I)$, $s > 0$ défini comme suit. Pour un entier $s \geqslant 0$,

$$H^s_{\omega_\alpha,A}(I) = \Big\{ f \text{ measurable on } I;\ \|f\|^2_{s,\omega_\alpha,A} = \sum_{k=0}^{\left[\frac{s-1}{2}\right]} \|(1-x^2)^{\frac{s}{2}-k}\partial_x^{s-k} f\|^2_{\omega_\alpha} + \|f\|^2_{[\frac{s}{2}],\omega_\alpha} < +\infty \Big\}. \tag{2.1.7}$$

Ici, $\|\cdot\|_{\omega_\alpha}$ et $\|\cdot\|_{s,\omega_\alpha}$ sont les normes usuelles de $L^2_{\omega_\alpha}(I)$ et $H^s_{\omega_\alpha}(I)$ (l'espace de Sobolev défini sur I). Pour un réel $s > 0$, les espaces $H^s_{\omega_\alpha,A}$ sont définis par interpolation en utilisant les espaces de Sobolev pondérés adaptés $H^{[s]}_{\omega_\alpha,A}(I)$ et $H^{[s]+1}_{\omega_\alpha,A}(I)$, (voir Guo (2000) pour plus de détails).

Soit π_N^α la projection orthogonale sur le sous-espace de dimension finie de $L^2_{\omega_\alpha}(I)$ engendré par

$\{\widetilde{C}_0^\alpha, \widetilde{C}_1^\alpha, \ldots, \widetilde{C}_N^\alpha\}$. *Autrement dit, pour* $f \in L^2_{\omega_\alpha}(I)$, *on a*

$$\pi_N^\alpha(f) \quad = \quad \sum_{k=0}^{N} < f, \widetilde{C}_k^\alpha >_{\omega_\alpha} \widetilde{C}_k^\alpha, \qquad avec \qquad < f, \widetilde{C}_k^\alpha >_{\omega_\alpha} = \int_I f(x)\widetilde{C}_k^\alpha(x)\,\omega_\alpha(x)\,dx. \tag{2.1.8}$$

Guo (2000) *a montré qu'il existe une constante uniforme* $C > 0$ *telle que:*

$$\|u - \pi_N^\alpha(u)\|_{L^2_{\omega_\alpha}(I)} \leqslant CN^{-s}\|u\|_{s,\omega_\alpha,A}, \qquad \forall\ u \in H^s_{\omega_\alpha,A}(I).$$

De plus, il a été récemment démontré dans Jaming et al. (2016) *que les polynômes de Gegenbauer sont bien adaptés à l'approximation des fonctions à c bandes limitées,* $c \in \mathbb{R}^+$. *L'espace de telles fonctions est connu sous le nom de l'espace de Paley-Wiener défini par:*

$$\mathcal{B}_c = \{f \in L^2(\mathbb{R}),\ supp(\mathcal{F}f) \in [-c, c]\}. \tag{2.1.9}$$

Ici $\mathcal{F}f$ *désigne la transformée de Fourier de* $f \in L^2(\mathbb{R})$. *Plus précisément,* Jaming et al. (2016) *ont montré que pour tout entier* $N \geqslant \max(3, \alpha, ec/2)$ *et pour tout* $f \in \mathcal{B}_c$, *on a:*

$$\|f - \pi_N^\alpha f\|_{L^2_{\omega_\alpha}(I)} \leqslant \frac{\gamma(\alpha)}{\sqrt{c}}\left(1 + \frac{1}{2\ln\left(\frac{2N+4}{ec}\right)}\right)\left(\frac{ec}{2N+2}\right)^{N+2}\|f\|_{L^2(\mathbb{R})},$$

où $\gamma(0) = 1$ *et* $\gamma(\alpha) = 2^{-\frac{3}{2}\alpha+\frac{1}{4}}e^{-\alpha-\frac{1}{4}}$, $\alpha \neq 0$.

Example 2.1.8. *(Noyau de Christoffel-Darboux Jacobi)*
Pour deux paramètres $\alpha, \beta > -1$ *et un entier positif* N, *le noyau des polynômes de Jacobi est donné par:*

$$K_N^{\alpha,\beta}(x,y) = \sum_{k=0}^{N}\left[\widetilde{P}_k^{(\alpha,\beta)}(X_i)\widetilde{P}_k^{(\alpha,\beta)}(x)\right] = A_N^{(\alpha,\beta)}\begin{cases} \dfrac{\widetilde{P}_{N+1}^{(\alpha,\beta)}(x)\widetilde{P}_N^{(\alpha,\beta)}(y) - \widetilde{P}_N^{(\alpha,\beta)}(x)\widetilde{P}_{N+1}^{(\alpha,\beta)}(y)}{x-y}, & x \neq y \\ \widetilde{P}_{N+1}^{(\alpha,\beta)}(x)'\widetilde{P}_N^{(\alpha,\beta)}(x) - \widetilde{P}_N^{(\alpha,\beta)}(x)'\widetilde{P}_{N+1}^{(\alpha,\beta)}(x), & x = y \end{cases}, \tag{2.1.10}$$

où $x, y \in I$, $A_N^{(\alpha,\beta)} = \dfrac{2}{2N+\alpha+\beta+2}\sqrt{\dfrac{(N+1)(N+\alpha+\beta+1)(N+\alpha+1)(N+\beta+1)}{(2N+\alpha+\beta+1)(2N+\alpha+\beta+3)}}$.

Les polynômes de Jacobi classiques $P_k^{(\alpha,\beta)}$ *sont définis pour* $x \in [-1, 1]$ *par la formule de Rodrigues suivante:*

$$P_k^{(\alpha,\beta)}(x) = \frac{(-1)^k}{2^k k!}\frac{1}{\omega_{\alpha,\beta}(x)}\frac{d^k}{dx^k}\Big(\omega_{\alpha,\beta}(x)(1-x^2)^k\Big), \quad \omega_{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta, \tag{2.1.11}$$

avec

$$P_k^{(\alpha,\beta)}(1) = \binom{k+\max(\alpha,\beta)}{k} = \frac{\Gamma(k+\max(\alpha,\beta)+1)}{k!\,\Gamma(\max(\alpha,\beta)+1)}.$$

Ici $\Gamma(\cdot)$ *désigne la fonction Gamma usuelle.*
Les polynômes de Jacobi normalisés de degré k *et de paramètres* α, β *et notés* $\widetilde{P}_k^{(\alpha,\beta)}$, *sont donnés par*

$$\widetilde{P}_k^{(\alpha,\beta)}(x) = \frac{1}{\sqrt{h_k^{\alpha,\beta}}} P_k^{(\alpha,\beta)}(x), \quad h_k^{\alpha,\beta} = \frac{2^{\alpha+\beta+1}\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}{k!(2k+\alpha+\beta+1)\Gamma(k+\alpha+\beta+1)}. \tag{2.1.12}$$

Dans ce cas, on a

$$\|\widetilde{P}_k^{(\alpha,\beta)}\|^2_{\omega_{\alpha,\beta}} = \int_{-1}^{1} (\widetilde{P}_k^{(\alpha,\beta)}(y))^2 \omega_{\alpha,\beta}(y)\, dy = 1. \tag{2.1.13}$$

Le lemme suivant nous fournit une borne supérieure des $\widetilde{P}_k^{(\alpha,\beta)}$ pour $k \geqslant 2$.

Lemma 2.1.9. *(Ben Saber and Karoui, 2021). Soient* $\alpha, \beta \geqslant -\frac{1}{2}$, $\mu = \max(\alpha,\beta) \geqslant -\frac{1}{2}$, $c_{\alpha,\beta} = \frac{\alpha+\beta+1}{2}$ *et*

$$\eta_{\alpha,\beta} = \frac{\sqrt{\pi}}{2^{\frac{\alpha+\beta}{2}}\Gamma(\mu+1)\sqrt{e}} \max\left[2^{\frac{1}{4}}, \left((\frac{3}{2}+\alpha+\beta)(\frac{3}{2}+\mu)\right)^{\frac{\mu}{2}}\right]. \tag{2.1.14}$$

Alors, on a

$$\begin{cases} \max\limits_{x\in[-1,1]} |\widetilde{P}_0^{(\alpha,\beta)}(x)| = (2^{\alpha+\beta+1}B(\alpha+1,\beta+1))^{-\frac{1}{2}}, & k=0, \\ \max\limits_{x\in[-1,1]} |\widetilde{P}_k^{(\alpha,\beta)}(x)| \leqslant \eta_{\alpha,\beta}(1+\sqrt{c_{\alpha,\beta}})k^{\mu+\frac{1}{2}}, & \forall k \geqslant 1. \end{cases} \tag{2.1.15}$$

Example 2.1.10. *(Noyau de convolution Sinus-cardinal)*
Nous rappelons que le noyau de convolution de Sinc est défini pour un réel fixé $c > 0$, *par*

$$K_c(x,y) = \frac{\sin(c(x-y))}{\pi(x-y)}, \quad x, y \in I. \tag{2.1.16}$$

Notons que $K_c(x,y) = \frac{c}{\pi}\widehat{\mu}_c(x-y)$ *où* μ_c *est la mesure de probabilité uniforme donnée par*

$$\mu_c(x) = \frac{1}{2c}\mathbf{1}_{[-c,c]}(x), \quad x \in \mathbb{R}.$$

Ici, $\widehat{\mu}_c(\cdot)$ *désigne la transformée de Fourier de* $\mu_c(\cdot)$. *Donc, par le théorème de Bochner, le noyau de convolution Sinc* $K_c(\cdot,\cdot)$ *est auto-reproduisant. Il est bien connu, voir par exemple Slepian and Pollak (1961), que le RKHS associé est l'espace de Paley-Wiener des fonctions à c bandes limitées* $\mathcal{B}_c$ *donné par (2.1.9). La restriction du noyau Sinc à* $I^2 = [-1,1]^2$, *est aussi un noyau de Mercer, dont la*

décomposition spectrale est donnée par:

$$K_c(x,y) = \frac{\sin(c(x-y))}{\pi(x-y)} = \sum_{n=0}^{\infty} \lambda_n(c)\psi_{n,c}(x)\psi_{n,c}(y), \quad \forall\, x,y \in I = [-1,1]. \tag{2.1.17}$$

Ici, les $\psi_{n,c}(\cdot)$ *et les* $\lambda_n(c)$ *sont les fonctions propres et les valeurs propres positives associées de l'opérateur de Hilbert-Schmidt* $\mathcal{Q}_c$ *défini sur* $L^2(-1,1)$ *par* $\mathcal{Q}_c f(x) = \int_{-1}^{1} \frac{\sin(c(x-y))}{\pi(x-y)} f(y)\,dy$. *Autrement dit, pour un entier* $n \geqslant 0$, *on a*

$$\int_{-1}^{1} \frac{\sin(c(x-y))}{\pi(x-y)} \psi_{n,c}(y)\,dy = \lambda_n(c)\psi_{n,c}(x), \quad x \in I.$$

Dans la littérature, les fonctions propres $\psi_{n,c}(\cdot)$ *sont connues sous le nom de fonctions d'onde sphéroïdales allongées (PSWF). Les valeurs propres* $\lambda_n(c)$, $n \geqslant 0$ *sont triées dans un ordre décroissant comme suit*

$$1 > \lambda_0(c) > \lambda_1(c) > \cdots > \lambda_n(c) > \cdots$$

La théorie et le calcul des PSWF et de leurs valeurs propres sont dus aux travaux de D. Slepian et de ses collaborateurs H. Landau et H. Pollack, voir par exemple Slepian and Pollak (1961). Il est à noter que les $\lambda_n(c)$ *décroissent super-exponentiellement vers zéro, voir par exemple Bonami and Karoui (2017). Enfin, dans Bonami et al. (2018), on a des estimations non asymptotiques suivantes des* $\lambda_n(c)$,

$$\lambda_n(c) \geqslant 1 - \frac{7}{\sqrt{c}} \frac{(2c)^n}{n\,!} e^{-c}, \quad \textit{for } 0 \leqslant n < \frac{c}{2.7}, \tag{2.1.18}$$

et

$$\lambda_n(c) \leqslant \exp\left(-(2n+1)\log\left(\frac{2}{ec}(n+1)\right)\right), \quad \forall\, n \geqslant \max\left(\frac{ec}{2}, 2\right).$$

2.1.3 Estimateur de régression par minimisation des moindres carrés régularisés de Tikhonov dans un RKHS

Dans ce paragraphe, nous décrivons d'abord un schéma d'apprentissage basé sur la minimisation régularisée de Tikhonov sur un RKHS, voir Smale and Zhou (2005, 2007). Le schéma de minimisation des moindres carrés régularisés de Tikhonov est utilisé comme un outil pour l'approximation d'une fonction de régression appropriée $f \in \mathcal{H}_K$ à partir des observations $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ tirées selon une mesure de probabilité conjointe ρ sur $\mathcal{X} \times \mathbb{R}$. Ici, les X_i sont des observations aléatoires à valeurs dans $\mathcal{X}$ et dont $\rho_{\mathcal{X}}$ est la distribution de probabilité sur $\mathcal{X}$.

Dans Smale and Zhou (2005, 2007), on a considéré l'algorithme de régularisation de Tikhonov, basé sur les données $(X_i, Y_i)_{1\leqslant i\leqslant n}$, pour estimer la vraie fonction de régression f_ρ définie sur $\mathcal{X}$ et donnée par (2.1.1). Pour un paramètre de régularisation $\lambda > 0$, cet algorithme donne la solution $f_\lambda \in \mathcal{H}_K$ du problème de minimisation suivant

$$f_\lambda = \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \Big(f(X_i) - Y_i\Big)^2 + \lambda \|f\|_K^2 \right\}, \tag{2.1.19}$$

où $\|\cdot\|_K$ est la norme associée au RKHS $\mathcal{H}_K$ généré par le noyau de Mercer $K(\cdot,\cdot)$. Smale and Zhou (2005, 2007) ont montré que la solution de (2.1.19) est donnée par l'estimateur

$$\widehat{f}_n^{\lambda}(x) = \sum_{i=1}^{n} c_{i,\lambda} K(x_i, x), \tag{2.1.20}$$

où les coefficients du vecteur $\mathbf{C}_\lambda = (c_{i,\lambda})_{1\leqslant i\leqslant n}$ sont une solution du système

$$\left[\Big[K(x_i,x_j)\Big]_{1\leqslant i,j\leqslant n} + n\lambda I_n\right] \mathbf{C}_\lambda = G_\lambda\, \mathbf{C}_\lambda = \mathbf{Y}, \quad \mathbf{Y} = (y_i)_{1\leqslant i\leqslant n}. \tag{2.1.21}$$

Ici I_n est la matrice identité de dimension n et $[K(x_i,x_j)]_{1\leqslant i,j\leqslant n}$ est la matrice de Gram aléatoire associée au noyau $K(\cdot,\cdot)$. Ce résultat est une conséquence du fameux théorème du représentant 2.1.6 qui est particulièrement utile. En effet, même si le RKHS associé au problème de minimisation (2.1.19) est de dimension infinie, la solution (2.1.20) se trouve toujours dans un espace de dimension finie. Les schémas basés sur les RKHS sont également utilisés pour résoudre d'autres types de problèmes de régression tel que la régression linéaire fonctionnelle, voir par exemple (Shin and Lee, 2016; Yuang and Cai, 2010). En outre, puisque la matrice de Gram aléatoire régularisée G_λ donnée par (2.1.21) est inversible, le vecteur des coefficients $\mathbf{C}_\lambda$ est donné par

$$\mathbf{C}_\lambda = G_\lambda^{-1}\mathbf{Y} = \left[\Big[K(x_i,x_j)\Big]_{1\leqslant i,j\leqslant n} + n\lambda I_n\right]^{-1} \mathbf{Y}, \quad \mathbf{Y} = (y_i)_{1\leqslant i\leqslant n}. \tag{2.1.22}$$

Remarque 2.1.11. *Le schéma de régularisation de Tikhonov avec le noyau de Mercer présente l'avantage de fonctionner avec des observations aléatoires $\{X_i,\, i = 1,\ldots,n\}$ tirées d'une mesure de probabilité $\rho_{\mathcal{X}}$ inconnue. Ceci est différent des schémas de projection empirique qu'on étudie par la suite et qui sont limités à des observations i.i.d. et suivant une loi bien précise. Cet aspect est important dans les applications où les X_i suivent une distribution de probabilité assez générale sur $\mathcal{X}$.*

Remarque 2.1.12. *Il existe autres méthodes d'estimation non paramétrique, à savoir la régression par splines de lissage. En effet, on cherche la fonction f qui appartient à l'espace de Sobolev $W_m^2[0,1], m \geqslant$*

1 *et qui minimise la quantité suivante*

$$\frac{1}{p}\sum_{j=1}^{p}\Big(\overline{y}_j - f(t_j)\Big)^2 + \lambda\int_0^1 (f^{(m)}(t))^2 dt, \tag{2.1.23}$$

où λ *est un paramètre de régularisation.*
Sous certaines hypothèses, Degras (2012) a montré que le problème (2.1.23) *admet une solution unique* $\hat{f}_\lambda$ *qui est une spline naturelle d'ordre* $2m$ *ayant ses noeuds aux points d'observation* $t_1, \ldots, t_p$*, pour un entier* m*.*
Il a fourni ensuite des erreurs de convergence de cet estimateur notamment l'erreur quadratique moyenne discrétisée de $\hat{f}_\lambda$ *et l'erreur quadratique moyenne intégrée (MISE).*

2.1.4 Régression linéaire fonctionnelle

Soit J un intervalle compact et $L^2(J)$ l'espace de Hilbert des fonctions de carrés intégrables. Pour $n \in \mathbb{N}$, soit $\{(X_i(\cdot), Y_i),\ i = 1, \ldots, n\}$ un ensemble d'observations, où Y_i sont des réponses scalaires et $X_i(\cdot) \in L^2(J)$ sont les prédicteurs fonctionnels que nous supposons être des copies i.i.d. d'un processus stochastique centré $X_t = X(t)$, $t \in J$, telle que $\mathbb{E}\big[X^2(t)\big] < \infty, \forall t \in J$. Le modèle de régression linéaire fonctionnelle (LFR) est alors donné par

$$Y_i = \int_J X_i(s)\beta_0(s)\, ds + \varepsilon_i. \tag{2.1.24}$$

où ε_i sont des variables aléatoires i.i.d centrées, de variance σ^2 et indépendantes des $X_i(\cdot)$ et $\beta_0(\cdot)$ est la fonction inconnue à estimer (Hall and Horowitz, 2007; Yuang and Cai, 2010).
Shin and Lee (2016) et Yuang and Cai (2010) ont étudié une classe d'estimateurs du problème précédent donnée par les moindres carrés pénalisés dans le cadre d'un RKHS.
Soit $\mathcal{H}_K$ un RKHS généré par un noyau de Mercer, alors ces estimateurs sont donnés par

$$\widehat{\beta}_n^\lambda = \arg\min_{\beta\in\mathcal{H}_K}\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{Y_i - \int_J X_i(s)\beta(s)\, ds}{\widehat{\sigma}}\right) + \lambda J(\beta)\right],$$

où $\rho(\cdot)$ une fonction de perte, $\widehat{\sigma}$ est un estimateur d'erreur, $J(\beta)$ est une fonction de pénalisation et λ est un paramètre de régularisation. Dans le cadre d'un RKHS, le problème de minimisation précédent a une solution unique grâce au théorème du représentant 2.1.6.
L'estimation de la fonction inconnue $\beta_0(\cdot)$ est un problème inverse et nécessite une certaine régularisation afin d'éviter une instabilité numérique. Plus précisément, si l'on considère une matrice aléatoire définie positive qu'on voudrait inverser, cette matrice est bien conditionnée si son nombre de

conditionnement donné par la définition suivante n'est pas grand.

Définition 2.1.13. *Soit A une matrice carrée de dimension n à coefficients dans $\mathbb{R}$ ou $\mathbb{C}$, on appelle nombre de conditionnement de A par rapport à la norme $\|.\|_2$, le nombre réel positif $\kappa_2(A)$ défini par:*

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2,$$

où $\|.\|_2$ est la norme matricielle L^2. D'une manière équivalente, on a

$$\kappa_2(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)},$$

où $\sigma_{\max}$ (resp. $\sigma_{\min}$) est la plus grande (resp. la plus petite) valeur singulière de A.

Théorème 2.1.14. ***(Théorème de Gerschgorin)*** *Soit $A \in \mathcal{M}_n(\mathbb{C})$, et soit*

$$R_i = \sum_{i \neq j} |a_{ij}| \quad , \quad H_i = \{z \in \mathbb{C},\ |z - a_{ii}| \leqslant R_i\}.$$

Alors chaque valeur propre de A se situe dans au moins un des disques de Gershgorin H_i.

2.1.5 Quelques inégalités de concentration

Le contenu de ce paragraphe est emprunté à la référence Tropp (2015).

Inégalité de Hoeffding

Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes bornées avec $X_i \in [a_i, b_i], \forall\, i$, où $-\infty < a_i < b_i < \infty$. Alors, $\forall\, t > 0$, on a

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geqslant t\Big) \leqslant \exp\Big(-\frac{2nt^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big),$$

et

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \leqslant -t\Big) \leqslant \exp\Big(-\frac{2nt^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big).$$

Inégalité de McDiarmid

Considérons des variables aléatoires indépendantes $X_1, \ldots, X_n \in \mathcal{X}$ et une application $f : \mathcal{X}^n \to \mathbb{R}$. Soit $i \in \{1, \ldots, n\}$, $x_1, \ldots, x_n, x_i' \in \mathcal{X}$ et $\phi : \mathcal{X}^n \to \mathbb{R}$, une fonction vérifiant

$$|\phi(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)-\phi(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)|\leqslant c_i;\quad 1\leqslant i\leqslant n.$$

Alors, $\forall\, t>0$, on a

$$\mathbb{P}\Big(\phi(X_1,X_2,\ldots,X_n)-\mathbb{E}[\phi(X_1,X_2,\ldots,X_n)]\geqslant t\Big)\leqslant \exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big).$$

Matrice de Chernoff

Soit $\{X_k, 1\leqslant k\leqslant n\}$ une suite finie de matrices Hermitiennes indépendantes, aléatoires et de dimension d. Supposons que

$$0\leqslant \lambda_{\min}(X_k)\quad\text{et}\quad \lambda_{\max}(X_k)\leqslant L\quad\text{pour chaque indice k.}$$

Nous introduisons la somme des matrices aléatoires suivante

$$Y=\sum_{k=1}^n X_k.$$

Soit $\mu_{\min}$ (resp. $\mu_{\max}$) la plus petite (resp. la plus grande) valeur propre de $\mathbb{E}[Y]$:

$$\mu_{\min}=\lambda_{\min}(\mathbb{E}[Y])=\lambda_{\min}\Big(\sum_k\mathbb{E}[X_k]\Big),\quad \mu_{\max}=\lambda_{\max}(\mathbb{E}[Y])=\lambda_{\max}\Big(\sum_{k=1}^n\mathbb{E}[X_k]\Big).$$

Alors, pour $\theta>0$, on a

$$\mathbb{E}[\lambda_{\min}(Y)]\geqslant\frac{1-e^{-\theta}}{\theta}\mu_{\min}-\frac{1}{\theta}L\log(d),\quad \mathbb{E}[\lambda_{\max}(Y)]\geqslant\frac{e^{\theta}-1}{\theta}\mu_{\max}+\frac{1}{\theta}L\log(d).$$

De plus, on a

$$\mathbb{P}[\lambda_{\min}(Y)\leqslant(1-\varepsilon)\mu_{\min}]\leqslant d\left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\mu_{\min}/L}\quad\text{pour }\varepsilon\in[0,1),$$

et

$$\mathbb{P}[\lambda_{\max}(Y)\geqslant(1+\varepsilon)\mu_{\max}]\leqslant d\left[\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right]^{\mu_{\max}/L}\quad\text{pour }\varepsilon\geqslant 0.$$

2.2 Modèles linéaires généralisés

Dans cette section, nous présentons tout d'abord les modèles linéaires généralisés (que nous désignerons par GLM dans la suite, acronyme de "generalized linear models"). Nous décrivons également la méthode d'estimation du maximum de vraisemblance (EMV), puis les propriétés asymptotiques de l'EMV, et ensuite l'inférence statistique qui en découle. À la fin de cette section, nous définissons le modèle de régression de Poisson ainsi que les données de comptage censurées à droite. La description de cette section est basée sur De Jong and Heller (2008), Dupuy (2018) et McCullagh and Nelder (1989).

2.2.1 Définition

Les modèles linéaires généralisés, qui ont été formalisés par Nelder and Wedderburn (1972), sont une classe de modèles statistiques de régression permettant de traiter des problèmes où la distribution de la variable réponse (variable à expliquer) n'est plus nécessairement gaussienne, comme dans le cas du modèle linéaire ou de l'ANOVA. Ces modèles sont une extension du modèle de régression linéaire standard, qui permettent de s'affranchir des hypothèses de linéarité de la relation entre la variable à expliquer et les variables explicatives et de la normalité des termes d'erreurs. Pour définir un modèle linéaire généralisé, il faut définir trois éléments : une composante aléatoire, un prédicteur linéaire et une fonction de lien.

Composante aléatoire

Supposons que l'on dispose d'un échantillon de n variables aléatoires $Y_1, Y_2, \ldots, Y_n$, copies indépendantes d'une variable aléatoire Y admettant une distribution issue d'une famille exponentielle (voir McCullagh and Nelder (1989) pour plus de détails). La densité de la variable réponse Y_i s'écrit sous la forme :

$$f_{Y_i}(y_i, \theta_i, \phi) = \exp\left(\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right), \quad i = 1, \ldots, n, \tag{2.2.1}$$

où $\theta_i \in \mathbb{R}$ est appelé paramètre canonique, $\phi \in \mathbb{R}_+^*$ est un paramètre de dispersion et les fonctions $a(\cdot)$, $b(\cdot)$ et $c(\cdot)$ sont spécifiques à chaque type de distribution.

Remarque 2.2.1. *L'espérance et la variance d'une famille exponentielle admettent des expressions remarquables. Elles peuvent s'exprimer en fonction de a, b et c. Soit Y une variable aléatoire ayant une densité de la forme (2.2.1), et définissons $\dot{b}(\theta) = \frac{\partial}{\partial\theta} b(\theta)$ et $\ddot{b}(\theta) = \frac{\partial^2}{\partial\theta^2} b(\theta)$. Alors:*

$$\mathbb{E}(Y) = \dot{b}(\theta) \quad et \quad Var(Y) = \ddot{b}(\theta) a(\phi). \tag{2.2.2}$$

Voir McCullagh and Nelder (1989) pour plus de détails concernant la preuve.

Le tableau 2.1 présente quelques exemples classiques de distributions qui appartiennent à des familles exponentielles. Nous précisons les trois fonctions a, b et c ainsi que les paramètres canonique et de dispersion. Pour simplifier la lecture du tableau, nous omettons l'indice i.

Distribution	$a(\phi)$	$b(\theta)$	$c(y,\phi)$	θ	ϕ
$\mathcal{P}(\mu)$	ϕ	e^θ	$-\log(y\,!)$	$\log(\mu)$	1
$\mathcal{B}(n,\mu)$	ϕ	$n\log(1+e^\theta)$	$\log(C_n^y)$	$\log\left(\frac{\mu}{1-\mu}\right)$	1
$\mathcal{N}(\mu,\sigma^2)$	ϕ	$\frac{\theta^2}{2}$	$-\frac{1}{2}[\log(2\pi\sigma^2)+\frac{y^2}{\sigma^2}]$	μ	σ^2
$\Gamma(\mu,\nu)$	ϕ	$-\log(-\theta)$	$(\nu-1)\log(y)+\nu\log(\nu)-\log(\Gamma(\nu)$	$-\frac{1}{\mu}$	$\frac{1}{\nu}$
$\mathcal{NB}(\mu,\kappa)$	ϕ	$-\frac{1}{\kappa}\log(1-e^\theta)$	$\log(\Gamma(y+\frac{1}{\kappa}))-\log(y\,!\Gamma(\frac{1}{\kappa}))$	$\log\left(\frac{\kappa\mu}{1+\kappa\mu}\right)$	1

Table 2.1: Quelques exemples de distributions de familles exponentielles

Prédicteur linéaire

Nous désignons par $\mathbf{X}_i$ le vecteur colonne de dimension p des variables explicatives observées sur l'individu i, $i=1,\ldots,n$. Nous notons X_{ij} ses composantes qui peuvent être quantitatives ou qualitatives, de sorte que $\mathbf{X}_i=(1,X_{i2},\ldots,X_{ip})^\top\in\mathbb{R}^p$. Nous notons également $\beta=(\beta_1,\ldots,\beta_p)^\top\in\mathbb{R}^p$ le vecteur de dimension p contenant les paramètres de régression inconnus. Le problème statistique consiste à estimer β à partir des observations $(Y_i,\mathbf{X}_i), i=1,\ldots,n$.
On appelle prédicteur linéaire (pour le $i-$ème individu) la combinaison linéaire suivante des variables explicatives :

$$\sum_{j=1}^{p}\beta_j X_{ij}=\beta^\top\mathbf{X}_i, i=1,\ldots,n.$$

Cette combinaison linéaire peut inclure des transformations des variables explicatives initiales (par exemple $\ln(X_{i2})$), ou des interactions (par exemple $X_{i2}\times X_{i3}$).

Fonction de lien

La fonction de lien explicite la relation entre la variable réponse Y_i et le prédicteur linéaire $\beta^\top\mathbf{X}_i$. Soit $\mu_i=\mathbb{E}(Y_i|\mathbf{X}_i)$ l'espérance conditionnelle de la variable à expliquer sachant les variables explicatives. Définissons:

$$g(\mu_i)=\beta^\top\mathbf{X}_i,\quad i=1,\ldots,n,$$

où g est une fonction monotone et différentiable, appelée fonction de lien. La fonction de lien canonique $g = (\partial b/\partial\theta_i)^{-1}$ est souvent utilisée et dans ce cas, on a $\theta_i = g(\mu_i) = \beta^\top \mathbf{X}_i$.
Dans le tableau 2.2, nous donnons les fonctions de liens canoniques associées à quelques lois classiques.

Distribution	Poisson $\mathcal{P}(\mu)$	Binomial $\mathcal{B}(n,\mu)$	Normal $\mathcal{N}(\mu,\sigma^2)$	Gamma $\Gamma(\kappa,\nu)$	Negative Binomial $\mathcal{NB}(\mu,\kappa)$
$g(x)$	$\log(x)$	$\log(\frac{x}{1-x})$	x	$-\frac{1}{x}$	$\log\left(\frac{\kappa x}{1+\kappa x}\right)$

Table 2.2: Quelques fonctions de lien canonique classiques

2.2.2 Estimation dans les modèles linéaires généralisés

Équations de vraisemblance

Supposons que l'on dispose d'un échantillon d'observations indépendantes $Y_1, \ldots, Y_n$ de densités

$$f_{Y_i}(y_i, \theta_i, \phi) = \exp\left(\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right), \quad i = 1, \ldots, n.$$

En se basant sur l'échantillon $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$, on calcule la vraisemblance du paramètre (β, ϕ), qui est donnée par

$$L_n(\beta, \phi) = \prod_{i=1}^{n} \exp\left(\frac{\theta_i Y_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi)\right).$$

La log-vraisemblance est donnée par $\ell_n(\beta, \phi) = \ln L_n(\beta, \phi)$:

$$\ell_n(\beta, \phi) = \sum_{i=1}^{n} \left(\frac{\theta_i Y_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi)\right) := \sum_{i=1}^{n} \ell_{n,i}(\beta, \phi).$$

L'EMV $\hat{\beta}_n$ de β est obtenu en résolvant le système à p équations suivant:

$$\frac{\partial}{\partial\beta}\ell_n(\beta, \phi)\Big|_{\beta=\hat{\beta}_n} = \sum_{i=1}^{n} \frac{\partial}{\partial\beta}\ell_{n,i}(\beta, \phi)\Big|_{\beta=\hat{\beta}_n} = 0.$$

2.2.3 Propriétés asymptotiques et inférence

Dans cette section, nous présentons les propriétés asymptotiques de l'EMV dans les GLM, ainsi que l'inférence statistique qui en découle et renvoyons le lecteur intéressé à McCullagh and Nelder (1989) pour plus de détails.

Dans le cadre général des GLM, Fahrmeir and Kaufmann (1981) démontrent différents résultats dont, en particulier, le théorème sur la normalité asymptotique de $\hat{\beta}_n$. Ce théorème repose principalement sur des hypothèses concernant les matrices hessiennes et d'information de Fisher.
Le théorème 2.2.2 suivant montre que l'EMV $\hat{\beta}_n$ est asymptotiquement gaussien, voir Antoniadis et al. (1992); Fahrmeir and Kaufmann (1981) pour plus de détails. Supposons que $\ell_n(\beta)$ est la log-vraisemblance et que le paramètre de dispersion est connu. Soit la matrice $\mathcal{I}_n(\beta) = -\partial^2 \ell_n(\beta)/\partial\beta\partial\beta^\top$ où β est le vrai paramètre.

Théorème 2.2.2. *Étant donné un GLM, supposons que :*

- *Les variables explicatives $X_{i1}, X_{i2}, \ldots, X_{ip}$ sont bornées.*
- *La plus petite valeur propre de la matrice $\mathbb{X}^\top\mathbb{X}$ tend vers l'infini quand n tend vers l'infini, où*

$$\mathbb{X} = \begin{pmatrix} 1 & X_{12} & \ldots & X_{1p} \\ 1 & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \ldots & X_{np} \end{pmatrix}.$$

Alors la suite $(\hat{\beta}_n)$ des estimateurs du maximum de vraisemblance converge en probabilité vers β et $\mathcal{I}_n(\hat{\beta}_n)^{\frac{1}{2}}(\hat{\beta}_n - \beta)$ converge en loi vers le vecteur gaussien $\mathcal{N}(0, I_p)$.

Dans la section suivante, nous décrivons quelques intervalles de confiance et tests d'hypothèses les plus utilisés.

Intervalles et régions de confiance

D'après le théorème 2.2.2, la loi de $\hat{\beta}_n$ peut être approchée, pour n grand, par le vecteur gaussien $\mathcal{N}(\beta, \mathcal{I}_n(\hat{\beta}_n)^{-1})$. Notons par $\hat{\sigma}_j^2$ le $j-$ième terme diagonal de $\mathcal{I}_n(\hat{\beta}_n)^{-1}$ et par $\hat{\beta}_{n,j}$ la $j-$ième composante de $\hat{\beta}_n$ (pour $j = 1, \ldots, p$). Alors la loi de $\hat{\beta}_{n,j}$ peut être approchée par la loi normale $\mathcal{N}(\beta_j, \hat{\sigma}_j^2)$. Un intervalle de confiance pour β_j au niveau de confiance asymptotique $(1-\alpha)$ est donné par

$$\left[\hat{\beta}_{n,j} - u_{1-\alpha/2}\hat{\sigma}_j \,; \hat{\beta}_{n,j} + u_{1-\alpha/2}\hat{\sigma}_j\right],$$

où

- $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$, défini par $\mathbb{P}(\mathcal{N}(0,1) \leqslant u_{1-\alpha/2}) = 1-\alpha/2$ (pour $\alpha \in]0,1[$).
- $\hat{\sigma}_j$ est souvent appelé "standard error".

Test sur une composante de β (test de Wald)

Une autre façon de mener l'inférence sur le paramètre β est de proposer des hypothèses portant sur des régions Θ susceptibles de contenir ce paramètre. Autrement dit, on se pose la question de savoir s'il est possible d'admettre que β appartienne à telle ou telle région Θ et ce au vu de ce que l'on a observé. Si l'on veut tester la significativité de la $j-$ième variable explicative dans le prédicteur linéaire $\beta^\top \mathbf{X}_i$, on teste les hypothèses $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$. Sous H_0, la statistique de Wald $\hat{\beta}_{n,j}/\hat{\sigma}_j$ converge en loi vers $\mathcal{N}(0,1)$. On prendra alors pour région de rejet de H_0, au niveau asymptotique α :

$$\mathcal{R}_\alpha = \left\{ \left| \frac{\hat{\beta}_{n,j}}{\hat{\sigma}_j} \right| \geqslant u_{1-\alpha/2} \right\}.$$

Remarque 2.2.3. *Il existe d'autres types de tests notamment le test du rapport de vraisemblance qui permet de tester la nullité de q coefficients de β.*

2.2.4 Algorithme de Newton-Raphson

L'algorithme de Newton-Raphson est une méthode itérative pour résoudre des équations non-linéaires. Il repose sur le principe suivant : on se donne une valeur initiale puis on obtient une seconde valeur en approchant la fonction de maximum de vraisemblance dans le voisinage de la valeur initiale par un polynôme du second degré et en trouvant la valeur maximisant ce polynôme. Cela fait appel à la matrice Hessienne et la matrice des dérivées secondes de la log-vraisemblance, (Lange, 2004). Puis on réitère le même procédé jusqu'à ce qu'un critère de convergence soit satisfait (par exemple, la norme de la différence entre deux approximations successives devient plus petite qu'un seuil fixé $\varepsilon > 0$.)

2.2.5 Modèle de régression de Poisson et données de comptage censurées à droite

Modèle de régression de Poisson

La régression de Poisson est un modèle qui s'applique lorsque la variable réponse Y est une variable de comptage (variable aléatoire à valeurs dans $\mathbb{N}$).
Le modèle de régression de Poisson est beaucoup utilisé dans le domaine des télécommunications (le nombre de communications dans un intervalle de temps donné), en assurance, en météorologie,...
Considérons n observations indépendantes $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$ du modèle de régression de Poisson, défini par

$$\forall i = 1, \ldots, n, \quad \begin{cases} Y_i & \sim & \mathcal{P}(\lambda(\mathbf{X}_i)), \\ \ln(\lambda(\mathbf{X}_i)) & = & \beta^\top \mathbf{X}_i, \end{cases} \tag{2.2.3}$$

La vraisemblance $L_n(\beta)$ ainsi que la log-vraisemblance $\ell_n(\beta)$ de ce modèle s'écrivent respectivement :

$$L_n(\beta) = \prod_{i=1}^{n} \left(e^{-\lambda(\mathbf{X}_i)} \lambda(\mathbf{X}_i)^{Y_i} \frac{1}{Y_i\,!} \right)$$

et

$$\ell_n(\beta) = \ln L_n(\beta = \sum_{i=1}^{n} \Big(Y_i \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \ln(Y_i\,!) \Big),$$

puis les équations du score :

$$\sum_{i=1}^{n} X_{ij}(Y_i - e^{\beta^\top \mathbf{X}_i}) = 0, \quad j = 1, \ldots, p. \tag{2.2.4}$$

L'estimateur du maximum de vraisemblance n'admet pas d'expression explicite. Donc la résolution de (2.2.4) sera faite par un algorithme numérique (Newton-Raphson). Les résultats généraux sur l'asymptotique de l'EMV dans les GLM assurent que si n est suffisamment grand, $\hat{\beta}_n$ est approximativement distribué comme le vecteur gaussien $\mathcal{N}(\beta, (\mathcal{I}_n(\beta))^{-1})$

Données censurées à droite

Les données de survie ont pour première particularité de ne concerner que des variables aléatoires positives. Une deuxième particularité de cette analyse est l'incomplétude des données qui équivaut à une perte d'information. Analysant la survenue d'un certain type d'évènement, nous qualifierons de « donnée complète » un temps correspondant à l'observation de la survenue de l'événement, et de « donnée incomplète » un temps correspondant à l'absence de son observation. En général, l'incomplétude prend la forme d'une censure.

Dans de nombreux contextes expérimentaux où des données de survie sont obtenues, comme dans les essais cliniques, toutes ces données ne sont pas disponibles pour les individus de l'étude et elles peuvent être censurées à droite. En d'autres termes, pour certains individus, nous pouvons seulement savoir qu'ils ont survécu jusqu'à un certain moment. Par exemple, dans un essai clinique où les patients entrent dans l'étude pendant une certaine période d'accumulation, l'étude est analysée avant que tous les patients ne meurent.

Un patient qui n'est pas décédé a un temps de survie qui est censuré à droite, c'est-à-dire que nous savons seulement que ce patient a survécu à la période de temps entre son entrée dans l'étude et le moment où l'étude a été analysée. Parmi les autres raisons de censure, on peut citer l'abandon d'un patient, où l'on sait seulement qu'il était encore en vie au moment où il a abandonné l'étude. Il existe d'autres types de censure notamment la censure à gauche et par intervalle.

Soit Y_i soit le temps de survie, qui peut ne pas être observé. Nous observons plutôt $Y_i^* = \min(Y_i, C_i)$,

où C_i est le temps de censure potentiel. Nous savons si les données ont été censurées, et avec Y_i^*, nous observons la variable indicatrice suivante :

$$\delta_i = \begin{cases} 1 & Y_i \leqslant C_i \quad \text{(pas de censure)}, \\ 0 & Y_i > C_i \quad \text{(sinon)}. \end{cases}$$

Les données de comptage peuvent être censurées également, et c'est ce à quoi nous allons nous intéresser.

2.3 Données manquantes

Cette section concerne les mécanismes de données manquantes et décrit quelques méthodes de traitement et prise en compte de ces données. Nous décrivons notamment la régression-calibration, l'imputation multiple et la pondération par probabilité inverse. La description de cette partie est basée sur Tsiatis (2007); Morisot (2015); Héraud (2012).

2.3.1 Introduction

Parmi les points essentiels de l'analyse des données est celui des données manquantes puisque l'existence de ces données est très fréquente, notamment en épidémiologie et en recherche clinique (déplacement du patient, question à laquelle le patient ne veut pas répondre, perte d'information, ...).
Avant les années 1970, l'analyse des cas complets (voir section 2.3.3) était la méthode standard pour traiter les données manquantes. Elle consiste à retirer de l'analyse statistique les individus ayant une ou plusieurs variables manquantes, ce qui peut entraîner un biais substantiel.
Madow et al. (1983) ont synthétisé les différentes méthodes d'imputation et en 1987, Little and Rubin (1987) ont présenté des méthodes basées sur des modèles de distribution des données, notamment les méthodes de maximum de vraisemblance et de l'imputation multiple.
Rubin (1978b) a proposé une approche populaire, à savoir l'imputation multiple, pour traiter les données manquantes, après avoir rendu compte qu'une seule imputation n'était généralement pas correcte. Puis Rubin (1987) a fourni les bases méthodologiques et statistiques de cette nouvelle approche. Afin de combiner en une seule valeur les estimations faites sur les différentes bases de données complétées par les données imputées, il utilise des règles, maintenant dites "de Rubin" (section 2.3.5.2). Enfin, il décrit les conditions de validité de l'inférence statistique de cette méthode.

Example 2.3.1. *(Tsiatis, 2007). Dans un essai clinique randomisé où l'on mène une étude pour comparer deux ou plusieurs traitements, nous réalisons l'inscription des sujets à cette étude et les assignons de manière aléatoire à l'un des traitements. Supposons que dans un tel essai clinique, les sujets sont censés revenir chaque semaine à la clinique pour fournir la mesure de la réponse Y_{ij} (pour le sujet i, semaine j). Cependant, certains sujets abandonnent l'étude, ne se présentant plus à aucune visite clinique après un certain temps. D'autres encore peuvent manquer des visites à la clinique ou cesser de prendre le traitement qui leur a été attribué.*

2.3.2 Mécanismes des données manquantes

Lorsque l'on analyse des données manquantes, le mécanisme qui induit l'observation ou l'absence de la donnée est souvent appelé mécanisme de données manquantes.
Nous indexons par i, $i = 1, \ldots, n$, un échantillon aléatoire de n individus qui proviennent d'une population donnée. Nous notons par Y_i^{obs} l'ensemble des données observées et par Y_i^{mis} celui des données manquantes. Pour chaque individu i, nous pouvons donc constituer l'ensemble des données complètes (celles que nous voudrions avoir recueillies sur tous les individus de l'échantillon) en partitionnant les composantes de Y_i comme suit

$$Y_i = (Y_i^{\text{obs}}, Y_i^{\text{mis}}). \tag{2.3.1}$$

Nous associons à l'individu i une variable R_i qui vaut 1 si Y_i est entièrement observé et 0 sinon. Soit $\mathbf{X} = (X_1, \ldots, X_l)$ un vecteur de l covariables complètement observées pour chaque individu. Un mécanisme de données manquantes peut être formulé comme la distribution conditionnelle de R sachant $\mathbf{X}$ et $Y = (Y_1, \ldots, Y_n)$. Sous l'hypothèse d'un modèle paramétrique pour cette distribution, on écrira :

$$\mathbb{P}(R|Y, \mathbf{X}, \phi) = \mathbb{P}(R|Y^{\text{obs}}, Y^{\text{mis}}, \mathbf{X}, \phi),$$

où ϕ est un vecteur de paramètres inconnus.

Little and Rubin (2002) ont classé les problèmes de données manquantes en trois catégories, en fonction du mécanisme de données manquantes. Les terminologies de ces mécanismes sont données par Rubin (1976).

Définition 2.3.2.

- ***Manquant complètement au hasard (Missing Completely At Random - MCAR):*** *les données sont "manquantes complètement au hasard" si :*

$$\mathbb{P}(R = 1|Y, \boldsymbol{X}, \phi) \text{ ne dépend pas de } Y \text{ et } \boldsymbol{X}. \tag{2.3.2}$$

L'assertion (2.3.2) *montre que* R *et* Y *sont indépendantes et la probabilité qu'une donnée soit manquante est une constante, c'est à dire,*

$$\mathbb{P}(R = 1|Y, \boldsymbol{X}, \phi) = \pi.$$

- ***Manquant au hasard (Missing At Random - MAR):*** *on dit que les données sont manquantes au hasard si la probabilité qu'une donnée soit manquante dépend uniquement des composantes de* Y *qui sont observées,* Y^{obs}*, et des covariables* $\boldsymbol{X}$*.*

$$\mathbb{P}(R = 1|Y, \boldsymbol{X}, \phi) = \mathbb{P}(R = 1|Y^{obs}, \boldsymbol{X}, \phi). \tag{2.3.3}$$

Cette probabilité ne dépend donc pas des valeurs non-observées des variables.

- ***Manquant non au hasard (Missing Non At Random (MNAR)):*** *on dit que les données sont manquantes non au hasard si la probabilité qu'une donnée soit manquante dépend des valeurs non-observées des variables ou bien des variables qui n'auraient pas été collectées, c'est-à-dire,*

$$\mathbb{P}(R = 1|Y, \boldsymbol{X}, \phi) = \mathbb{P}(R = 1|Y^{obs}, Y^{miss}, \boldsymbol{X}, \phi). \tag{2.3.4}$$

2.3.3 Analyse des cas-complets

Une méthode simple et populaire d'analyse des données manquantes est l'analyse des cas-complets, qui consiste à restreindre l'analyse aux individus pour lesquels toutes les variables sont entièrement renseignées. Elle est appliquée par défaut par les logiciels statistiques dans le traitement des données manquantes.
Miettinen (1985) propose de retirer de l'analyse tous les sujets pour lesquels les informations sont manquantes, ajoutant que cette approche est la seule qui garantisse l'absence de biais tandis que selon Enders (2010), les inconvénients de cette méthode l'emportent sur les avantages dans la plupart des situations.
Dans la littérature, l'utilisation de cette méthode induit une perte de puissance et donc de précision puisqu'elle n'utilise pas toute l'information disponible dans la base de données. Dans le cas d'une analyse multivariée, ce type d'analyse peut également biaiser le processus de sélection des variables puisqu'il se fera en faveur des variables les mieux renseignées. Enfin, Little and Rubin (2002) ont montré que cette méthode peut induire un biais dans les estimations en fonction du mécanisme de données manquantes puisque l'analyse des cas-complets sélectionne un sous-échantillon de la base de données initiale qui n'est généralement pas aléatoire.

2.3.4 Régression-calibration

Carroll and Stefanski (1990) ont été parmi les premiers à développer et explorer la méthode de régression-calibration pour traiter les données manquantes. Dans la littérature, elle a été beaucoup appliquée sur les données non censurées. Cette méthode a été utilisée notamment dans le cadre des GLMs et des modèles à risques proportionnels (Armstrong, 1985; Prentice, 1982; Hughes, 1993; Spiegelman et al., 1997). Rosner et al. (1989) et Reeves et al. (1998) ont développé une approximation dans le cadre des modèles d'analyse non linéaire, qui s'est avérée performante dans de nombreuses circonstances pour la régression logistique. La méthode régression-calibration consiste à remplacer les données manquantes par leur espérance conditionnelle sachant les données observées. Le lecteur est renvoyé à Wang et al. (1997) et Hardin et al. (2003) pour plus de détails.

2.3.5 Imputation multiple

La méthode d'imputation multiple consiste à remplacer les données manquantes par des données générées suivant un modèle d'imputation. Cette imputation est répétée M fois, ce qui génère M ensembles de données complets. Pour obtenir un estimateur global, nous combinons les estimations des M échantillons complétés.
L'imputation multiple, qui repose sur l'hypothèse MAR, est parmi les meilleures méthodes de traitement des données manquantes parce qu'elle permet de prendre en compte la variabilité autour de chaque imputation ce que mène à des variances correctes pour les estimations.
En 1987, Little and Rubin (1987) ont été parmi les premiers à proposer et développer une telle méthode dans le cadre des sciences sociales. Elle a aussi été appliquée dans le contexte de grandes bases de données issues d'enquêtes complexes (Rubin and Schenker, 1991).
Dans la littérature, les livres de Little and Rubin (1987, 2002) et Schafer (1997) sont les références principales.
Grâce au développement des programmes et des machines à haute performance, cette méthode est devenue plus accessible (Kenward and Carpenter, 2007). L'imputation multiple a été aussi appliquée dans le domaine de la santé publique (Arnold and Kronmal, 2003; Cattle et al., 2011; Molenberghs and Kenward, 2007), à la conception des essais cliniques (Wood et al., 2005) et dans la construction de modèles prédictifs (Janssen et al., 2010; Vergouw et al., 2010). Les articles de Arnold and Kronmal (2003); Barnard and Meng (1999); Taylor et al. (2002) et Nur et al. (2009) ont présenté une application de cette méthode à des données de surveillance et ceux de (Raghunathan, 2004; Schafer and Graham, 2002; Sinharay et al., 2001; Rubin, 1996; Reiter, 2007; Graham, 2009) ont synthétisé la théorie de l'imputation multiple.
Le choix du nombre d'imputations a été discuté par Royston (2004); Graham et al. (2007); Bodner

(2008); Van Buuren et al. (1999).

2.3.5.1 Procédure générale

Dans cette partie, on résume les trois étapes principales de l'imputation multiple.

- 1ère **étape: imputation**. L'imputation multiple génère m ensembles complets de données en remplaçant les valeurs manquantes par des valeurs plausibles qui sont tirées suivant une distribution spécifique.
- 2ème **étape: analyse sur chaque ensemble de données imputées**. On estime les paramètres d'intérêt à partir de chacun des m ensembles imputés, sur lesquels on applique la méthode qui aurait été utilisée si les données avaient été complètes.
- 3ème **étape: combinaison des m estimateurs**. On combine les m estimations du paramètre en une seule estimation et on estime sa variance par les règles de Rubin (à partir des variances "intra-imputation" et "inter-imputation").

2.3.5.2 Règles de Rubin

En appliquant la procédure générale de l'imputation multiple (section 2.3.5.1), on obtient m estimateurs $\hat{\theta}_i$, $i = 1, \ldots, m$, du paramètre θ et une matrice de variance-covariance $\hat{M}_i$ estimée pour chaque ensemble de données imputées. En utilisant les règles de Rubin (Little and Rubin, 1987), les estimations $\hat{\theta}_i$ sont regroupées en une seule estimation $\bar{\theta}$, définie comme la moyenne des $\hat{\theta}_i$ pour les m imputations :

$$\bar{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i.$$

La variance totale pour l'estimateur combiné $\bar{\theta}$ est donnée par

$$V = \bar{M} + \left(1 + \frac{1}{m}\right) B.$$

où

- $\bar{M} = \frac{1}{m} \sum_{i=1}^{m} \hat{M}_i$ est la variance "intra-imputation", qui est une mesure classique de la variabilité puisque nous prenons un échantillon plutôt que la population,
- $B = \frac{1}{m-1} \sum_{i=1}^{m} \left(\hat{\theta}_i - \bar{\theta}\right) \left(\hat{\theta}_i - \bar{\theta}\right)^{\top}$ est la variance "inter-imputation" qui mesure la variabilité due

au fait que l'on impute des échantillons aléatoirement. Elle rend compte de la variabilité due au processus aléatoire d'imputation.

La description de la section 2.3.5 a été basée sur Héraud (2012).

2.3.6 Pondération par l'inverse de la probabilité de sélection (IPW)

Les cas-complets constituent souvent un sous-ensemble non représentatif de l'échantillon donc une analyse qui n'utilise que ce sous-ensemble est potentiellement biaisée. Les méthodes d'imputation consistent à spécifier un modèle d'imputation pour les variables inobservées. Les valeurs manquantes sont ensuite remplacées par des valeurs simulées sur la base de ce modèle. La mauvaise spécification de ce modèle peut induire un biais important.
La méthode d'estimation par pondération par l'inverse de la probabilité de sélection (IPW, pour "inverse probability weighting") est une méthode couramment utilisée pour analyser les données manquantes. Cette méthode spécifie plutôt un modèle pour le processus de données manquantes, voir Little and Rubin (2002) pour plus de détails. Désignons la probabilité d'observer un cas complet par

$$\mathbb{P}(R = 1|W) = \pi(W).$$

où W_i un vecteur qui contient les variables observées Y_i et X_i et éventuellement certaines variables de substitution observées pour R. Pour $1 \leqslant i \leqslant n$, $\pi(W_i)$ est inconnue et doit être estimée. Nous supposons que cette probabilité peut être spécifiée par un modèle paramétrique $\pi(W, \gamma)$, où γ est un paramètre inconnu à estimer. Soit $\hat{\gamma_n}$ l'estimation du maximum de vraisemblance du vrai paramètre γ. Alors $\hat{\gamma_n}$ est obtenu comme :

$$\hat{\gamma_n} = \underset{\gamma}{\operatorname{argmax}} \prod_{i=1}^{n} \pi(W_i, \gamma)^{R_i} (1 - \pi(W_i, \gamma))^{1-R_i}.$$

La méthode IPW consiste alors à estimer les paramètres du modèle d'intérêt en pondérant les cas complets par $1/\pi(W_i, \hat{\gamma_n})$.

Estimateur IPW augmenté (AIPW)

Pour l'estimateur IPW, nous avons besoin de spécifier un modèle pour la probabilité $\mathbb{P}(R = 1|W) = \pi(W, \gamma)$. Si ce modèle est mal spécifié, l'estimateur IPW risque d'être biaisé.
La notion d'estimateurs doublement robustes a été introduite pour la première fois par Scharfstein et al. (1999). Ces estimateurs ont été étudiés par Lipsitz et al. (1999); Robins (1999); Robins et al. (2000); Lunceford and Davidian (2004); Neugebauer and van der Laan (2005); Robins and Rotnitzky

(2001); van der Laan and Robins (2003). Un excellent aperçu est également donné par Bang and Robins (2005). Pour obtenir un tel estimateur, on spécifie les deux modèles suivants:

$$\mathbb{E}[Y|W] = \mu(W, \alpha) \quad \text{et} \quad \mathbb{P}(R = 1|W) = \pi(W, \gamma). \tag{2.3.5}$$

On calcule alors un estimateur, dit AIPW, qui intègre à la fois le modèle pour les données manquantes et le modèle pour le mécanisme des données manquantes. Cet estimateur est doublement robuste, au sens où il est consistant même si l'un des deux modèles de (2.3.5) est mal spécifié (plus de détails sont donnés dans Scharfstein et al. (1999); Robins et al. (1994); Robins (1999)).

2.4 Processus empiriques

Soit $X_1, X_2, \ldots$ une suite de variables aléatoires réelles i.i.d, de loi P, définies sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et qui admettent une fonction de répartition F donnée par $F(t) = P((-\infty, t])$. La fonction de répartition empirique $\mathbb{F}_n$ de l'échantillon $X_1, X_2, \ldots, X_n$ est définie par :

$$\forall t \in \mathbb{R},\ \forall \omega \in \Omega,\ \mathbb{F}_n(\omega)(t) = \mathbb{P}_n(\omega)(]-\infty, t]) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i(\omega) \leqslant t\}}.$$

Pour tout ω, $t \to F_n(\omega)(t)$ est une fonction de répartition (fonction de répartition de la loi de probabilité uniforme sur l'ensemble $\{X_1(\omega), \ldots, X_n(\omega)\}$). Pour tout t, $1_{\{X_i \leqslant t\}}$ est une variable aléatoire de Bernoulli, de paramètre $F(t)$.

On omettra ω dans les notations et on notera $\mathbb{F}_n(t) = \mathbb{P}_n(]-\infty, t]) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leqslant t\}}$. Par la loi forte des grands nombres, on a

$$\forall t, \quad \mathbb{F}_n(t) \xrightarrow{\text{p.s}} F(t).$$

Par le théorème central limite, $\sqrt{n}(\mathbb{F}_n(t) - F(t))$ converge en loi vers la loi normale $\mathcal{N}(0, F(t)(1-F(t)))$. Définissons maintenant le processus empirique réel par la quantité :

$$\alpha_n = \sqrt{n}(\mathbb{F}_n - F). \tag{2.4.1}$$

Les théorèmes de Glivenko-Cantelli et Donsker formulent des versions "uniformes" des deux résultats asymptotiques que nous venons d'énoncer.

Théorème 2.4.1. ***(Théorème de Glivenko-Cantelli)*** *Soit $X_1, X_2, \ldots$ une suite de variables aléa-*

toires i.i.d de fonction de répartition F. Alors la statistique de Kolmogorov-Smirnov $\|\mathbb{F}_n - F\|_\infty$ *vérifie:*

$$\sup_{t\in\mathbb{R}}|\mathbb{F}_n(t) - F(t)| = \|\mathbb{F}_n - F\|_\infty \xrightarrow{p.s} 0.$$

Théorème 2.4.2. ***(Théorème de Donsker)*** *Soit* $X_1, X_2, \ldots$ *une suite de variable aléatoires i.i.d de fonction de répartition* F*. Alors la suite* α_n*, donnée par (2.4.1), converge en loi dans* $D[-\infty,\infty]$ *(espace des fonctions càdlàg) vers un processus gaussien* G_F *centré et de fonction de covariance*

$$cov(G_F(s)G_F(t)) = F(s\wedge t) - F(s)F(t), \quad \forall s,t\in\mathbb{R}.$$

Soit $(\mathcal{X},\mathcal{A})$ un espace mesurable et $\mathcal{F}\subset L_1(P)$ une classe de fonctions mesurables de $\mathcal{X}$ dans $\mathbb{R}$. Pour $f\in\mathcal{F}$, nous notons par Pf l'espérance de F sous P i.e. $Pf = \mathbb{E}[f(X)]$ et nous définissons également $\mathbb{P}_n(f)$ comme suit

$$\mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) = \int f d\mathbb{P}_n.$$

Le processus empirique indexé par $\mathcal{F}$ est donné par

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - Pf), \quad f\in\mathcal{F}. \tag{2.4.2}$$

Définition 2.4.3. ***(Classes de Glivenko-Cantelli et de Donsker)***

- *Si* $\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}|\mathbb{P}_n f - Pf| \xrightarrow{p.s.} 0$*, alors la classe* $\mathcal{F}\subset L_1(P)$ *de fonctions mesurables* $f:\mathcal{X}\to\mathbb{R}$ *est dite* $P-$*Glivenko-Cantelli.*

- *si* $\{\mathbb{G}_n f, f\in\mathcal{F}\}$ *converge en loi dans* $l^\infty(\mathcal{F})$*(espace des fonctions définies sur* $\mathcal{F}$*, à valeurs réelles et bornées) vers un processus* $\{Gf, f\in\mathcal{F}\}$ *alors la classe* $\mathcal{F}\subset L_2(P)$ *de fonctions mesurables* $f:\mathcal{X}\to\mathbb{R}$ *est dite* $P-$*Donsker. Dans ce cas,* G *est un processus gaussien centré de fonction de covariance* $cov(Gf_1, Gf_2) = P(f_2f_2) - Pf_1Pf_2$*, appelé* $P-$*pont brownien.*

Définition 2.4.4.

- *Pour une fonction mesurable* f*, on appelle entropie sans crochet de* $\mathcal{F}$ *la quantité* $\log\Big(N(\varepsilon,\mathcal{F},\|.\|)\Big)$ *où* $N(\varepsilon,\mathcal{F},\|.\|)$ *est le nombre minimum de boules de rayon* ε *nécessaires pour recouvrir* $\mathcal{F}$*.*

- *Pour deux fonctions données* l *et* u*, on appelle* $\varepsilon-$*crochets* $[l,u]$*, l'ensemble des fonctions* f *vérifiant* $l\leqslant f\leqslant u$ *et* $\|l-u\|\leqslant\varepsilon$*.*

- *Pour toute classe de fonctions mesurables $\mathcal{F}$, on définit $N_{[]}(\varepsilon, \mathcal{F}, \|.\|)$ comme le nombre minimum de $\varepsilon-$crochets $[l, u]$ nécessaires pour recouvrir $\mathcal{F}$. On définit également comme entropie à crochet la quantité $\log\Big(N_{[]}(\varepsilon, \mathcal{F}, \|.\|)\Big)$.*

Théorème 2.4.5. *(van der Vaart, 2000) Soit $\mathcal{F}$ une classe de fonctions mesurables.*

1. *Si pour tout $\varepsilon > 0$, $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$, alors $\mathcal{F}$ est $P-$Glivenko-Cantelli.*
2. *Soit $F : \mathcal{X} \to \mathbb{R}^+$ une fonction mesurable. Si*
   - $\sup_{f\in\mathcal{F}}|f(x)| \leqslant F(x)$ *et* $PF^2 < \infty$,
   - *l'intégrale entropique à crochet donnée par* $J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log(N_{[]}(\varepsilon, \mathcal{F}, L_2(P)))}\, d\varepsilon$ *est finie,*

   alors $\mathcal{F}$ est is P-Donsker.

Propriété 2.4.6. *Si $\mathcal{F}$ et $\mathcal{G}$ sont des classes de Donsker, alors $\mathcal{F} + \mathcal{G}$ l'est aussi.*

3 Random orthogonal projections based schemes for solving nonparametric and linear functional regressions problems

In the first part of this work, we develop and study some estimators based on empirical orthogonal projections for solving nonparametric regression problems. The projection kernels investigated here are given by the Gegenbauer Christoffel-Darboux kernel and the convolution Sinc kernel. We provide mean integrated squared error analysis of the proposed estimators, under the assumption that the regression function belongs to some suitable functional spaces. These estimators have the advantages of being stable without the need of an extra regularization process. Moreover, their convergence rates are similar to the optimal convergence rate of a nonparametric regression estimator, when this later is applied to a class of functions with some smoothness conditions. We briefly describe how our estimators can be adapted to handle more general reproducing kernels. In the second part of this work, we study an orthogonal projection based estimator for the fast and stable solution of a linear functional regression (LFR) problem. This problem is solved under the usual assumption that the unknown slope function to be estimated belongs to a finite dimensional subspace of the Hilbert space $L^2(J)$, where J is an interval of $\mathbb{R}$. Let N be the dimension of this subspace, then our estimator is based on the use of the pseudo-inverse of a random matrix. The random matrix to be inverted is positive definite and has a fairly reduced dimension N, compared to the size n of the training data set used to build an LFR estimator. More importantly, we show that this random matrix is well conditioned. Consequently, our estimator for solving the LFR problem is stable. Moreover, unlike the classical KRR schemes for LFR problem, our estimator does not need any regularization process. Also, it is a fast estimator and provides us with more accurate approximation of the unknown true slope function. Finally, we illustrate the different results of this work by some numerical simulations.

Contents

3.1 Introduction

In the first part of this work, we consider the construction of kernel projection based schemes for approximating the regression function in the nonparametric regression problem. The problem is as follows. Let $\mathcal{X}$ be a complete metric space and $\mathcal{Y}$ be an output space. The main issue of learning theory is to develop algorithms that take a training set $\{(X_i, Y_i), 1 \leqslant i \leqslant n\}$ in $\mathcal{X} \times \mathcal{Y}$ and return a function f such that for $x \in \mathcal{X}$, $f(x)$ is a good estimate (or prediction) of the corresponding output $y := y(x)$.

The observations (X_i, Y_i) are assumed to be drawn from a joint probability measure ρ on $\mathcal{X} \times \mathcal{Y}$. In the special case where $\mathcal{Y}$ is a measurable subset of $\mathbb{R}$, this learning problem is known as a nonparametric regression problem. In this work, we shall restrict ourselves to this case. Using standard notations (see for example Smale and Zhou (2005)) and letting ρ_X denote the marginal probability measure over $\mathcal{X}$, the true regression function associated with this regression problem is given by

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x), \quad x \in \mathcal{X},$$

where $d\rho(y|x)$ is the conditional distribution of Y given $X = x$. It is well known (see for example Smale and Zhou (2005)) that f_ρ minimizes the mean square error $\int_{\mathcal{X}\times\mathcal{Y}}(y - f(x))^2 d\rho$. In practice, the outputs Y_i are generally noised observations of the true regression function, which we will simply denote, from now on, by f. Therefore, we consider the following nonparametric regression model:

$$Y_i = f(X_i) + \eta_i, \quad 1 \leqslant i \leqslant n,$$

where $(X_i)_{1\leqslant i\leqslant n}$ are random variables (or inputs) with distribution ρ_X and the noise terms $(\eta_i)_{1\leqslant i\leqslant n}$ are i.i.d. real-valued random variables with mean zero and variance σ^2. For simplicity, we will assume that the X_i are drawn from a known distribution supported on a compact of $\mathbb{R}$ (say $I = [-1, 1]$). But we will also show how the proposed estimators can be generalized to the case where the X_i are i.i.d. observations from an unknown distribution on $\mathbb{R}$. Moreover, we will briefly indicate how these estimators can be generalized to the multivariate case where $X_i \in \mathbb{R}^d$, $d \geqslant 2$.

In the literature, a popular solution to the nonparametric regression problem is given by the Tikhonov regularized least-square algorithms class. More precisely, for an appropriate choice of a Hilbert space $\mathcal{H}$ and a given regularization parameter $\lambda > 0$, this scheme solves the minimization problem

$$\widehat{f}_n^{\lambda} = \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \Big(f(X_i) - Y_i\Big)^2 + \lambda\|f\|_{\mathcal{H}}^2 \right\},$$

where $\|\cdot\|_{\mathcal{H}}$ is the usual norm of $\mathcal{H}$. We refer the interested reader to Smale and Zhou (2005) for a detailed account of these procedures in the special and interesting case of a reproducing kernel Hilbert space generated by a reproducing kernel $K(\cdot,\cdot)$. There exist other kinds of schemes for solving the nonparametric regression problem. For example, in Comte and Catalot (2020), the authors investigate various estimators of the regression function, based on projections associated with different orthogonal polynomials, such as Laguerre and Hermite polynomials. In Baraud (2002), the author considers a convenient collection of finite dimensional subspaces $\{S_m,\, m \in \mathcal{M}_n\}$ of $L^2(\mathcal{A},\nu)$, where ν is a fixed measure supported on a subset $\mathcal{A}$ of $\mathbb{R}^d$. Then, under some technical assumptions involving the dimension of S_m, the moments of the η_i and the structure of the considered orthonormal basis of $L^2(\mathcal{A},\nu)$, the author obtains the following optimal L_2-risk error of a penalized least squares type regression estimator $\widetilde{f}$ of f, based on the projection spaces:

$$\mathbb{E}\Big[\|f-\widetilde{f}\|_{\nu}^2\Big] \leqslant C \inf_{m\in\mathcal{M}_n}\left(\inf_{g\in S_m}\|f-g\|_{\nu}^2 + pen(m) + \varepsilon_n(f)\right), \quad \varepsilon_n(f) = O\Big(\frac{1}{n}\Big).$$

Here, $pen(\cdot)$ is a penalization function and C is a uniform constant. A similar type of L_2-risk estimator error was recently given in Asin and Johannes (2017), where a nonparametric regression estimator is given in terms of approximate projections of f over a finite dimensional subspace of a weighted normed space generated by an orthonormal family $\{\varphi_j, j \in \mathbb{N}\}$ of $L^2([0,1])$. As in Baraud (2002), the appropriate dimension of the projection subspace is given as a solution of a minimization problem that involves the expansion coefficients of the estimator, as well as a sequence of non-negative and non-decreasing penalty functions. In Cohen et al. (2013), authors consider a subset $\mathcal{X}$ of $\mathbb{R}^d$ and a measure ρ supported on $\mathcal{X}$, as well as an appropriate orthonormal set $\{L_j(\cdot),\, j \in \mathbb{N}\}$ of $L^2(\mathcal{X},\rho)$. They provide a similar L_2-risk error for a least squares nonparametric estimator based on the projections over finite dimensional subspaces spanned by the first m orthonormal basis functions $L_j(\cdot)$. The L_2-risk error given in Cohen et al. (2013) depends on the quantity $K(m) = \sup_{x\in\mathcal{X}}\sum_{j=1}^{m}|L_j(x)|^2$. Another class of schemes for solving the regression problem is based on the construction of an estimator belonging to a RKHS generated by a special convolution kernel with variable bandwidth, see for example Benelmadani et al. (2019) and references therein.

Here, we mention that it is a known fact (see for example Bauer and Kohler (2019) and the references

therein) that one has to restrict the class of regression functions in order to get a non trivial L_2-risk error. For a positive real number $p = q + s$, $q \in \mathbb{N}_0$ and $s \in (0, 1)$, a class of such functions is given by the $(p, C)-$smooth functions defined on $\mathbb{R}^d$ with partial derivatives up to order q being $s-$Hölderian. In Stone (1982) (see also Bauer and Kohler (2019)), the optimal minimax rate of convergence for the estimation of a $(p, C)-$smooth regression function is of order $O(n^{-\frac{2p}{2p+d}})$.

Our aim is to develop some kernels projections based schemes that provide convenient and stable estimates of f, when this latter satisfies some smoothness property. We restrict ourselves to the Sinc kernel (assuming that the X_i follow a uniform distribution over I) and Gegenbauer Christoffel-Darboux kernel (assuming that the X_i are drawn from a scaled Beta distribution defined on I and with parameter $a = b = \alpha + \frac{1}{2}$, $\alpha > -\frac{1}{2}$).

In the second part of this work, we are interested in the construction of a random pseudo-inverse based estimator for solving a linear functional regression problem. The LFR model is described as follows, see for example Hall and Horowitz (2007); Cardot and Johannes (2010); Yuang and Cai (2010); Comte and Johannes (2012).
Let J be a compact interval and $L^2(J)$ be the Hilbert space of square integrable functions on J with its usual inner product $< \cdot, \cdot >$. For a positive integer n, let $\{(X_i(\cdot), Y_i),\ i = 1, \dots, n\}$ be the set of observations, where the Y_i are scalar responses and the $X_i(\cdot) \in L^2(J)$ are the functional predictors, which we assume to be i.i.d. copies of a centered second order stochastic process $X_t = X(t), t \in J$. Then the LFR model is given by

$$Y_i = \int_J X_i(s)\beta_0(s)\, ds + \varepsilon_i.$$

Here, the noise terms ε_i are i.i.d. centered real valued random variables, with variance σ^2, independent of the $X_i(\cdot)$ and $\beta_0(\cdot)$ is the unknown slope function to be estimated. In practice, we assume that $X_i(\cdot)$ lies in a finite dimensional subspace $\mathcal{H}_N$ of $L^2(J)$ and is given by

$$X_i(s) = \sum_{k=1}^{N} \xi_k Z_{i,k} \varphi_k(s), \tag{3.1.1}$$

where $\{\varphi_k(\cdot),\ k = 1, \dots, N\}$ is an orthonormal family of $L^2(J)$, the $Z_{i,k}$ are i.i.d centered random variables with variance σ_Z^2 and $(\xi_k)_{1 \leqslant k \leqslant N}$ is a finite deterministic sequence of $\mathbb{R}\backslash\{0\}$. Let $n > N$ be a positive integer and consider n values $X_i(\cdot)$ of the functional predictor, then the LFR problem reduces to solving in a stable way, the following over-determined linear system

$$Y_i = \int_J X_i(s)\, \beta_0(s)\, ds + \varepsilon_i, \quad i = 1, \dots, n. \tag{3.1.2}$$

Our random pseudo-inverse based estimator $\widehat{\beta}_{n,N}(\cdot)$ of $\beta_0(\cdot)$ is given by

$$\widehat{\beta}_{n,N}(s) = \sum_{k=1}^{N} \widehat{c}_k \varphi_k(s), \qquad \widehat{\boldsymbol{c}} = [\widehat{c}_1, \ldots, \widehat{c}_N]' = G_N^{-1} \cdot \Big(F_N' \cdot \frac{1}{\sqrt{n}} \big[Y_i\big]'_{1 \leqslant i \leqslant n} \Big). \tag{3.1.3}$$

Here, the reduced size and positive definite $N \times N$ random matrix G_N is given by

$$G_N = F_N' F_N, \qquad F_N = \frac{1}{\sqrt{n}} \Big[\xi_j Z_{i,j} \Big]_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant N}.$$

Here, the ξ_j and $Z_{i,j}$ are as given by (3.1.1). Note that the choice and analysis of our estimator $\widehat{\beta}_{n,N}(\cdot)$ are based on the assumption that the functional predictors $X_i(\cdot)$ and the slope function $\beta_0(\cdot)$ are expanded in the same orthonormal basis $\{\varphi_k(\cdot),\, 1 \leqslant k \leqslant N\}$. Nonetheless, we show how to adapt this estimator in order to the more realistic scenario where $X_i(\cdot)$ and $\beta_0(\cdot)$ are expanded in two different orthonormal bases. It is interesting to note that the study of this general case follows in a straightforward manner from the special case of the expansion in the same orthonormal basis.

Before giving our plan, let us highlight our main findings. First, we show that the empirical projection operators associated with the Gegenbauer and Sinc kernels provide stable and fairly accurate approximations to the true regression function f. To establish these results, we need to assume some regularity conditions on f. Precisely, we will assume that f belongs to an adapted Sobolev space $H^s_{\omega_\alpha,A}(I)$ for some $s > 0$ or f is the restriction to I of a $c-$bandlimited function $\widetilde{f}$, for some $c > 0$ (that is, $\widetilde{f}$ belongs to the Paley-Wiener space $\mathcal{B}_c$, defined as the set of functions of $L^2(\mathbb{R})$ with Fourier transforms supported on the interval $[-c, c]$).

More precisely, for $\alpha > -\frac{1}{2}$, we define our estimator of the regression function f, based on Gegenbauer kernel, by:

$$\widehat{f}^{\alpha}_{N,n}(x) \quad = \quad \frac{A_\alpha}{n} \sum_{i=1}^{n} Y_i K^{\alpha}_N(X_i, x), \quad x \in I, \tag{3.1.4}$$

where

$$A_\alpha = 2^{2\alpha} B\Big(\alpha + \frac{1}{2}, \alpha + \frac{1}{2}\Big). \tag{3.1.5}$$

Here, $B(\cdot,\cdot)$ is the beta function. Recall that the X_i have values in I and follow the Beta distribution with parameter $\alpha + \frac{1}{2}$. Next, for $s \geqslant 0$, let $H^s_{\omega_\alpha,A}(I)$ be an adapted weighted Sobolev space given by (2.1.7). Then, by using some spectral approximation properties of the Gegenbauer kernel, we prove that if $f \in H^s_{\omega_\alpha,A}(I)$ for some $s > 0$, and if $|f(x)| \leqslant M$, $a.e. x \in I$, then there exists a uniform constant

c_1 such that we have the following inequality:

$$\mathbb{E}\left[\|f-\widehat{f}^{\alpha}_{N,n}\|^2_{\omega_\alpha}\right]\leqslant A_\alpha\frac{N+1}{n}\Big(M^2+\sigma^2\Big)+c_1N^{-2s}\|f\|^2_{s,\omega_\alpha,A}. \tag{3.1.6}$$

Here, $\sigma^2=\mathbb{E}\big[\eta_i^2\big]$. Moreover, under the assumption that for some $c>0$, f is the restriction to I of a $c-$bandlimited function $\widetilde{f}$, we prove that:

$$\mathbb{E}\left[\|f-\widehat{f}^{\alpha}_{N,n}\|^2_{\omega_\alpha}\right]\leqslant A_\alpha\frac{N+1}{n}\Big(M^2+\sigma^2\Big)+\frac{\gamma^2(\alpha)}{c}\left(1+\frac{1}{2\ln\left(\frac{2N+4}{ec}\right)}\right)^2\left(\frac{ec}{2N+2}\right)^{2N+4}\left\|\widetilde{f}\right\|^2_{L^2(\mathbb{R})}, \tag{3.1.7}$$

where $\gamma(0)=1$ and $\gamma(\alpha)=2^{-\frac{3}{2}\alpha+\frac{1}{4}}e^{-\alpha-\frac{1}{4}}$ for $\alpha\neq 0$. Moreover, for a positive real number $c>0$, we consider a second estimator of the regression function f, based on the Sinc kernel K_c with bandwidth c. This estimator is given by:

$$\widehat{f}_{c,n}(x)=\frac{2}{n}\sum_{i=1}^{n}Y_iK_c(X_i,x),$$

where $K_c(\cdot,\cdot)$ is given by (2.1.16). Under the assumption that $c\geqslant 6$, the X_i are uniformly distributed on I and f belongs to some weighted Sobolev space $\widetilde{H}^s(I)$, $s>0$, we prove that

$$\mathbb{E}\left[\left\|f-\widehat{f}_{c,n}\right\|^2\right]\leqslant\frac{8c}{n\pi}\Big(\sigma^2+M^2\Big)+\frac{98}{3}\left(\frac{e^2}{6}\right)^{-2[c/3]}\|f\|^2+4\left[\frac{c}{3}\right]^{-2s}\|f\|^2_{\widetilde{H}^s}. \tag{3.1.8}$$

Next, for the LFR problem (3.1.2), we prove that our estimator (3.1.3) is stable. In particular, by using some techniques from the spectral theory of positive definite random matrices, we prove that G_N is well conditioned. More precisely, we show that if $\kappa_2(G_N)$ is the $2-$norm condition number of G_N, then for any $\eta>0$, we have with high probability,

$$\kappa_2(G_N)\leqslant\frac{1.72\max_{k\geqslant 1}\sigma_Z^2\xi_k^2+\frac{M_{\boldsymbol{\xi},N}}{n}\log(N)+\eta}{0.63\min_{k\geqslant 1}\sigma_Z^2\xi_k^2-\frac{M_{\boldsymbol{\xi},N}}{n}\log(N)-\eta}, \tag{3.1.9}$$

for some constant $M_{\boldsymbol{\xi},N}$ depending on the deterministic sequence $(\xi_k)_k$ and on N. Also, we show that for any $\eta>0$, the following squared L^2-error of our estimator $\widehat{\beta}_{n,N}$ holds with high probability:

$$\|\widehat{\beta}_{n,N}(\cdot)-\beta_0(\cdot)\|^2_{L_2}\leqslant\kappa_2(G_N)\frac{\frac{1}{n}\|\boldsymbol{\varepsilon}\|^2_{\ell_2}\|\beta_0(\cdot)\|^2_2}{\sigma_Z^2\max_{k\geqslant 1}\xi_k^2c_k^2-\eta},$$

where $\boldsymbol{\xi}(s)=\sum_{k\geqslant 1}\xi_k\varphi_k(s)$ and $\beta_0(s)=\sum_{k\geqslant 1}c_k\varphi_k(s)$. This work is organized as follows. In section 2, we give some mathematical preliminaries that will be useful in this work. In section 3, we define our

projection estimators based on Gegenbauer polynomials kernel. In section 4, we extend the previous study to the case of the non-degenerate and reproducing Sinc kernel types. In particular, we provide L^2-risk errors for the previous estimators. In section 5, we describe our random pseudo-inverse based estimator for the slope function associated with the LFR problem (3.1.2). Then, we prove that under some decay condition of the deterministic sequence $(\xi_k)_k$ associated with the functional predictor, our estimator is stable and provides a high approximation to the true slope function associated with the LFR problem. Finally, in section 6, we provide various numerical simulations to illustrate our results.

3.2 Nonparametric regression by Gegenbauer projection kernel

Let $(X_i, Y_i)_{1\leqslant i\leqslant n}$ be independent observations of the nonparametric regression model

$$Y_i = f(X_i) + \eta_i, \quad i = 1, \ldots, n. \tag{3.2.1}$$

The random design variables $(X_i)_{1\leqslant i\leqslant n}$ are assumed to be independent and distributed according to a re-scaled Beta distribution with parameter $a = b = \alpha + \frac{1}{2}$, $\alpha > -\frac{1}{2}$ and defined on $I = [-1, 1]$. The noise terms $(\eta_i)_{1\leqslant i\leqslant n}$ are i.i.d. copies of a centered random variable, with variance σ^2. We assume that the X_i and η_i are independent.
The problem is to estimate the function $f : I \to \mathbb{R}$ from the observations $(X_i, Y_i)_{1\leqslant i\leqslant n}$. We assume that $f(\cdot)$ lies in a subspace of the Hilbert space $L^2_{\omega_\alpha}(I)$. Let N be a positive integer and let $K^\alpha_N(x, y) = \sum_{k=0}^{N} \widetilde{C}^\alpha_k(x)\widetilde{C}^\alpha_k(y)$ be the Gegenbauer polynomials kernel given by (2.1.4). Let π^α_N be the projection operator given by (2.1.8). Moreover, for a positive integer n, let $\widehat{f}^\alpha_{N,n}$ be the empirical projection operator given by (3.1.4). In the sequel, we will need some of the following assumptions for the regression function f.

$[H_0]$: f is almost everywhere bounded by a uniform constant M on I, that is $|f(x)| \leqslant M$, $a.e.\, x \in I$.

$[H_1]$: f belongs to an adapted weighted Sobolev space $H^s_{\omega_\alpha,A}(I)$, given by (2.1.7) for some $s > 0$.

$[H_2]$: f is the restriction to I of a function $\widetilde{f}$, belonging to the Paley-Wiener space $\mathcal{B}_c$ of $c-$bandlimited functions, defined by (2.1.9), for some $c > 0$.

In Theorem 3.2.1 below, we show that Gegenbauer kernel is well adapted for nonparametric regression of functions satisfying $[H_0]$ and $[H_1]$ or $[H_2]$..

Theorem 3.2.1. *Let $\alpha \geqslant 0$. Under the above notations, we have*

$$\mathbb{E}\big[\widehat{f}^{\alpha}_{N,n}(x)\big] = \pi^{\alpha}_{N}(f)(x), \quad x \in I. \tag{3.2.2}$$

Moreover,

- *If f satisfies hypothesis $[H_0]$ and $[H_1]$, then an estimate of the MISE($\widehat{f}^{\alpha}_{N,n}$) is given by*

$$\mathbb{E}\Big[\|f - \widehat{f}^{\alpha}_{N,n}\|^2_{\omega_\alpha}\Big] \leqslant A_\alpha \frac{N+1}{n}\Big(M^2 + \sigma^2\Big) + c_1 N^{-2s}\|f\|^2_{s,\omega_\alpha,A}, \tag{3.2.3}$$

  where A_α is given by (3.1.5) and c_1 is a uniform constant.

- *If f satisfies hypotheses $[H_0]$ and $[H_2]$, then for any integer $N \geqslant \max(3, \alpha, ec/2)$ we have*

$$\mathbb{E}\Big[\|f - \widehat{f}^{\alpha}_{N,n}\|^2_{\omega_\alpha}\Big] \leqslant A_\alpha \frac{N+1}{n}\Big(M^2 + \sigma^2\Big) + \frac{\gamma^2(\alpha)}{c}\left(1 + \frac{1}{2\ln\left(\frac{2N+4}{ec}\right)}\right)^2 \left(\frac{ec}{2N+2}\right)^{2N+4} \Big\|\widetilde{f}\Big\|^2_{L^2(\mathbb{R})}, \tag{3.2.4}$$

  where $\gamma(0) = 1$ and $\gamma(\alpha) = 2^{-\frac{3}{2}\alpha+\frac{1}{4}}e^{-\alpha-\frac{1}{4}}$ for $\alpha \neq 0$.

Proof. First, we note that:

$$\begin{aligned}
\mathbb{E}_X\Big[f(X_i)\widetilde{C}^{\alpha}_{k}(X_i)\Big] &= \frac{1}{A_\alpha}\int_I f(y)\widetilde{C}^{\alpha}_{k}(y)\,\omega_\alpha(y)\,dy \\
&= \frac{1}{A_\alpha} < f, \widetilde{C}^{\alpha}_{k} >_{\omega_\alpha}, \quad 1 \leqslant i \leqslant n.
\end{aligned}$$

By independence of X_i and η_i, and using the fact that $\mathbb{E}(\eta_i) = 0$, we also note that $\mathbb{E}(\eta_i C^{\alpha}_{k}(X_i)) = 0$, $1 \leqslant i \leqslant n$. Thus, we have:

$$\begin{aligned}
\mathbb{E}\Big[\widehat{f}^{\alpha}_{N,n}(x)\Big] &= \sum_{k=0}^{N}\frac{A_\alpha}{n}\sum_{i=1}^{n}\mathbb{E}\Big[(f(X_i)+\eta_i)\widetilde{C}^{\alpha}_{k}(X_i)\Big]\widetilde{C}^{\alpha}_{k}(x) \\
&= \sum_{k=0}^{N}\left(\frac{A_\alpha}{n}\sum_{i=1}^{n}\mathbb{E}_X\Big[f(X_i)\widetilde{C}^{\alpha}_{k}(X_i)\Big]\widetilde{C}^{\alpha}_{k}(x)\right) \\
&= \sum_{k=0}^{N} < f, \widetilde{C}^{\alpha}_{k} >_{\omega_\alpha} \widetilde{C}^{\alpha}_{k}(x) = \pi^{\alpha}_{N}(f)(x).
\end{aligned}$$

Now, we have

$$\begin{aligned}\widehat{f}^{\alpha}_{N,n}(x) - \mathbb{E}\big[\widehat{f}^{\alpha}_{N,n}(x)\big] &= \frac{A_\alpha}{n}\sum_{i=1}^{n} Y_i K^{\alpha}_N(X_i,x) - \pi^{\alpha}_N(f)(x)\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{N}\Big[A_\alpha(f(X_i)+\eta_i)\widetilde{C}^{\alpha}_k(X_i)\widetilde{C}^{\alpha}_k(x) - < f,\widetilde{C}^{\alpha}_k >_{\omega_\alpha} \widetilde{C}^{\alpha}_k(x)\Big]\\ &= \frac{1}{n}\sum_{i=1}^{n}\xi_i(x). \end{aligned} \tag{3.2.5}$$

By using Parseval's equality, one gets

$$\begin{aligned}\|\xi_i\|^2_{\omega_\alpha} &= \left\|\sum_{k=0}^{N}\Big(A_\alpha\,(f(X_i)+\eta_i)\widetilde{C}^{\alpha}_k(X_i) - < f,\widetilde{C}^{\alpha}_k >_{\omega_\alpha}\Big)\widetilde{C}^{\alpha}_k\right\|^2_{\omega_\alpha},\\ &= \sum_{k=0}^{N}\Big(A_\alpha\,(f(X_i)+\eta_i)\widetilde{C}^{\alpha}_k(X_i) - < f,\widetilde{C}^{\alpha}_k >_{\omega_\alpha}\Big)^2 \end{aligned} \tag{3.2.6}$$

By using the fact that the X_i are independent of the η_i as well as the facts that $\mathbb{E}[\eta_i]=0,\ \mathbb{E}[\eta_i^2]=\sigma^2$, together with the orthonormality of the set $(\widetilde{C}^{\alpha}_k)_{0\leqslant k\leqslant N}$, it is easy to check that

$$\mathbb{E}\Big[A_\alpha\, f(X_i)\widetilde{C}^{\alpha}_k(X_i)\Big] =< f,\widetilde{C}^{\alpha}_k >_{\omega_\alpha},\quad \mathbb{E}\Big[\eta_i f(X_i)\widetilde{C}^{\alpha}_k(X_i)\Big]=0, \tag{3.2.7}$$

and

$$\mathbb{E}\Big[A^2_\alpha\,\eta_i^2\Big(\widetilde{C}^{\alpha}_k(X_i)\Big)^2\Big] = A_\alpha\mathbb{E}\Big[\eta_i^2\Big]\cdot\mathbb{E}\Big[A_\alpha\Big(\widetilde{C}^{\alpha}_k(X_i)\Big)^2\Big] = A_\alpha\sigma^2. \tag{3.2.8}$$

Hence, by using (3.2.6)–(3.2.7), together with Parseval's equality and hypothesis $[H_0]$, one gets

$$\begin{aligned}\mathbb{E}\Big[\|\xi_i\|^2_{\omega_\alpha}\Big] &= (N+1)A_\alpha\sigma^2 + \sum_{k=0}^{N}A_\alpha\mathbb{E}\Big[A_\alpha f^2(X_i)\Big(\widetilde{C}^{\alpha}_k(X_i)\Big)^2\Big] - \sum_{k=0}^{N}\Big(< f,\widetilde{C}^{\alpha}_k >_{\omega_\alpha}\Big)^2\\ &= A_\alpha(N+1)\sigma^2 + \sum_{k=0}^{N}A_\alpha\mathbb{E}\Big[A_\alpha f^2(X_i)\Big(\widetilde{C}^{\alpha}_k(X_i)\Big)^2\Big] - \|\pi^{\alpha}_N f\|^2_{\omega_\alpha}\\ &\leqslant A_\alpha(N+1)\big(\sigma^2+M^2\big). \end{aligned} \tag{3.2.9}$$

Next, for $1\leqslant i,j\leqslant n$, with $i\neq j$, we have by Parseval's equality,

$$< \xi_i,\xi_j >_{\omega_\alpha} = \sum_{k=0}^{N}\Big[A_\alpha\Big(f(X_i)+\eta_i\Big)\widetilde{C}^{\alpha}_k(X_i) - < f,\widetilde{C}^{\alpha}_k >_{\omega_\alpha}\Big]\Big[A_\alpha\Big(f(X_j)+\eta_j\Big)\widetilde{C}^{\alpha}_k(X_j) - < f,\widetilde{C}^{\alpha}_k >_{\omega_\alpha}\Big].$$

Hence, by using (3.2.7), one gets

$$\mathbb{E}\Big[< \xi_i, \xi_j >_{\omega_\alpha} \Big] = 0, \quad \forall\, i \neq j. \tag{3.2.10}$$

On the other hand, from (3.2.5), we have

$$\Big\| \widehat{f}^{\alpha}_{N,n} - \mathbb{E}\big[\widehat{f}^{\alpha}_{N,n}\big] \Big\|^2_{\omega_\alpha} = \frac{1}{n^2} \sum_{i=1}^{n} \|\xi_i\|^2_{\omega_\alpha} + \frac{1}{n^2} \sum_{i,j=1, i\neq j}^{n} < \xi_i, \xi_j >_{\omega_\alpha} . \tag{3.2.11}$$

Consequently, by using (3.2.9) and (3.2.11), we obtain

$$\mathbb{E}\left[\Big\| \widehat{f}^{\alpha}_{N,n} - \mathbb{E}\big[\widehat{f}^{\alpha}_{N,n}\big] \Big\|^2_{\omega_\alpha} \right] = \frac{1}{n^2} \mathbb{E}\left[\sum_{i=1}^{n} \|\xi_i\|^2_{\omega_\alpha} \right] \leqslant A_\alpha \frac{N+1}{n} \big(\sigma^2 + M^2\big). \tag{3.2.12}$$

Next, since

$$\widehat{f}^{\alpha}_{N,n} - \mathbb{E}\Big[\widehat{f}^{\alpha}_{N,n} \Big] = \widehat{f}^{\alpha}_{N,n} - \pi^{\alpha}_{N} f \in F_N = \mathrm{Span}\{\widetilde{C}^{\alpha}_{k},\, 0 \leqslant k \leqslant N\}$$

and since $f - \pi^{\alpha}_{N} f \in F_N^{\perp}$, then we have

$$\Big\| f - \widehat{f}^{\alpha}_{N,n} \Big\|^2_{\omega_\alpha} = \Big\| \widehat{f}^{\alpha}_{N,n} - \mathbb{E}\big[\widehat{f}^{\alpha}_{N,n}\big] \Big\|^2_{\omega_\alpha} + \Big\| f - \pi^{\alpha}_{N} f \Big\|^2_{\omega_\alpha}. \tag{3.2.13}$$

On the other hand, from the mathematical preliminaries section, if f belongs to the adapted weighted Sobolev space $H^{s}_{\omega_\alpha, A}(I)$, then

$$\|f - \pi^{\alpha}_{N}(f)\|^2_{\omega_\alpha} \leqslant c_1 N^{-2s} \|f\|^2_{s,\omega_\alpha,A}, \tag{3.2.14}$$

for some uniform constant c_1. Hence, by combining (3.2.12)–(3.2.14), we get the desired result (3.2.3). Finally, to prove (3.2.4), it suffices to note that if $f \in \mathcal{B}_c$, we have, for $N \geqslant \max\big(3, \alpha, \frac{ec}{2}\big)$:

$$\Big\| f - \pi^{\alpha}_{N} f \Big\|^2_{\omega_\alpha} \leqslant \frac{\gamma^2(\alpha)}{c} \left(1 + \frac{1}{2\ln\big(\frac{2N+4}{ec}\big)} \right)^2 \left(\frac{ec}{2N+2} \right)^{2N+4} \Big\| \widetilde{f} \Big\|^2_{L^2(\mathbb{R})}.$$

By using similar techniques as above, we obtain (3.2.4). □

Remark 3.2.2. *From the error bound given by (3.2.3), we conclude that if $H^{s}_{\omega_\alpha, A}(I)$, $s > 0$, then the squared bias is of order N^{-2s}, which is the standard order, and that of the variance is equal to $\frac{N}{n}$, which is the order of a nonparametric regression estimator in the univariate case and when the regression function satisfies a similar smoothness property. Also, the minimum L^2-risk error of the estimator $\widehat{f}^{\alpha}_{N,n}$ is obtained when the two error terms in (3.2.3) are of the same order. Straightforward*

computation shows that this is the case if $N = O(n^{\frac{1}{2s+1}})$. *Consequently, the* L^2-*risk error or the mean integrated squared error of our estimator* $\widehat{f}^{\alpha}_{N,n}$ *is* $O(n^{\frac{-2s}{2s+1}})$, *where* s *is the Sobolev smoothness of* f, *see Stone (1982) or Bauer and Kohler (2019) for more details.*

Remark 3.2.3. *Our Gegenbauer polynomials kernel projection estimator* $\widehat{f}^{\alpha}_{N,n}$ *was developed and studied in the univariate case. It is interesting to note that it can be straightforwardly generalized to the multivariate case, where the random sampling sets* $\{X_i, 1 \leqslant i \leqslant n\} \subset \mathbb{R}^d$. *In fact, it suffices to replace each Gegenbauer polynomial* $\widetilde{C}^{\alpha}_{k}$ *by its tensor product* $d-$*dimensional version*

$$\Phi^{\alpha}_{\boldsymbol{m}}(\boldsymbol{x}) = \prod_{j=1}^{d} \widetilde{C}^{\alpha}_{m_j}(x_j), \quad \boldsymbol{x} = (x_1, \ldots, x_d) \in I^d, \quad \boldsymbol{m} = (m_1, \ldots, m_d) \in \{0, 1, \ldots, N\}^d.$$

Note that the $\Phi^{\alpha}_{\boldsymbol{m}}$ *give rise to an orthonormal basis of* $L^2(I^d, \boldsymbol{\omega}_{\alpha})$, *where* $\boldsymbol{\omega}_{\alpha}(\boldsymbol{x}) = \prod_{j=1}^{d} \omega_{\alpha}(x_j)$. *For simplicity and readability, we restrict ourselves to the case* $d = 1$ *in this work.*

Finally, we briefly describe two different schemes that can be used to handle random sampling sets where the marginal distribution ρ_X is unknown.

Scheme 1: This method is based on the well-known transformation technique of random sampling laws. This technique is briefly described as follows. We first assume that the random sampling points X_i are i.i.d. copies of a random variable with known cumulative distribution function (CDF) $F_X(\cdot)$. Then it is easy to see that the $F_X(X_i)$ follow the uniform law over $(0, 1)$. Since the CDF of the Beta distribution $B(\alpha, \alpha)$ is given by the regularized incomplete Beta function $I_x(\alpha, \alpha)$, and this later is invertible, then the transformed sampling points

$$Z_i = I_x^{-1}(\alpha, \alpha)\big(F_X(X_i)\big), \quad 1 \leqslant i \leqslant n$$

follow the $Beta(\alpha, \alpha)-$distribution. For the case of an unknown sampling law ρ_X, one may replace the true CDF $F_X(\cdot)$ by an accurate estimate in the previous equality. There exists a rich literature devoted to the subject of fairly efficient CDF estimators, see for example Funke and Palmes (2017) and references therein. Finally, one should use a convenient interpolation scheme in order to get sufficient accurate approximations of the data values at the transformed sampling points Z_i. A deep study of the modified estimators obtained by combining our estimators with the transformed sampling strategy is beyond the scope of this work.

Scheme 2: This second method does not require any knowledge or estimate of the CDF $F_X(\cdot)$. It is rather based on the use of a Shepard's like interpolation of the random sampling points X_i at the neighboring new sampling points following a $B(\alpha, \alpha)$ law. A comprehensive review on this type of univariate and multivariate scattered data interpolation techniques is given in Dell'Accio and Di Tommaso (2016). More precisely, let $\mu > 0$ be a positive real number and $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set of n distinct points of a domain $\mathcal{D} \subset \mathbb{R}^d$, $d \geqslant 1$. Then, for the n associated real valued function evaluations $f_i = f(\boldsymbol{x}_i)$ and for a given radius of influence $R > 0$, the modified Shepard's algorithm is given by

$$S_\mu^R(f)(\boldsymbol{x}) = \sum_{i=1}^{n} W_{\mu,i}(\boldsymbol{x}) f_i, \qquad W_{\mu,i}(\boldsymbol{x}) = \frac{\left(\frac{1}{\mathrm{d}(\boldsymbol{x},\boldsymbol{x}_i)} - \frac{1}{R}\right)_+^\mu}{\sum_{i=1}^n \left(\frac{1}{\mathrm{d}(\boldsymbol{x},\boldsymbol{x}_i)} - \frac{1}{R}\right)_+^\mu},$$

where $(t)_+ = \max(t, 0)$ and $d(.,.)$ the Euclidean distance. Moreover, as described in Dell'Accio and Di Tommaso (2016), a higher order Shepard's like interpolation is given by the following combined Shepard-Multivariate Taylor interpolation polynomial $S_{\mu,r}$, where for an integer $r \geqslant 0$,

$$S_{\mu,r}(f)(\boldsymbol{x}) = \sum_{i=1}^{n} A_{\mu,i}(\boldsymbol{x}) T_{r,\boldsymbol{x}_i}(f)(\boldsymbol{x}), \quad T_{r,\boldsymbol{x}_i}(f)(\boldsymbol{x}) = \sum_{\nu=0}^{r} \frac{D^\nu f(\boldsymbol{x}_i)}{\nu!} (\boldsymbol{x} - \boldsymbol{x}_i)^\nu, \quad A_{\mu,i}(\boldsymbol{x}) = \frac{\left(\mathrm{d}(\boldsymbol{x}, \boldsymbol{x}_i)\right)^{-\mu}}{\sum_{i=1}^n \left(\mathrm{d}(\boldsymbol{x}, \boldsymbol{x}_i)\right)^{-\mu}}.$$

In this case, we have

$$\|S_{\mu,r}(f) - f\| = \begin{cases} O(h^{r+1}) & \text{if } \mu - d > r + 1 \\ O(h^{\mu-d} |\log h|) & \text{if } \mu - d = r + 1. \end{cases} \tag{3.2.15}$$

Here, h is the mesh step size, that is, the largest distance between the neighboring points $\boldsymbol{x}_i$. For more details, the reader is referred to Dell'Accio and Di Tommaso (2016) and the references therein.

3.3 Nonparametric regression using Sinc-type kernels

3.3.1 Sinc kernel case

In this section, we extend the result of Theorem 3.2.1 to the case of the Sinc projection kernel. For this purpose, we consider a real number $c > 0$ and the Sinc kernel $K_c(x, y) = \frac{\sin c(x-y)}{\pi(x-y)}$, $x, y \in I$. Let $(\psi_{n,c}(x))_{n\geqslant 0}$ and $(\lambda_n(c))_{n\geqslant 0}$ be the sets of orthonormal eigenfunctions and associated eigenvalues of the Sinc kernel operator $\mathcal{Q}_c$, given by $\mathcal{Q}_c(f)(x) =< K_c(x, \cdot), f >$. Let π_c denote the projection operator over the subspace spanned by the $\psi_{n,c}$. Then, by using the fact that the $(\psi_{n,c}(\cdot))_{n\geqslant 0}$ form

an orthonormal basis of $L^2(I)$, we obtain

$$f(x) = \pi_c(f)(x) = \sum_{n=0}^{\infty} < f, \psi_{n,c} > \psi_{n,c}(x), \quad \forall\, f \in L^2(I). \tag{3.3.1}$$

The regression estimator of f in model (3.2.1), given by the empirical projection operator associated with the Sinc kernel, is defined by:

$$\widehat{f}_{c,n}(x) = \widetilde{\pi}_{c,n}(f)(x) = \frac{2}{n}\sum_{i=1}^{n} Y_i K_c(x, X_i) = \frac{2}{n}\sum_{i=1}^{n}(f(X_i) + \eta_i)K_c(x, X_i). \tag{3.3.2}$$

The Sinc kernel defined on I^2 is a Mercer's kernel. Then assuming that f is bounded on $I = [-1, 1]$, $\mathbb{E}(\eta_i K(X_i, x)) = 0$, $\mathbb{E}\left(f(X_i)K_c(x, X_i)\right) = \int_I f(y)K_c(x, y)dP(y) = \frac{1}{2}\int_I f(y)K_c(x, y)\,dy = \frac{1}{2} < f, K_c >$ and by using (2.1.17), we calculate:

$$\begin{aligned}
\mathbb{E}(\widehat{f}_{c,n}(x)) &= \mathbb{E}\Big[\widetilde{\pi}_{c,n}(f)(x)\Big] = \int_I K_c(x, y)f(y)\,dy = \int_I \sum_{k=0}^{\infty} \lambda_k(c)\psi_{k,c}(x)\psi_{k,c}(y)f(y)\,dy \\
&= \sum_{k=0}^{\infty} \lambda_k(c)\left(\int_I \psi_{k,c}(y)f(y)\,dy\right)\psi_{k,c}(x) = \sum_{k=0}^{\infty} \lambda_k(c) < f, \psi_{k,c} > \psi_{k,c}(x).
\end{aligned} \tag{3.3.3}$$

Note that $\mathbb{E}(\widehat{f}_{c,n}) \neq \pi_c(f)$. Consequently, the approach we used for analysing the Gegenbauer Christoffel kernel based regression scheme cannot be applied to the Sinc kernel. To overcome this difficulty, we first substitute the usual Sobolev space $H^s(I)$ with the weighted Sobolev space $\widetilde{H}^s(I)$, defined by

$$\widetilde{H}^s(I) = \left\{ f \in L^2(I),\ \|f\|^2_{\widetilde{H}^s} = \sum_{k\geqslant 0}(1 + k^2)^s |< f, \psi_{k,c} >|^2 < +\infty \right\}.$$

It is easy to check that if $f \in \widetilde{H}^s(I)$, then for any positive integer N,

$$\sum_{n=N+1}^{\infty} |< f, \psi_{n,c} >|^2 = \|f - \pi_{c,N}(f)\|^2 \leqslant N^{-2s}\|f\|^2_{\widetilde{H}^s}. \tag{3.3.4}$$

We first establish two technical lemmas that will be needed to provide an error analysis of the Sinc kernel based regression scheme for a regression function belonging to the weighted Sobolev space $\widetilde{H}^s(I)$.

Lemma 3.3.1. *For any real number $c \geqslant 6$ and any positive integer N such that $N+1 \leqslant \frac{c}{3}$, we have*

$$\frac{e^{-c}}{\sqrt{c}} \sum_{k=0}^{N} \frac{(2c)^k}{k!} \leqslant \frac{1}{\sqrt{6}} \left(\frac{e^2}{6}\right)^{-c/3}. \tag{3.3.5}$$

Proof. First, note from Batir (2008) that $k! = \Gamma(k+1) \geqslant \sqrt{2e}\left(\frac{2k+1}{2e}\right)^{k+\frac{1}{2}}$. Therefore:

$$\frac{e^{-c}}{\sqrt{c}} \sum_{k=0}^{N} \frac{(2c)^k}{k!} \leqslant e^{-c} \sum_{k=0}^{N} \frac{1}{\sqrt{(2k+1)c}} \left(\frac{4ec}{2k+1}\right)^k. \tag{3.3.6}$$

If $0 \leqslant k \leqslant N$ with $N+1 \leqslant \frac{c}{3}$, and since $k \mapsto \frac{1}{\sqrt{(2k+1)c}}\left(\frac{4ec}{2k+1}\right)^k$ is an increasing function for $0 \leqslant k \leqslant 2c - \frac{1}{2}$ and in particular, for $k \leqslant \frac{c}{3}$, then (3.3.6) is simply bounded as follows:

$$\begin{aligned}
e^{-c} \sum_{k=0}^{N} \frac{1}{\sqrt{(2k+1)c}} \left(\frac{4ec}{2k+1}\right)^k &\leqslant e^{-c} \sum_{k=0}^{N} \frac{1}{\sqrt{(2c/3+1)c}} \left(\frac{4ec}{2c/3+1}\right)^{c/3} \\
&\leqslant e^{-c} \frac{N+1}{\sqrt{(2c/3)c}} \left(\frac{4ec}{2c/3}\right)^{c/3} \\
&\leqslant e^{-c} \frac{c/3}{\sqrt{(2c/3)c}} (6e)^{c/3} \\
&\leqslant e^{3(-c/3)} \frac{1}{\sqrt{6}} \left(\frac{1}{6e}\right)^{-c/3} = \frac{1}{\sqrt{6}} \left(\frac{e^2}{6}\right)^{-c/3}.
\end{aligned}$$

□

Lemma 3.3.2. *Let $f \in \widetilde{H}^s(I)$, $s > 0$. Then for any real $c \geqslant 6$, we have (with notations as above):*

$$\|f - \mathbb{E}\big[\widehat{f}_{c,n}\big]\|^2 \leqslant \frac{49}{3} \left(\frac{e^2}{6}\right)^{-2[c/3]} \|f\|^2 + 2\left[\frac{c}{3}\right]^{-2s} \|f\|^2_{\widetilde{H}^s}. \tag{3.3.7}$$

Proof. Let N be a positive integer such that $N \leqslant \left[\frac{c}{3}\right]$. It follows from (3.3.1) and (3.3.3) that:

$$\begin{aligned}
\Big\|f - \mathbb{E}\big[\widehat{f}_{c,n}\big]\Big\| &= \Big\|f - \sum_{n=0}^{\infty} \lambda_n(c) < f, \psi_{n,c} > \psi_{n,c}\Big\| = \Big\| \sum_{n=0}^{\infty} \big(1 - \lambda_n(c)\big) < f, \psi_{n,c} > \psi_{n,c}\Big\| \\
&\leqslant \Big\| \sum_{n=0}^{N} \big(1 - \lambda_n(c)\big) < f, \psi_{n,c} > \psi_{n,c}\Big\| + \Big\| \sum_{n=N+1}^{\infty} (1 - \lambda_n(c)) < f, \psi_{n,c} > \psi_{n,c}\Big\|.
\end{aligned} \tag{3.3.8}$$

To bound the first term in the right-hand side of (3.3.8), we proceed as follows. Cauchy-Schwarz

inequality and the fact that $\|\psi_{n,c}\| = 1$ imply $| < f, \psi_{n,c} > | \leqslant \|f\|\|\psi_{n,c}\| \leqslant \|f\|$. Moreover, $1-\lambda_n(c) > 0$ for $n \geqslant 0$. Thus, by using Minkowski inequality, we obtain:

$$\Big\| \sum_{n=0}^{N} \big(1 - \lambda_n(c)\big) < f, \psi_{n,c} > \psi_{n,c} \Big\| \leqslant \sum_{n=0}^{N} \big(1 - \lambda_n(c)\big)\|f\|.$$

From (2.1.18), for any $0 \leqslant n \leqslant N \leqslant c/2.7$, we have $1 - \lambda_n(c) \leqslant \dfrac{7}{\sqrt{c}} e^{-c} \dfrac{(2c)^n}{n!}$. Combining this with the previous inequality and the inequality (3.3.5) of Lemma 3.3.1, we obtain, for $N = [\frac{c}{3}]$:

$$\Big\| \sum_{n=0}^{N} \big(1 - \lambda_n(c)\big) < f, \psi_{n,c} > \psi_{n,c} \Big\| \leqslant \frac{7}{\sqrt{6}} \left(\frac{e^2}{6}\right)^{-[c/3]} \|f\|. \tag{3.3.9}$$

To bound the second term in the right-hand side of (3.3.8), we use again the fact that $0 < \lambda_n(c) < 1$ for $n \geqslant 0$. Then, using Parseval's equality yields:

$$\Big\| \sum_{n=N+1}^{\infty} \big(1 - \lambda_n(c)\big) < f, \psi_{n,c} > \psi_{n,c} \Big\|^2 = \sum_{n=N+1}^{\infty} \big(1 - \lambda_n(c)\big)^2 | < f, \psi_{n,c} > |^2 \leqslant \sum_{n=N+1}^{\infty} | < f, \psi_{n,c} > |^2.$$

Hence, by using (3.3.4) with $N = \left[\frac{c}{3}\right]$, we obtain

$$\Big\| \sum_{n=N+1}^{\infty} (1 - \lambda_n(c)) < f, \psi_{n,c} > \psi_{n,c} \Big\| \leqslant \left[\frac{c}{3}\right]^{-s} \|f\|_{\widetilde{H}^s}. \tag{3.3.10}$$

Finally, by combining (3.3.8), (3.3.9) and (3.3.10) and using the inequality $(a+b)^2 \leqslant 2\big(a^2 + b^2\big)$, one gets the desired inequality (3.3.7). □

The next theorem quantifies the quality of the Sinc kernel based regression scheme in the weighted Sobolev space.

Theorem 3.3.3. *Let $c \geqslant 6$ be a positive real number. Assume that the regression function f in model (3.2.1) belongs to $\widetilde{H}^s(I)$, for some $s > 0$ and $|f(x)| \leqslant M$, a.e. on I. Then, under the previous hypotheses on the noises η_i, we have*

$$\mathbb{E}\left[\Big\| f - \widehat{f}_{c,n} \Big\|^2\right] \leqslant \frac{8c}{n\pi}\big(\sigma^2 + M^2\big) + \frac{98}{3}\left(\frac{e^2}{6}\right)^{-2[c/3]} \|f\|^2 + 4\left[\frac{c}{3}\right]^{-2s} \|f\|^2_{\widetilde{H}^s}. \tag{3.3.11}$$

Proof. We first check that

$$\big\|K_c(\cdot,\cdot)\big\|_2^2 \leqslant \frac{2c}{\pi}, \quad \forall\, x \in I. \tag{3.3.12}$$

To see this, note that

$$\int_I \left(\frac{\sin c(x-y)}{\pi(x-y)}\right)^2 dy = \int_{x-1}^{x+1} \left(\frac{\sin ct}{\pi t}\right)^2 dt \leqslant \int_{\mathbb{R}} \left(\frac{\sin ct}{\pi t}\right)^2 dt.$$

Since $\frac{\sin ct}{\pi t} = \frac{1}{2\pi}\mathcal{F}(\mathbf{1}_{[-c,c]}(\cdot))(t)$ (with $\mathcal{F}$ the usual Fourier transform), then Plancherel's equality implies

$$\int_{\mathbb{R}} \left(\frac{\sin ct}{\pi t}\right)^2 dt = \frac{1}{4\pi^2} 2\pi \int_{\mathbb{R}} \left(\mathbf{1}_{[-c,c]}(t)\right)^2 dt = \frac{c}{\pi},$$

which gives us (3.3.12). Next, we have

$$\|f - \widehat{f}_{c,n}\| \leqslant \|f - \mathbb{E}[\widehat{f}_{c,n}]\| + \|\mathbb{E}[\widehat{f}_{c,n}] - \widehat{f}_{c,n}\|. \tag{3.3.13}$$

Now, using (3.3.2) and (3.3.3), we can write:

$$\widehat{f}_{c,n}(x) - \mathbb{E}(\widehat{f}_{c,n}(x)) = \frac{1}{n}\sum_{i=1}^n \Big(2K_c(X_i,x)(f(X_i)+\eta_i) - < f, K_c(\cdot,x) > \Big) = \frac{1}{n}\sum_{i=1}^n \xi_i(x).$$

For the sake of simplification, we adopt the notations

$$h_c(X_i,x) = 2K_c(X_i,x)(f(X_i)+\eta_i), \qquad a_f(x) =< f, K_c(\cdot,x) > .$$

Then, we have

$$\begin{aligned}\Big(\sum_{i=1}^n \xi_i(x)\Big)^2 &= \sum_{i=1}^n \big(h_c(X_i,x)\big)^2 + a_f(x)\Big(a_f(x) - 2h_c(X_i,x)\Big) \\ &\quad + \sum_{i,j=1, i\neq j}^n h_c(X_i,x)h_c(X_j,x) - a_f(x)\Big(h_c(X_i,x) + h_c(X_j,x) - a_f(x)\Big)\end{aligned} \tag{3.3.14}$$

By using the hypotheses on the η_i and the independence of the η_i and the X_i, as well as Fubini's theorem, together with (3.3.12), one gets

$$\begin{aligned}\mathbb{E}\Big[\int_I \big(h_c(X_i,x)\big)^2 dx\Big] &= \int_I \mathbb{E}\Big[\big(h_c(X_i,x)\big)^2\Big] dx \leqslant \int_I \mathbb{E}\Big[4K_c^2(X_i,x)(f^2(X_i) + 2\eta_i f(X_i) + \eta_i^2)\Big] dx \\ &= 2\int_{I^2} K_c^2(y,x) f^2(y)\, dy\, dx + 2\sigma^2 \int_{I^2} K_c^2(y,x)\, dy\, dx \\ &\leqslant 2(M^2+\sigma^2)\|K_c(\cdot,\cdot)\|_2^2 \leqslant \frac{4c}{\pi}(M^2+\sigma^2).\end{aligned} \tag{3.3.15}$$

Also, we have $\mathbb{E}\Big[a_f(x) - 2h_c(X_i,x)\Big] = 0$, so that:

$$\mathbb{E}\Big[\int_I a_f(x)\Big(a_f(x) - 2h_c(X_i,x)\Big)\,dx\Big] = \int_I a_f(x)\mathbb{E}\Big[a_f(x) - 2h_c(X_i,x)\Big]\,dx = 0. \tag{3.3.16}$$

Similarly, one can easily check that

$$\mathbb{E}\Big[\int_I h_c(X_i,x)h_c(X_j,x)\,dx\Big] = \mathbb{E}\Big[\int_I a_f(x)h_c(X_j,x)\,dx\Big] = \int_I a_f^2(x)\,dx.$$

That is, for $i \neq j$, one gets

$$\mathbb{E}\Big[\int_I \Big(h_c(X_i,x)h_c(X_j,x) - a_f(x)\Big(h_c(X_i,x) + h_c(X_j,x) - a_f(x)\Big)\,dx\Big] = 0 \tag{3.3.17}$$

Consequently, by using (3.3.14)–(3.3.17), one concludes that

$$\mathbb{E}\Big[\Big\|\widehat{f}_{c,n} - \mathbb{E}(\widehat{f}_{c,n})\Big\|^2\Big] = \mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i=1}^n \xi_i\Big\|^2\Big] \leqslant \frac{4c}{n\pi}\Big(\sigma^2 + M^2\Big). \tag{3.3.18}$$

For the Sinc kernel, $f - \mathbb{E}(\widehat{f}_{c,n})$ and $\mathbb{E}(\widehat{f}_{c,n}) - \widehat{f}_{c,n}$ are not orthogonal, unlike the Gegenbauer projection kernel. Nonetheless, from (3.3.13), together with (3.3.7) and (3.3.18), one gets

$$\begin{aligned}
\mathbb{E}\Big[\Big\|f - \widehat{f}_{c,n}\Big\|^2\Big] &\leqslant \mathbb{E}\Big[\Big(\Big\|f - \mathbb{E}\big[\widehat{f}_{c,n}\big]\Big\| + \Big\|\mathbb{E}\big[\widehat{f}_{c,n}\big] - \widehat{f}_{c,n}\Big\|\Big)^2\Big] \\
&\leqslant \mathbb{E}\Big[2\Big(\Big\|\mathbb{E}\big[\widehat{f}_{c,n}\big] - \widehat{f}_{c,n}\Big\|^2 + \Big\|f - \mathbb{E}\big[\widehat{f}_{c,n}\big]\Big\|^2\Big)\Big] \\
&\leqslant 2\mathbb{E}\Big[\Big\|\mathbb{E}\big[\widehat{f}_{c,n}\big] - \widehat{f}_{c,n}\Big\|^2\Big] + 2\mathbb{E}\Big[\Big\|f - \mathbb{E}\big[\widehat{f}_{c,n}\big]\Big\|^2\Big] \\
&\leqslant \frac{8c}{n\pi}\Big(\sigma^2 + M^2\Big) + 2\Big(\frac{49}{3}\Big(\frac{e^2}{6}\Big)^{-2[c/3]}\|f\|^2 + 2\Big[\frac{c}{3}\Big]^{-2s}\|f\|^2_{\widehat{H}^s}\Big).
\end{aligned}$$

□

Under the assumption that the bandwidth $c > 0$ is large enough and using (3.3.11), we conclude that the Sinc kernel based empirical projection operator attains its minimum error when the first and third terms in (3.3.11) are of the same order. Straightforward computation shows that this is the case when $c = O(n^{\frac{1}{2s+1}})$. Hence, the estimation error of $\widehat{f}_{c,n}$ is of $O(n^{\frac{-s}{2s+1}})$, where $s > 0$ denotes the Sobolev smoothness of f.

Remark 3.3.4. *As for the Gegenbauer polynomials kernel, our nonparametric Sinc-kernel projection kernel estimator $\widehat{f}_c$ can be generalized to the multivariate setting by different ways. One way is to*

consider the tensor product of univariate Sinc-kernels. A second way is to consider the well known $d-$variate version of the Sinc-kernel, known as the wave kernel, which is also a spectral kernel. This last kernel is a popular kernel in many applications, including applications arising in machine learning. For a real $c > 0$, this kernel is given by

$$\boldsymbol{K}_c(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sin(c\|\boldsymbol{x} - \boldsymbol{y}\|_2)}{\pi\|\boldsymbol{x} - \boldsymbol{y}\|_2}, \quad \boldsymbol{x}, \boldsymbol{y} \in I^d,$$

where $\|\cdot\|_2$ denotes the usual $2-$norm of $\mathbb{R}^d$.

3.3.2 Extension to other kernels

Now, we briefly describe how our random kernel projection based estimators can be applied when the inputs X_i are drawn from some more general marginal probability measure ρ_X. For this purpose, we assume that $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel and also a Mercer's kernel. In this case, the associated integral operator L_K, defined on $L^2(\mathcal{X}, \rho_X) = L^2_{\rho_X}$ by $L_K(f)(x) = \displaystyle\int_{\mathcal{X}} K(x, y) f(y)\, d\rho_X(y)$ is a self-adjoint compact operator. We assume that the countable set of eigenfunctions φ_n of L_K form an orthonormal set of a weighted $L^2(\mathcal{X}, \omega_K(x)dx)-$space. If $C_K = \displaystyle\int_{\mathcal{X}} \omega_K(x)\, dx$, then $\mu_K(\cdot) = \frac{1}{C_K}\omega_K(\cdot)$ is a probability measure on $\mathcal{X}$. We assume that the X_i are i.i.d. copies of a random variable with distribution associated with the probability measure μ_X. The associated general kernel projection estimator for problem (3.2.1) is given by

$$\widehat{f}_{n,K} = \frac{C_K}{n} \sum_{i=1}^{n} Y_i K_{X_i}(x), \quad \text{where} \quad K_{X_i}(x) = K(X_i, x), \quad x \in \mathcal{X}.$$

Let $\mathcal{H}_K$ be the reproducing Hilbert space associated with the kernel $K(\cdot, \cdot)$. Assume that the true regression function f lies in a larger subspace of the Hilbert space $L^2_{\rho_X}$. It is easy to see that in this case, we have

$$\mathbb{E}\big[\widehat{f}_{n,K}\big] = \int_{\mathcal{X}} K(x, y) f(y)\, d\mu_X(y) = L_K(f)(x).$$

Then, we use the classical variance-bias decomposition

$$\mathbb{E}\Big[\|\widehat{f}_{n,K} - f\|^2_{L^2_{\mu_X}}\Big] \leqslant 2\,\mathbb{E}\Big[\|\widehat{f}_{n,K} - L_K f\|^2_{L^2_{\mu_X}}\Big] + 2\,\mathbb{E}\Big[\|L_K f - f\|^2_{L^2_{\mu_X}}\Big]. \tag{3.3.19}$$

Let $\kappa^2 = \|K(\cdot, \cdot)\|^2_{\mu_X}$ and consider the auxiliary function

$$h(X_i, x) = C_K K(X_i, x)(f(X_i) + \eta_i), \qquad a_f(x) = < f, K(\cdot, x) > .$$

We assume that $|f(x)| \leqslant M$, *a.e.* on $\mathcal{X}$. Then, by repeating the steps of the proof of Theorem 3.3.3, one can easily get the following general estimate of the L^2-risk variance term

$$\mathbb{E}\Big[\big\|\widehat{f}_{n,K} - L_K f\big\|^2_{L^2_{\mu_X}}\Big] \leqslant C_K \frac{\kappa^2}{n}\big(M^2 + \sigma^2\big). \tag{3.3.20}$$

Unlike the previous general estimate of the variance term, the estimate of the bias term $\mathbb{E}[\|L_K f - f\|^2_{L^2_{\mu_X}}]$ is much more involved. It depends heavily on more involved spectral properties of the operator L_K, as well as on the specific properties of the subspace to which the true regression function f belongs. For more details on this issue, the reader is referred to the following reference works on the subject: Cucker and Smale (2002); Smale and Zhou (2007, 2005); Steinwart and Christmann (2008), as well as the references therein.

Example 3.3.5. *(Bessel kernel): For two real numbers $\alpha > -1$ and $c > 0$, we consider the Bessel kernel $K_{\alpha,c}(\cdot,\cdot)$, defined by*

$$K_{\alpha,c}(x,y) = 2^{\alpha}\frac{c}{\pi}\Gamma(\alpha+1)\frac{J_{\alpha+1/2}\big(c(x-y)\big)}{\big(c(x-y)\big)^{\alpha+1/2}}, \quad x,y \in I^2, \quad I = [-1,1].$$

Here, $J_\alpha(\cdot)$ is the Bessel function of the first kind and of order α and $\Gamma(\cdot)$ is the Gamma function. In the special case $\alpha = 0$, the kernel $K_{0,c}$ reduces to the Sinc kernel of the previous section. It has been shown in Karoui and Souabni (2016) that for the Bessel kernel, we have

$$\kappa^2 = \|K_{\alpha,c}(\cdot,\cdot)\|_2^2 = \frac{c}{2}\frac{\Gamma^2(\alpha+1)}{\Gamma^2(\alpha+3/2)}. \tag{3.3.21}$$

The eigenfunctions of the integral operator with the Bessel kernel form an infinite and countable orthonormal set of the weighted L^2-space $L^2(I, \omega_{\alpha+1/2}(x)dx)$, where $\omega_{\alpha+1/2}\alpha$ is the weight function associated with the Gegenbauer polynomials, that is $\omega_{\alpha+1/2}\alpha(x) = (1-x^2)^{\alpha-1/2}$. In this case, the associated probability measure is given by

$$\mu_K(x) = \frac{1}{2^{2\alpha+1}B(\alpha+1,\alpha+1)}\,\omega_{\alpha+1/2}(x) = \frac{1}{C_K}\,\omega_{\alpha+1/2}(x), \quad x \in I. \tag{3.3.22}$$

Our kernel projection based estimator for nonparametric regression is then given by

$$\widehat{f}_{n,K} = \frac{C_K}{n}\sum_{i=1}^{n} Y_i K_{\alpha,c}(X_i, x),$$

where the X_i are i.i.d. observations from the Beta distribution on I with both parameters equal to

$\alpha+1$. *Then, using* (3.3.20), (3.3.21) *and* (3.3.22), *one gets the following bound for the variance term of the* L^2-*error:*

$$\mathbb{E}\Big[\big\|\widehat{f}_{n,K}(x)- < f(\cdot), K_{\alpha,c}(\cdot,x) >_{L^2_{\mu_K}} \big\|^2_{L^2_{\mu_X}}\Big] \leqslant 2^{2\alpha+1}B(\alpha+1,\alpha+1)\frac{c}{2}\frac{\Gamma^2(\alpha+1)}{\Gamma^2(\alpha+3/2)}\frac{1}{n}\big(M^2+\sigma^2\big),$$

where $|f(x)| \leqslant M$ *a.e. for* $x \in I$. *Also, since from Karoui and Souabni (2016), the eigenvalues* $\lambda_n^\alpha(c)$ *of the integral operator associated with the Bessel kernel have very similar behaviour as the eigenvalues of the Sinc kernel, then we expect that the bias error term for the estimator* $\widehat{f}_{n,K}(\cdot)$ *is similar to the one given in Theorem 3.3.3.*

3.4 Random pseudo-inverse based estimator for LFR problem.

It is well known that the linear functional regression (LFR) model is used in a wide range of applications from different research area, such as medicine, agriculture and finance. The LFR model is described as follows. Let J be a compact interval and $L^2(J)$ be the Hilbert space of square integrable functions with its usual inner product $< \cdot,\cdot >$. For a positive integer n, let $\{(X_i(\cdot), Y_i),\ i = 1,\ldots,n\}$ be the set of observations, where the Y_i are scalar responses and the $X_i(\cdot) \in L^2(J)$ are the functional predictors, which we assume to be i.i.d. copies of a centered second order stochastic process $X_t = X(t), t \in J$, that is $\mathbb{E}\big[X^2(t)\big] < \infty$, for all $t \in J$. Then the LFR model is given by:

$$Y_i = \int_J X_i(s)\beta_0(s)\,ds + \varepsilon_i, \tag{3.4.1}$$

see for example Hall and Horowitz (2007); Yuang and Cai (2010). Here, the noise terms ε_i are i.i.d. centered real valued random variables with variance σ^2, independent of the random functional predictors $X_i(\cdot)$ and $\beta_0(\cdot)$ is the unknown slope function to be estimated. In general, we assume that the $X_i(\cdot)$ have the following orthogonal series expansion $X_i(s) = \sum_{k\geqslant 1} \xi_k Z_{i,k}\varphi_k(s)$, where the φ_k form an orthonormal family of $L^2(J)$, the $Z_{i,k}$ are i.i.d. centered random variables with variance σ_Z^2 and $(\xi_k)_{k\geqslant 1}$ is a deterministic sequence of $\mathbb{R}\backslash\{0\}$. This last sequence is closely related to the eigenvalues of the covariance operator. In fact, from the previous expression of $X_i(\cdot)$ and the fact that the $\varphi_k(\cdot)$ are the orthonormal eigenfunctions of the covariance operator associated with the eigenvalues λ_k, it is easy to see that

$$\xi_k^2\sigma_Z^2 = Cov(\xi_k Z_{i,k}, \xi_{k'} Z_{i,k'}) = \lambda_k\delta_{kk'}.$$

In practice, we require that the λ_k decay with a certain rate and consequently, we may assume that

$$X_i(s) = \sum_{k=1}^{N} \xi_k Z_{i,k} \varphi_k(s), \quad 1 \leqslant i \leqslant n, \tag{3.4.2}$$

for an appropriate positive integer N. In the sequel, we are interested in the approximation of the solution $\beta_0(\cdot) \in L^2(J)$ with minimal $L^2(J)-$norm. Let $n > N$ be a positive integer and consider n values $X_i(\cdot)$ of the functional predictor, then the LFR problem reduces to solving in a stable way the following over-determined linear system

$$Y_i = \int_J X_i(s)\, \beta_0(s)\, ds + \varepsilon_i, \quad i = 1, \ldots, n. \tag{3.4.3}$$

By combining (3.4.2) and (3.4.3) and using the projection theorem, the minimum $L^2(J)-$norm solution of the LFR problem is given by $\pi_N \beta_0(\cdot)$, where $\beta_0(\cdot)$ is any solution of (3.4.3) and π_N is the orthogonal projection over $\mathcal{H}_N$. Hence, in the sequel we will simply write $\beta_0(\cdot)$ instead of $\pi_N \beta_0(\cdot)$.

Note that a popular class of estimators for the previous LFR problem is given by the penalized least squares in the framework of an RKHS. More precisely, see for example Shin and Lee (2016); Yuang and Cai (2010), let $\mathcal{H}_K$ be a reproducing kernel Hilbert space generated by a Mercer's kernel. Then, such estimators are given by

$$\widehat{\beta}_n^{\lambda} = \arg \min_{\beta \in \mathcal{H}_K} \left[\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{Y_i - \int_J X_i(s)\beta(s)}{\widehat{\sigma}} \right) + \lambda J(\beta) \right].$$

Here, $\rho(\cdot)$ is a loss function, $\widehat{\sigma}$ is a scale estimate of errors, $J(\beta)$ is a penalization function and λ is a regularization parameter. The special case of $\rho(x) = x^2$ corresponds to the RKHS based least squared error estimator. In Yuang and Cai (2010) and thanks to the representer theorem that holds in the RKHS framework, the solution of the previous minimization problem has the expansion $\widehat{\beta}_n^{\lambda}(s) = \sum_{j=1}^{M} d_j \psi_j(s) + \sum_{i=1}^{n} c_i < K(\cdot, s), X_i(\cdot) >$. Here, the $\psi_j(\cdot)$ form an orthonormal basis of the null space of the functional penalty J. The expansion coefficients vectors $\boldsymbol{d} = [d_j]'$ and $\boldsymbol{c} = [c_i]'$ are given by

$$\boldsymbol{d} = \left(T' W^{-1} T\right)^{-1} T' W^{-1} \boldsymbol{Y}, \quad \boldsymbol{c} = W^{-1} \Big[I_n - T \left(T' W^{-1} T\right)^{-1} T' W^{-1} \Big] \boldsymbol{Y},$$

where

$$W = \Sigma + n\lambda I_n, \quad \Sigma_{ij} = \int_{J^2} X_i^c(u) K(u,v) X_j^c(v) du dv, \quad T_{ij} = \int_J X_i^c(u) \psi_j(u)\, du.$$

Here, the $X_i^c(\cdot)$ are centered samples of the prediction functional. Let ρ_k and μ_k denote the eigenvalues

of the integral operators associated with the reproducing kernel $K(\cdot,\cdot)$ and the covariance kernel $C(\cdot,\cdot)$, respectively. In Yuang and Cai (2010) and under some technical assumptions including the decay assumptions $\rho_k \asymp k^{-2r}$ and $\mu_k \asymp k^{-2s}$, for some $r, s > \frac{1}{2}$, it has been shown that the optimal mean squared prediction error of the RKHS penalized scheme is of order $n^{-\frac{2(r+s)}{2(r+s)+1}}$.
In the sequel, we describe a special case of a more general random pseudo-inverse LFR estimator that have been given in Ben Saber and Karoui (2021). The estimate of the N expansion coefficients of $\pi_N\beta_0(\cdot)$ with respect to the basis functions $\varphi_j(\cdot)$ is based on the use of a pseudo-inverse of a random projection matrix. More precisely, since $\beta_0(s) = \sum_{j=1}^{N} c_j\varphi_j(s), \quad \forall\, s \in J$, then under a mild condition on $n \geqslant N$, we have

$$\boldsymbol{c} = [c_1, \ldots, c_N]' = G_N^{-1} \cdot \Big(F_N' \cdot \frac{1}{\sqrt{n}}\big[Y_i - \varepsilon_i\big]'_{1\leqslant i\leqslant n}\Big),$$

where F_N is an $n \times N$ random projection matrix and G_N is an $N \times N$ positive definite random matrix, given respectively by

$$G_N = F_N'F_N, \qquad F_N = \frac{1}{\sqrt{n}}\Big[\xi_j Z_{i,j}\Big]_{1\leqslant i\leqslant n, 1\leqslant j\leqslant N}. \tag{3.4.4}$$

The random pseudo-inverse based estimator of $\boldsymbol{c}$ and our estimator $\widehat{\beta}_{n,N}(\cdot)$ of $\beta_0(\cdot)$ are given by

$$\widehat{\beta}_{n,N}(s) = \sum_{k=1}^{N} \widehat{c}_k\varphi_k(s), \qquad \widehat{\boldsymbol{c}} = [\widehat{c}_1, \ldots, \widehat{c}_N]' = G_N^{-1} \cdot \Big(F_N' \cdot \frac{1}{\sqrt{n}}\big[Y_i\big]'_{1\leqslant i\leqslant n}\Big). \tag{3.4.5}$$

From Ben Saber and Karoui (2021), one gets the important result that with high probability, the matrix G_N is indeed positive definite and hence invertible. Also, we show that under a slow decay rate condition of the sequence $(\xi_k)_k$, the matrix G_N is also well conditioned. This last property is crucial for the stability of our estimator. By stability, we mean that the integrated squared error (ISE) of the estimator is proportional to $\|(\varepsilon_i)_i\|_{\ell_2}^2$, the squared ℓ_2-norm of the added noise vector. More precisely, if

$$\kappa_2(G_N) = \frac{\lambda_{\max}(G_N)}{\lambda_{\min}(G_N)}$$

is the $2-$norm condition number of the random matrix G_N, then we prove that for any $\eta > 0$, we have with high probability,

$$\kappa_2(G_N) \leqslant \frac{1.72\max_{k\geqslant 1}\sigma_Z^2\xi_k^2 + \frac{M_{\boldsymbol{\xi},N}}{n}\log(N) + \eta}{0.63\min_{k\geqslant 1}\sigma_Z^2\xi_k^2 - \frac{M_{\boldsymbol{\xi},N}}{n}\log(N) - \eta}.$$

Here, for $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$, $\|\boldsymbol{\xi}\|_{l_1} = \sum_{j=1}^{N} |\xi_j|$ and $|Z_{i,k}| \leqslant M_Z$ almost surely, we have

$$M_{\boldsymbol{\xi},N} = M_Z^2 \max_{1 \leqslant j \leqslant N} |\xi_j| \cdot \|\boldsymbol{\xi}\|_{l_1}.$$

In fact, from the expression of the predictor functional given by (3.4.2), we have $X_i(s) = \sum_{k=1}^{N} \xi_k Z_{i,k} \varphi_k(s)$, $s \in J$. Taking into account the orthonormality of the $\varphi_k(\cdot)$ and by substituting the previous expansion in (3.4.3), the scalar LFR model re-scaled by a factor $\frac{1}{\sqrt{n}}$ is written in the following simpler form

$$\frac{1}{\sqrt{n}} Y_i = \frac{1}{\sqrt{n}} \sum_{k=1}^{N} \xi_k Z_{i,k} c_k + \frac{1}{\sqrt{n}} \varepsilon_i, \quad i = 1, \ldots, n.$$

In matrix form, the previous over-determined system is written as

$$\frac{1}{\sqrt{n}} \boldsymbol{Y} = \frac{1}{\sqrt{n}} \Big[\xi_k Z_{i,k} \Big]_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant k \leqslant N+1}} \cdot \boldsymbol{c}' + \frac{1}{\sqrt{n}} \boldsymbol{\varepsilon}' = F_N \cdot \boldsymbol{c}' + \frac{1}{\sqrt{n}} \boldsymbol{\varepsilon}'. \tag{3.4.6}$$

Multiplying the previous system by F_N' and under the fact that the positive definite random matrix of reduced dimension N, given by $G_N = F_N' F_N$ is well conditioned, a stable approximate solution of system (3.4.6) is given by the estimator

$$\widehat{\boldsymbol{c}}_{n,N}' = G_N^{-1} \cdot \left(F_N' \frac{1}{\sqrt{n}} \boldsymbol{Y} \right). \tag{3.4.7}$$

The following theorem, which is a special case of a more general theorem given in Ben Saber and Karoui (2021), is important and it shows that the random matrix G_N is invertible and well conditioned. For the sake of convenience, we give some steps of the proof given in more details in the previous theorem.

Theorem 3.4.1. *Assume that* $\max_{k \geqslant 1}(|Z_{i,k}|) \leqslant M_Z$ *almost surely. Then, under the previous notations and assumptions, for any* $\eta > 0$, *we have with probability at least* $1 - 2\exp\left(\dfrac{-n\eta^2}{2M_{\boldsymbol{\xi},N}^2}\right)$,

$$\kappa_2(G_N) \leqslant \frac{1.72 \max_{k \geqslant 1} \sigma_Z^2 \xi_k^2 + \frac{M_{\boldsymbol{\xi},N}}{n} \log(N) + \eta}{0.63 \min_{k \geqslant 1} \sigma_Z^2 \xi_k^2 - \frac{M_{\boldsymbol{\xi},N}}{n} \log(N) - \eta}. \tag{3.4.8}$$

Here, $M_{\boldsymbol{\xi},N} = M_Z^2 \max_{1 \leqslant j \leqslant N} |\xi_j| \cdot \|\boldsymbol{\xi}\|_{l_1}$.

Proof: We first note that the random matrix G_N is explicitly given by $G_N = \frac{1}{n}\Big[\sum_{i=1}^{n} \xi_j \xi_k Z_{i,j} Z_{i,k}\Big]_{\substack{1\leqslant j\leqslant N\\ 1\leqslant k\leqslant N}}$. Since the i.i. d. random variables $Z_{i,j},\ i,j \geqslant 1$ are centered with variances σ_Z^2, then it is easy to see that

$$\mathbb{E}\big(G_N\big) = \begin{bmatrix} \sigma_Z^2\xi_1^2 & & \\ & \ddots & \\ & & \sigma_Z^2\xi_N^2 \end{bmatrix}.$$

Consequently, the minimum and maximum eigenvalues of $\mathbb{E}\big(G_N\big)$ are given by

$$\Lambda_{\min} = \lambda_{\min}\big(\mathbb{E}\big(G_N\big)\big) = \sigma_Z^2 \min_{k\geqslant 1} \xi_k^2, \quad \Lambda_{\max} = \lambda_{\max}\big(\mathbb{E}\big(G_N\big)\big) = \sigma_Z^2 \max_{k\geqslant 1} \xi_k^2.$$

On the other hand, the random matrix G_N is written in the following form

$$G_N = \frac{1}{n}\sum_{i=1}^{n} D_i, \quad D_i = \Big[\xi_j \xi_k Z_{i,j} Z_{i,k}\Big]_{1\leqslant j,k\leqslant N.}$$

Note that each D_i is a positive semi definite matrix, since for any $\boldsymbol{x} \in \mathbb{R}^N$, we have

$$\boldsymbol{x}' D_i \boldsymbol{x} = \boldsymbol{x}' B_i' \cdot B_i \boldsymbol{x} \geqslant 0, \qquad B_i = \frac{1}{\sqrt{n}}\big[\xi_j Z_{i,j}\big]_{1\leqslant j\leqslant N}.$$

From Gershgorin circle theorem, we have $\Big|\lambda_{\max}\left(\frac{D_i}{n}\right) - \frac{1}{n}\xi_j^2 Z_{i,j}^2\Big| \leqslant \frac{1}{n}\sum_{k\neq j} |\xi_j \xi_k Z_{i,j} Z_{i,k}|$, so that

$$\begin{aligned}
\lambda_{\max}\left(\frac{D_i}{n}\right) &\leqslant \frac{1}{n}|\xi_j^2||Z_{i,j}^2| + \frac{1}{n}\sum_{k\neq j} |\xi_j Z_{i,j}||\xi_k Z_{i,k}| \\
&\leqslant \frac{1}{n}|\xi_j Z_{i,j}| \sum_{k=1}^{N} |\xi_k Z_{i,k}| \\
&\leqslant \frac{M_Z^2}{n} \max_{1\leqslant j\leqslant N} |\xi_j| \sum_{k=1}^{N} |\xi_k| = \frac{M_Z^2}{n}\Big(\max_{1\leqslant j\leqslant N} |\xi_j|\Big)\|\boldsymbol{\xi}\|_{l_1} = \frac{M_{\boldsymbol{\xi},N}}{n},
\end{aligned}$$

Hence, from Tropp (2015), we have

$$\mathbb{E}\big(\lambda_{\min}(G_N)\big) \geqslant 0.63 \min_{k\geqslant 1} \sigma_Z^2 \xi_k^2 - \frac{M_{\boldsymbol{\xi},N}}{n}\log(N)$$

and

$$\mathbb{E}\big(\lambda_{\max}(G_N)\big) \leqslant 1.72 \max_{k\geqslant 1} \sigma_Z^2 \xi_k^2 + \frac{M_{\boldsymbol{\xi},N}}{n}\log(N).$$

To conclude the proof of (3.4.8), it suffices to use the following inequalities from Ben Saber and Karoui (2021). These inequalities are based on the use of McDiarmid's concentration inequality. For our matrix G_N, they are given as follows. For any $\eta > 0$, we have

$$\lambda_{\min}(G_N) \geqslant \mathbb{E}\big(\lambda_{\min}(G_N)\big) - \eta, \quad \lambda_{\max}(G_N) \leqslant \mathbb{E}\big(\lambda_{\max}(G_{N+1})\big) + \eta$$

where each inequality holds with probability at least $1 - \exp \frac{-n\eta^2}{2M^2_{\boldsymbol{\xi},N}}$. By combining the previous two inequalities, one gets the desired inequality (3.4.8).

In Ben Saber and Karoui (2021), the authors have provided an L^2-risk error of a truncated version of the LFR estimator $\widehat{\beta}_{n,N}(\cdot)$, defined as follows. Under the assumption that $|\beta_0(x)| \leqslant L$, $a.e.x \in J$, the truncated estimator $\widetilde{\beta}_{N,L}(\cdot)$ is given by

$$\widetilde{\beta}_{N,L}(x) = \mathrm{Sign}\big(\widehat{\beta}_{n,N}(x)\big) \min\Big(L, |\widehat{\beta}_{n,N}(x)|\Big). \tag{3.4.9}$$

Let $r > 0$ and $\eta_N > 0$ be such that

$$\mathbb{P}\big(\lambda_{\min}(G_N) \geqslant \eta_N\big) \geqslant 1 - n^{-r}. \tag{3.4.10}$$

Then, it is shown in Ben Saber and Karoui (2021) that

$$E\Big[\|\tilde{\beta}_{N,M}(\cdot) - \beta_0(\cdot)\|_2^2\Big] \leqslant \frac{\sigma^2\sigma_Z^2}{n^2}\frac{N}{\eta_N^2}\big\|(\xi_i)_i\big\|^2_{\ell_2} + \frac{4L^2}{n^r}, \tag{3.4.11}$$

where $\sigma^2 = \mathbb{E}\Big[\varepsilon_i^2\Big]$.

Remark 3.4.2. *In general, we assume that the coefficients of the deterministic sequence $(\xi_i)_i$ have a certain decay rate to zero of the form $\xi_k \asymp k^{-s}$, for some $s > 0$. In this case, we have $\|(\xi_i)_i\|^2_{\ell_2} \lesssim \max\big(N^{1-2s}, 1\big)$. Consequently, from the previous L2−risk error , one concludes that our proposed LFR estimator has a fairly fast convergence rate.*

Next, if $\kappa_2(G_N)$ is large, our estimator (3.4.5) will suffer from numerical instability. To overcome this problem, we propose to use a partition of the set $[[1, N]]$, given by p subsets $I_k = [[N_k, N_{k+1} - 1]]$, $1 \leqslant k \leqslant p$. For each subset I_k, we consider $X_{i,k}(\cdot) = S_k X_i(\cdot)$, the projection of the functional predictor $X_i(\cdot)$ on $\mathcal{H}_k = \mathrm{Span}\{\varphi_j(\cdot),\, j = N_k, \ldots, N_{k+1} - 1\}$. With these p components of the original prediction functionals $X_i(\cdot)$, the solution of the LFR problem is obtained by summing the p solutions of the p

reduced LFR problems

$$Y_{i,k} = \int_J X_{i,k}(s)\beta_0(s)\,ds + \varepsilon_{i,k}, \quad i = 1, \ldots, n, \quad 1 \leqslant k \leqslant p. \tag{3.4.12}$$

Here, the $\varepsilon_{i,k}$ are i.i.d. random variables with mean zero and variance $\mathbb{E}\big[\varepsilon_{i,k}^2\big] = \sigma_k^2$. For each $1 \leqslant k \leqslant p$, let $\widehat{\beta}_{n,N_k}(\cdot)$ be our pseudo-inverse based estimator associated with problem (3.4.12) with fixed value of k. Then, $\widehat{\beta}_{n,N_k}(\cdot)$ is an estimator of $S_k\beta_0(\cdot)$, the projection of $\beta_0(\cdot)$ over the subspace $\mathcal{H}_k$. It is interesting to note that if $G(N_k)$ denotes the random positive definite matrix used to compute the pseudo-inverse matrix associated with (3.4.12), then its 2−norm condition number is proportional to $\dfrac{\max_{l\in I_k}\xi_l^2}{\min_{l\in I_k}\xi_l^2}$, which is much smaller than the original condition number $\kappa_2(G_N)$. Hence, the new estimator $\widehat{\beta}_{n,N}(\cdot)$, given by

$$\widehat{\beta}_{n,N}(\cdot) = \sum_{k=1}^{p} \widehat{\beta}_{n,N_k}(\cdot) \tag{3.4.13}$$

has smaller prediction and estimation errors than the estimator (3.4.5). Note that the prediction and estimation errors of the estimator are given respectively by $\|\widehat{\beta}_{n,N}(\cdot) - \widehat{\beta}_0(\cdot)\|_0^2$ and $\|\widehat{\beta}_{n,N}(\cdot) - \widehat{\beta}_0(\cdot)\|_{L_2}^2$. By Parseval's equality, these errors are simply given by

$$\|\widehat{\beta}_{n,N}(\cdot) - \widehat{\beta}_0(\cdot)\|_0^2 = \sum_k \mu_k|\widehat{c}_k - c_k|^2, \quad \|\widehat{\beta}_{n,N}(\cdot) - \widehat{\beta}_0(\cdot)\|_{L_2}^2 = \sum_k |\widehat{c}_k - c_k|^2,$$

where the μ_k are the eigenvalue of the positive definite covariance operator.

Next, we show how our previous scheme can be adapted to handle the more realistic and practical case where the unknown slope function $\beta_0(\cdot)$ is given (or efficiently approximated) by an expansion with respect to another orthonormal set $\{\psi_j(\cdot),\, 1 \leqslant j \leqslant M\}$ of $L^2(J)$, that is

$$\beta_0(s) = \sum_{j=1}^{M} d_j\psi_j(s), \quad s \in J,$$

see for example Cai and Yuan (2012), Typically, we assume that $M \leqslant N$ in the sense that the second basis is better adapted for the approximation of $\beta_0(\cdot)$. By using the previous expansion together with the expansion of the $X_i(\cdot)$ given by (3.4.2), it is easy to see that a random pseudo-inverse estimator for the approximate solution of (3.4.3) is given by

$$\widehat{\beta}_{n,M}(s) = \sum_{j=1}^{M} \widehat{d}_j\psi_j(s), \qquad \widehat{\boldsymbol{d}} = [\widehat{d}_1, \ldots, \widehat{d}_M]' = \mathcal{G}_M^{-1} \cdot \Big(\mathcal{F}_M' \cdot \frac{1}{\sqrt{n}}\big[Y_i\big]_{1\leqslant i\leqslant n}'\Big). \tag{3.4.14}$$

Here,

$$\mathcal{G}_M = \mathcal{F}'_M \mathcal{F}_M, \quad \mathcal{F}_M = F_N \cdot T_{N,M}, \quad T_{N,M} = \Big[< \varphi_k(\cdot), \psi_j(\cdot) > \Big]_{\substack{1 \leqslant k \leqslant N \\ 1 \leqslant j \leqslant M}} \tag{3.4.15}$$

and the random matrix F_N is as given by (3.4.4). We assume that the transformation matrix $T_{N,M} \in \mathbb{R}^{N \times M}$ has a 2−condition number $\kappa_2(T_{N,M})$ of reasonable magnitude. It is well known that since $n \geqslant N \geqslant M$, then $\kappa_2(\mathcal{F}_M) = \kappa_2(F_N \cdot T_{N,M}) \leqslant \kappa_2(F_N)\kappa_2(T_{N,M})$, so that

$$\kappa_2(\mathcal{G}_M) \leqslant \kappa_2(G_N)\big(\kappa_2(T_{N,M})\big)^2.$$

Remark 3.4.3. *The previous upper bound for $\kappa_2(\mathcal{G}_M)$ is not optimal, it is used to justify that $\mathcal{G}_M$ is invertible. In practice, the actual value of $\kappa_2(\mathcal{G}_M)$ is smaller than this upper bound and even smaller than $\kappa_2(G_N)$, for small enough positive integer M.*

In the third example of the next paragraph, we illustrate the performance of this more general estimator $\widehat{\beta}_{n,M}(\cdot)$ in the case where of the two orthonormal sets of $L^2([0,1])$, given by $\{\varphi_1 = 1,\ \varphi_k(s) = \sqrt{2}\cos(\pi(k-1)s),\ 2 \leqslant k \leqslant N\}$ and $\{\psi_j(s) = \sqrt{2}\widetilde{C}_{j-1}^{1/2}(2s-1),\ 1 \leqslant j \leqslant M\}$ with $N = 50$ and $M = 5$. Here, the special case of the Gegenbauer polynomials φ_k are just the orthonormal Legendre polynomials over $J = [0,1]$. Consider the special value of $\alpha = \frac{1}{2}$ in the general formula of the finite Fourier transform of the Gengenbauer polynomials, given in [Olver et al. (2010), p.456] (see also Jaming et al. (2016)), the different transformation matrix entries $T_{N,M}$ are exactly given by the following formula

$$2 < \cos(\pi k s), \widetilde{C}_j^{1/2}(2s-1) >= \begin{cases} (-1)^{\frac{j+k}{2}} \sqrt{\frac{4j+2}{k}} J_{j+\frac{1}{2}}\Big(\frac{\pi k}{2}\Big) & \text{if } j+k \text{ is even} \\ \\ 0 & \text{if } j+k \text{ is odd,} \end{cases}$$

where $J_\mu(\cdot)$ is the Bessel function of the first type and order μ.

In the last part of this paragraph, we give a brief comparison between our LFR estimator $\widehat{\beta}_{n,N}(\cdot)$ and the well-known LFR Functional Principal Component Analysis (FPCA) based estimator, see for example Cai and Yuan (2012); Hall and Horowitz (2007); Ramsay and Silverman (2005). This last estimator is briefly described as follows. The n random observations $\int_J X_i(s)\beta_0(s)\,ds$ are used to build the associated empirical covariance kernel, given by

$$\widehat{K}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\big(X_i(s) - \overline{X}(s)\big)\big(X_i(t) - \overline{X}(t)\big), \quad \overline{X}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} X_i(\cdot).$$

Let $L_{\widehat{K}}$ be the associated positive definite integral operator and let $(\widehat{\lambda}_j)_{j\geqslant 1}$ and $(\widehat{\varphi}_j(\cdot))_{j\geqslant 1}$ be the sequences of positive eigenvalues and associated eigenfunctions of $L_{\widehat{K}}$. The LFR FPCA-based estimator, which we denote by $\widehat{\beta}(\cdot)$ is then given as a solution of the integral equation

$$L_{\widehat{K}}\widehat{\beta}(s) = \widehat{g}(s), \quad \widehat{g}(s) = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \overline{Y}\big)\big(X_i(s) - \overline{X}(s)\big) = \sum_{j\geqslant 1}\widehat{g}_j\widehat{\varphi}(s).$$

By using the spectral decomposition of the operator $L_{\widehat{K}}$, the estimator $\widehat{\beta}(\cdot)$ is given by $\widehat{\beta}(s) = \sum_{j\geqslant 1}\frac{\widehat{g}_j}{\widehat{\lambda}_j}\widehat{\varphi}_j(s)$. Note that since the $\widehat{\lambda}_j$ are approximate values of the eigenvalues λ_j of the true covariance operator, then typically the sequence $(\widehat{\lambda}_j)_j$ has a fast polynomial decay rate. Consequently, the previous expansion of $\widehat{\beta}(\cdot)$ suffers from numerical instability. Hence, a regularization procedure has to be further performed on $\widehat{\beta}(\cdot)$. Among the possible regularization schemes, we cite the truncation of the previous expansion of $\widehat{\beta}(\cdot)$ to an appropriate order m. A second regularization scheme is the use of an appropriate penalized parameter $\rho > 0$ on the spectrum of $L_{\widehat{K}}$, so that to replace the $\widehat{\lambda}_j$ by $\widehat{\lambda}_j + \rho$ in the previous expansion of $\widehat{\beta}(\cdot)$.

Note that our LFR estimator $\widehat{\beta}_{n,N}(\cdot)$ is based on a totally different methodology. It transforms the LFR observation into an over-determined system with unknowns, the expansion coefficients of $\beta_0(\cdot)$ with respect to the orthonormal eigenfunctions of the true covariance operator. This system is given by (3.4.7). By using some techniques from the spectral theory of positive definite random matrices, we are able to get a convenient upper bound for the 2−norm condition number of the random matrix associated with this system. This allows us to apply and prove that the pseudo-inverse technique computes in a stable way approximate values $\widehat{c}_j$ of the true expansion coefficients of $\widehat{\beta}_0(\cdot)$. These $\widehat{c}_j$ are taken as the expansion coefficients of $\widehat{\beta}_{n,N}(\cdot)$ with respect to the true $\varphi_j(\cdot)$.

3.5 Numerical results

In this section, we give various numerical tests that illustrate the nonparametric and functional regressions schemes of section 3, section 4 and section 5.

Example 3.5.1. *In this first example, we illustrate the Gegenbauer and Sinc kernels projection estimators, as well as Tikhonov regularization scheme with Sinc kernel, for approximating the regression function f in model (3.2.1). Here, f is the restriction to I of a bandlimited function with bandwidth $c = 20$. Thus, the model is*

$$Y_i = f(X_i) + \eta_i, \quad \textit{with} \quad f(x) = \frac{\sin(20x)}{20x}, \quad x \in I.$$

The X_i's are independently drawn from the uniform distribution on I. The random errors η_i are distributed as $0.1 \times Z_i$, where Z_i follows the standard normal distribution. We compute the two empirical projection estimators $\widehat{f}^{\alpha}_{N,n}$ and $\widehat{f}_{c,n}$ with $\alpha = 0$ and $N = 20$ (for Gegenbauer kernel) and $c = 20$ (for the Sinc kernel). For comparison, we also calculate the kernel ridge regression estimator $\widehat{f}^{\lambda}_{n}$ (with Sinc kernel $K_c(\cdot,\cdot)$). Comparisons are made for various sample sizes, namely $n = 100, 500, 1000$. We use generalized cross validation (GCV) to choose the appropriate value λ_{GCV} of λ in the kernel ridge regression estimator, that is, the value which provides an error close to the minimum regression error. Note that λ_{GCV} is given by

$$\lambda_{GCV} = \arg\min \frac{\sum_{i=1}^{n} \left(\widehat{f}^{\lambda}_{n}(X_i) - Y_i\right)^2}{n\left(1 - \frac{Trace(Q_\lambda^{-1} Q_0)}{n}\right)^2},$$

where Q_λ is the penalized Gram matrix. Under the simulation scenario described above, λ_{GCV} is approximately equal to 0.01.

A Gaussian quadrature scheme is used to compute the squared $L^2(I)$-regression errors $\|f - \widehat{f}^{\alpha}_{N,n}\|^2$, $\|f - \widehat{f}_{c,n}\|^2$ and $\|f - \widehat{f}^{\lambda}_{n}\|^2$. The simulation scenario is repeated 500 times and the empirical MISE is calculated by averaging the 500 squared $L^2(I)$-regression errors. Results are given in Table 3.1.

These results indicate that the three estimators perform similarly, for the respective values of their parameters, with slightly better performance for $\widehat{f}_{c,n}$ as soon as the sample size exceeds a few hundreds. However, the estimators $\widehat{f}_{c,n}$ and $\widehat{f}^{\alpha}_{N,n}$ are much faster to compute than the estimator $\widehat{f}^{\lambda}_{n}$. For example, we found that computing $\widehat{f}_{c,n}$ with $n = 1000$ is 15 times faster than computing $\widehat{f}^{\lambda}_{n}$ with $n = 150$. This is coherent with the well known fact (see for example Yang et al. (2017)) that the time complexity for computing the estimator $\widehat{f}^{\lambda}_{n}$ is $O(n^3)$. On the other hand, the empirical projection estimator, which requires a vector-matrix multiplication, has an $O(n^2)$ complexity. See also Rudi et al. (2016) for considerations about the time complexity of the kernel ridge estimator.

n	$\mathbb{E}\|f - \widehat{f}_{c,n}\|^2$	$\mathbb{E}\|f - \widehat{f}^{\alpha}_{N,n}\|^2$	$\mathbb{E}\|f - \widehat{f}^{\lambda}_{n}\|^2$
100	2.06e − 02	4.01e − 02	**5.54e − 03**
500	**4.46e − 03**	5.01e − 03	4.62e − 03
1000	**2.57e − 03**	2.78e − 03	3.01e − 03

Table 3.1: Empirical MISE (example (3.5.1)). The smallest values are indicated in bold.

Example 3.5.2. *In this example, we illustrate the accuracy of the three estimators $\widehat{f}_{c,n}$, $\widehat{f}^{\alpha}_{N,n}$ and $\widehat{f}^{\lambda}_{n}$ in the case of a regression function belonging to the Sobolev space $H^{\gamma}(I), \gamma > 0$. We consider for f the*

general periodic Brownian motion function $f^s(x)$ *given by:*

$$f^s(x) = \sum_{k=1}^{30} \frac{a_k}{k^s} \cos(k\pi x), \quad -1 \leqslant x \leqslant 1,$$

where s *is a positive real number and the* a_k*'s are coefficients generated from the standard normal distribution. It is well known that* $f^s \in H^\gamma(I)$ *almost surely for any* $\gamma < s - \frac{1}{2}$*. Here,* $H^\gamma(I)$ *denotes the usual Sobolev space over* I *with Sobolev regularity exponent* s*. For more details on smoothness properties of the periodic Brownian motions, the reader is referred to Bényi and Oh (2011).*
Random noise terms η_i *in the model* $Y_i = f^s(X_i) + \eta_i$ *are taken as* $\eta_i = 0.1 \times Z_i$*, with* Z_i *distributed as a standard normal law. We consider two values for* s*, namely* $s = 1, 2$*. Then, we calculate* $\widehat{f}_{c,n}$*,* $\widehat{f}^{\alpha}_{N,n}$ *and* $\widehat{f}^{\lambda}_{n}$*, with* $\alpha = 0$*,* $c = 30$*,* $N = 20$*,* $n = 100, 500, 1000$*. We consider the following sample sizes:* $n = 100, 500, 1000$*, and 500 samples are simulated for each simulation scenario. The empirical MISE is obtained by averaging the squared* $L^2(I)$*-regression errors over the 500 samples. Results are reported in Table 3.2. They suggest, again, similar performance of the three estimators, with slightly better performance for* $\widehat{f}_{c,n}$ *as the sample size grows, with a dramatic improvement of computational time when using the empirical projection estimators.*

s	n	$\mathbb{E}\lVert f^s - \widehat{f}_{c,n}\rVert^2$	$\mathbb{E}\lVert f^s - \widehat{f}^{\alpha}_{N,n}\rVert^2$	$\mathbb{E}\lVert f^s - \widehat{f}^{\lambda}_{n}\rVert^2$
1	100	5.55e − 01	4.58e − 01	**1.35e − 01**
	500	**1.02e − 01**	1.29e − 01	1.04e − 01
	1000	**8.55e − 02**	8.69e − 02	8.99e − 02
2	100	3.54e − 02	4.52e − 02	**1.54e − 02**
	500	7.75e − 03	8.01e − 03	**4.84e − 03**
	1000	**2.25e − 03**	2.47e − 03	2.78e − 03

Table 3.2: Empirical MISE (example (3.5.2)). The smallest values are indicated in bold.

Example 3.5.3. *In this example, we illustrate the results of Theorem 3.4.1 concerning the stability of our estimator* $\widehat{\beta}_{n,N}(\cdot)$ *for the slope function, associated with the LFR problem. Also, we illustrate the performance of our more general LFR estimator, given by (3.4.14) and (3.4.15). For this purpose, we use the following simulation test initially given in Hall and Horowitz (2007). The interval* $J = [0, 1]$

and the slope function is given by

$$\beta_0(t) = \sum_{j=1}^{50} 4\frac{(-1)^{j+1}}{j^2}\varphi_j(t), \quad \varphi_j(t) = \begin{cases} 1 & \text{if } j = 1 \\ \sqrt{2}\cos(\pi(j-1)t) & \text{if } j \geqslant 2. \end{cases}$$

The random predictor functional is given by $X(\cdot) = \sum_{k=1}^{50} \xi_k Z_k \varphi_k(\cdot)$, $\xi_k = \frac{(-1)^{k+1}}{k^{s/2}}$, $s \geqslant 0$, *where the* Z_k *are independent samples following the uniform law* $U(-\sqrt{3}, \sqrt{3})$. *To illustrate the result of Theorem 3.4.1, we have computed the average over* 10 *realizations of the condition number* $\kappa_2(G_N)$ *with the different values of* $s = 0.25, 0.5, 1.0$, $N = 20, 50$ *and* $n = 100, 200, 300$. *Also, we have computed the* $2-$*condition number of the positive definite matrix* $\mathcal{G}_M$, *given by (3.4.15), associated with a transformation matrix* $T_{N,M}$, *corresponding to the two orthonormal sets* $\{\varphi_1 = 1, \varphi_k(t) = \sqrt{2}\cos(\pi(k-1)t),\ 2 \leqslant k \leqslant N\}$ *and* $\{\psi_j(t) = \sqrt{2}\widetilde{C}_{j-1}^{1/2}(2t-1),\ 1 \leqslant j \leqslant M\}$ *with* $N = 50$ *and* $M = 5$. *The obtained numerical results are given by Table 3.3. The numerical values given by Table 3.3 are coherent with the theoretical estimates given by Theorem 3.4.1 for* $\kappa_2(G_N)$ *and with remark 3.4.3 for* $\kappa_2(\mathcal{G}_M)$.

s	N	n	$\kappa_2(G_N)$	s	N	n	$\kappa_2(G_N)$	$\kappa_2(\mathcal{G}_5)$
0.25	20	100	6.18	0.25	50	100	35.84	1.51
–		200	4.09	–		200	10.29	1.39
–		300	3.46	–		300	6.92	1.33
0.5		100	11.11	0.5		100	66.67	1.72
–		200	7.29	–		200	20.92	1.65
–		300	6.12	–		300	14.56	1.61
1.0		100	35.09	1.0		100	300.78	6.56
–		200	29.10	–		200	114.08	6.01
–		300	25.02	–		300	88.48	5.90

Table 3.3: The 2−norm condition numbers of the matrices G_N and $\mathcal{G}_M$, $M = 5$.

Moreover, to illustrate the prediction and estimation errors of our estimator $\widehat{\beta}_{n,N}$ of the slope function $\beta_0(\cdot)$, we have considered the model (3.4.1) with $N = 50$ and with white noise variance $\sigma_\varepsilon = 0.5$. Then, we have considered the different values of $s = 0.25, 1.1, 1.5, 2.0$. We have computed our estimator $\widehat{\beta}_{n,N}(\cdot)$ given by (3.4.5) for $s = 0.25$ and different values of n. For the larger values of s, we have computed $\widehat{\beta}_{n,N}(\cdot)$ by the decomposition technique given by (3.4.12) and (3.4.13) with $p = 3$ and

the partition of $[[1, 50]]$, given by $[[1, 5]], [[6, 15]], [[16, 50]]$. We have computed the average of the prediction and estimation error over 10 realizations. The obtained prediction and estimation errors for the different values of n are given by Table 3.4.

By comparing our results with those obtained in Yuang and Cai (2010) (Figures 1. and 2.) or some of those given Shin and Lee (2016) for this same test example, we conclude that our estimators given by (3.4.5) (for small values of s) or by (3.4.13) (for larger values of s) outperforms the RKHS penalized estimators. More precisely, even when these last estimators are applied with the optimal choices of the the penalization parameter λ and loss function $\rho(\cdot)$, our estimators $\widehat{\beta}_{n,N}(\cdot)$ seem to be more accurate. Besides, they are fast and can handle the LFR problem with very small value as well as with larger values of the exponent s, the decay rate exponent of the eigenvalues of the covariance operator. Also, from the numerical results of Table 3.4, it is interesting to note that our estimators $\widehat{\beta}_{n,N}(\cdot)$ behave like the KRR estimator given in Yuang and Cai (2010) in the case of closely spaced eigenvalues of the covariance operator. In this case, the the prediction and the estimator errors decay with respect to exponent s, see Figure 4. of Yuang and Cai (2010).

Finally, we illustrate the performance of the more general LFR estimator $\widehat{\beta}_{n,M}(\cdot)$, given by (3.4.14) and (3.4.15). For this purpose, we use the following expansion $\widehat{\beta}_{n,M}(\cdot)$ with respect to the first $M = 5$ orthonormal Legendre polynomials over $J = [0, 1]$, that is $\widehat{\beta}_{n,M}(t) = \sum_{j=1}^{5} \hat{d}_j \sqrt{2} \widetilde{C}_{j-1}^{1/2}(2t - 1)$, $t \in J$. Then, we have applied this estimator to the previous LFR problem with fixed $N = 50$ and different values of $100 \leqslant n \leqslant 500$. Table 3.5 lists the averages over 10 realizations of the squared L^2-errors associated with the estimator $\widehat{\beta}_{n,M}(\cdot)$. From these numerical simulations, one concludes that the general estimator $\widehat{\beta}_{n,M}(\cdot)$ is fast, accurate and it is particularly adapted for the LFR problem considered in this example.

s	n	$\|\beta_0 - \widehat{\beta}_{n,N}\|_0^2$	$\|\beta_0 - \widehat{\beta}_{n,N}\|_{L_2}^2$	s	n	$\|\beta_0 - \widehat{\beta}_{n,N}\|_0^2$	$\|\beta_0 - \widehat{\beta}_{n,N}\|_{L_2}^2$
0.25	100	2.89e-1	4.97e-1	1.5	100	1.93e-3	1.19e-2
–	200	8.22e-2	1.76e-1	–	200	1.25e-3	6.00e-3
–	300	5.11e-2	1.09e-1	–	300	9.37e-4	4.80e-3
–	500	3.25e-2	6.65e-2	–	500	2.49e-4	1.81e-3
1.1	100	1.36e-3	6.38e-3	2.0	100	1.68e-3	2.00e-2
–	200	3.11e-4	1.71e-3	–	200	5.16e-3	3.50e-3
–	300	1.32 e-4	4.00e-3	–	300	3.76e-4	4.39e-3
–	500	1.03e-4	3.98e-3	–	500	1.31e-4	1.17e-3

Table 3.4: Prediction and squared L^2-errors associated with the estimator $\widehat{\beta}_{n,N}$ with $N = 50$ and different values of s, n.

s	n	$\|\beta_0 - \widehat{\beta}_{n,M}\|_{L_2}^2$	s	n	$\|\beta_0 - \widehat{\beta}_{n,M}\|_{L_2}^2$	s	n	$\|\beta_0 - \widehat{\beta}_{n,M}\|_{L_2}^2$
0.25	100	2.95e-2	1.1	100	3.45e-2	2	100	5.28e-1
–	200	1.25e-2	–	200	2.54e-2	–	200	3.94e-2
–	300	8.22e-3	–	300	1.25e-2	–	300	2.88e-2
–	500	1.54e-3	–	500	7.32e-3	–	500	1.38e-2

Table 3.5: Squared L^2-errors associated with the estimator $\widehat{\beta}_{n,M}$ with $M = 5$ and different values of s, n.

4 Censored count data regression with missing censoring information

We investigate estimation in Poisson regression model when the count response is right-censored and the censoring indicators are missing at random. We propose several estimators based on the regression calibration, multiple imputation and augmented inverse probability weighting methods. Under appropriate regularity conditions, we prove the consistency of our estimators and we derive their asymptotic distributions. Simulation experiments are carried out to investigate the finite sample behaviour and relative performance of the proposed estimates. These estimates are illustrated on a real data set.

Contents

Ce chapitre fait l'objet d'un article intitulé "Censored count data regression with missing censoring information" (auteurs: Bilel Bousselmi, Jean-François Dupuy et Abderrazek Karoui), paru dans la revue "Electronic Journal of Statistics (2021)".

4.1 Introduction

Poisson regression is a popular tool for modeling the relationship between a count response (such as the number of cases of a specific disease in epidemiology, or the number of insurance claims within a given period of time) and a set of predictors or covariates. Over the past years, Poisson regression has been extended to accommodate censored count data. Although censoring is usually associated to lifetime data analysis, count data can also be censored, the most common type being right-censoring, which occurs when it is only known that the true count is higher than the observed one. For example, consider a study investigating the smoking habits of some population, where people report their number of cigarettes smoked per day. If one possible answer is "20 cigarettes or more", all cigarettes counts greater than 20 are right-censored at 20. Ignoring censoring is known to yield biased estimates and thus, incorrect inferences. Statistical inference in censored Poisson regression and extensions was therefore addressed by several authors ; see, for example, Terza (1985), Caudill and Mixon (1995), Famoye and Wang (2004), Xie and Wei (2007), Mahmoud and Alderiny (2010) for censored generalized Poisson regression, Karlis et al. (2016) for finite mixtures of censored Poisson regressions and Saffari and Adnan (2011), Nguyen and Dupuy (2020) for zero-inflated censored Poisson regression.
Censored models for count data can be conveniently specified by introducing a censoring indicator which is set to 1 if the observed count is not censored and 0 otherwise. In this paper, we consider the situation where the censoring indicator is missing for some sample individuals. In the context of survival analysis, this issue has been considered by several authors. For example, Van Der Lann and McKeague (1998) and Subramanian (2006, 2011) address estimation of the survival function of a random survival time with missing censoring indicators. McKeague and Subramanian (1998), Chen and Cai (2018) and Brownstein et al. (2021) consider estimation in the proportional hazards and additive hazard regression models with missing censoring indicators. Wang and Shen (2008), Wang et al. (2012) and Brunel et al. (2013) propose various nonparametric estimates of the hazard and conditional hazard functions with censoring indicators missing at random. Wang and Dinse (2011) and Zou et al. (2020) estimate the linear regression and partially linear single-index models for survival data with missing censoring indicators. A similar issue arises in competing risks data analysis with missing cause of failure, see for example Bakoyannis et al. (2010, 2020); Zheng et al. (2016); Nevo et al. (2018).

Estimation in censored Poisson regression with missing censoring information is still an open problem. Our aim in this paper is to provide and compare several estimates adapted to this setting.

Missing data problems have given rise to a rich literature and several adapted estimation methods have been proposed. A common and simple approach, called complete-case analysis, is to exclude individuals with missing data. This method can induce bias and substantial variance increase. Two alternatives are regression calibration and multiple imputation. Regression calibration is a general

method for handling missing or mismeasured variables. It consists in replacing missing data by their conditional expectation given observed data. We refer to Carroll et al. (2006) for a detailed account of this method, which has been used in a variety of contexts, including the linear regression model (Huang, 2005), proportional hazards regression model for survival data (Wang et al., 1997; Dupuy and Leconte, 2009; Liao et al., 2011), generalized linear models (Hardin et al., 2003; Weller et al., 2007). In multiple imputation, missing data are replaced by data generated from an imputation model. This imputation is repeated M times, generating M completed data sets. Each of them is analysed and an overall estimator is obtained by combining the estimates of the M completed samples. Multiple imputation was also used in a number of settings, including linear regression (Horton and Lipsitz, 2001), generalized linear models (Ibrahim et al., 2005), proportional hazards regression (White and Royston, 2009; Hsu and Yu, 2009) and count data models (Kleinke and Reinecke, 2013). Both regression calibration and multiple imputation require a model for the missing data given the observed data. Inverse probability weighting constitutes another alternative method for dealing with missing data (see for example Seaman and White (2013) for a review of this method). Similarly to the complete-case analysis, inverse probability weighting only uses complete cases, but weights are used to rebalance the set of complete cases. Calculating these weights requires a model for the probability that an individual has complete data. Augmented inverse probability weighting was then proposed to ensure robustness against misspecification of the missingness model (see, for example, Tsiatis (2007) for a detailed account on the method).

In this paper, we investigate, both theoretically and numerically, the regression calibration, multiple imputation and augmented inverse probability weighting estimators of the regression parameter in the censored Poisson regression model with missing censoring indicators. Our analysis of these estimates will be based on parametric assumptions for the conditional models for missing data and the missingness mechanism. The plan of the paper is as follows. In Section 4.2, we describe the model setup and we introduce the notations that will be used throughout the paper. In Section 4.3, we introduce our regression calibration estimator and we establish its consistency and asymptotic normality. In Sections 4.4 and 4.5, we propose our multiple imputation and augmented inverse probability weighted estimators, and we derive their asymptotic properties. All our theoretical derivations are based on an incomplete gamma function formulation of the distribution function of the Poisson regression model. Consistent asymptotic variance estimates are also proposed for the regression calibration, multiple imputation and augmented inverse probability weighted estimators. In Section 4.6, we conduct a simulation study to assess the finite sample performance and robustness to parametric assumptions of the proposed estimates. We also illustrate the proposed estimates on a real data set. Discussion and perspectives are given in Section 4.7. All proofs are deferred to appendices.

4.2 Model, data, notations

Let Y denote the count of interest and $\mathbf{X} = (1, X_2, \ldots, X_p)^\top$ be a p-vector of covariates ($\top$ denotes the transpose operator). We assume that the conditional distribution of Y given $\mathbf{X}$ is given by a Poisson regression model with parameter $\lambda = \exp(\beta^\top \mathbf{X})$, where $\beta \in \mathbb{R}^p$ is a vector of unknown parameters.

We consider the situation where Y can be right-censored, that is, instead of the true Y, we eventually observe a value which is smaller than Y. This can be formalised by introducing a finite random variable C such that we observe either Y if $Y < C$ or C if $Y \geqslant C$, and an indicator δ (called censoring indicator thereafter) which is equal to 1 if $Y < C$ and 0 if $Y \geqslant C$. In what follows, we assume that Y and C are independent conditionally on $\mathbf{X}$ and that the distribution of C does not depend on β. These conditions are reminiscent of survival analysis, where they are called the independent censoring and non-informative censoring hypotheses respectively. These hypotheses are reasonable if censoring is due to an external event (i.e., the value of C is not directly driven by the value of Y) or is fixed by design. Otherwise, one needs to model the joint distribution of (Y, C). We denote by Y^* the observed count value (that is, $Y^* = \min(Y, C)$).

Assume that n independent individuals are available and that for each of them, we observe the triplet $(Y_i^*, \mathbf{X}_i, \delta_i)$ (with $i \in \{1, \ldots, n\}$). Under the above hypotheses, the likelihood of β is calculated as:

$$L_n(\beta) = \prod_{i=1}^{n} \mathbb{P}(Y_i = Y_i^* | \mathbf{X}_i)^{\delta_i} \mathbb{P}(Y_i \geqslant Y_i^* | \mathbf{X}_i)^{1-\delta_i},$$

from which we easily deduce the loglikelihood $\ell_n(\beta) = \log L_n(\beta)$:

$$\ell_n(\beta) = \sum_{i=1}^{n} \left\{ \delta_i \left(Y_i^* \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Y_i^*!) \right) + (1 - \delta_i) \log \left(1 - \sum_{k=0}^{Y_i^* - 1} \frac{e^{-\exp(\beta^\top \mathbf{X}_i) + k\beta^\top \mathbf{X}_i}}{k!} \right) \right\} \quad (4.2.1)$$

By standard asymptotic theory, the maximum likelihood estimator $\hat{\beta}_n = \arg\max_\beta \ell_n(\beta)$ is consistent and asymptotically normal with variance $-\mathbb{E}[\partial^2 \ell_1(\beta)/\partial\beta\partial\beta^\top]$.

Remark 4.2.1. *The censoring variable C does not have to be a count. For example, if $Y_i = 4$ and $C_i = 2.5$, then $Y_i^* = 2.5$ (and $\delta_i = 0$) and the contribution of the observation $(Y_i^*, \delta_i) = (2.5, 0)$ to the likelihood is $\mathbb{P}(Y_i \geqslant 2.5 | \mathbf{X}_i)$. But Y_i is discrete, hence $\mathbb{P}(Y_i \geqslant 2.5 | \mathbf{X}_i) = \mathbb{P}(Y_i \geqslant 3 | \mathbf{X}_i)$ and (2.1) is still valid if we write the contribution of a censored observation as $\mathbb{P}(Y_i \geqslant \lceil Y_i^* \rceil | \mathbf{X}_i)$, where $\lceil Y_i^* \rceil$ denotes the smallest integer not less than Y_i^*. If C is continuous, contributions of uncensored observations are unchanged since these observations are integer values.*
Overall, the loglikelihood (4.2.1) remains valid when C is continuous, with the appropriate change of

notation $\mathbb{P}(Y_i \geqslant \lceil Y_i^* \rceil | \mathbf{X}_i)$ *for censored observations. In order to keep notations simple, and without loss of generality, we assume that* C *is discrete (as is also the case in most applications).*

Now, we consider the situation where some additional uncertainty can arise in the observations. Precisely, we consider the situation where the censoring indicator δ_i is missing for some individuals. Let ξ be a missingness indicator, that is, $\xi = 1$ if δ is observed and $\xi = 0$ otherwise. Then, for individual $i \in \{1, \ldots, n\}$, the observed data are

$$(Y_i^*, \mathbf{X}_i, \delta_i, \xi_i = 1) \quad \text{or} \quad (Y_i^*, \mathbf{X}_i, \xi_i = 0). \tag{4.2.2}$$

We consider a missing at random (MAR) mechanism, which means that ξ and δ are independent given all other observed variables (a more restrictive assumption is that ξ and δ are independent, which is called "missing completely at random"). In the next sections, we propose, investigate and compare several estimators of β in this context.

4.3 Regression calibration estimation

4.3.1 The proposed estimator

Our first estimator is based on the regression calibration idea. It consists in replacing any missing δ_i in (4.2.1) by its conditional expectation $\mathbb{E}(\delta_i|\mathbf{W}_i)$, where $\mathbf{W}_i$ contains the observed variables Y_i^* and $\mathbf{X}_i$ and eventually (if available) some observed surrogate variables $\mathbf{V}_i$ for δ_i. Thus, we let $\mathbf{W}_i = (Y_i^*, \mathbf{X}_i^\top, \mathbf{V}_i^\top)^\top$ (we denote by q the dimension of $\mathbf{W}_i$). An approximated version of δ_i can then be defined as:

$$\hat{\delta}_i = \xi_i \delta_i + (1 - \xi_i)\mathbb{E}(\delta_i|\mathbf{W}_i).$$

The conditional expectation $\mathbb{E}(\delta_i|\mathbf{W}_i)$ (or conditional probability $\mathbb{P}(\delta_i = 1|\mathbf{W}_i)$) will generally be unknown and will have to be estimated. As is usual with the regression calibration approach, we assume that $\mathbb{E}(\delta_i|\mathbf{W}_i)$ can be specified by a parametric model $m(\mathbf{W}_i, \theta)$, where θ is an unknown q-dimensional parameter with true value θ_0.

Remark 4.3.1. *A convenient candidate for* $m(\cdot, \cdot)$ *is the logistic regression model* $m(\mathbf{W}_i, \theta) = logit^{-1}(\theta^\top \mathbf{W}_i)$ *but other choices, such as the probit, are possible. One may also allow for polynomial, spline and interaction terms in these models, in order to make them as flexible as desired. In what follows, we assume a general model* $m(\mathbf{W}_i, \theta)$ *with some regularity conditions stated in section 4.3.2.*

At a first stage, we estimate θ_0 by maximizing a likelihood based on complete cases $i \in \{1, \ldots, n | \xi_i =$

1} only:

$$\hat{\theta}_n = \arg\max_{\theta} \prod_{i=1}^{n} m(\mathbf{W}_i, \theta)^{\xi_i \delta_i} (1 - m(\mathbf{W}_i, \theta))^{\xi_i (1-\delta_i)}. \tag{4.3.1}$$

Let

$$\dot{m}(\mathbf{W}_i, \theta) = \frac{\partial m(\mathbf{W}_i, \theta)}{\partial \theta}, \quad \widetilde{m}_i(\theta) = \frac{\dot{m}(\mathbf{W}_i, \theta)}{m(\mathbf{W}_i, \theta)(1 - m(\mathbf{W}_i, \theta))},$$

and

$$\Theta(\theta) = \mathbb{E}\left[\frac{\dot{m}^{\otimes 2}(\mathbf{W}, \theta)}{m(\mathbf{W}, \theta)(1 - m(\mathbf{W}, \theta))} \xi \right],$$

where for any column vector u, $u^{\otimes 2} = uu^\top$. Then it is rather straightforward to see that $\hat{\theta}_n$ is asymptotically linear with influence function $\Theta^{-1}(\theta_0)\widetilde{m}_i(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i, \theta_0))$, that is :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Theta^{-1}(\theta_0)\widetilde{m}_i(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i, \theta_0)) + o_{\mathbb{P}}(1). \tag{4.3.2}$$

Finally, it will be useful to note that if Y is distributed as Poisson with parameter λ, then for any $u \in \mathbb{N}$, $\mathbb{P}(Y \leqslant u) = \sum_{k=0}^{u} \exp(-\lambda)\lambda^k/k! = \Gamma(u+1, \lambda)/u!$ where $\Gamma(u, \lambda) = \int_\lambda^\infty t^{u-1} \exp(-t)dt$ is the incomplete gamma function, whose derivative with respect to λ is given by $\partial\Gamma(u, \lambda)/\partial\lambda = -\exp(-\lambda)\lambda^{u-1}$.

Now, letting $\hat{\delta}_i(\theta) = \xi_i\delta_i + (1 - \xi_i)m(\mathbf{W}_i, \theta)$ be the approximation of δ_i based on model $m(\mathbf{W}_i, \theta)$, we define our regression calibration estimator of β as

$$\tilde{\beta}_n = \arg\max_{\beta} \widetilde{\ell}_n(\beta, \hat{\theta}_n),$$

where

$$\widetilde{\ell}_n(\beta, \hat{\theta}_n) = \sum_{i=1}^{n} \left\{ \hat{\delta}_i(\hat{\theta}_n) \left(Y_i^* \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Y_i^*!) \right) + (1 - \hat{\delta}_i(\hat{\theta}_n)) \log\left(1 - \frac{\Gamma\left(Y_i^*, e^{\beta^\top \mathbf{X}_i}\right)}{(Y_i^* - 1)!} \right) \right\}$$

is an approximated version of (4.2.1).

4.3.2 Regularity conditions and asymptotic results

The following regularity conditions are needed to establish the asymptotic properties of the regression calibration estimator. We assume:

C1 The covariates vectors $\mathbf{X}_i$ and $\mathbf{V}_i$ are bounded, for every $i = 1, 2, \ldots$

C2 The true parameter values β_0 and θ_0 lie in the interior of some bounded sets $\mathcal{B} \subset \mathbb{R}^p$ and $\Theta \subset \mathbb{R}^q$ respectively.

C3 We have $\mathbb{P}(Y^* \geqslant 1 | \xi\delta = 0) = 1$ and $\mathbb{P}(\delta = 1) > 0$.

C4 The function $m(\mathbf{w}, \theta)$ is differentiable with respect to θ, for every $\mathbf{w}$. For every $\theta, \tilde{\theta} \in \Theta$, $|m(\mathbf{w}, \theta) - m(\mathbf{w}, \tilde{\theta})| \leqslant h(\mathbf{w})\|\theta - \tilde{\theta}\|$ for some bounded function h, with $\mathbb{E}[h(\mathbf{W})] = v$.

Remark 4.3.2. *Condition C3 requires that a minimum amount of information is available on the count response when it is either censored ($\delta = 0$) or its censoring status is unknown ($\xi = 0$). Intuitively, the observation $\{Y^* = 0\}$ carries no information if it is unknown that $\delta = 1$ (i.e., that it is a "genuine" zero count), since all counts are non-negative.*

Before stating the asymptotics of $\tilde{\beta}_n$, we introduce some further notations. Let h_β be the function defined by:

$$h_\beta(y, x) = \frac{e^{-e^{\beta^\top x} + \beta^\top xy}}{(y-1)! - \Gamma(y, e^{\beta^\top x})} \tag{4.3.3}$$

for any $\beta \in \mathbb{R}^p$, $x \in \mathbb{R}^p$ and $y \in \mathbb{N}\backslash\{0\}$. Let also $\pi(\mathbf{W}) = \mathbb{P}(\xi = 1 | \mathbf{W})$ and define the matrices

$$\begin{aligned}
\Sigma_1(\beta) &= \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top \left(\delta e^{\beta^\top \mathbf{X}} + (\delta - 1)\left\{Y^* - e^{\beta^\top \mathbf{X}} - h_\beta(Y^*, \mathbf{X})\right\} h_\beta(Y^*, \mathbf{X})\right)\right], \\
\Sigma_2(\beta, \theta) &= \mathbb{E}\left[\mathbf{X}\dot{m}^\top(\mathbf{W}, \theta)\left(Y^* - e^{\beta^\top \mathbf{X}} - h_\beta(Y^*, \mathbf{X})\right)(1 - \pi(\mathbf{W}))\right], \\
\Sigma_3(\beta, \theta) &= \mathbb{E}\left[\mathbf{X}\dot{m}^\top(\mathbf{W}, \theta)\left(Y^* - e^{\beta^\top \mathbf{X}} - h_\beta(Y^*, \mathbf{X})\right)\right].
\end{aligned}$$

We are now in position to state our first theorem. The proof is given in Appendix A.

Theorem 4.3.3. *Assume that conditions C1-C4 hold. Then $\tilde{\beta}_n \xrightarrow{\mathbb{P}} \beta_0$ as $n \to \infty$ and $\sqrt{n}(\tilde{\beta}_n - \beta_0)$ is asymptotically normal with mean zero and variance Σ, where*

$$\Sigma = \Sigma_1^{-1}(\beta_0)\left\{\Sigma_1(\beta_0) + (2\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0))\,\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0)\right\}\Sigma_1^{-1}(\beta_0).$$

Remark 4.3.4. *If $\pi(\mathbf{W})$ is identically equal to 1 (that is, if there is no missing data), Σ reduces to the asymptotic variance of the maximum likelihood estimator $\hat{\beta}_n$ in (4.2.1), which in turn reduces to the usual asymptotic variance $(\mathbb{E}[\mathbf{X}\mathbf{X}^\top e^{\beta_0^\top \mathbf{X}}])^{-1}$ in Poisson regression if $m(\mathbf{W}, \theta_0)$ is identically equal to 1 (that is, no censoring can affect the data).*

A consistent estimator of Σ is given by

$$\Sigma_n = \Sigma_{1,n}^{-1}(\tilde{\beta}_n, \hat{\theta}_n) \left\{ \Sigma_{1,n}(\tilde{\beta}_n, \hat{\theta}_n) + \left(2\Sigma_{3,n}(\tilde{\beta}_n, \hat{\theta}_n) - \Sigma_{2,n}(\tilde{\beta}_n, \hat{\theta}_n) \right) \Theta_n^{-1}(\hat{\theta}_n) \Sigma_{2,n}^{\top}(\tilde{\beta}_n, \hat{\theta}_n) \right\} \Sigma_{1,n}^{-1}(\tilde{\beta}_n, \hat{\theta}_n),$$

where

$$\begin{aligned}
\Sigma_{1,n}(\beta, \theta) &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^{\top} \left(\hat{\delta}_i(\theta) e^{\beta^{\top} \mathbf{X}_i} + (\hat{\delta}_i(\theta) - 1) \left\{ Y_i^* - e^{\beta^{\top} \mathbf{X}_i} - h_{\beta}(Y_i^*, \mathbf{X}_i) \right\} h_{\beta}(Y_i^*, \mathbf{X}_i) \right), \\
\Sigma_{2,n}(\beta, \theta) &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \dot{m}^{\top}(\mathbf{W}_i, \theta) \left(Y_i^* - e^{\beta^{\top} \mathbf{X}_i} - h_{\beta}(Y_i^*, \mathbf{X}_i) \right) (1 - \xi_i), \\
\Sigma_{3,n}(\beta, \theta) &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \dot{m}^{\top}(\mathbf{W}_i, \theta) \left(Y_i^* - e^{\beta^{\top} \mathbf{X}_i} - h_{\beta}(Y_i^*, \mathbf{X}_i) \right), \\
\Theta_n(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \frac{\dot{m}^{\otimes 2}(\mathbf{W}_i, \theta)}{m(\mathbf{W}_i, \theta)(1 - m(\mathbf{W}_i, \theta))} \xi_i.
\end{aligned}$$

The consistency proof of the variance estimator uses similar arguments as the proof of consistency of $\tilde{\beta}_n$, it is thus omitted. The estimator $\tilde{\beta}_n$ will be evaluated in the simulation study of Section 4.6.

Several methods have been proposed to address missing data problems in regression. Among them is the multiple imputation, which provides an alternative, popular and widely-used approach. The basic idea is to create several (say M) completed data sets, by filling in plausible values for the missing data. Then, each filled sample is analysed as if it were the complete data set. Finally, the M imputed-samples inferences are combined into a single overall inference. In the next section, we investigate this approach for estimating β in our problem.

4.4 Multiple imputation

In this section, we assume, as in Section 4.3, that the conditional expectation $\mathbb{E}(\delta_i|\mathbf{W}_i)$ can be specified by a parametric model $m(\mathbf{W}_i, \theta_0)$, and we denote by $\hat{\theta}_n$ the maximum likelihood estimator of θ_0 based on the complete cases $i \in \{1, \ldots, n | \xi_i = 1\}$.

The imputation procedure is as follows. Each missing δ_i is replaced by a random draw from the Bernoulli distribution $\mathcal{B}(m(\mathbf{W}_i, \hat{\theta}_n))$. We obtain a completed data set. This procedure is repeated M times to form M imputed data sets. For a given θ, let $D_{i,j}(\theta) \sim \mathcal{B}(m(\mathbf{W}_i, \theta))$ denote the imputation of δ_i in the j-th completed data set ($j = 1, \ldots, M$). Let also

$$\delta_{i,j}^*(\theta) = \xi_i \delta_i + (1 - \xi_i) D_{i,j}(\theta)$$

be the random variable which is equal to δ_i is $\xi_i = 1$ (that is, if δ_i is observed) and to $D_{i,j}(\theta)$ if $\xi_i = 0$ (that is, if δ_i is missing) (note the difference between the imputation method, where $\delta^*_{i,j}(\theta) \in \{0,1\}$, and the regression calibration approach, where $\hat{\delta}_i(\theta) \in [0,1]$). A single-imputation estimator $\hat{\beta}^*_{n,j}$ of β_0 is obtained by maximizing the imputed log-likelihood

$$\ell^*_{n,j}(\beta, \hat{\theta}_n) = \sum_{i=1}^{n} \left\{ \delta^*_{i,j}(\hat{\theta}_n) \left(Y_i^* \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Y_i^*!) \right) + (1 - \delta^*_{i,j}(\hat{\theta}_n)) \log \left(1 - \frac{\Gamma\left(Y_i^*, e^{\beta^\top \mathbf{X}_i}\right)}{(Y_i^* - 1)!} \right) \right\}.$$

The final multiple imputation estimator $\hat{\beta}^*_n$ is obtained by averaging the M estimators $\hat{\beta}^*_{n,j}$ as:

$$\hat{\beta}^*_n = \frac{1}{M} \sum_{j=1}^{M} \hat{\beta}^*_{n,j}.$$

The next theorem gives the asymptotic properties of $\hat{\beta}^*_n$. Its proof is given in Appendix B.

Theorem 4.4.1. *For $j = 1, \ldots, M$, let $f_{\beta,\theta,j}(\mathcal{O}_i) = \mathbf{X}_i\{\delta^*_{i,j}(\theta)[Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i)] + h_\beta(Y_i^*, \mathbf{X}_i)\}$, where $\mathcal{O}_i$ denotes the observation (4.2.2). Let also $\Sigma_1^*(\beta,\theta) = var(\frac{1}{M}\sum_{j=1}^{M} f_{\beta,\theta,j}(\mathcal{O}_1))$. If conditions C1-C4 hold, then $\hat{\beta}^*_n \xrightarrow{\mathbb{P}} \beta_0$ as $n \to \infty$ and $\sqrt{n}(\hat{\beta}^*_n - \beta_0)$ is asymptotically normal with mean zero and variance Σ^*, where*

$$\Sigma^* = \Sigma_1^{-1}(\beta_0) \left\{ \Sigma_1^*(\beta_0, \theta_0) + (2\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)) \, \Theta^{-1}(\theta_0) \Sigma_2^\top(\beta_0, \theta_0) \right\} \Sigma_1^{-1}(\beta_0).$$

A consistent estimator of Σ^* can be obtained as

$$\Sigma^*_n = \bar{\Sigma}^{-1}_{1,n}(\hat{\beta}^*_n, \hat{\theta}_n) \left\{ \Sigma^*_{1,n}(\hat{\beta}^*_n, \hat{\theta}_n) + \left(2\Sigma_{3,n}(\hat{\beta}^*_n, \hat{\theta}_n) - \Sigma_{2,n}(\hat{\beta}^*_n, \hat{\theta}_n) \right) \Theta_n^{-1}(\hat{\theta}_n) \Sigma^\top_{2,n}(\hat{\beta}^*_n, \hat{\theta}_n) \right\} \bar{\Sigma}^{-1}_{1,n}(\hat{\beta}^*_n, \hat{\theta}_n),$$

where $\Sigma^*_{1,n}(\beta,\theta)$ is the empirical covariance of the vectors $\frac{1}{M}\sum_{j=1}^{M} f_{\beta,\theta,j}(\mathcal{O}_i)$ $(i = 1, \ldots, n)$, $\bar{\Sigma}_{1,n}(\beta,\theta)$ is the average $\frac{1}{M}\sum_{j=1}^{M} \Sigma_{1,n,j}(\beta,\theta)$, with

$$\Sigma_{1,n,j}(\beta,\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \left(\delta^*_{i,j}(\theta) e^{\beta^\top \mathbf{X}_i} + (\delta^*_{i,j}(\theta) - 1) \left\{ Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i) \right\} h_\beta(Y_i^*, \mathbf{X}_i) \right),$$

and $\Sigma_{2,n}, \Sigma_{3,n}$ and Θ_n are as given in Section 4.3.

Regression calibration and multiple imputation rely on the ability of the investigator to formulate an appropriate model for $\mathbb{E}(\delta|\mathbf{W})$. Misspecifying this model is likely to yield biased estimates of the parameters of interest. An alternative approach is to specify the selection probabilities $\pi(\mathbf{W}_i) = \mathbb{P}(\xi_i = 1|\mathbf{W}_i)$ and to use the inverse probability weighting (IPW) of complete-case technique of Horvitz and

Thompson (1952). The basic idea of IPW is to adjust a complete-case analysis by weighting individuals with no missing data by the inverse of their selection probability. Selection probabilities are generally unknown and have to be estimated. Again, misspecifying the $\pi(\mathbf{W}_i), i = 1, \ldots, n$ is likely to yield biased inference. Moreover, by discarding individuals with missing data, IPW is also known to yield loss of efficiency.

For these reasons, the augmented IPW approach (AIPW henceforth, see Robins et al., 1994) was proposed to improve the basic IPW. Since its introduction, the method has been shown to be doubly robust in several models, such as the proportional hazards model (Wang and Chen, 2001), the single-index model (Guo et al., 2015), the additive hazards model (Sun et al., 2017) and the accelerated failure time model (Steingrimsson and Strawderman, 2017). Double robustness refers to the fact that the AIPW estimates are consistent as long as either the selection probability model or the conditional expectation of the missing data is correctly specified. In the next section, we propose an augmented IPW estimating equation adapted to our problem, and we investigate the asymptotic properties of the resulting estimator.

4.5 Augmented inverse probability weighted estimation

Inspired by Horvitz and Thompson (1952), the inverse probability weighting of complete cases has become a classical estimation method in missing data problems. One drawback of the method is that the observed variables of subjects with missing data are not fully used, except through the estimation of the unknown selection probabilities. The AIPW method improves IPW by introducing an additional term involving contributions from individuals with some missing data (we refer to Tsiatis (2007) for a detailed account on the method and numerous references). Adapting this idea, we propose the following augmented IPW estimating equation for β:

$$\sum_{i=1}^{n} \mathbf{X}_i \left[\left\{ \frac{\xi_i \delta_i}{\pi(\mathbf{W}_i)} + \left(1 - \frac{\xi_i}{\pi(\mathbf{W}_i)}\right) \mathbb{E}(\delta_i | \mathbf{W}_i) \right\} \left(Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i) \right) + h_\beta(Y_i^*, \mathbf{X}_i) \right].$$

The quantities $\mathbb{E}(\delta_i|\mathbf{W}_i)$ and $\pi(\mathbf{W}_i)$ are unknown and have to be estimated. We assume that they can be specified by some parametric models $m(\mathbf{W}_i, \theta)$ and $\pi(\mathbf{W}_i, \gamma)$ respectively, where θ and γ are unknown q-dimensional parameters with true values θ_0 and γ_0. Let $\hat{\theta}_n$ and $\hat{\gamma}_n$ be the maximum likelihood estimates of θ_0 and γ_0. $\hat{\theta}_n$ is given by (4.3.1). Similarly, $\hat{\gamma}_n$ can be obtained as

$$\hat{\gamma}_n = \arg\max_{\gamma} \prod_{i=1}^{n} \pi(\mathbf{W}_i, \gamma)^{\xi_i} (1 - \pi(\mathbf{W}_i, \gamma))^{1-\xi_i}.$$

Finally, our AIPW estimator $\breve{\beta}_n$ of β solves the estimating equation $\breve{\ell}_n(\beta, \hat{\theta}_n, \hat{\gamma}_n) = 0$, where

$$\breve{\ell}_n(\beta, \hat{\theta}_n, \hat{\gamma}_n) = \sum_{i=1}^{n} \mathbf{X}_i \left[\left\{ \frac{\xi_i \delta_i}{\pi(\mathbf{W}_i, \hat{\gamma}_n)} + \left(1 - \frac{\xi_i}{\pi(\mathbf{W}_i, \hat{\gamma}_n)}\right) m(\mathbf{W}_i, \hat{\theta}_n) \right\} \left(Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i)\right) + h_\beta(Y_i^*, \mathbf{X}_i) \right].$$

Before stating the asymptotic properties of $\breve{\beta}_n$, we introduce some further notations and regularity conditions. For any $\theta, \gamma \in \mathbb{R}^q$, we let

$$\breve{\delta}_i(\theta, \gamma) = \frac{\xi_i \delta_i}{\pi(\mathbf{W}_i, \gamma)} + \left(1 - \frac{\xi_i}{\pi(\mathbf{W}_i, \gamma)}\right) m(\mathbf{W}_i, \theta).$$

Assuming the parametric model $\pi(\mathbf{W}_i, \gamma)$ for the selection probabilities, the maximum likelihood estimator $\hat{\gamma}_n$ is asymptotically linear with influence function $\Sigma_4^{-1}(\gamma_0)\widetilde{\pi}_i(\gamma_0)(\xi_i - \pi(\mathbf{W}_i, \gamma_0))$, where

$$\dot{\pi}(\mathbf{W}_i, \gamma) = \frac{\partial \pi(\mathbf{W}_i, \gamma)}{\partial \gamma}, \quad \widetilde{\pi}_i(\gamma) = \frac{\dot{\pi}(\mathbf{W}_i, \gamma)}{\pi(\mathbf{W}_i, \gamma)(1 - \pi(\mathbf{W}_i, \gamma))},$$

and

$$\Sigma_4(\gamma) = \mathbb{E}\left[\frac{\dot{\pi}^{\otimes 2}(\mathbf{W}, \gamma)}{\pi(\mathbf{W}, \gamma)(1 - \pi(\mathbf{W}, \gamma))} \right].$$

That is :

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma_4^{-1}(\gamma_0)\widetilde{\pi}_i(\gamma_0)(\xi_i - \pi(\mathbf{W}_i, \gamma_0)) + o_{\mathbb{P}}(1). \tag{4.5.1}$$

If the models $m(\mathbf{W}_i, \theta)$ and $\pi(\mathbf{W}_i, \gamma)$ are misspecified, then by White (2012), there exists θ^* and γ^* such that $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$ and $\hat{\gamma}_n \xrightarrow{\mathbb{P}} \gamma^*$. Moreover, the asymptotic linear expansions for $\hat{\theta}_n$ and $\hat{\gamma}_n$ are given by (4.3.2) and (4.5.1), with θ_0 and γ_0 replaced by θ^* and γ^* respectively. If the model $m(\mathbf{W}_i, \theta)$ (respectively $\pi(\mathbf{W}_i, \gamma)$) is correctly specified, then $\theta^* = \theta_0$ (respectively $\gamma^* = \gamma_0$).

Finally, let

$$\begin{aligned}
\Sigma_5(\beta, \theta, \gamma) &= \mathbb{E}\left[\mathbf{X} \left(Y^* - e^{\beta^\top \mathbf{X}} - h_\beta(Y^*, \mathbf{X})\right) \left(1 - \frac{\xi}{\pi(\mathbf{W}, \gamma)}\right) \dot{m}^\top(\mathbf{W}, \theta) \right], \\
\Sigma_6(\beta, \theta, \gamma) &= \mathbb{E}\left[\mathbf{X} \left(Y^* - e^{\beta^\top \mathbf{X}} - h_\beta(Y^*, \mathbf{X})\right) \xi \frac{\dot{\pi}^\top(\mathbf{W}, \gamma)}{\pi^2(\mathbf{W}, \gamma)} (m(\mathbf{W}, \theta) - \delta) \right], \\
\Sigma_7(\beta, \theta, \gamma) &= \Sigma_1(\beta) + (2\Sigma_3(\beta, \theta) - \Sigma_5(\beta, \theta, \gamma))\, \Theta^{-1}(\theta)\Sigma_5^\top(\beta, \theta, \gamma),
\end{aligned}$$

and

$$\Sigma_8(\beta,\theta,\gamma) = \Sigma_1(\beta) - \Sigma_6(\beta,\theta,\gamma)\Sigma_4^{-1}(\gamma)\Sigma_6^{\top}(\beta,\theta,\gamma),$$

We assume the following additional regularity conditions:

C5 The parameter space for γ is a bounded set $\mathcal{G} \subset \mathbb{R}^q$ and the true parameter value γ_0 lies in the interior of $\mathcal{G}$.

C6 The function $\pi(\mathbf{w},\gamma)$ is strictly greater than 0 for all value of $\mathbf{w}$ in the support of $\mathbf{W}$ and all $\gamma \in \mathcal{G}$.

C7 The function $\pi(\mathbf{w},\gamma)$ is differentiable with respect to γ, for every $\mathbf{w}$. For every $\gamma, \tilde{\gamma} \in \mathcal{G}$, $|\pi(\mathbf{w},\gamma) - \pi(\mathbf{w},\tilde{\gamma})| \leqslant g(\mathbf{w})\|\gamma - \tilde{\gamma}\|$ for some bounded function g with $\mathbb{E}[g(\mathbf{W})] = u$.

Conditions C5 and C7 for γ and $\pi(\cdot,\cdot)$ are similar to conditions C2 and C4 for θ and $m(\cdot,\cdot)$. We are now in position to state the asymptotic properties of our AIPW estimator of β.

Theorem 4.5.1. *Assume that conditions C1-C7 hold. If either or both of the models $m(\mathbf{W}_i,\theta)$ and $\pi(\mathbf{W}_i,\gamma)$ are well specified, then $\breve{\beta}_n \xrightarrow{\mathbb{P}} \beta_0$ as $n \to \infty$.*

From this result, the proposed estimator $\breve{\beta}_n$ is doubly robust, in the sense that it estimates consistently β_0 as long as one of $m(\mathbf{W}_i,\theta)$ and $\pi(\mathbf{W}_i,\gamma)$ is correctly modeled.

Remark 4.5.2. *The basic idea of regression calibration and multiple imputation is to replace a missing δ_i by an approximation whose conditional expectation given observed variables is equal to $\mathbb{E}(\delta_i|\mathbf{W}_i)$ (one can check that $\mathbb{E}(\hat{\delta}_i(\theta_0)|\mathbf{W}_i) = \mathbb{E}(\delta^*_{i,j}(\theta_0)|\mathbf{W}_i) = \mathbb{E}(\delta_i|\mathbf{W}_i)$), so that the expectation of the corresponding estimating equations coincide with the expectation of the estimating equation with no missing data. Similarly, one can easily check that if $m(\mathbf{W}_i,\theta)$ (respectively $\pi(\mathbf{W}_i,\gamma)$) is correctly specified, then $\mathbb{E}(\breve{\delta}_i(\theta_0,\gamma^*)|\mathbf{W}_i) = \mathbb{E}(\delta_i|\mathbf{W}_i)$ (respectively $\mathbb{E}(\breve{\delta}_i(\theta^*,\gamma_0)|\mathbf{W}_i) = \mathbb{E}(\delta_i|\mathbf{W}_i)$). Here is the intuition underlying the AIPW method, and the seemingly complicated expression of $\breve{\delta}_i(\theta,\gamma)$.*

The proof of Theorem 4.5.1 is given in Appendix C. The next theorem describes the asymptotic distribution of $\breve{\beta}_n$. Its proof is given in Appendix D.

Theorem 4.5.3. *Assume that conditions C1-C7 hold. Then, as $n \to \infty$, $\sqrt{n}(\breve{\beta}_n - \beta_0)$ converges in distribution to the Gaussian random vector $\mathcal{N}(0,\mathbf{J})$, where*

$$\mathbf{J} = \begin{cases} \Sigma_1^{-1}(\beta_0)\Sigma_7(\beta_0,\theta_0,\gamma^*)\Sigma_1^{-1}(\beta_0) \text{ if } m(\mathbf{W}_i,\theta) \text{ is correctly specified,} \\ \Sigma_1^{-1}(\beta_0)\Sigma_8(\beta_0,\theta^*,\gamma_0)\Sigma_1^{-1}(\beta_0) \text{ if } \pi(\mathbf{W}_i,\gamma) \text{ is correctly specified,} \\ \Sigma_1^{-1}(\beta_0) \text{ if both } m(\mathbf{W}_i,\theta) \text{ and } \pi(\mathbf{W}_i,\gamma) \text{ are correctly specified.} \end{cases}$$

In order to estimate the asymptotic variance of $\breve{\beta}_n$, let:

$$\breve{\Sigma}_{1,n}(\beta,\theta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i^\top \left(\breve{\delta}_i(\theta,\gamma)e^{\beta^\top \mathbf{X}_i} + (\breve{\delta}_i(\theta,\gamma)-1)\left\{Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*,\mathbf{X}_i)\right\} h_\beta(Y_i^*,\mathbf{X}_i)\right),$$

$$\Sigma_{4,n}(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \frac{\dot{\pi}^{\otimes 2}(\mathbf{W}_i,\gamma)}{\pi(\mathbf{W}_i,\gamma)(1-\pi(\mathbf{W}_i,\gamma))},$$

$$\Sigma_{5,n}(\beta,\theta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\left(Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*,\mathbf{X}_i)\right)\left(1-\frac{\xi_i}{\pi(\mathbf{W}_i,\gamma)}\right)\dot{m}^\top(\mathbf{W}_i,\theta),$$

$$\Sigma_{6,n}(\beta,\theta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\left(Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*,\mathbf{X}_i)\right)\xi_i\frac{\dot{\pi}^\top(\mathbf{W}_i,\gamma)}{\pi^2(\mathbf{W}_i,\gamma)}\left(m(\mathbf{W}_i,\theta)-\delta_i\right),$$

$$\Sigma_{7,n}(\beta,\theta,\gamma) = \breve{\Sigma}_{1,n}(\beta,\theta,\gamma) + \left(2\Sigma_{3,n}(\beta,\theta) - \Sigma_{5,n}(\beta,\theta,\gamma)\right)\Theta_n^{-1}(\theta)\Sigma_{5,n}^\top(\beta,\theta,\gamma),$$

and

$$\Sigma_{8,n}(\beta,\theta,\gamma) = \breve{\Sigma}_{1,n}(\beta,\theta,\gamma) - \Sigma_{6,n}(\beta,\theta,\gamma)\Sigma_{4,n}^{-1}(\gamma)\Sigma_{6,n}^\top(\beta,\theta,\gamma),$$

where $\Sigma_{3,n}$ and Θ_n are as given in Section 4.3. Then a consistent estimator of $\mathbf{J}$ is given by:

$$\mathbf{J}_n = \begin{cases} \breve{\Sigma}_{1,n}^{-1}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n)\Sigma_{7,n}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n)\breve{\Sigma}_{1,n}^{-1}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n) \text{ if } m(\mathbf{W}_i,\theta) \text{ is correctly specified,} \\ \breve{\Sigma}_{1,n}^{-1}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n)\Sigma_{8,n}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n)\breve{\Sigma}_{1,n}^{-1}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n) \text{ if } \pi(\mathbf{W}_i,\gamma) \text{ is correctly specified,} \\ \breve{\Sigma}_{1,n}^{-1}(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n) \text{ if both } m(\mathbf{W}_i,\theta) \text{ and } \pi(\mathbf{W}_i,\gamma) \text{ are correctly specified.} \end{cases} \quad (4.5.2)$$

The proof of consistency of $\mathbf{J}_n$ is omitted.

4.6 Numerical results

4.6.1 A simulation study

4.6.1.1 Simulation design

In this section, we investigate the finite sample performance of the regression calibration (RC), multiple imputation (MI) and AIPW estimators. The simulation design is as follows. For each of n individuals, the count response Y is simulated from a Poisson regression model with parameter

$$\lambda = \exp(\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5),$$

where $\beta = (0.2, -0.1, 0.4, 0.3, 0.5)$, $X_2 \sim \mathcal{N}(0,1)$, $X_3 \sim$ Bernoulli(0.3), $X_4 \sim \mathcal{N}(0, 1.5)$ and $X_5 \sim$ uniform$[2,5]$. The censoring and missingness mechanisms are set to be $\text{logit}(m(\mathbf{W},\theta)) = \theta_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_4 + \theta_5 X_5 + \theta_6 Y$ and $\text{logit}(\pi(\mathbf{W},\gamma)) = \gamma_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 Y^*$ respectively, where θ and γ are chosen to yield the desired fractions of censored and missing data. We run a series of experiments in order to assess the effect of the sample size, censoring rate (CR) and missing rate (MR) on estimation:

- experiment 1: we take $n = 250$, CR $= 20\%$, MR $= 20\%$,
- experiment 2: we take $n = 500$, CR $= 20\%$, MR $= 20\%$,
- experiment 3: we take $n = 500$, CR $= 20\%$, MR $= 40\%$,
- experiment 4: we take $n = 500$, CR $= 40\%$, MR $= 20\%$,

We can assess the effect of the sample size by comparing results of experiments 1 and 2. Similarly, by comparing experiments 2 and 3 (respectively 2 and 4), we can assess the effect of the missing rate (respectively censoring rate). Within each experiment, we compare the RC, MI and AIPW estimates under three scenario : (i) only $m(\mathbf{W},\theta)$ is correctly modeled, (ii) only $\pi(\mathbf{W},\gamma)$ is correctly modeled, (iii) both $m(\mathbf{W},\theta)$ and $\pi(\mathbf{W},\gamma)$ are correctly modeled. In the first scenario, $\pi(\mathbf{W},\gamma)$ is incorrectly modeled as $\text{logit}(\pi(\mathbf{W},\gamma)) = \gamma_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 Y^*$. In the second scenario, $m(\mathbf{W},\theta)$ is incorrectly modeled as $\text{logit}(m(\mathbf{W},\theta)) = \theta_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 Y^*$.

Our simulation results are based on $N = 1000$ simulated samples. For each estimator, we report the average bias, average standard error (SE), empirical root mean square error (RMSE) and empirical coverage probability (CP) of 95%-level confidence intervals. MI estimates are obtained with $M = 50$ (from our numerical experiments, this is large enough to ensure stability of the estimates). To establish a benchmark for comparisons, we also include an estimator based on the full data set with no missing censoring indicators and the complete-case (CC) estimator which maximizes the log-likelihood (4.2.1) on the subsample of complete cases only. Results of experiment j are summarized in Table j, for $j = 1, \ldots, 4$. All estimates are obtained using the Newton-Raphson algorithm, implemented in `R` (R Core Team, 2020).

Remark 4.6.1. *The EM algorithm is a popular tool for calculating maximum likelihood estimates in missing data problems. In the context of Poisson regression with missing data, it has been used by several authors. For example, Faria and Soromenho (2012), Bermúdez and Karlis (2012) and Karlis et al. (2016) use EM in finite mixtures of Poisson, bivariate Poisson and censored Poisson regression models. In these works, the EM algorithm is motivated by the missing data formulation of mixture models, where the unknown mixture component indicator is treated as the missing data. EM was*

also used in zero-inflated Poisson regression (Hall, 2000). Here, the missing data is the unobserved state variable (zero state vs Poisson state). Adamids and Loukas (1994) use EM in bivariate Poisson regression with missing outcome. The EM algorithm could also be used in our setting, and it would be interesting to investigate the convergence rate of the sequence of EM estimates. This, however, falls beyond the scope of our paper and constitutes a topic for future work.

4.6.1.2 Results

As expected, the performance of the estimators improve when sample size increases. In the first scenario of each experiment, the RC, MI and AIPW methods appear to have similar performance. Coverage probabilities are close to the nominal confidence level, indicating that the asymptotic variances are appropriately estimated.

In the second scenario, the AIPW method generally achieves the smallest SE and RMSE, while the bias of the RC and MI estimates increase substantially, resulting in coverage probabilities smaller than desired (this is particularly noticeable when the censoring rate is large, see Table 4.4). This result was expected since $m(\mathbf{W}, \theta)$ is misspecified. On the other hand, when censoring is moderate (Table 4.1-Table 4.3), the bias of the AIPW estimate stays moderate and of the same order of magnitude (but generally slightly larger, see our explanation below) as in the first scenario, which is also expected due to the double robustness property stated in Theorem 4.5.1. When the censoring rate is high, the bias of the AIPW estimate is more important and coverage probabilities can be affected (but less than for the RC and MI methods). This suggests that in finite samples, the AIPW estimator is more sensitive to a misspecification of $m(\mathbf{W}, \theta)$ than of $\pi(\mathbf{W}, \gamma)$. This can be explained by the expression of $\breve{\delta}_i(\theta, \gamma)$, which is equal to $m(\mathbf{W}_i, \theta)$ if $\xi_i = 0$ and to $m(\mathbf{W}_i, \theta) + \frac{(\delta_i - m(\mathbf{W}_i, \theta)}{\pi(\mathbf{W}_i, \gamma)}$ if $\xi_i = 1$. Indeed, when $m(\mathbf{W}, \theta)$ is wrong, every individual i contributes to the likelihood with a misspecified term, whatever ξ_i is. On the other hand, when $\pi(\mathbf{W}, \gamma)$ is wrong, only individuals with $\xi_i = 1$ contribute to the likelihood with a misspecified term, since $\pi(\mathbf{W}, \gamma)$ does not appear in the contribution of individuals with $\xi_i = 0$. This unbalance may explain the greater sensitivity of the AIPW estimate to a misspecification of $m(\mathbf{W}, \theta)$. Finally, when both models are correct (third scenario), all three methods perform similarly (results for the RC and MI methods are the same as for the first scenario).

Overall, this simulation study confirms the theoretical results stated in the previous sections. The regression calibration, multiple imputation and robust IPW methods provide similar results when either $m(\mathbf{W}, \theta)$ or both $m(\mathbf{W}, \theta)$ and $\pi(\mathbf{W}, \gamma)$ are correctly specified. When $m(\mathbf{W}, \theta)$ is misspecified, the AIPW approach performs better than RC and MI, in particular in terms of point estimation (with substantially smaller bias for AIPW). The CC estimates are outperformed by the three methods in all scenarios.

4.6.1.3 Asymptotic variance estimation

Estimation of Σ, Σ^* and $\mathbf{J}$ (the asymptotic variances of the RC, MI and AIPW estimates respectively) is a crucial issue for statistical inference purpose. In this section, we investigate the accuracy of their respective estimates Σ_n, Σ_n^* and $\mathbf{J}_n$.

First, note that although Σ, Σ^* and $\mathbf{J}$ have explicit expressions, they cannot be calculated analytically, due to the complex expectations involved in the Σ_j, $j = 1, \ldots, 8$. Therefore, we propose to compare Σ_n (respectively $\Sigma_n^*, \mathbf{J}_n$) to some "oracle" estimate Σ^{or} (respectively $\Sigma^{*,or}, \mathbf{J}^{or}$), which is obtained as follows: we simulate a very large number (here, 15000) of observations $(Y_i^*, \mathbf{X}_i, \delta_i, \xi_i)$, and we calculate empirical versions of the Σ_j where all expectations are replaced by sample averages and parameters are fixed at their true value (hence the name oracle). We expect these oracles to be as close as possible of the true unknown asymptotic variances. Comparisons between $\Sigma_n, \Sigma_n^*, \mathbf{J}_n$ and the oracles are based on the results of the above simulation study.

For each experiment and each of the RC, MI and AIPW method, we calculate the relative differences $100 \times |\Sigma_{n,(j,j)}^{1/2} - (\Sigma_{(j,j)}^{or})^{1/2}|/(\Sigma_{(j,j)}^{or})^{1/2}$, $100 \times |(\Sigma_{n,(j,j)}^{*})^{1/2} - (\Sigma_{(j,j)}^{*,or})^{1/2}|/(\Sigma_{(j,j)}^{*,or})^{1/2}$ and $100 \times |\mathbf{J}_{n,(j,j)}^{1/2} - (\mathbf{J}_{(j,j)}^{or})^{1/2}|/(\mathbf{J}_{(j,j)}^{or})^{1/2}$ (where $A_{(j,j)}$ is the j-th diagonal element of a matrix A) between the estimated and oracle standard deviations of β_j $(j = 1, \ldots, 5)$ (we calculate the relative error for standard deviations rather than for variance, since standard deviations are used to obtain confidence intervals and Wald test statistics, which are the cornerstones of statistical inference in regression models). Results are averaged over the N simulated samples and reported in Table 4.5. We also report the RMSE of the RC, MI and AIPW variance estimates (the corresponding oracle variance estimates are used as reference). For example, the RMSE of the RC variance estimate of β_j is calculated as

$$\sqrt{\frac{1}{N}\sum_{\ell=1}^{N}\left(\Sigma_{n,(j,j)}^{(\ell)} - \Sigma_{(j,j)}^{or}\right)^2},$$

where $\Sigma_{n,(j,j)}^{(\ell)}$ denotes the RC variance estimate of β_j in the ℓ-th simulated sample. Results are given in Table 4.6. Regarding AIPW, we evaluate the three variance estimates given in (4.5.2) (results are reported at line "AIPW1" for misspecified $\pi(\mathbf{W}, \gamma)$, "AIPW2" for misspecified $m(\mathbf{W}, \theta)$, "AIPW3" when both models are correct, in both Tables 4.5 and 4.6).

From these results, it appears that both RMSE and relative differences between estimated and oracle standard deviations decrease when sample size increases and censoring and missing rates decrease. Relative errors and RMSE are the smallest for the AIPW estimate when both $m(\mathbf{W}, \theta)$ and $\pi(\mathbf{W}, \gamma)$ are correctly specified (line AIPW3). In each scenario, the relative errors and RMSE of AIPW are smaller when $\pi(\mathbf{W}, \gamma)$ is misspecified than when $m(\mathbf{W}, \theta)$ is misspecified. In fact, relative errors and RMSE show little sensitivity to misspecification of $\pi(\mathbf{W}, \gamma)$ (lines AIPW1 and AIPW3 are close to

each other). On the other hand, when $m(\mathbf{W}, \theta)$ is misspecified and censoring is high, the AIPW relative error can be substantial (around 20%, yielding the low coverage probabilities - around 75% - reported for $\beta_1, \beta_4, \beta_5$ in Table 4.4). But overall, all variance estimates essentially provide low relative errors (most of them being less than 6%). When $m(\mathbf{W}, \theta)$ is well specified and $\pi(\mathbf{W}, \gamma)$ is misspecified : *i*) the RC (respectively MI) variance estimate performs slightly better (respectively a little less well) than AIPW in terms of relative error, *ii*) AIPW variance estimate performs better than RC and MI in terms of RMSE. These observations suggest that the AIPW variance estimator is superior to RC and MI when both $m(\mathbf{W}, \theta)$ and $\pi(\mathbf{W}, \gamma)$ are correct, and that AIPW and RC variance estimates have similar performance when $\pi(\mathbf{W}, \gamma)$ is misspecified.

4.6.2 A real data analysis

We apply the proposed estimates to a data set from a survey of daily fruits and vegetables intake. The data were collected by the Office for National Statistics (UK), as part of a larger opinion survey. Respondents were asked about their usual daily intake of fruits and vegetables. Precisely, we have the number of portions of fruits and vegetables eaten by each respondent the day before the survey, and we know whether this number coincides with the respondent's usual intake or whether it is less than the usual intake. In this latter case, the usual intake is right-censored. The total sample size is $n = 928$. The censoring information is missing for 228 respondents (that is, 24.6% of the sample) and 29.6% of the respondents with known censoring information have a right-censored daily intake. Covariates are gender, age, marital status (married vs single/divorced/separated), educational level (with three levels: "General Certificate of Secondary Education (GCSE) or no qualification", "A-level or equivalent", "higher education") and a factor coding respondents appreciation of their daily intake of fruits and vegetables ("enough", "not enough", "more than enough"). We use logistic regression models (with covariates the number of portions reported by the respondents and the five variables mentioned above) for the conditional expectation of the censoring indicator and the selection probabilities. A forward-and-backward elimination strategy based on the AIC is used to select the final models. Finally, we calculate the RC, MI (with $M = 50$ completed data sets) and AIPW estimates in a Poisson regression model for the usual daily intake of fruits and vegetables.

Results are presented in Table 4.7 (in this table, *gender*=1 for male and 0 for female ; *single*=1 for a single/divorced/separated respondent and 0 for a married respondent ; *GCSE/no qualif.*=1 if the respondent has either no qualification or has obtained a GCSE, and 0 otherwise, *A-level or equiv.*=1 if the respondent has obtained a A-level or an equivalent diploma ; *more than enough*=1 if the respondent considers that her/his daily intake of fruits and vegetables is more than enough and 0 otherwise, *not enough*=1 if the respondent considers that her/his daily intake is not enough and 0 otherwise).

All methods conclude that age has a significant effect on the daily intake of fruits and vegetables,

with older people consuming more than younger ones. The gender effect is not significant (at level 5%) for the CC, RC and AIPW analysis but is significant for the MI method, with women consuming more fruits and vegetables than men. All methods find that being married is associated with increased fruits and vegetables intakes, while being single, separated or divorced is associated with lower consumption. For example, using the MI estimate, we find that on average, being single yields a $(1-e^{-0.0944})\times 100 \approx$ 9% decrease in the daily intake (holding fixed all the other effects). Our results also suggest that individuals with higher education have a higher consumption of fruits and vegetables than those with lower education (the reference level in Table 4.7 is "higher education"). The difference in daily intake between respondents with an A-level or equivalent diploma and respondents with a higher degree is not significant (although not being far from it, for all methods except the complete-case analysis) but there is a very significant difference between respondents with a GCSE or no qualification and those with a high degree. Finally, respondents who perceive their intake as more than enough (respectively not enough) indeed consume more (respectively less) fruits and vegetables, which may reflect the fact that respondents are well-informed on the usual recommandations about fruits and vegetables intake. Our results are coherent with the findings of previous studies that investigated factors that influence fruits and vegetables consumption, see Pollard et al. (2002) for example. Although here, the complete-case analysis yields the same conclusions as the RC, MI and AIPW methods, we observe some differences between the CC estimates and the RC, MI and AIPW estimates, which might reflect the bias of the CC method observed in the simulation study. Moreover, the CC estimates have usually larger standard errors, which reflects the loss of efficiency of the method.

In this example, the three AIPW variance estimates given in (4.5.2) are equal up to 3 digits (for this reason, only one standard error is reported in Table 4.7). This suggests that both the missingness model and conditional model for the censoring indicators given observed variables are correctly specified. For this reason and in view of the conclusions of the simulation study, we would recommand to use the AIPW estimate for further statistical inference on this data set.

Remark 4.6.2. *A naive (but easy to implement, using standard statistical softwares) estimation method consists in fitting an uncensored Poisson regression model to the data (that is, the censored intakes are treated as if they were uncensored). Using this method, the estimated constant is 1.338 (from Table 4.7, a consensus estimate is around 1.6). As expected, this naive method underestimates the baseline level of fruits and vegetables consumption.*

4.7 Discussion

In this article, we have investigated several estimators of the regression parameter of the censored Poisson regression model when censoring indicators are partially missing. The regression calibration

and multiple imputation estimates and their asymptotic variance estimators lead to reliable inferences when the model for the missing data given the observed variables is correctly specified, while the augmented inverse probability weighted estimator is asymptotically robust against misspecification of either the model for the missing data or the missingness mechanism. In finite samples, the AIPW estimator seems to be be more sensitive to a misspecification of the censoring mechanism than of the missingness mechanism.

Now, several issues deserve attention. First, in this work, we considered missing censoring indicators in the Poisson regression model, which assumes equidispersion. A similar issue may arise with under- or over-dispersed counts. The generalized Poisson regression model (see Famoye and Wang (2004) for example) is an appealing model for such data. The negative binomial regression model is an other option for modeling over-dispersed counts. When over-dispersion is due to zero-inflation, zero-inflated regression models (such as zero-inflated Poisson, zero-inflated generalized Poisson or zero-inflated negative binomial models) are appropriate. The estimates proposed in our paper may be adapted to these models and similar techniques could be used to investigate their asymptotic properties.

An other topic for further research is as follows. Our estimators rely on parametric models for the missing data and missingness mechanism. It is important to assess the sensitivity of the statistical inference to deviations to these models. An alternative estimation strategy may use semiparametric or nonparametric estimation of the models for missing data and missingness mechanism, and is also the topic for our future work.

Acknowledgements

Authors are grateful to two referees and the Associate Editor for their comments and suggestions that led substantial improvements of this paper. Authors acknowledge financial support from the Hubert Curien "PHC-Utique" program (CMCU number : 20G1503 - Campus France number : 44172SL), implemented by Campus France.

Appendix A: Proof of Theorem 4.3.3

Consistency. The consistency of $\tilde{\beta}_n$ can be proved by verifying the conditions of the inverse function theorem (Foutz, 1977). We describe the main steps of the proof and omit calculation details.

Let $\dot{\widetilde{\ell}}_n(\beta,\theta) := \partial\widetilde{\ell}_n(\beta,\theta)/\partial\beta$. Straightforward calculations yield:

$$\dot{\widetilde{\ell}}_n(\beta,\theta) = \sum_{i=1}^{n} \mathbf{X}_i \left[\hat{\delta}_i(\theta)\left(Y_i^* - e^{\beta^\top \mathbf{X}_i}\right) + (1-\hat{\delta}_i(\theta))h_\beta(Y_i^*, \mathbf{X}_i)\right],$$

where $h_\beta(y,x)$ is given by (4.3.3). We first need to show that $\partial\dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)/\partial\beta^\top$ exists and is continuous in a neighborhood of β_0. The map $\beta \mapsto \dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)$ is trivially differentiable with respect to β and its derivative is given by:

$$\frac{\partial\dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)}{\partial\beta^\top} = \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top \left(-\hat{\delta}_i(\hat{\theta}_n)e^{\beta^\top\mathbf{X}_i} + (1-\hat{\delta}_i(\hat{\theta}_n))\left\{Y_i^* - e^{\beta^\top\mathbf{X}_i} - h_\beta(Y_i^*,\mathbf{X}_i)\right\} h_\beta(Y_i^*,\mathbf{X}_i)\right),$$

which is continuous in β.

Secondly, we need to show that $n^{-1}\dot{\tilde{\ell}}_n(\beta_0,\hat{\theta}_n) = o_{\mathbb{P}}(1)$. To see this, we decompose $n^{-1}\dot{\tilde{\ell}}_n(\beta_0,\hat{\theta}_n)$:

$$\frac{1}{n}\dot{\tilde{\ell}}_n(\beta_0,\hat{\theta}_n) = \frac{1}{n}\left(\dot{\tilde{\ell}}_n(\beta_0,\hat{\theta}_n) - \dot{\tilde{\ell}}_n(\beta_0,\theta_0)\right) + \frac{1}{n}\dot{\tilde{\ell}}_n(\beta_0,\theta_0).$$

By the weak law of large numbers, $n^{-1}\dot{\tilde{\ell}}_n(\beta_0,\theta_0)$ converges in probability to

$$\begin{aligned}&\mathbb{E}\left[\mathbf{X}\left(\hat{\delta}(\theta_0)(Y^* - e^{\beta_0^\top\mathbf{X}}) + (1-\hat{\delta}(\theta_0))h_{\beta_0}(Y^*,\mathbf{X})\right)\right]\\ &\quad= \mathbb{E}\left[\mathbf{X}\left(\mathbb{E}(\hat{\delta}(\theta_0)|\mathbf{W})(Y^* - e^{\beta_0^\top\mathbf{X}}) + (1-\mathbb{E}(\hat{\delta}(\theta_0)|\mathbf{W}))h_{\beta_0}(Y^*,\mathbf{X})\right)\right], \qquad (4.7.1)\end{aligned}$$

where the second line follows by taking the conditional expectation given $\mathbf{W}$. Under the missing at random assumption,

$$\begin{aligned}\mathbb{E}(\hat{\delta}(\theta_0)|\mathbf{W}) &= \mathbb{E}(\xi\delta + (1-\xi)\mathbb{E}(\delta|\mathbf{W})|\mathbf{W})\\ &= \mathbb{E}(\xi|\mathbf{W})\mathbb{E}(\delta|\mathbf{W}) + (1-\mathbb{E}(\xi|\mathbf{W}))\mathbb{E}(\delta|\mathbf{W})\\ &= \mathbb{E}(\delta|\mathbf{W}).\end{aligned}$$

Therefore, (4.7.1) is equal to

$$\begin{aligned}&\mathbb{E}\left[\mathbf{X}\left(\mathbb{E}(\delta|\mathbf{W})(Y^* - e^{\beta_0^\top\mathbf{X}}) + (1-\mathbb{E}(\delta|\mathbf{W}))h_{\beta_0}(Y^*,\mathbf{X})\right)\right]\\ &\quad= \mathbb{E}\left[\mathbf{X}\left(\delta(Y^* - e^{\beta_0^\top\mathbf{X}}) + (1-\delta)h_{\beta_0}(Y^*,\mathbf{X})\right)\right],\end{aligned}$$

which is equal to 0 (this can be seen by taking successively the conditional expectations given $\{\delta = 1\}$ and $\mathbf{X}$). Convergence to 0 of $n^{-1}(\dot{\tilde{\ell}}_n(\beta_0,\hat{\theta}_n) - \dot{\tilde{\ell}}_n(\beta_0,\theta_0))$ is a consequence of the consistency of $\hat{\theta}_n$ and of assumptions C1, C2, C4. Details are omitted.

Thirdly, we need to show that $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)/\partial\beta^\top$ converges in probability to a fixed matrix, uni-

formly in an open neighborhood of β_0. We have:

$$\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta,\theta)}{\partial\beta^\top} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\left(-\hat{\delta}_i(\theta)e^{\beta^\top\mathbf{X}_i} + (1-\hat{\delta}_i(\theta))\left\{Y_i^* - e^{\beta^\top\mathbf{X}_i} - h_\beta(Y_i^*,\mathbf{X}_i)\right\}h_\beta(Y_i^*,\mathbf{X}_i)\right).$$

We proceed as above and decompose $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)/\partial\beta^\top$ as

$$\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)}{\partial\beta^\top} = \frac{1}{n}\left(\frac{\partial\dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)}{\partial\beta^\top} - \frac{\partial\dot{\tilde{\ell}}_n(\beta,\theta_0)}{\partial\beta^\top}\right) + \frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta,\theta_0)}{\partial\beta^\top}.$$

The first term converges to 0 (by the consistency of $\hat{\theta}_n$ and assumptions C1, C2, C4) and $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta,\theta_0)/\partial\beta^\top$ converges in probability to $-\Sigma_1(\beta)$ (by the weak law of large numbers). Therefore, $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)/\partial\beta^\top$ converges in probability to $-\Sigma_1(\beta)$. Under conditions C1 and C2, the derivative of $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)/\partial\beta^\top$ with respect to β is bounded, for every n. Hence the sequence $(n^{-1}\partial\dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)/\partial\beta^\top)_n$ is equicontinuous. It follows from Ascoli theorem that the convergence of $n^{-1}\partial\dot{\tilde{\ell}}_n(\beta,\hat{\theta}_n)/\partial\beta^\top$ to $-\Sigma_1(\beta)$ is uniform around β_0.

Having proved the conditions of the inverse function theorem, we conclude that $\tilde{\beta}_n$ converges in probability to β_0.

Asymptotic normality. A Taylor's expansion of $\dot{\tilde{\ell}}_n(\tilde{\beta}_n,\hat{\theta}_n)$ around (β_0,θ_0) yields

$$\sqrt{n}(\tilde{\beta}_n-\beta_0) = \left(-\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta_0,\theta_0)}{\partial\beta^\top}\right)^{-1}\left(\frac{1}{\sqrt{n}}\dot{\tilde{\ell}}_n(\beta_0,\theta_0) + \frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta_0,\theta_0)}{\partial\theta^\top}\sqrt{n}(\hat{\theta}_n-\theta_0)\right) + o_{\mathbb{P}}(1).$$

We have:

$$\begin{aligned}\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta,\theta)}{\partial\theta^\top} &= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\left(Y_i^* - e^{\beta^\top\mathbf{X}_i} - h_\beta(Y_i^*,\mathbf{X}_i)\right)(1-\xi_i)\dot{m}^\top(\mathbf{W}_i,\theta)\\ &= \Sigma_2(\beta,\theta) + o_{\mathbb{P}}(1).\end{aligned}$$

Combining this and (4.3.2), we can write:

$$\begin{aligned}\sqrt{n}(\tilde{\beta}_n-\beta_0) &= \left(-\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta_0,\theta_0)}{\partial\beta^\top}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\mathbf{X}_i\left\{\hat{\delta}_i(\theta_0)(Y_i^*-e^{\beta_0^\top\mathbf{X}_i})+(1-\hat{\delta}_i(\theta_0))h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right\}\\ &\qquad+\Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\widetilde{m}_i(\theta_0)\xi_i(\delta_i-m(\mathbf{W}_i,\theta_0))\Big]+o_{\mathbb{P}}(1)\\ &:= \left(-\frac{1}{n}\frac{\partial\dot{\tilde{\ell}}_n(\beta_0,\theta_0)}{\partial\beta^\top}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{U}_i+o_{\mathbb{P}}(1).\end{aligned}$$

Now, note that

$$\text{var}\left(\mathbf{X}_i\left\{\hat{\delta}_i(\theta_0)(Y_i^*-e^{\beta_0^\top\mathbf{X}_i})+(1-\hat{\delta}_i(\theta_0))h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right\}\right)=\Sigma_1(\beta_0),$$

and

$$\begin{aligned}\text{var}\left(\Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\widetilde{m}_i(\theta_0)\xi_i(\delta_i-m(\mathbf{W}_i,\theta_0))\right) &= \Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\mathbb{E}\left[\widetilde{m}_i^{\otimes 2}(\theta_0)\xi_i(\delta_i-m(\mathbf{W}_i,\theta_0))^2\right]\\ &\quad\times\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0)\\ &= \Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0),\end{aligned}$$

since under the missing at random assumption, we have :

$$\begin{aligned}\mathbb{E}\left[\widetilde{m}_i^{\otimes 2}(\theta_0)\xi_i(\delta_i-m(\mathbf{W}_i,\theta_0))^2\right] &= \mathbb{E}\left[\frac{\dot{m}^{\otimes 2}(\mathbf{W}_i,\theta_0)}{\{m(\mathbf{W}_i,\theta_0)(1-m(\mathbf{W}_i,\theta_0))\}^2}\mathbb{E}\left[\xi_i(\delta_i-m(\mathbf{W}_i,\theta_0))^2|\mathbf{W}_i\right]\right]\\ &= \mathbb{E}\left[\frac{\dot{m}^{\otimes 2}(\mathbf{W}_i,\theta_0)}{\{m(\mathbf{W}_i,\theta_0)(1-m(\mathbf{W}_i,\theta_0))\}^2}\mathbb{E}\left[\xi_i|\mathbf{W}_i\right]\mathbb{E}\left[\delta_i-2\delta_i m(\mathbf{W}_i,\theta_0)\right.\right.\\ &\qquad\left.\left.+m^2(\mathbf{W}_i,\theta_0)|\mathbf{W}_i\right]\right]\\ &= \mathbb{E}\left[\frac{\dot{m}^{\otimes 2}(\mathbf{W}_i,\theta_0)}{m(\mathbf{W}_i,\theta_0)(1-m(\mathbf{W}_i,\theta_0))}\pi(\mathbf{W}_i)\right]\\ &= \Theta(\theta_0).\end{aligned}$$

We consider now the covariance structure of $\mathcal{U}_i$. We have

$$\begin{aligned}&\text{cov}\left(\mathbf{X}_i\left\{\hat{\delta}_i(\theta_0)(Y_i^*-e^{\beta_0^\top\mathbf{X}_i})+(1-\hat{\delta}_i(\theta_0))h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right\},\Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\widetilde{m}_i(\theta_0)\xi_i(\delta_i-m(\mathbf{W}_i,\theta_0))\right)\\ =\ &\mathbb{E}\left[\mathbf{X}_i\widetilde{m}_i^\top(\theta_0)\mathbb{E}\left[\left(\hat{\delta}_i(\theta_0)(Y_i^*-e^{\beta_0^\top\mathbf{X}_i}-h_{\beta_0}(Y_i^*,\mathbf{X}_i))+h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right)\xi_i(\delta_i-m(\mathbf{W}_i,\theta_0))|\mathbf{W}_i\right]\right]\\ &\times\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0,\theta_0),\end{aligned}$$

and

$$\begin{aligned}
& \mathbb{E}\left[\left(\hat{\delta}_i(\theta_0)(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i)) + h_{\beta_0}(Y_i^*, \mathbf{X}_i)\right) \xi_i(\delta_i - m(\mathbf{W}_i, \theta_0))|\mathbf{W}_i\right] \\
= \; & \mathbb{E}\left[\xi_i\delta_i(1 - m(\mathbf{W}_i, \theta_0))(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i)) + \xi_i h_{\beta_0}(Y_i^*, \mathbf{X}_i)(\delta_i - m(\mathbf{W}_i, \theta_0))|\mathbf{W}_i\right] \\
= \; & (1 - m(\mathbf{W}_i, \theta_0))(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i))m(\mathbf{W}_i, \theta_0)\pi(\mathbf{W}_i),
\end{aligned}$$

therefore,

$$\begin{aligned}
& \operatorname{cov}\left(\mathbf{X}_i\left\{\hat{\delta}_i(\theta_0)(Y_i^* - e^{\beta_0^\top \mathbf{X}_i}) + (1 - \hat{\delta}_i(\theta_0))h_{\beta_0}(Y_i^*, \mathbf{X}_i)\right\}, \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\widetilde{m}_i(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i, \theta_0))\right) \\
= \; & \mathbb{E}\left[\mathbf{X}_i\widetilde{m}_i^\top(\theta_0)(1 - m(\mathbf{W}_i, \theta_0))(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i))m(\mathbf{W}_i, \theta_0)\pi(\mathbf{W}_i)\right]\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0) \\
= \; & \mathbb{E}\left[\mathbf{X}_i\dot{m}_i^\top(\mathbf{W}_i, \theta_0)(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i))\pi(\mathbf{W}_i)\right]\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0) \\
= \; & \left(\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0).
\end{aligned}$$

It follows that

$$\operatorname{var}(\mathcal{U}_i) = \Sigma_1(\beta_0) + \left(2\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0).$$

Finally, Theorem 4.3.3 follows from the multivariate central limit theorem and Slutsky's theorem. □

Appendix B: Proof of Theorem 4.4.1

Consistency can be proved in much the same way as $\tilde{\beta}_n$; the proof is therefore omitted. We turn to asymptotic normality. A technical lemma is needed. For $j = 1, \ldots, M$, let

$$\begin{aligned}
\dot{\ell}_{n,j}^*(\beta, \theta) &= \frac{\partial \ell_{n,j}^*(\beta, \theta)}{\partial \beta} \\
&= \sum_{i=1}^n \mathbf{X}_i\left(\delta_{i,j}^*(\theta)\left[Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i)\right] + h_\beta(Y_i^*, \mathbf{X}_i)\right) \qquad (4.7.1) \\
&:= \sum_{i=1}^n f_{\beta,\theta,j}(\mathcal{O}_i).
\end{aligned}$$

Then the following holds:

Lemma 4.7.1. *Under conditions C1, C2 and C4 :*

$$\frac{1}{\sqrt{n}}\left[\dot{\ell}_{n,j}^*(\beta_0, \hat{\theta}_n) - n\mathbb{E}[\dot{\ell}_{1,j}^*(\beta_0, \hat{\theta}_n)] - \left(\dot{\ell}_{n,j}^*(\beta_0, \theta_0) - n\mathbb{E}[\dot{\ell}_{1,j}^*(\beta_0, \theta_0)]\right)\right] \xrightarrow{\mathbb{P}} 0 \qquad (4.7.2)$$

as $n \to \infty$.

Proof of Lemma 4.7.1. In this proof, for notational simplicity, we will write f_θ instead of $f_{\beta_0,\theta,j}$. First, note that

$$\begin{aligned}\frac{1}{\sqrt{n}}\left[\dot{\ell}^*_{n,j}(\beta_0,\theta) - n\mathbb{E}\big[\dot{\ell}^*_{1,j}(\beta_0,\theta)\big]\right] &= \frac{1}{\sqrt{n}}\left[\sum_{i=1}^n f_\theta(\mathcal{O}_i) - n\mathbb{E}\big[f_\theta(\mathcal{O}_1)\big]\right] \\ &= \mathbb{G}_n f_\theta,\end{aligned}$$

where $\mathbb{G}_n f_\theta$ denotes the empirical process evaluated at f_θ. To prove the lemma, we first prove that the class of functions $\{f_\theta : \theta \in \Theta\}$ is Donsker (see, for example, van der Vaart (2000) for a detailed account on empirical processes and Donsker classes). For that purpose, we decompose f_θ in (4.7.1) as $f_\theta(\mathcal{O}_i) = \mathbf{X}_i(f_{1,\theta}(\mathcal{O}_i) + f_{2,\theta}(\mathcal{O}_i) + f_{3,\theta}(\mathcal{O}_i))$, where $f_{1,\theta}(\mathcal{O}_i) = -\delta^*_{i,j}(\theta)e^{\beta_0^\top \mathbf{X}_i} + h_{\beta_0}(Y_i^*, \mathbf{X}_i)$, $f_{2,\theta}(\mathcal{O}_i) = \delta^*_{i,j}(\theta)Y_i^*$ and $f_{3,\theta}(\mathcal{O}_i) = -\delta^*_{i,j}(\theta)h_{\beta_0}(Y_i^*, \mathbf{X}_i)$ and we show that the classes $\mathcal{F}_1 := \{f_{1,\theta} : \theta \in \Theta\}$, $\mathcal{F}_2 := \{f_{2,\theta} : \theta \in \Theta\}$ and $\mathcal{F}_3 := \{f_{3,\theta} : \theta \in \Theta\}$ are Donsker.

For illustration purpose, we show that $\mathcal{F}_1$ is Donsker. Here, it is useful to see $D_{i,j}(\theta) \sim \mathcal{B}(m(\mathbf{W}_i,\theta))$ as the random variable $1_{\{U_i \leqslant m(\mathbf{W}_i,\theta)\}}$, where U_i is a uniform random variable on $[0,1]$, independent of $\mathcal{O}_i$.

Let $d := \text{diam}(\Theta)$ denote the diameter of $\Theta \subset \mathbb{R}^q$. Then the size of Θ in every direction is at most d and thus, we can cover Θ with fewer than $(d/\kappa)^q$ cubes of length κ. The circumscribed balls have radius a multiple $\kappa^* := \alpha\kappa$ of κ ($\alpha > 0$) and these balls also cover Θ. Now, for a given $\theta \in \Theta$, consider the set

$$\{f_{1,\tilde{\theta}} : \tilde{\theta} \in \Theta \cap \mathcal{B}(\theta,\kappa^*)\},$$

where $\mathcal{B}(\theta,\kappa^*) = \{\tilde{\theta} \in \mathbb{R}^q : \|\theta - \tilde{\theta}\| \leqslant \kappa^*\}$ is the ball of radius κ^* and center θ. If $\tilde{\theta} \in \mathcal{B}(\theta,\kappa^*)$, condition C4 implies that

$$|m(\mathbf{w},\theta) - m(\mathbf{w},\tilde{\theta})| \leqslant h(\mathbf{w})\kappa^*,$$

hence $m(\mathbf{w},\theta) - h(\mathbf{w})\kappa^* \leqslant m(\mathbf{w},\tilde{\theta}) \leqslant m(\mathbf{w},\theta) + h(\mathbf{w})\kappa^*$ and thus $1_{\{U_i \leqslant m(\mathbf{w},\theta) - h(\mathbf{w})\kappa^*\}} \leqslant 1_{\{U_i \leqslant m(\mathbf{w},\tilde{\theta})\}}$ $\leqslant 1_{\{U_i \leqslant m(\mathbf{w},\theta) + h(\mathbf{w})\kappa^*\}}$. From this, we can see that

$$f^L_\theta(\mathcal{O}_i) \leqslant f_{1,\tilde{\theta}}(\mathcal{O}_i) \leqslant f^U_\theta(\mathcal{O}_i),$$

where

$$\begin{aligned} f_\theta^L(\mathcal{O}_i) &= h_{\beta_0}(Y_i^*, \mathbf{X}_i) - (\xi_i\delta_i + (1-\xi_i)1_{\{U_i \leqslant m(\mathbf{W}_i,\theta)+h(\mathbf{W}_i)\kappa^*\}})e^{\beta_0^\top \mathbf{X}_i}, \\ f_\theta^U(\mathcal{O}_i) &= h_{\beta_0}(Y_i^*, \mathbf{X}_i) - (\xi_i\delta_i + (1-\xi_i)1_{\{U_i \leqslant m(\mathbf{W}_i,\theta)-h(\mathbf{W}_i)\kappa^*\}})e^{\beta_0^\top \mathbf{X}_i}. \end{aligned}$$

Moreover, under conditions C1, C2 and C4, there exists a finite positive constant c_1 such that

$$\mathbb{E}\left[\left(f_\theta^U(\mathcal{O}_i) - f_\theta^L(\mathcal{O}_i)\right)^2\right] \leqslant 2c_1\kappa^* v.$$

Therefore, $[f_\theta^L, f_\theta^U]$ is an ε-bracket for $\{f_{1,\tilde{\theta}} : \tilde{\theta} \in \Theta \cap \mathcal{B}(\theta, \kappa^*)\}$, with $\varepsilon^2 = 2c_1\kappa^* v$. Since we can cover Θ with fewer than $(d/\kappa)^q$ balls of radius κ^*, we can cover $\mathcal{F}_1 = \{f_{1,\tilde{\theta}} : \tilde{\theta} \in \Theta\}$ with fewer than $(d/\kappa)^q$ ε-brackets $[f_\theta^L, f_\theta^U]$, with $\varepsilon = \sqrt{2c_1\kappa^* v}$. The number of such ε-brackets is thus bounded by $(\alpha d/\kappa^*)^q = (2\alpha c_1 dv/\varepsilon^2)^q$, which is order ε^{-2q}. Hence, the bracketing integral is of order $\int_0^1 \sqrt{-2q\log\varepsilon}\, d\varepsilon$, which is finite. Therefore, the class of functions $\mathcal{F}_1$ is Donsker, by Theorem 19.5 of van der Vaart (2000).

By using similar arguments, we can prove that $\mathcal{F}_2$ and $\mathcal{F}_3$ are also Donsker classes. It follows that the class of functions $\{f_{1,\theta} + f_{2,\theta} + f_{3,\theta} : \theta \in \Theta\}$ is Donsker (sums of Donsker classes are Donsker). Finally, $\mathbf{X}$ is bounded (by condition C1), thus the class of functions $\{f_\theta : \theta \in \Theta\}$ is Donsker.

It follows that the sequence of processes $\{\mathbb{G}_n f_\theta : \theta \in \Theta\}$ converges in distribution to a tight limit process, and as such, is stochastically equicontinuous. Thus, Lemma 14.3 of Tsiatis (2007) and the consistency of $\hat{\theta}_n$ imply that $\mathbb{G}_n f_{\hat{\theta}_n} - \mathbb{G}_n f_{\theta_0} \xrightarrow{\mathbb{P}} 0$, which is exactly (4.7.2). This concludes the proof. □

We come back to the proof of asymptotic normality. By a Taylor expansion of $\dot{\ell}_{n,j}^*(\hat{\beta}_{n,j}^*, \hat{\theta}_n)$ around β_0 (for $j = 1, \ldots, M$), we have:

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\dot{\ell}_{n,j}^*(\hat{\beta}_{n,j}^*, \hat{\theta}_n) \\ &= \frac{1}{\sqrt{n}}\dot{\ell}_{n,j}^*(\beta_0, \hat{\theta}_n) + \frac{1}{n}\frac{\partial \dot{\ell}_{n,j}^*(\beta_0, \hat{\theta}_n)}{\partial \beta^\top}\sqrt{n}(\hat{\beta}_{n,j}^* - \beta_0) + o_{\mathbb{P}}(1). \end{aligned}$$

Then, using Lemma 4.7.1, we obtain:

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\dot{\ell}_{n,j}^*(\beta_0, \theta_0) - \sqrt{n}\mathbb{E}[\dot{\ell}_{1,j}^*(\beta_0, \theta_0)] + \sqrt{n}\mathbb{E}[\dot{\ell}_{1,j}^*(\beta_0, \hat{\theta}_n)] + \frac{1}{n}\frac{\partial \dot{\ell}_{n,j}^*(\beta_0, \hat{\theta}_n)}{\partial \beta^\top}\sqrt{n}(\hat{\beta}_{n,j}^* - \beta_0) + o_{\mathbb{P}}(1) \\ &= \frac{1}{\sqrt{n}}\dot{\ell}_{n,j}^*(\beta_0, \theta_0) + \sqrt{n}\left(\frac{\partial \mathbb{E}[\dot{\ell}_{1,j}^*(\beta_0, \theta_0)]}{\partial \theta^\top}(\hat{\theta}_n - \theta_0) + o_{\mathbb{P}}(\|\hat{\theta}_n - \theta_0\|)\right) + \frac{1}{n}\frac{\partial \dot{\ell}_{n,j}^*(\beta_0, \hat{\theta}_n)}{\partial \beta^\top}\sqrt{n}(\hat{\beta}_{n,j}^* - \beta_0) \\ &\quad + o_{\mathbb{P}}(1), \end{aligned} \tag{4.7.3}$$

where the second line follows from a Taylor expansion of $\mathbb{E}[\dot{\ell}^*_{1,j}(\beta_0, \hat{\theta}_n)]$ around θ_0. Two technical lemmas are now needed :

Lemma 4.7.2. *For $j = 1, \ldots, M$, we have*

$$\frac{\partial \mathbb{E}[\dot{\ell}^*_{1,j}(\beta, \theta)]}{\partial \theta^\top} = \Sigma_2(\beta, \theta).$$

Proof of Lemma 4.7.2. First, we note that

$$\begin{aligned} \mathbb{E}[\delta^*_{1,j}(\theta)|\mathbf{W}_1] &= \mathbb{E}[\xi_1\delta_1 + (1-\xi_1)D_{1,j}(\theta)] \\ &= \pi(\mathbf{W}_1)m(\mathbf{W}_1, \theta_0) + (1-\pi(\mathbf{W}_1))m(\mathbf{W}_1, \theta). \end{aligned} \tag{4.7.4}$$

Hence, using (4.7.1) and iterating the expectation with conditioning on $\mathbf{W}_1$, we obtain:

$$\begin{aligned} \mathbb{E}[\dot{\ell}^*_{1,j}(\beta, \theta)] &= \mathbb{E}\left[\mathbf{X}_1\left(\delta^*_{1,j}(\theta)\left[Y_1^* - e^{\beta^\top \mathbf{X}_1} - h_\beta(Y_1^*, \mathbf{X}_1)\right] + h_\beta(Y_1^*, \mathbf{X}_1)\right)\right] \\ &= \mathbb{E}\Big[\mathbf{X}_1\Big((\pi(\mathbf{W}_1)m(\mathbf{W}_1, \theta_0) + (1-\pi(\mathbf{W}_1))m(\mathbf{W}_1, \theta))\left[Y_1^* - e^{\beta^\top \mathbf{X}_1} - h_\beta(Y_1^*, \mathbf{X}_1)\right] \\ &\qquad + h_\beta(Y_1^*, \mathbf{X}_1)\Big)\Big]. \end{aligned}$$

Finally, straightforward calculations yield

$$\begin{aligned} \frac{\partial \mathbb{E}[\dot{\ell}^*_{1,j}(\beta, \theta)]}{\partial \theta^\top} &= \mathbb{E}\left[\mathbf{X}_1(1-\pi(\mathbf{W}_1))\dot{m}^\top(\mathbf{W}_1, \theta)\left(Y_1^* - e^{\beta^\top \mathbf{X}_1} - h_\beta(Y_1^*, \mathbf{X}_1)\right)\right] \\ &= \Sigma_2(\beta, \theta). \end{aligned}$$

□

Lemma 4.7.3. *For $j = 1, \ldots, M$,*

$$\frac{1}{n}\frac{\partial \dot{\ell}^*_{n,j}(\beta_0, \hat{\theta}_n)}{\partial \beta^\top} \xrightarrow{\mathbb{P}} -\Sigma_1(\beta_0).$$

Proof of Lemma 4.7.3. Let $j = 1, \ldots, M$. Straightforward calculations yield:

$$\frac{\partial \dot{\ell}^*_{n,j}(\beta_0, \hat{\theta}_n)}{\partial \beta^\top} = \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\left[-\delta^*_{i,j}(\hat{\theta}_n)e^{\beta_0^\top \mathbf{X}_i} + (1-\delta^*_{i,j}(\hat{\theta}_n))h_{\beta_0}(Y_i^*, \mathbf{X}_i)\left(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i)\right)\right].$$

Then we decompose $n^{-1}\partial\dot{\ell}^*_{n,j}(\beta_0,\hat{\theta}_n)/\partial\beta^\top$ as:

$$\begin{aligned}
\frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\hat{\theta}_n)}{\partial\beta^\top} &= \left(\frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\hat{\theta}_n)}{\partial\beta^\top}-\frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\theta_0)}{\partial\beta^\top}\right)+\frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\theta_0)}{\partial\beta^\top}\\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\Big[-e^{\beta_0^\top\mathbf{X}_i}(1-\xi_i)(D_{i,j}(\hat{\theta}_n)-D_{i,j}(\theta_0))\\
&\qquad +h_{\beta_0}(Y_i^*,\mathbf{X}_i)\left(Y_i^*-e^{\beta_0^\top\mathbf{X}_i}-h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right)(1-\xi_i)(D_{i,j}(\theta_0)-D_{i,j}(\hat{\theta}_n))\Big]\\
&\qquad +\frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\theta_0)}{\partial\beta^\top}\\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top(1-\xi_i)\Big[e^{\beta_0^\top\mathbf{X}_i}+h_{\beta_0}(Y_i^*,\mathbf{X}_i)\left(Y_i^*-e^{\beta_0^\top\mathbf{X}_i}-h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right)\Big]\\
&\qquad \times(D_{i,j}(\theta_0)-D_{i.j}(\hat{\theta}_n))+\frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\theta_0)}{\partial\beta^\top}\\
&= \frac{1}{n}\sum_{i=1}^n \mathcal{Z}_i(1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\theta_0)\}}-1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\hat{\theta}_n)\}})+\frac{1}{n}\frac{\partial\dot{\ell}^*_{n,j}(\beta_0,\theta_0)}{\partial\beta^\top}, \qquad (4.7.5)
\end{aligned}$$

where $\mathcal{Z}_i := \mathbf{X}_i\mathbf{X}_i^\top(1-\xi_i)[e^{\beta_0^\top\mathbf{X}_i}+h_{\beta_0}(Y_i^*,\mathbf{X}_i)(Y_i^*-e^{\beta_0^\top\mathbf{X}_i}-h_{\beta_0}(Y_i^*,\mathbf{X}_i))]$ (in what follows, we will denote by $\mathcal{Z}_{i,(\ell,k)}$ the (ℓ,k)-th element of $\mathcal{Z}_i$) and $U_{i,j}$ is a uniform random variable on $[0,1]$, independent of all other random variables.

Consider the first term in the right-hand side of (4.7.5). The random variable $|1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\theta_0)\}}-1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\hat{\theta}_n)\}}|$ is equal to 0 or 1 and takes the value 1 with probability $|m(\mathbf{W}_i,\theta_0)-m(\mathbf{W}_i,\hat{\theta}_n)|$. Let $\varepsilon>0$. Then, for $\ell,k\in\{1,\ldots,p\}$, Markov's inequality implies that

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{i,(\ell,k)}(1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\theta_0)\}}-1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\hat{\theta}_n)\}})\right|>\varepsilon\right)\\
&\qquad\leqslant\frac{1}{\varepsilon}\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \mathcal{Z}_{i,(\ell,k)}(1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\theta_0)\}}-1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\hat{\theta}_n)\}})\right|\right].
\end{aligned}$$

Under conditions C1 and C2, there exists a finite positive constant c_2 such that $|\mathcal{Z}_{i,(\ell,k)}| \leqslant c_2$. Thus,

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathcal{Z}_{i,(\ell,k)}(1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\theta_0)\}}-1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\hat{\theta}_n)\}})\right|>\varepsilon\right)&\\
\leqslant \frac{c_2}{\varepsilon n}\sum_{i=1}^{n}|m(\mathbf{W}_i,\theta_0)-m(\mathbf{W}_i,\hat{\theta}_n)|&\\
\leqslant \frac{c_2}{\varepsilon n}\sum_{i=1}^{n}h(\mathbf{W}_i)\|\theta_0-\hat{\theta}_n\|&\\
\leqslant \frac{c_2}{\varepsilon}\|\theta_0-\hat{\theta}_n\|(v+o_{\mathbb{P}}(1)),&
\end{aligned}$$

where the last two lines follow from the condition C4. Finally, consistency of $\hat{\theta}_n$ implies that $\frac{1}{n}\sum_{i=1}^{n}\mathcal{Z}_{i,(\ell,k)}(1_{\{U_{i,j}\leqslant m(}$ $1_{\{U_{i,j}\leqslant m(\mathbf{W}_i,\hat{\theta}_n)\}})$ converges in probability to 0, and the first term in the right-hand side of (4.7.5) also converges to 0.

We consider now the second term in the right-hand side of (4.7.5). By the weak law of large numbers, $n^{-1}\partial\dot{\ell}^*_{n,j}(\beta_0,\theta_0)/\partial\beta^\top$ converges in probability to

$$\mathbb{E}\left[\mathbf{X}_1\mathbf{X}_1^\top\left[-\delta^*_{1,j}(\theta_0)e^{\beta_0^\top\mathbf{X}_1}+(1-\delta^*_{1,j}(\theta_0))h_{\beta_0}(Y_1^*,\mathbf{X}_1)\left(Y_1^*-e^{\beta_0^\top\mathbf{X}_1}-h_{\beta_0}(Y_1^*,\mathbf{X}_1)\right)\right]\right]. \tag{4.7.6}$$

Using the fact that $\mathbb{E}[\delta^*_{1,j}(\theta_0)|\mathbf{W}_1]=m(\mathbf{W}_1,\theta_0)$ (see (4.7.4)), and iterating the expectation in (4.7.6) with conditioning on $\mathbf{W}_1$, we easily show that (4.7.6) is equal to $-\Sigma_1(\beta_0)$.

Thus, we have shown that $n^{-1}\partial\dot{\ell}^*_{n,j}(\beta_0,\hat{\theta}_n)/\partial\beta^\top$ converges in probability to $-\Sigma_1(\beta_0)$, which concludes the proof. □

By combining (4.7.3) with Lemmas 4.7.2 and 4.7.3, we obtain the following approximation of $\hat{\beta}^*_{n,j}$:

$$\begin{aligned}
\sqrt{n}(\hat{\beta}^*_{n,j}-\beta_0) \;=\; \Sigma_1^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\mathbf{X}_i\left\{\delta^*_{i,j}(\theta_0)[Y_i^*-e^{\beta_0^\top\mathbf{X}_i}-h_{\beta_0}(Y_i^*,\mathbf{X}_i)]+h_{\beta_0}(Y_i^*,\mathbf{X}_i)\right\}&\\
+\Sigma_2(\beta_0,\theta_0)\Theta^{-1}(\theta_0)\widetilde{m}_i(\theta_0)\xi_i(\delta_i-m(\mathbf{W}_i,\theta_0))\Big]+o_{\mathbb{P}}(1),&
\end{aligned}$$

which in turn implies the approximation of the multiple imputation estimator $\hat{\beta}_n^*$:

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_n^* - \beta_0) &= \frac{1}{M}\sum_{j=1}^{M}\left(\sqrt{n}(\hat{\beta}_{n,j}^* - \beta_0)\right) \\
&= \Sigma_1^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{1}{M}\sum_{j=1}^{M}\mathbf{X}_i\left\{\delta_{i,j}^*(\theta_0)[Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i)] + h_{\beta_0}(Y_i^*, \mathbf{X}_i)\right\}\right. \\
&\qquad \left. + \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\widetilde{m_i}(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i, \theta_0))\right] + o_{\mathbb{P}}(1) \\
&:= \Sigma_1^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{1}{M}\sum_{j=1}^{M} f_{\beta_0,\theta_0,j}(\mathcal{O}_i) + \mathcal{V}_i\right] + o_{\mathbb{P}}(1), \qquad (4.7.7)
\end{aligned}$$

where $\mathcal{V}_i := \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\widetilde{m_i}(\theta_0)\xi_i(\delta_i - m(\mathbf{W}_i, \theta_0))$. We have already shown (see proof of Theorem 4.3.3) that

$$\text{var}(\mathcal{V}_i) = \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0).$$

Similar calculations as in the proof of Theorem 4.3.3 yield:

$$\text{cov}(f_{\beta_0,\theta_0,j}(\mathcal{O}_i), \mathcal{V}_i) = \left(\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0).$$

Therefore,

$$\begin{aligned}
\text{var}\left(\frac{1}{M}\sum_{j=1}^{M} f_{\beta_0,\theta_0,j}(\mathcal{O}_i) + \mathcal{V}_i\right) &= \text{var}\left(\frac{1}{M}\sum_{j=1}^{M} f_{\beta_0,\theta_0,j}(\mathcal{O}_i)\right) + \text{var}(\mathcal{V}_i) + \frac{2}{M}\sum_{j=1}^{M}\text{cov}(f_{\beta_0,\theta_0,j}(\mathcal{O}_i), \mathcal{V}_i) \\
&= \Sigma_1^*(\beta_0, \theta_0) + \Sigma_2(\beta_0, \theta_0)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0) \\
&\qquad + 2\left(\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0) \\
&= \Sigma_1^*(\beta_0, \theta_0) + \left(2\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0). \qquad (4.7.8)
\end{aligned}$$

Finally, it follows from (4.7.7), (4.7.8) and the multivariate central limit theorem that $\sqrt{n}(\hat{\beta}_n^* - \beta_0)$ converges in distribution to a Gaussian vector with mean zero and variance

$$\Sigma_1^{-1}(\beta_0)\left\{\Sigma_1^*(\beta_0, \theta_0) + \left(2\Sigma_3(\beta_0, \theta_0) - \Sigma_2(\beta_0, \theta_0)\right)\Theta^{-1}(\theta_0)\Sigma_2^\top(\beta_0, \theta_0)\right\}\Sigma_1^{-1}(\beta_0),$$

which concludes the proof. □

Appendix C: Proof of Theorem 4.5.1

Assume that the model $m(\mathbf{W}_i, \theta)$ is correctly specified. It is straightforward to check that the map $\beta \mapsto \partial\breve{\ell}_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)/\partial\beta$ exists and is continuous in a neighborhood of β_0 (condition i).

Now, we show that $n^{-1}\breve{\ell}_n(\beta_0, \hat{\theta}_n, \hat{\gamma}_n) = o_{\mathbb{P}}(1)$ (condition ii). To see this, decompose $n^{-1}\breve{\ell}_n(\beta_0, \hat{\theta}_n, \hat{\gamma}_n)$ as:

$$\begin{aligned}
\frac{1}{n}\breve{\ell}_n(\beta_0, \hat{\theta}_n, \hat{\gamma}_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\frac{\xi_i}{\pi(\mathbf{W}_i, \hat{\gamma}_n)}\left(\delta_i - m(\mathbf{W}_i, \hat{\theta}_n)\right)\left(Y_i^* - e^{\beta_0^\top\mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i)\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\left\{m(\mathbf{W}_i, \hat{\theta}_n)\left(Y_i^* - e^{\beta_0^\top\mathbf{X}_i}\right) + \left(1 - m(\mathbf{W}_i, \hat{\theta}_n)\right)h_{\beta_0}(Y_i^*, \mathbf{X}_i)\right\}, \\
&:= Q_n^{(1)}(\hat{\theta}_n, \hat{\gamma}_n) + Q_n^{(2)}(\hat{\theta}_n).
\end{aligned}$$

First, we consider the term $Q_n^{(1)}(\hat{\theta}_n, \hat{\gamma}_n)$. Let $\mathcal{Q}_i \equiv \mathbf{X}_i\xi_i(Y_i^* - e^{\beta_0^\top\mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i))$. We have:

$$\begin{aligned}
Q_n^{(1)}(\hat{\theta}_n, \hat{\gamma}_n) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi(\mathbf{W}_i, \hat{\gamma}_n)}(\delta_i - m(\mathbf{W}_i, \hat{\theta}_n))\mathcal{Q}_i, \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\pi(\mathbf{W}_i, \hat{\gamma}_n)} - \frac{1}{\pi(\mathbf{W}_i, \gamma^*)} + \frac{1}{\pi(\mathbf{W}_i, \gamma^*)}\right)(\delta_i - m(\mathbf{W}_i, \hat{\theta}_n))\mathcal{Q}_i, \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\pi(\mathbf{W}_i, \hat{\gamma}_n)} - \frac{1}{\pi(\mathbf{W}_i, \gamma^*)}\right)(\delta_i - m(\mathbf{W}_i, \hat{\theta}_n))\mathcal{Q}_i + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi(\mathbf{W}_i, \gamma^*)}(\delta_i - m(\mathbf{W}_i, \theta_0))\mathcal{Q}_i \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi(\mathbf{W}_i, \gamma^*)}(m(\mathbf{W}_i, \theta_0) - m(\mathbf{W}_i, \hat{\theta}_n))\mathcal{Q}_i \\
&:= Q_{n,1}^{(1)} + Q_{n,2}^{(1)} + Q_{n,3}^{(1)}.
\end{aligned}$$

Now, letting $Q_{n,1,\ell}^{(1)}$ and $\mathcal{Q}_{i,\ell}$ denote the ℓ-th component of the vectors $Q_{n,1}^{(1)}$ and $\mathcal{Q}_i$ respectively (for $\ell = 1, \ldots, p$), we have:

$$|Q_{n,1,\ell}^{(1)}| \leqslant \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\pi(\mathbf{W}_i, \gamma^*) - \pi(\mathbf{W}_i, \hat{\gamma}_n)}{\pi(\mathbf{W}_i, \hat{\gamma}_n)\pi(\mathbf{W}_i, \gamma^*)}\right||\delta_i - m(\mathbf{W}_i, \hat{\theta}_n)||\mathcal{Q}_{i,\ell}|.$$

Conditions C1 and C6 ensure that there exists a finite positive constant c_3 such that

$$|Q_{n,1,\ell}^{(1)}| \leqslant \frac{c_3}{n}\sum_{i=1}^{n}\left|\pi(\mathbf{W}_i, \gamma^*) - \pi(\mathbf{W}_i, \hat{\gamma}_n)\right|,$$

and the condition C7 implies that

$$\begin{aligned} |Q_{n,1,\ell}^{(1)}| &\leqslant \frac{c_3}{n}\sum_{i=1}^{n} g(\mathbf{W}_i)\|\gamma^* - \hat{\gamma}_n\|, \\ &\leqslant c_3(u + o_{\mathbb{P}}(1))\|\gamma^* - \hat{\gamma}_n\|. \end{aligned}$$

Finally, the convergence of $\hat{\gamma}_n$ to γ^* implies that $Q_{n,1,\ell}^{(1)}$ $(\ell = 1, \ldots, p)$, and thus $Q_{n,1}^{(1)}$, converge to 0 as $n \to \infty$. Similarly, under conditions C1 and C6, there exists a finite positive constant c_4 such that

$$|Q_{n,3,\ell}^{(1)}| \leqslant \frac{c_4}{n}\sum_{i=1}^{n}\left|m(\mathbf{W}_i, \theta_0) - m(\mathbf{W}_i, \hat{\theta}_n)\right|, \quad \ell = 1, \ldots, p,$$

and condition C4 implies

$$|Q_{n,3,\ell}^{(1)}| \leqslant c_4(v + o_{\mathbb{P}}(1))\|\theta_0 - \hat{\theta}_n\|.$$

If the model $m(\mathbf{W}_i, \theta)$ is correctly specified (that is, if $\hat{\theta}_n$ is consistent for θ_0), $Q_{n,3,\ell}^{(1)}$ $(\ell = 1, \ldots, p)$, and thus $Q_{n,3}^{(1)}$, converge to 0 as $n \to \infty$. Finally, by the law of large numbers, $Q_{n,2}^{(1)}$ converges in probability to

$$\mathbb{E}\left[\frac{1}{\pi(\mathbf{W}_i, \gamma^*)}(\delta_i - m(\mathbf{W}_i, \theta_0))\mathcal{Q}_i\right] = \mathbb{E}\left[\frac{\mathbf{X}_i(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i))}{\pi(\mathbf{W}_i, \gamma^*)}\mathbb{E}(\xi_i|\mathbf{W}_i)(\mathbb{E}(\delta_i|\mathbf{W}_i) - m(\mathbf{W}_i, \theta_0))\right],$$

which equals 0 if the model $m(\mathbf{W}_i, \theta)$ is correctly specified (in this case, $m(\mathbf{W}_i, \theta_0) = \mathbb{E}(\delta_i|\mathbf{W}_i)$). It follows that $Q_n^{(1)}(\hat{\theta}_n, \hat{\gamma}_n) = o_{\mathbb{P}}(1)$. Therefore,

$$\frac{1}{n}\breve{\ell}_n(\beta_0, \hat{\theta}_n, \hat{\gamma}_n) = Q_n^{(2)}(\hat{\theta}_n) + o_{\mathbb{P}}(1).$$

With obvious notations, we have $Q_n^{(2)}(\hat{\theta}_n) = Q_n^{(2)}(\hat{\theta}_n) - Q_n^{(2)}(\theta_0) + Q_n^{(2)}(\theta_0)$. By the law of large numbers, $Q_n^{(2)}(\theta_0)$ converges in probability to $\mathbb{E}[\mathbf{X}\{m(\mathbf{W}, \theta_0)(Y^* - e^{\beta_0^\top \mathbf{X}}) + (1 - m(\mathbf{W}, \theta_0))h_{\beta_0}(Y^*, \mathbf{X})\}]$, which is equal to 0 (see proof of Theorem 4.3.3). We also have

$$Q_n^{(2)}(\hat{\theta}_n) - Q_n^{(2)}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(m(\mathbf{W}_i, \hat{\theta}_n) - m(\mathbf{W}_i, \theta_0))(Y_i^* - e^{\beta_0^\top \mathbf{X}_i} - h_{\beta_0}(Y_i^*, \mathbf{X}_i)),$$

and using similar arguments as for $Q_{n,3}^{(1)}$, we can show that this converges to 0 if model $m(\mathbf{W}_i, \theta)$ is correctly specified. Finally, $Q_n^{(2)}(\hat{\theta}_n) = o_{\mathbb{P}}(1)$, which concludes the proof of condition *ii*.

Now, we prove that $n^{-1}\partial\breve{\ell}_n(\beta, \hat{\theta}_n, \hat{\gamma}_n)/\partial\beta^\top$ converges to $-\Sigma_1(\beta)$, uniformly in a neighborhood of β_0

(condition *iii*). Letting $\mathcal{Q}_{i,\beta} = (Y_i^* - e^{\beta^\top \mathbf{X}_i} - h_\beta(Y_i^*, \mathbf{X}_i))h_\beta(Y_i^*, \mathbf{X}_i)$, some easy calculations yield:

$$\begin{aligned}
\frac{1}{n}\frac{\partial \breve{\ell}_n(\beta,\theta,\gamma)}{\partial \beta^\top} &= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\left[-\breve{\delta}_i(\theta,\gamma)(e^{\beta^\top\mathbf{X}_i}+\mathcal{Q}_{i,\beta})+\mathcal{Q}_{i,\beta}\right], \\
&= -\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top(e^{\beta^\top\mathbf{X}_i}+\mathcal{Q}_{i,\beta})m(\mathbf{W}_i,\theta)+\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\mathcal{Q}_{i,\beta} \\
&\quad +\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\frac{\xi_i}{\pi(\mathbf{W}_i,\gamma)}(m(\mathbf{W}_i,\theta)-\delta_i)(e^{\beta^\top\mathbf{X}_i}+\mathcal{Q}_{i,\beta}).
\end{aligned}$$

Now, decompose $n^{-1}\partial\breve{\ell}_n(\beta,\hat{\theta}_n,\hat{\gamma}_n)/\partial\beta^\top$ as

$$\begin{aligned}
\frac{1}{n}\frac{\partial \breve{\ell}_n(\beta,\hat{\theta}_n,\hat{\gamma}_n)}{\partial \beta^\top} &= \frac{1}{n}\frac{\partial \breve{\ell}_n(\beta,\hat{\theta}_n,\hat{\gamma}_n)}{\partial \beta^\top}-\frac{1}{n}\frac{\partial \breve{\ell}_n(\beta,\theta_0,\gamma^*)}{\partial \beta^\top}+\frac{1}{n}\frac{\partial \breve{\ell}_n(\beta,\theta_0,\gamma^*)}{\partial \beta^\top}, \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top(e^{\beta^\top\mathbf{X}_i}+\mathcal{Q}_{i,\beta})(m(\mathbf{W}_i,\theta_0)-m(\mathbf{W}_i,\hat{\theta}_n)) \\
&\quad +\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\xi_i(e^{\beta^\top\mathbf{X}_i}+\mathcal{Q}_{i,\beta})\frac{m(\mathbf{W}_i,\hat{\theta}_n)-\delta_i}{\pi(\mathbf{W}_i,\hat{\gamma}_n)} \\
&\quad -\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\xi_i(e^{\beta^\top\mathbf{X}_i}+\mathcal{Q}_{i,\beta})\frac{m(\mathbf{W}_i,\theta_0)-\delta_i}{\pi(\mathbf{W}_i,\gamma^*)}+\frac{1}{n}\frac{\partial \breve{\ell}_n(\beta,\theta_0,\gamma^*)}{\partial \beta^\top}, \\
&\equiv T_n^{(1)}+T_n^{(2)}+T_n^{(3)}+\frac{1}{n}\frac{\partial \breve{\ell}_n(\beta,\theta_0,\gamma^*)}{\partial \beta^\top}.
\end{aligned}$$

Using similar arguments as for $Q_{n,3}^{(1)}$ (respectively $Q_n^{(1)}$ and $Q_{n,2}^{(1)}$), we can show that $T_n^{(1)}$ (respectively $T_n^{(2)}$ and $T_n^{(3)}$) converge to 0 as $n \to \infty$. Details are omitted. Now, $n^{-1}\partial\breve{\ell}_n(\beta,\theta_0,\gamma^*)/\partial\beta^\top$ converges in probability to $\mathbb{E}[\mathbf{X}_i\mathbf{X}_i^\top(-\breve{\delta}_i(\theta_0,\gamma^*)(e^{\beta^\top\mathbf{X}_i}+\mathcal{Q}_{i,\beta})+\mathcal{Q}_{i,\beta})]$. If the model $m(\mathbf{W}_i,\theta)$ is correctly specified (that is, $m(\mathbf{W}_i,\theta_0)=\mathbb{E}(\delta_i|\mathbf{W}_i)$), we have

$$\begin{aligned}
\mathbb{E}[\breve{\delta}_i(\theta_0,\gamma^*)|\mathbf{W}_i] &= \frac{\mathbb{E}[\xi_i|\mathbf{W}_i]\mathbb{E}[\delta_i|\mathbf{W}_i]}{\pi(\mathbf{W}_i,\gamma^*)}+\left(1-\frac{\mathbb{E}[\xi_i|\mathbf{W}_i]}{\pi(\mathbf{W}_i,\gamma^*)}\right)m(\mathbf{W}_i,\theta_0), \\
&= \mathbb{E}[\delta_i|\mathbf{W}_i],
\end{aligned}$$

thus

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\top\left(-\breve{\delta}_i(\theta_0,\gamma^*)(e^{\beta^\top\mathbf{X}_i}+\mathcal{Q}_{i,\beta})+\mathcal{Q}_{i,\beta}\right)\right] &= \mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\top\left(-\mathbb{E}[\delta_i|\mathbf{W}_i](e^{\beta^\top\mathbf{X}_i}+\mathcal{Q}_{i,\beta})+\mathcal{Q}_{i,\beta}\right)\right] \\
&= -\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\top\left(\delta_i e^{\beta^\top\mathbf{X}_i}+(\delta_i-1)\mathcal{Q}_{i,\beta}\right)\right] \\
&= -\Sigma_1(\beta).
\end{aligned}$$

It follows that $n^{-1}\partial\breve{\ell}_n(\beta,\hat{\theta}_n,\hat{\gamma}_n)/\partial\beta^\top$ converges in probability to $-\Sigma_1(\beta)$. Uniformity of the convergence follows by the same arguments as in the proof of Theorem 4.3.3.

Finally, having proved conditions *i*, *ii* and *iii*, we apply the inverse function theorem of Foutz (1977) and conclude that $\breve{\beta}_n$ converges in probability to β_0 if $m(\mathbf{W}_i,\theta)$ is correctly specified. The consistency proof of $\breve{\beta}_n$ when model $\pi(\mathbf{W}_i,\gamma)$ is correctly specified proceeds along the same lines and is omitted. □

Appendix D: Proof of Theorem 4.5.3

First, we have

$$\frac{\partial\breve{\delta}_i(\theta,\gamma)}{\partial\theta^\top}=\left(1-\frac{\xi_i}{\pi(\mathbf{W}_i,\gamma)}\right)\dot{m}^\top(\mathbf{W}_i,\theta)\quad\text{and}\quad\frac{\partial\breve{\delta}_i(\theta,\gamma)}{\partial\gamma^\top}=(m(\mathbf{W}_i,\theta)-\delta_i)\xi_i\frac{\dot{\pi}^\top(\mathbf{W}_i,\gamma)}{\pi^2(\mathbf{W}_i,\gamma)}.$$

Using this, it is straightforward to see that

$$\frac{1}{n}\frac{\partial\breve{\ell}_n(\beta_0,\theta^*,\gamma^*)}{\partial\theta^\top}\xrightarrow{\mathbb{P}}\Sigma_5(\beta_0,\theta^*,\gamma^*)\quad\text{and}\quad\frac{1}{n}\frac{\partial\breve{\ell}_n(\beta_0,\theta^*,\gamma^*)}{\partial\gamma^\top}\xrightarrow{\mathbb{P}}\Sigma_6(\beta_0,\theta^*,\gamma^*)\tag{4.7.1}$$

as $n\to\infty$ (calculations are omitted). Moreover, if the model $m(\mathbf{W}_i,\theta)$ is correctly specified (that is, $\theta^*=\theta_0$), then $\Sigma_6(\beta_0,\theta_0,\gamma^*)=0$. Similarly, if model $\pi(\mathbf{W}_i,\gamma)$ is correctly specified (and thus, $\gamma^*=\gamma_0$), then $\Sigma_5(\beta_0,\theta_0,\gamma^*)=0$. Now, taking Taylor's expansion of $\breve{\ell}_n(\breve{\beta}_n,\hat{\theta}_n,\hat{\gamma}_n)$ around $(\beta_0,\theta^*,\gamma^*)$ gives

$$\begin{aligned}\sqrt{n}(\breve{\beta}_n-\beta_0)\;=\;&\left(-\frac{1}{n}\frac{\partial\breve{\ell}_n(\beta_0,\theta^*,\gamma^*)}{\partial\beta^\top}\right)^{-1}\left(\frac{1}{\sqrt{n}}\breve{\ell}_n(\beta_0,\theta^*,\gamma^*)+\frac{1}{n}\frac{\partial\breve{\ell}_n(\beta_0,\theta^*,\gamma^*)}{\partial\theta^\top}\sqrt{n}(\hat{\theta}_n-\theta^*)\right.\\&\left.+\frac{1}{n}\frac{\partial\breve{\ell}_n(\beta_0,\theta^*,\gamma^*)}{\partial\gamma^\top}\sqrt{n}(\hat{\gamma}_n-\gamma^*)\right)+o_{\mathbb{P}}(1).\end{aligned}\tag{4.7.2}$$

Finally, combining (4.3.2), (4.5.1), (4.7.1) and (4.7.2) and using the limit central theorem yield the asymptotic distribution of $\sqrt{n}(\breve{\beta}_n-\beta_0)$ when either $m(\mathbf{W}_i,\theta)$ or $\pi(\mathbf{W}_i,\gamma)$ is correctly specified. Formulas for the asymptotic variance follow from easy albeit tedious calculations.

If both $m(\mathbf{W}_i,\theta)$ and $\pi(\mathbf{W}_i,\gamma)$ are correctly specified, $\Sigma_7(\beta_0,\theta_0,\gamma_0)=\Sigma_8(\beta_0,\theta_0,\gamma_0)=\Sigma_1(\beta_0)$ and the asymptotic variance of $\breve{\beta}_n$ reduces to $\Sigma_1^{-1}(\beta_0)$, which concludes the proof. □

estimator		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
full data	bias	-0.0097	-0.0001	0.0005	0.0009	0.0021	-0.0097	-0.0001	0.0005	0.0009	0.0021	-0.0097	-0.0001	0.0005	0.0009	0.0021
	SE	0.1092	0.0213	0.0459	0.0164	0.0267	0.1092	0.0213	0.0459	0.0164	0.0267	0.1092	0.0213	0.0459	0.0164	0.0267
	RMSE	0.1549	0.0307	0.0643	0.0232	0.0380	0.1549	0.0307	0.0643	0.0232	0.0380	0.1549	0.0307	0.0643	0.0232	0.0380
	CP	0.9571	0.9397	0.9510	0.9540	0.9581	0.9571	0.9397	0.9510	0.9540	0.9581	0.9571	0.9397	0.9510	0.9540	0.9581
CC	bias	0.0624	-0.0033	-0.0062	-0.0096	-0.0082	0.0624	-0.0033	-0.0062	-0.0096	-0.0082	0.0624	-0.0033	-0.0062	-0.0096	-0.0082
	SE	0.1235	0.0233	0.0500	0.0187	0.0295	0.1235	0.0233	0.0500	0.0187	0.0295	0.1235	0.0233	0.0500	0.0187	0.0295
	RMSE	0.1842	0.0334	0.0698	0.0281	0.0423	0.1842	0.0334	0.0698	0.0281	0.0423	0.1842	0.0334	0.0698	0.0281	0.0423
	CP	0.9326	0.9418	0.9571	0.9142	0.9540	0.9326	0.9418	0.9571	0.9142	0.9540	0.9326	0.9418	0.9571	0.9142	0.9540
RC	bias	-0.0062	0.0001	-0.0003	0.0004	0.0014	0.0256	0.0026	-0.0069	-0.0042	-0.0061	-0.0062	0.0001	-0.0003	0.0004	0.0014
	SE	0.1111	0.0217	0.0469	0.0167	0.0273	0.1095	0.0217	0.0470	0.0164	0.0268	0.1111	0.0217	0.0469	0.0167	0.0273
	RMSE	0.1568	0.0311	0.0653	0.0235	0.0386	0.1615	0.0316	0.0666	0.0242	0.0396	0.1568	0.0311	0.0653	0.0235	0.0386
	CP	0.9540	0.9438	0.9510	0.9540	0.9540	0.9387	0.9336	0.9428	0.9234	0.9305	0.9540	0.9438	0.9510	0.9540	0.9540
AIPW	bias	-0.0119	-0.0002	0.0008	0.0012	0.0026	-0.0133	-0.0002	0.0014	0.0016	0.0029	-0.0100	-0.0001	0.0005	0.0010	0.0021
	SE	0.1089	0.0213	0.0458	0.0162	0.0267	0.1058	0.0211	0.0453	0.0159	0.0260	0.1092	0.0213	0.0460	0.0164	0.0268
	RMSE	0.1557	0.0308	0.0646	0.0232	0.0383	0.1609	0.0316	0.0660	0.0243	0.0395	0.1557	0.0308	0.0648	0.0233	0.0383
	CP	0.9510	0.9397	0.9428	0.9499	0.9459	0.9152	0.9183	0.9275	0.9122	0.9142	0.9520	0.9397	0.9459	0.9520	0.9489
MI	bias	-0.0069	0.0000	-0.0001	0.0005	0.0015	0.0241	0.0024	-0.0065	-0.0040	-0.0058	-0.0069	0.0000	-0.0001	0.0005	0.0015
	SE	0.1091	0.0212	0.0457	0.0163	0.0267	0.1124	0.0219	0.0476	0.0168	0.0274	0.1091	0.0212	0.0457	0.0163	0.0267
	RMSE	0.1556	0.0308	0.0646	0.0233	0.0383	0.1633	0.0317	0.0671	0.0245	0.0399	0.1556	0.0308	0.0646	0.0233	0.0383
	CP	0.9520	0.9418	0.9479	0.9489	0.9489	0.9459	0.9397	0.9510	0.9356	0.9408	0.9520	0.9418	0.9479	0.9489	0.9489

Table 4.1: Simulation results for $n = 250$, censoring rate = 20%, missing rate = 20%. SE: average standard error. RMSE : root mean square error. CP: empirical coverage probability of 95%-level confidence intervals.

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
full data	bias	-0.0060	-0.0008	0.0014	0.0005	0.0013	-0.0060	-0.0008	0.0014	0.0005	0.0013	-0.0060	-0.0008	0.0014	0.0005	0.0013
	SE	0.0766	0.0150	0.0323	0.0115	0.0188	0.0766	0.0150	0.0323	0.0115	0.0188	0.0766	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1099	0.0215	0.0453	0.0162	0.0269	0.1099	0.0215	0.0453	0.0162	0.0269	0.1099	0.0215	0.0453	0.0162	0.0269
	CP	0.9460	0.9490	0.9500	0.9560	0.9420	0.9460	0.9490	0.9500	0.9560	0.9420	0.9460	0.9490	0.9500	0.9560	0.9420
CC	bias	0.0688	-0.0038	-0.0051	-0.0109	-0.0092	0.0688	-0.0038	-0.0051	-0.0109	-0.0092	0.0688	-0.0038	-0.0051	-0.0109	-0.0092
	SE	0.0866	0.0163	0.0350	0.0131	0.0207	0.0866	0.0163	0.0350	0.0131	0.0207	0.0866	0.0163	0.0350	0.0131	0.0207
	RMSE	0.1421	0.0238	0.0493	0.0213	0.0310	0.1421	0.0238	0.0493	0.0213	0.0310	0.1421	0.0238	0.0493	0.0213	0.0310
	CP	0.8676	0.9388	0.9519	0.8656	0.9188	0.8676	0.9388	0.9519	0.8656	0.9188	0.8676	0.9388	0.9519	0.8656	0.9188
RC	bias	-0.0022	-0.0006	0.0008	0.0000	0.0005	0.0311	0.0021	-0.0065	-0.0048	-0.0073	-0.0022	-0.0006	0.0008	0.0000	0.0005
	SE	0.0780	0.0153	0.0329	0.0117	0.0192	0.0769	0.0153	0.0330	0.0115	0.0188	0.0780	0.0153	0.0329	0.0117	0.0192
	RMSE	0.1114	0.0218	0.0463	0.0163	0.0273	0.1177	0.0220	0.0476	0.0174	0.0286	0.1114	0.0218	0.0463	0.0163	0.0273
	CP	0.9490	0.9490	0.9520	0.9570	0.9430	0.9100	0.9450	0.9330	0.9160	0.9120	0.9490	0.9490	0.9520	0.9570	0.9430
AIPW	bias	-0.0078	-0.0009	0.0020	0.0007	0.0017	-0.0069	-0.0008	0.0018	0.0009	0.0015	-0.0060	-0.0008	0.0017	0.0006	0.0013
	SE	0.0765	0.0150	0.0322	0.0114	0.0187	0.0747	0.0149	0.0319	0.0112	0.0183	0.0766	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1106	0.0217	0.0459	0.0161	0.0271	0.1151	0.0221	0.0474	0.0170	0.0280	0.1106	0.0217	0.0459	0.0162	0.0271
	CP	0.9410	0.9450	0.9430	0.9500	0.9390	0.8990	0.9280	0.9290	0.9150	0.9070	0.9420	0.9420	0.9450	0.9520	0.9420
MI	bias	-0.0026	-0.0006	0.0009	0.0000	0.0006	0.0301	0.0019	-0.0062	-0.0047	-0.0071	-0.0026	-0.0006	0.0009	0.0000	0.0006
	SE	0.0772	0.0150	0.0324	0.0115	0.0189	0.0805	0.0155	0.0340	0.0120	0.0196	0.0772	0.0150	0.0324	0.0115	0.0189
	RMSE	0.1109	0.0216	0.0460	0.0162	0.0272	0.1199	0.0222	0.0482	0.0177	0.0290	0.1109	0.0216	0.0460	0.0162	0.0272
	CP	0.9410	0.9450	0.9430	0.9460	0.9380	0.9340	0.9500	0.9470	0.9330	0.9300	0.9410	0.9450	0.9430	0.9460	0.9380

Table 4.2: Simulation results for $n = 500$, censoring rate = 20%, missing rate = 20%.

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
	bias	-0.0015	-0.0001	-0.0001	0.0006	0.0003	-0.0015	-0.0001	-0.0001	0.0006	0.0003	-0.0015	-0.0001	-0.0001	0.0006	0.0003
full data	SE	0.0765	0.0150	0.0323	0.0115	0.0188	0.0765	0.0150	0.0323	0.0115	0.0188	0.0765	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1100	0.0213	0.0447	0.0162	0.0270	0.1100	0.0213	0.0447	0.0162	0.0270	0.1100	0.0213	0.0447	0.0162	0.0270
	CP	0.9370	0.9450	0.9520	0.9540	0.9440	0.9370	0.9450	0.9520	0.9540	0.9440	0.9370	0.9450	0.9520	0.9540	0.9440
	bias	0.1224	0.0121	-0.0128	-0.0177	-0.0152	0.1224	0.0121	-0.0128	-0.0177	-0.0152	0.1224	0.0121	-0.0128	-0.0177	-0.0152
CC	SE	0.0993	0.0186	0.0391	0.0151	0.0233	0.0993	0.0186	0.0391	0.0151	0.0233	0.0993	0.0186	0.0391	0.0151	0.0233
	RMSE	0.1864	0.0289	0.0562	0.0275	0.0363	0.1864	0.0289	0.0562	0.0275	0.0363	0.1864	0.0289	0.0562	0.0275	0.0363
	CP	0.7500	0.8940	0.9400	0.7800	0.8920	0.7500	0.8940	0.9400	0.7800	0.8920	0.7500	0.8940	0.9400	0.7800	0.8920
	bias	0.0065	0.0004	-0.0021	-0.0006	-0.0013	0.0779	0.0010	-0.0173	-0.0113	-0.0181	0.0065	0.0004	-0.0021	-0.0006	-0.0013
RC	SE	0.0793	0.0154	0.0336	0.0119	0.0196	0.0764	0.0153	0.0335	0.0115	0.0187	0.0793	0.0154	0.0336	0.0119	0.0196
	RMSE	0.1134	0.0217	0.0465	0.0167	0.0281	0.1387	0.0220	0.0504	0.0204	0.0335	0.1134	0.0217	0.0465	0.0167	0.0281
	CP	0.9430	0.9480	0.9580	0.9490	0.9460	0.7750	0.9420	0.9200	0.8080	0.7920	0.9430	0.9480	0.9580	0.9490	0.9460
	bias	-0.0052	0.0000	0.0004	0.0010	0.0012	-0.0061	-0.0005	0.0008	0.0014	0.0013	-0.0017	0.0000	-0.0002	0.0007	0.0003
AIPW	SE	0.0765	0.0150	0.0322	0.0112	0.0188	0.0698	0.0149	0.0310	0.0105	0.0173	0.0765	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1115	0.0215	0.0455	0.0163	0.0276	0.1185	0.0225	0.0480	0.0177	0.0291	0.1113	0.0216	0.0455	0.0164	0.0275
	CP	0.9390	0.9400	0.9410	0.9470	0.9370	0.8501	0.9080	0.8925	0.8273	0.8635	0.9410	0.9410	0.9430	0.9550	0.9420
	bias	0.0056	0.0004	-0.0018	-0.0005	-0.0011	0.0763	0.0010	-0.0168	-0.0110	-0.0177	0.0056	0.0004	-0.0018	-0.0005	-0.0011
MI	SE	0.0777	0.0150	0.0326	0.0116	0.0191	0.0822	0.0156	0.0352	0.0125	0.0200	0.0777	0.0150	0.0326	0.0116	0.0191
	RMSE	0.1124	0.0215	0.0458	0.0165	0.0278	0.1412	0.0222	0.0514	0.0209	0.0340	0.1124	0.0215	0.0458	0.0165	0.0278
	CP	0.9440	0.9420	0.9440	0.9470	0.9420	0.8330	0.9450	0.9390	0.8690	0.8440	0.9440	0.9420	0.9440	0.9470	0.9420

Table 4.3: Simulation results for $n = 500$, censoring rate = 20%, missing rate = 40%.

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
full data	bias	0.0039	0.0002	-0.0005	-0.0002	-0.0007	0.0039	0.0002	-0.0005	-0.0002	-0.0007	0.0039	0.0002	-0.0005	-0.0002	-0.0007
	SE	0.0907	0.0180	0.0402	0.0144	0.0233	0.0907	0.0180	0.0402	0.0144	0.0233	0.0907	0.0180	0.0402	0.0144	0.0233
	RMSE	0.1285	0.0256	0.0573	0.0199	0.0327	0.1285	0.0256	0.0573	0.0199	0.0327	0.1285	0.0256	0.0573	0.0199	0.0327
	CP	0.9587	0.9518	0.9420	0.9676	0.9538	0.9587	0.9518	0.9420	0.9676	0.9538	0.9587	0.9518	0.9420	0.9676	0.9538
CC	bias	0.0764	-0.0033	-0.0062	-0.0116	-0.0115	0.0764	-0.0033	-0.0062	-0.0116	-0.0115	0.0764	-0.0033	-0.0062	-0.0116	-0.0115
	SE	0.1049	0.0202	0.0451	0.0168	0.0263	0.1049	0.0202	0.0451	0.0168	0.0263	0.1049	0.0202	0.0451	0.0168	0.0263
	RMSE	0.1668	0.0289	0.0645	0.0261	0.0387	0.1668	0.0289	0.0645	0.0261	0.0387	0.1668	0.0289	0.0645	0.0261	0.0387
	CP	0.8702	0.9440	0.9479	0.8899	0.9272	0.8702	0.9440	0.9479	0.8899	0.9272	0.8702	0.9440	0.9479	0.8899	0.9272
RC	bias	0.0221	0.0012	-0.0037	-0.0027	-0.0051	0.1838	0.0123	-0.0361	-0.0271	-0.0464	0.0221	0.0012	-0.0037	-0.0027	-0.0051
	SE	0.0960	0.0190	0.0429	0.0152	0.0250	0.0903	0.0192	0.0441	0.0142	0.0231	0.0960	0.0190	0.0429	0.0152	0.0250
	RMSE	0.1356	0.0269	0.0605	0.0208	0.0349	0.2279	0.0299	0.0720	0.0346	0.0577	0.1356	0.0269	0.0605	0.0208	0.0349
	CP	0.9548	0.9548	0.9469	0.9676	0.9587	0.4808	0.9036	0.8673	0.5152	0.5034	0.9548	0.9548	0.9469	0.9676	0.9587
AIPW	bias	-0.0036	-0.0004	0.0022	0.0010	0.0011	0.0199	0.0017	-0.0038	-0.0028	-0.0047	0.0046	0.0002	0.0003	-0.0002	-0.0010
	SE	0.0899	0.0178	0.0398	0.0138	0.0231	0.0681	0.0169	0.0366	0.0110	0.0173	0.0907	0.0180	0.0402	0.0144	0.0233
	RMSE	0.1296	0.0263	0.0589	0.0198	0.0332	0.1332	0.0273	0.0617	0.0216	0.0340	0.1302	0.0264	0.0589	0.0201	0.0333
	CP	0.9508	0.9292	0.9272	0.9489	0.9489	0.7443	0.8741	0.8348	0.7443	0.7345	0.9479	0.9292	0.9361	0.9626	0.9459
MI	bias	0.0203	0.0010	-0.0031	-0.0025	-0.0046	0.1781	0.0117	-0.0344	-0.0262	-0.0449	0.0203	0.0010	-0.0031	-0.0025	-0.0046
	SE	0.0933	0.0183	0.0412	0.0147	0.0241	0.1011	0.0199	0.0467	0.0162	0.0258	0.0933	0.0183	0.0412	0.0147	0.0241
	RMSE	0.1337	0.0264	0.0594	0.0205	0.0343	0.2281	0.0301	0.0728	0.0348	0.0577	0.1337	0.0264	0.0594	0.0205	0.0343
	CP	0.9430	0.9390	0.9381	0.9587	0.9508	0.5821	0.9145	0.9046	0.6332	0.5929	0.9430	0.9390	0.9381	0.9587	0.9508

Table 4.4: Simulation results for $n = 500$, censoring rate = 40%, missing rate = 20%.

	experiment 1					experiment 2				
	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
RC	3.5064	4.5296	3.5551	4.3983	3.6484	2.4176	3.3005	2.4370	3.0294	2.4596
MI	5.8060	7.7976	6.0465	6.9825	6.0021	3.9850	5.7020	4.3539	5.0137	4.1130
AIPW1	3.5505	4.6329	3.5394	4.5765	3.6785	2.4617	3.3519	2.4448	3.1441	2.5167
AIPW2	4.4184	4.6137	3.7427	5.2623	4.2667	2.8580	3.4715	2.5475	3.6508	2.7213
AIPW3	3.2736	4.4388	3.4035	4.2700	3.3965	2.2022	3.2673	2.3002	2.9241	2.2328
	experiment 3					experiment 4				
	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
RC	2.5299	3.2919	2.7640	3.1681	2.6332	3.1954	3.7548	3.4727	3.6572	3.3338
MI	4.1904	5.4829	4.4390	5.0557	4.2105	4.9915	5.6509	5.0467	5.6463	5.0811
AIPW1	2.5886	3.3626	2.7770	3.6558	2.7379	3.1904	4.0247	3.5270	4.1102	3.4390
AIPW2	6.2583	3.2878	3.7213	7.3639	5.5936	21.6887	5.0476	6.9504	20.6814	21.9974
AIPW3	2.2943	3.2301	2.4348	2.9428	2.3428	2.7317	3.5375	3.0144	3.3137	2.8459

Table 4.5: Relative errors (in %) of estimated standard deviations for the RC, MI and AIPW methods (with oracle standard deviations as reference values).

	experiment 1					experiment 2				
	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
RC	0.2761	0.0141	0.0512	0.0078	0.0173	0.1837	0.0098	0.0342	0.0052	0.0114
MI	0.4431	0.0227	0.0835	0.0117	0.0277	0.3012	0.0162	0.0584	0.0084	0.0186
AIPW1	0.2710	0.0138	0.0483	0.0078	0.0169	0.1801	0.0096	0.0328	0.0051	0.0111
AIPW2	0.3556	0.0133	0.0515	0.0091	0.0206	0.2327	0.0100	0.0353	0.0065	0.0128
AIPW3	0.2467	0.0129	0.0464	0.0072	0.0153	0.1615	0.0094	0.0309	0.0048	0.0099
	experiment 3					experiment 4				
	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
RC	0.2039	0.0100	0.0398	0.0059	0.0128	0.3637	0.0173	0.0829	0.0106	0.0261
MI	0.3242	0.0158	0.0599	0.0089	0.0198	0.5520	0.0241	0.1112	0.0155	0.0376
AIPW1	0.1927	0.0096	0.0371	0.0059	0.0122	0.3254	0.0163	0.0725	0.0098	0.0234
AIPW2	0.4568	0.0094	0.0531	0.0114	0.0249	0.9898	0.0185	0.1106	0.0248	0.0650
AIPW3	0.1694	0.0093	0.0323	0.0050	0.0105	0.2801	0.0146	0.0637	0.0085	0.0193

Table 4.6: Root mean square errors of the RC, MI and AIPW variance estimates (with oracle variances as reference values).

	CC			RC			MI			AIPW		
	est	se	p-value	est	se	p-value	est	se	p-value	est	se	p-value
constant	1.6372	0.0897	0.0000	1.6042	0.0785	0.0000	1.5980	0.0632	0.0000	1.6187	0.0766	0.0000
gender	-0.0721	0.0452	0.1107	-0.0699	0.0402	0.0818	-0.0701	0.0353	0.0470	-0.0715	0.0385	0.0636
age	0.0027	0.0013	0.0343	0.0027	0.0011	0.0163	0.0028	0.0009	0.0033	0.0026	0.0011	0.0239
single	-0.1175	0.0444	0.0082	-0.0979	0.0382	0.0104	-0.0944	0.0329	0.0041	-0.1039	0.0382	0.0065
GCSE/no qualif.	-0.2600	0.0535	0.0000	-0.2392	0.0477	0.0000	-0.2389	0.0405	0.0000	-0.2389	0.0455	0.0000
A-level or equiv.	-0.1196	0.0797	0.1335	-0.1182	0.0710	0.0961	-0.1195	0.0630	0.0577	-0.1236	0.0675	0.0669
more than enough	0.3749	0.0679	0.0000	0.3746	0.0610	0.0000	0.3753	0.0582	0.0000	0.3712	0.0574	0.0000
not enough	-0.5611	0.0546	0.0000	-0.5045	0.0478	0.0000	-0.5029	0.0424	0.0000	-0.5072	0.0459	0.0000

Table 4.7: Analysis results for the daily fruits and vegetables intake.

Bibliography

K. Adamids and S. Loukas. Ml estimation in the bivariate poisson distribution in the presence of missing values via the em algorithm. *Journal of Statistical Computation and Simulation*, **50**(12):163–172, 1994. doi: 10.1080/00949659408811608. URL https://doi.org/10.1080/00949659408811608%.

A. Antoniadis, J. Berruyer, and R. Carmona. *Régression non linéaire et applications*. Collection "Economie et statistiques avancées.": Série Ecole nationale de la statistique et de l'administration économique et Centre d'études des programmes économiques. Economica, 1992.

A. Argyriou, C. Micchelli, and M. Pontil. When is there a representer theorem? vector versus matrix regularizers. *J. Mach. Learn. Res.*, **10**:2507–2529, 2009.

B. Armstrong. Measurement error in generalized linear models. *Commun. Stat. - Theory Methods*, **14**(2):529–544, 1985. doi: 10.1080/03610918508812457. URL https://doi.org/10.1080/03610918508812457%.

A. M. Arnold and R. A. Kronmal. Multiple imputation of baseline data in the cardiovascular health study. *Am. J. Epidemiol.*, **157**(1):74–84, 2003. doi: 10.1093/aje/kwf156. URL https://doi.org/10.1093/aje/kwf156%.

N. Aronszajn. Theory of Reproducing Kernels. *Trans. Amer. Math. Soc.*, **68**(3):337–404, 1950. doi: 10.2307/1990404. URL https://doi.org/10.2307/1990404%.

N. Asin and J. Johannes. Adaptive nonparametric estimation in the presence of dependence. *J. Nonparametr. Stat.*, **29**(4):694–730, 2017. doi: 10.1080/10485252.2017.1367788. URL https://doi.org/10.1080/10485252.2017.1367788%.

G. Bakoyannis, F. Siannis, and G. Touloumi. Modelling competing risks data with missing cause of failure. *Biometrics*, **29**(30):3172–3185, 2010. doi: 10.1002/sim.4133. URL https://doi.org/10.1002/sim.4133%.

G. Bakoyannis, Y. Zhang, and C.T. Yiannoutsos. Semiparametric regression and risk prediction with competing risks data under missing cause of failure. *Lifetime Data Analysis*, **26**:659–684, 2020. doi: 10.1007/s10985-020-09494-1. URL https://doi.org/10.1007/s10985-020-09494-1%.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x. URL https://doi.org/10.1111/j.1541-0420.2005.00377.x.

Y. Baraud. Model selection for regression on a random design. *ESAIM: Probability and Statistics*, **6**: 127–146, 2002. doi: 10.1051/ps:2002007. URL https://doi.org/10.1051/ps:2002007%.

J. Barnard and X. L. Meng. Applications of multiple imputation in medical studies: from aids to nhanes. *Stat. Methods Med. Res*, **8**(1):17–36, 1999. doi: 10.1177/096228029900800103. URL https://doi.org/10.1177/096228029900800103%.

N. Batir. Inequalities for the gamma function. *Arch. Math*, **91**:554–563, 2008. doi: 10.1007/s00013-008-2856-9. URL https://doi.org/10.1007/s00013-008-2856-9%.

B. Bauer and M. Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.*, **47**(4):2261—2285, 2019. doi: 10.1214/18-AOS1747. URL https://doi.org/10.1214/18-AOS1747%.

A. Ben Saber and A. Karoui. Spectral analysis of some random pseudo-inverse based schemes for nonparametric and functional regressions estimators. 2021. Work in progress.

D. Benelmadani, K. Benhenni, and S. Louhichi. The reproducing kernel Hilbert space approach in nonparametric regression problems with correlated observations. *Ann. Inst. Stat. Math.*, **72**:1479–1500, 2019. doi: 10.1080/03610926.2019.1676442. URL https://doi.org/10.1007/s10463-019-00733-3.

L. Bermúdez and D. Karlis. A finite mixture of bivariate poisson regression models with an application to insurance ratemaking. *Computational Statistics & Data Analysis*, **56**(12):3988–3999, 2012. doi: 10.1016/j.csda.2012.05.016. URL https://doi.org/10.1016/j.csda.2012.05.016%.

A. Bényi and T. Oh. Modulation spaces, wiener amalgam spaces, and brownian motions. *Adv. Math.*, **228**(5):2943–2981, 2011. doi: 10.1016/j.aim.2011.07.023. URL https://doi.org/10.1016/j.aim.2011.07.023%.

T. E. Bodner. What improves with increased missing data imputations? *Struct. Equ. Modeling*, **15**(4):651–675, 2008. doi: 10.1080/10705510802339072. URL https://doi.org/10.1080/10705510802339072%.

A. Bonami and A. Karoui. Spectral Decay of Time and Frequency Limiting Operator. *Appl. Comput. Harmon. Anal.*, **42**(1):1–20, 2017. doi: 10.1016/j.acha.2015.05.003. URL https://doi.org/10.1016/j.acha.2015.05.003.

A. Bonami, P. Jaming, and A. Karoui. Non-Asymptotic Behaviour of the Sinc-Kernel Operator and Related Applications. 2018. URL https://arxiv.org/abs/1804.01257.

N.C. Brownstein, V. Bunn, L.M. Castro, and D. Sinha. Bayesian analysis of survival data with missing censoring indicators. *Biometrics*, **77**(1):305–315, 2021. doi: 10.1111/biom.13280. URL https://doi.org/10.1111/biom.13280%.

E. Brunel, F. Comte, and A. Guilloux. Nonparametric estimation for survival data with censoring indicators missing at random. *Journal of Statistical Planning and Inference*, **143**(10):1653–1671, 2013. doi: 10.1016/j.jspi.2013.04.010. URL https://doi.org/10.1016/j.jspi.2013.04.010%.

T. Cai and M. Yuan. Minimax and adaptive prediction for functional linear regression. *J. Amer. Statist. Assoc.*, **107**(499):1201–1216, 2012. doi: 10.1080/01621459.2012.716337. URL https://doi.org/10.1080/01621459.2012.716337%.

H. Cardot and J. Johannes. Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis*, **101**(2):395–408, 2010. doi: 10.1016/j.jmva.2009.03.001. URL https://doi.org/10.1016/j.jmva.2009.03.001%.

R. J. Carroll and L. A. Stefanski. Approximate quasilikelihood estimation in models with surrogate predictors. *J. Am. Stat. Assoc.*, **85**(411):652–663, 1990. doi: 10.2307/2290000. URL https://doi.org/10.2307/2290000%.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective (2nd ed.)*. Chapman and Hall/CRC, second edition, 2006. ISBN 978-1584886334. doi: 10.1201/9781420010138. URL https://doi.org/10.1201/9781420010138%.

B. A. Cattle, P. D. Baxter, D. C. Greenwood, C. P. Gale, and R. M. West. Multiple imputation for completion of a national clinical audit dataset. *Stat. Med.*, **30**(22):74–84, 2011. doi: 10.1002/sim.4314. URL https://doi.org/10.1002/sim.4314%.

S. B. Caudill and F. G. Mixon. Modeling household fertility decisions: Estimation and testing of censored regression models for count data. *Empirical Economics*, **20**(2):183–196, 1995. doi: 10.1007/BF01205434. URL https://doi.org/10.1007/BF01205434%.

X. Chen and J. Cai. Reweighted estimators for additive hazard model with censoring indicators missing at random. *Lifetime Data Analysis*, **24**(2):224–249, 2018. doi: 10.1007/s10985-017-9398-z. URL https://doi.org/10.1007/s10985-017-9398-z%.

A. Cohen, M. A. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Found. Comput. Math.*, **13**(5):819–834, 2013. doi: 10.1007/s10208-013-9142-3. URL https://doi.org/10.1007/s10208-013-9142-3%.

F. Comte and V. G. Catalot. Regression function estimation as a partly inverse problem. *Ann. Inst. Stat. Math.*, **72**:1023–1054, 2020. doi: 10.1007/s10463-019-00718-2. URL https://doi.org/10.1007/s10463-019-00718-2.

F. Comte and J. Johannes. Adaptive functional linear regression. *Ann. Statist.*, **40**(6):2765–2797, 2012. doi: 10.1214/12-AOS1050. URL https://doi.org/10.1214/12-AOS1050%.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, **39**(1):1–49, 2002. doi: 10.1090/S0273-0979-01-00923-5. URL https://doi.org/10.1090/S0273-0979-01-00923-5%.

P. De Jong and G. Z. Heller. *Generalized linear models for insurance data*. International Series on Actuarial Science. Cambridge University Press, 2008. doi: 10.1017/CBO9780511755408. URL https://doi.org/10.1017/CBO9780511755408%.

D. Degras. *Régression non paramétrique en présence de corrélation: Applications aux données fonctionnelles*. Presses Académiques Francophones, 2012. ISBN 9783838171760. URL https://books.google.fr/books?id=-vikMQEACAAJ.

F. Dell'Accio and F. Di Tommaso. Scattered data interpolation by shepard's like methods: classical results and recent advances. *Dolomites Research Notes on Approximation*, **9**:32–44, 2016.

J-F. Dupuy. *Statistical methods for the analysis of overdispersed count data*. ISTE Press, France, first edition, 2018. ISBN 9781785482663.

J.-F. Dupuy and E. Leconte. A study of regression calibration in a partially observed stratified cox model. *Journal of Statistical Planning and Inference*, **139**(2):317–328, 2009. doi: 10.1016/j.jspi.2008.04.024. URL https://doi.org/10.1016/j.jspi.2008.04.024%.

C. K. Enders. *Applied missing data analysis*. Guilford Press, New York, 2010.

L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.*, **13**(1):342–368, 1981. doi: 10.1214/aos/1176346597. URL https://doi.org/10.1214/aos/1176346597%.

F. Famoye and W. Wang. Censored generalized poisson regression model. *Computational Statistics & Data Analysis*, **46**(3):547–560, 2004. doi: 10.1016/j.csda.2003.08.007. URL https://doi.org/10.1016/j.csda.2003.08.007%.

S. Faria and G. Soromenho. Comparison of em and sem algorithms in poisson regression models: A simulation study. *Communications in Statistics - Simulation and Computation*, **41**(4):497–509, 2012. doi: 10.1080/03610918.2011.594534. URL https://doi.org/10.1080/03610918.2011.594534%.

R. V. Foutz. On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association*, **72**(357):147–148, 1977. doi: 10.1080/01621459.1977.10479926. URL https://doi.org/10.1080/01621459.1977.10479926.

B. Funke and C. Palmes. A note on estimating cumulative distribution functions by the use of convolution power kernels. *Stat. Probabil. Lett.*, **121**:90–98, 2017. doi: 10.1016/j.spl.2016.10.004. URL https://doi.org/10.1016/j.spl.2016.10.004%.

J. W. Graham. Missing data analysis: making it work in the real world. *Annu. Rev. Psychol.*, **60**(1): 549–576, 2009. doi: 10.1146/annurev.psych.58.110405.085530. URL https://doi.org/10.1146/annurev.psych.58.110405.085530%.

J. W. Graham, A. E. Olchowski, and T. D. Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prev. Sci.*, **8**(3):206–213, 2007. doi: 10.1007/s11121-007-0070-9. URL https://doi.org/10.1007/s11121-007-0070-9%.

B. Y. Guo. Jacobi Approximations in Certain Hilbert Spaces and Their Applications to Singular Differential Equations. *J. Math. Anal. Appl.*, **243**(2):373–408, 2000. doi: 10.1006/jmaa.1999.6677. URL https://doi.org/10.1006/jmaa.1999.6677.

X. Guo, C. Niu, Y. Yang, and W. Xu. Empirical likelihood for single index model with missing covariates at random. *Statistics*, **49**(3):588–601, 2015. doi: 10.1080/02331888.2014.881826. URL https://doi.org/10.1080/02331888.2014.881826.

D.B. Hall. Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, **56**(4):1030–1039, 2000. doi: 10.1111/j.0006-341X.2000.01030.x. URL https://doi.org/10.1111/j.0006-341X.2000.01030.x%.

P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. *Ann. Statist.*, **35**(1):70–91, 2007. doi: 10.1214/009053606000000957. URL https://doi.org/10.1214/009053606000000957.

J.W. Hardin, H. Schmiediche, and R.J. Carroll. The regression calibration method for fitting generalized linear models with additive measurement error. *The Stata Journal*, **3**(4):361–372, 2003. doi: 10.1177/1536867X0400300406. URL https://doi.org/10.1177/1536867X0400300406%.

N. J. Horton and S. R. Lipsitz. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am. Stat.*, **55**(3):244–254, 2001. doi: 10.1198/000313001317098266. URL https://doi.org/10.1198/000313001317098266%.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, **47**(260):663–685, 1952. doi: 10.2307/2280784. URL https://doi.org/10.2307/2280784.

B. V. Héraud. *Traitement des données manquantes en épidémiologie : application de l'imputation multiple à des données de surveillance et d'enquêtes*. Theses, Université Paris Sud - Paris XI, April 2012. URL https://tel.archives-ouvertes.fr/tel-00713926.

C.H. Hsu and M. Yu. Cox regression analysis with missing covariates via nonparametric multiple imputation. *Statistical Methods in Medical Research*, **28**(6):1676–1688, 2009. doi: 10.1177/0962280218772592. URL https://doi.org/10.1177/0962280218772592%.

S. Y. H. Huang. Regression calibration using response variables in linear models. *Statistica Sinica*, **15**:685–696, 2005.

M. D. Hughes. Regression dilution in the proportional hazards model. *Biometrics*, **49**(4):1056–1066, 1993. doi: 10.2307/2532247. URL https://doi.org/10.2307/2532247%.

J.G. Ibrahim, M.-H. Chen, S.R. Lipsitz, and A.H. Herring. Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*, **100**(469):332–346, 2005. doi: 10.1198/016214504000001844. URL https://doi.org/10.1198/016214504000001844%.

P. Jaming, A. Karoui, and S. Spektor. The approximation of almost time- and band-limited functions by their expansion in some orthogonal polynomials bases. *J. Approx. Theory*, **212**(3):41–65, 2016. doi: 10.1016/j.jat.2016.08.002. URL https://doi.org/10.1016/j.jat.2016.08.002.

K. J. Janssen, A. R. Donders, F. E. Jr. Harrell, Y. Vergouwe, Q. Chen, Grobbee D. E., and K. G. Moons. Missing covariate data in medical research: to impute is better than to ignore. *J. Clin. Epidemiol.*, **63**(7):589–595, 2010. doi: 10.1016/j.jclinepi.2009.12.008. URL https://doi.org/10.1016/j.jclinepi.2009.12.008%.

D. Karlis, P. Papatla, and S. Roy. Finite mixtures of censored poisson regression models. *Statistica Neerlandica*, **70**(2):100–122, 2016. doi: 10.1111/stan.12079. URL https://doi.org/10.1111/stan.12079%.

A. Karoui and A. Souabni. Generalized prolate spheroidal wave functions: Spectral analysis and approximation of almost band-limited functions. *J. Fourier Anal. Appl.*, **22**:383–412, 2016. doi: 10.1007/s00041-015-9420-3. URL https://doi.org/10.1007/s00041-015-9420-3%.

M. G. Kenward and J. Carpenter. Multiple imputation: current perspectives. *Stat. Methods Med. Res*, **16**(3):199–218, 2007. doi: 10.1177/0962280206075304. URL https://doi.org/10.1177/0962280206075304%.

K. Kleinke and J. Reinecke. Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica*, **67**(3):311–336, 2013. doi: 10.1111/stan.12009. URL https://doi.org/10.1111/stan.12009%.

K. Lange. *Elementary Optimization*. Springer, New York, NY, 2004. ISBN 978-1-4419-1910-6. doi: 10.1007/978-1-4757-4182-7_1. URL https://doi.org/10.1007/978-1-4757-4182-7_1%.

X. Liao, D.M. Zucker, Y. Li, and D. Spiegelman. Survival analysis with error-prone time-varying covariates: a risk set calibration approach. *Biometrics*, **67**(1):50–58, 2011. doi: 10.1111/j.1541-0420.2010.01423.x. URL https://doi.org/10.1111/j.1541-0420.2010.01423.x%.

S. R. Lipsitz, J. G. Ibrahim, and L. P. Zhao. A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Am. Stat. Assoc.*, **94**(448):1147–1160, 1999. doi: 10.2307/2669931. URL https://doi.org/10.2307/2669931.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, New York., 1987.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., second edition, 2002. ISBN 9780471183860. doi: 10.1002/9781119013563. URL https://doi.org/10.1002/9781119013563%.

J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.*, **23**(19):2937–2960, 2004. doi: 10.1002/sim.1903. URL https://doi.org/10.1002/sim.1903.

W.G. Madow, H. Nisselson, I. Olkin, D.B. Rubin, National Research Council (U.S.). Panel on Incomplete Data, Assembly of Behavioral, and Social Sciences (U.S.). Panel on Incomplete Data.

Incomplete Data in Sample Surveys: Theory and bibliographies. Incomplete Data in Sample Surveys. Academic Press, 1983. ISBN 9780123639028. doi: 10.1002/bimj.4710290615. URL https://doi.org/10.1002/bimj.4710290615%.

M. M. Mahmoud and M. M. Alderiny. On estimating parameters of censored generalized poisson regression model. *Applied Mathematical Sciences*, **4**(13-16):623–635, 2010.

P. McCullagh and J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman Hall, London, 1989. ISBN 0-412-31760-5. Second edition [of MR0727836].

I.W. McKeague and S. Subramanian. Product-limit estimators and cox regression with missing censoring information. *Scandinavian Journal of Statistics*, **25**(4):589–601, 1998. doi: 10.1111/1467-9469.00123. URL https://doi.org/10.1111/1467-9469.00123%.

O.S. Miettinen. *Theoretical epidemiology*. John Wiley & Sons, Inc., 1985. doi: 10.1002/sim.4780060213. URL https://doi.org/10.1002/sim.47800602136%.

G. Molenberghs and M. G. Kenward. *Missing data in clinical studies*. John Wiley & Sons, Inc., 2007. ISBN 9780470849811. doi: 10.1002/9780470510445. URL http://dx.doi.org/10.1002/9780470510445%.

A. Morisot. *Méthodes d'analyse de survie, valeurs manquantes et fractions attribuables temps dépendantes : application aux décès par cancer de la prostate*. Theses, Université Montpellier, December 2015. URL https://tel.archives-ouvertes.fr/tel-01408070.

J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society, Series A General*, **135**(3):370–384, 1972. doi: 10.2307/2344614. URL https://doi.org/10.2307/2344614%.

R. Neugebauer and M. J. van der Laan. Why prefer double robust estimators in causal inference? *J. Stat. Plan. Inference*, **129**(1-2):405–426, 2005. doi: 10.1016/j.jspi.2004.06.060. URL https://doi.org/10.1016/j.jspi.2004.06.060.

D. Nevo, R. Nishihara, S. Ogino, and M. Wang. The competing risks cox model with auxiliary case covariates under weaker missing-at-random cause of failure. *Lifetime Data Analysis*, **24**(3):425–442, 2018. doi: 10.1007/s10985-017-9401-8. URL https://doi.org/10.1007/s10985-017-9401-8%.

V. T. Nguyen and J.-F. Dupuy. Asymptotic results in censored zero-inflated poisson regression. *Communications in Statistics - Theory and Methods*, 2020. doi: 10.1080/03610926.2019.1676442. URL https://doi.org/10.1080/03610926.2019.1676442%. to appear.

U. Nur, N. T. Longford, and J. E. Cade. The impact of handling missing data on alcohol consumption estimates in the uk women cohort study. *Eur. J. Epidemiol.*, **24**(10):589–595, 2009. doi: 10.1007/s10654-009-9384-1. URL https://doi.org/10.1007/s10654-009-9384-1%.

F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark. NIST Handbook of Mathematical Functions. *Cambridge University Press, New York*, 2010.

J. Pollard, S. Kirk, and J. Cade. Factors affecting food choice in relation to fruit and vegetable intake: A review. *Nutrition Research Reviews*, **15**(2):373–387, 2002. doi: 10.1079/NRR200244. URL https://doi.org/10.1079/NRR200244%.

R. L. Prentice. Covariate measurement errors and parameter estimation in failure time regression models. *Biometrika*, **69**(2):331–342, 1982. doi: 10.1093/biomet/69.2.331. URL https://doi.org/10.1093/biomet/69.2.331%.

R. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, (Vienna, Austria), 2020. URL https://www.R-project.org/%.

T. E. Raghunathan. What do we do with missing data? some options for analysis of incomplete data. *Annu. Rev. Public Health*, **25**:99–117, 2004. doi: 10.1146/annurev.publhealth.25.102802.124410. URL https://doi.org/10.1146/annurev.publhealth.25.102802.124410%.

J. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag New York, second edition, 2005. ISBN 978-0-387-40080-8. doi: 10.1007/b98888. URL https://doi.org/10.1007/b98888%.

G. Reeves, D. Cox, S. Darby, and E. Whitley. Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Stat. Med.*, **17**(19):2157–2177, 1998. doi: 10.1002/(SICI)1097-0258(19981015)17:19<2157::AID-SIM916>3.0.CO;2-F. URL https://doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2157::AID-SIM916>3.0.CO;2-F%.

J. P. Reiter. Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika*, **94**(2):502–508, 2007. doi: 10.1093/biomet/asm028. URL https://doi.org/10.1093/biomet/asm028%.

J. M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. *American Statistical Association; Section on Bayesian Statistical Science, PROCEEDINGS- SECTION ON BAYESIAN STATISTICAL SCIENCE AMERICAN STATISTICAL ASSOCIATION*, : 6–10, 1999.

J. M. Robins and A. Rotnitzky. Comment on "inference for semiparametric models: Some questions and an answer.". *Stat. Sin.*, **11**:920–936, 2001. doi: 10.1016/j.jspi.2004.06.060. URL https://doi.org/10.1016/j.jspi.2004.06.060.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.*, **89**(427):846–866, 1994. doi: 10.2307/2290910. URL https://doi.org/10.2307/2290910.

J. M. Robins, A. Rotnitzky, and M. van der Laan. Comment on "on profile likelihood.". *J. Am. Stat. Assoc.*, **95**(450):477–482, 2000. doi: 10.2307/2669391. URL https://doi.org/10.2307/2669391%.

B. Rosner, W. C. Willett, and D. Spiegelman. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat. Med.*, **8**(9):1051–1069, 1989. doi: 10.1002/sim.4780080905. URL https://doi.org/10.1002/sim.4780080905%.

P. Royston. Multiple imputation of missing values. *Stata. J.*, **27**:227–241, 2004. doi: 10.1177/1536867X0400400301. URL https://doi.org/10.1177/1536867X0400400301%.

D. B. Rubin. Inference and missing data. *Biometrika*, **63**(3):581—-592, 1976. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581%.

D. B. Rubin. Multiple imputations in sample surveys — a phenomenological bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association,*, pages 20–34, 1978b.

D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1987. ISBN 9780470316696. doi: 10.1002/9780470316696. URL https://doi.org/10.1002/9780470316696%.

D. B. Rubin. Multiple imputation after 18+ years. *Psychol. Methods*, **91**(434):473–489, 1996. doi: 10.2307/2291635. URL https://doi.org/10.2307/2291635%.

D. B. Rubin and N. Schenker. Multiple imputation in health-care databases: an overview and some applications. *Stat. Med.*, **10**(4):585–598, 1991. doi: 10.1002/sim.4780100410. URL https://doi.org/10.1002/sim.4780100410%.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. *NIPS*, **22**:1657–1665, 2016.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, first edition, 2003. ISBN 978-0521785167. doi: 10.1017/CBO9780511755453. URL https://doi.org/10.1017/CBO9780511755453%.

S. E. Saffari and R. Adnan. Zero-inflated poisson regression models with right censored count data. *Matematika*, **27**(1):21–29, 2011. doi: 10.17265/2161-6221/2011.09.020. URL https://doi.org/10.17265/2161-6221/2011.09.020%.

J. L. Schafer. *Analysis of incomplete multivariate data*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, first edition, 1997. ISBN 9780367803025. doi: 10.1201/9781439821862. URL http://dx.doi.org/10.1201/9781439821862%.

J. L. Schafer and J. W. Graham. Missing data: Our view of the state of the art. *Psychol. Methods*, **7**(2):147–177, 2002. doi: 10.1037/1082-989X.7.2.147. URL https://doi.org/10.1037/1082-989X.7.2.147%.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Stat. Assoc.*, **94**(448):1135–1146, 1999. doi: 10.2307/2669923. URL https://doi.org/10.2307/2669923%.

S. R. Seaman and I. R. White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, **22**(3):278–295, 2013. doi: 10.1177/0962280210395740. URL https://doi.org/10.1177/0962280210395740%.

H. Shin and S. Lee. An RKHS approach to robust functional linear regression. *Stat. Sinica*, **26**(1):255–272, 2016.

S. Sinharay, H. S. Stern, and D. Russell. The use of multiple imputation for the analysis of missing data. *Psychol. Methods*, **6**(4):317–329, 2001. doi: 10.1037/1082-989X.6.4.317. URL https://doi.org/10.1037/1082-989X.6.4.317%.

D. Slepian and H. O. Pollak. Prolate spheroidal wave functions, Fourier analysis and uncertainty I. *Bell System Tech. J.*, **40**(1):43–64, 1961. doi: 10.1002/j.1538-7305.1961.tb03976.x. URL https://doi.org/10.1002/j.1538-7305.1961.tb03976.x.

S. Smale and D. X. Zhou. Shannon sampling II: Connections to learning theory. *Appl. Comput. Harmon. Anal.*, **19**(3):285–302, 2005. doi: 10.1016/j.acha.2005.03.001. URL https://doi.org/10.1016/j.acha.2005.03.001.

S. Smale and D. X. Zhou. Learning Theory Estimates via Integral Operators and Their Approximations. *Constr. Approx.*, **26**(2):153—-172, 2007. doi: 10.1007/s00365-006-0659-y. URL https://doi.org/10.1007/s00365-006-0659-y.

D. Spiegelman, A. McDermott, and Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am. J. Clin. Nutr.*, **65**(4):1179S–1186S, 1997. doi: 10.1093/ajcn/65.4.1179S. URL https://doi.org/10.1093/ajcn/65.4.1179S%.

J. A. Steingrimsson and R. L. Strawderman. Estimation in the semiparametric accelerated failure time model with missing covariates: improving efficiency through augmentation. *Journal of the American Statistical Association*, **112**(519):1221–1235, 2017. doi: 10.1080/01621459.2016.1205500. URL https://doi.org/10.1080/01621459.2016.1205500%.

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer-Verlag New York, first edition, 2008. doi: 10.1007/978-0-387-77242-4. URL https://doi.org/10.1007/978-0-387-77242-4%.

J. C. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**(4):1040–1053, 1982. doi: 10.1214/aos/1176345969. URL https://doi.org/10.1214/aos/1176345969%.

S. Subramanian. Survival analysis for the missing censoring indicator model using kernel density estimation techniques. *Statistical methodology*, **3**(2):125–136, 2006. doi: 10.1016/j.stamet.2005.09.014. URL https://doi.org/10.1016/j.stamet.2005.09.014%.

S. Subramanian. Multiple imputations and the missing censoring indicator model. *Journal of Multivariate Analysis*, **102**(1):105–117, 2011. doi: 10.1016/j.jmva.2010.08.005. URL https://doi.org/10.1016/j.jmva.2010.08.005%.

Y. Sun, X. Qian, Q. Shou, and P. B. Gilbert. Analysis of two-phase sampling data with semiparametric additive hazards models. *Lifetime Data Analysis*, **23**(3):377–399, 2017. doi: 10.1007/s10985-016-9363-2. URL https://doi.org/10.1007/s10985-016-9363-2%.

J. M. Taylor, K. L. Cooper, J. T. Wei, A. V. Sarma, T. E. Raghunathan, and S. G. Heeringa. Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among african-american men. *Am. J. Epidemiol.*, **8**(1):774–782, 2002. doi: 10.1093/aje/kwf110. URL https://doi.org/10.1093/aje/kwf110%.

J. V. Terza. A tobit-type estimator for the censored poisson regression model. *Economics Letters*, **18**(4):361–365, 1985. doi: 10.1016/0165-1765(85)90053-9. URL https://doi.org/10.1016/0165-1765(85)90053-9%.

J. A. Tropp. An Introduction to Matrix Concentration Inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015. doi: 10.1561/2200000048. URL http://dx.doi.org/10.1561/2200000048.

A. A. Tsiatis. *A Semiparametric Theory and Missing Data*. Springer New York, 2007. doi: 10.1007/0-387-37345-4. URL https://doi.org/10.1007/0-387-37345-4%.

S. Van Buuren, H. C. Boshuizen, and D. L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat. Med.*, **18**(6):681–694, 1999. doi: 10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R. URL https://doi.org/10.1002(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R%.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer-Verlag New York, first edition, 2003. ISBN 978-0-387-21700-0. doi: 10.1007/978-0-387-21700-0. URL https://doi.org/10.1007/978-0-387-21700-0%.

M. J. Van Der Lann and I. W. McKeague. Efficient estimation from right-censored data when failure indicators are missing at random. *Annals of Statistics*, **26**(1):164–182, 1998. doi: 10.1214/aos/1030563981. URL https://doi.org/10.1214/aos/1030563981%.

A. W. van der Vaart. *Asymtotics Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000. doi: 10.1017/CBO9780511802256. URL https://doi.org/10.1017/CBO9780511802256%.

D. Vergouw, M. W. Heymans, G. M. Peat, T. Kuijpers, P. R. Croft, H. C. de Vet, H. E. van der Horst, and D. A. van der Windt. The search for stable prognostic models in multiple imputed data sets. *BMC Med. Res. Methodol.*, **10**:81, 2010. doi: 10.1186/1471-2288-10-81. URL https://doi.org/10.1186/1471-2288-10-81%.

C. Y. Wang and H. Y. Chen. Augmented inverse probability weighted estimator for cox missing covariate regression. *Biometrics*, **57**(2):414–419, 2001. doi: 10.1111/j.0006-341x.2001.00414.x. URL https://doi.org/10.1111/j.0006-341x.2001.00414.x%.

C.Y. Wang, L. Hsu, Z.D. Feng, and R.L. Prentice. Regression calibration in failure time regression. *Biometrics*, **53**(1):131–145, 1997. doi: 10.2307/2533103. URL https://doi.org/10.2307/2533103%.

Q. Wang and G. E. Dinse. Linear regression analysis of survival data with missing censoring indicators. *Lifetime data analysis*, **17**(2):256–279, 2011. doi: 10.1007/s10985-010-9175-8. URL https://doi.org/10.1007/s10985-010-9175-8%.

Q. Wang and J. Shen. Estimation and confidence bands of a conditional survival function with censoring indicators missing at random. *Journal of Multivariate Analysis*, **99**(5):928–3948, 2008. doi: 10.1016/j.jmva.2007.06.002. URL https://doi.org/10.1016/j.jmva.2007.06.002%.

Q. Wang, G. E. Dinse, and C. Liu. Hazard function estimation with cause-of-death data missing at random. *Annals of the Institute of Statistical Mathematics*, **64**(2):415–438, 2012. doi: 10.1007/s10463-010-0317-2. URL https://doi.org/10.1007/s10463-010-0317-2%.

E.A. Weller, D.K. Milton, E.A. Eisen, and D. Spiegelman. Regression calibration for logistic regression with multiple surrogates for one exposure. *Journal of Statistical Planning and Inference*, **137**(2):449–461, 2007. doi: 10.1016/j.jspi.2006.01.009. URL https://doi.org/10.1016/j.jspi.2006.01.009%.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, **50**(1):1–25, 2012. doi: 10.2307/1912526. URL https://doi.org/10.2307/1912526%.

I. White and P. Royston. Imputing missing covariate values for the cox model. *Statistics in Medicine*, **28**(15):1982–1998, 2009. doi: 10.1002/sim.3618. URL https://doi.org/10.1002/sim.3618%.

A. M. Wood, I. R. White, M. Hillsdon, and J. Carpenter. Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. *Int. J. Epidemiol.*, **34**(1):89–99, 2005. doi: 10.1093/ije/dyh297. URL https://doi.org/10.1093/ije/dyh297%.

F.-C. Xie and B.-C. Wei. Diagnostics analysis in censored generalized poisson regression model. *Journal of Statistical Computation and Simulation*, **77**(8):1–25, 2007. doi: 10.1080/10629360600581316. URL https://doi.org/10.1080/10629360600581316%.

Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *Ann. Statist.*, **45**(3):991–1023, 2017. doi: 10.1214/16-AOS1472. URL https://doi.org/10.1214/16-AOS1472.

M. Yuang and T. T. Cai. A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression. *Ann. Stat.*, **38**(6):3412–3444, 2010. doi: 10.1214/09-AOS772. URL https://doi.org/10.1214/09-AOS772.

M. Zheng, R. Lin, and W. Yu. Competing risks data analysis under the accelerated failure time model with missing cause of failure. *Annals of the Institute of Statistical Mathematics*, **68**(4):855–876, 2016. doi: 10.1007/s10463-015-0516-y. URL https://doi.org/10.1007/s10463-015-0516-y%.

Y. Zou, G. Fan, and R. Zhang. Quantile regression and variable selection for partially linear single-index models with missing censoring indicators. *Journal of Statistical Planning and Inference*, **204**: 80–95, 2020. doi: 10.1016/j.jspi.2019.04.008. URL https://doi.org/10.1016/j.jspi.2019.04.008%.

Titre : Problèmes d'estimation dans des modèles de régression non paramétrique et de Poisson.

Mots clés : Régression non paramétrique, régression linéaire fonctionnelle, analyse des données de comptage, données manquantes, propriétés asymptotiques, simulations.

Résumé: Cette thèse a deux contributions principales. La première contribution porte sur l'étude des estimateurs basés sur la projection aléatoire pour résoudre des problèmes de régression non paramétrique ou de régression fonctionnelle linéaire. Plus précisément, les noyaux de projection que nous considérons dans la construction de nos estimateurs sont donnés par le noyau des polynômes de Gegenbauer, Christoffel-Darboux et le noyau de convolution Sinc. En particulier, nous fournissons des analyses d'erreur et de convergence des estimateurs proposés sous certaines hypothèses de régularité sur la classe des fonctions de régression. Nous étudions également un estimateur basé sur la projection orthogonale pour une résolution stable d'un problème de régression fonctionnelle linéaire. Ce problème est résolu sous l'hypothèse habituelle que la fonction de pente inconnue à estimer est bien approximée par sa projection sur un sous-espace de dimension finie d'un espace de Hilbert. Enfin, nous menons une étude de simulation pour évaluer les propriétés de ces estimateurs.
La deuxième contribution de ce travail concerne la régression de poisson censurée en présence d'indicatrices de censure manquantes.

Notons que la régression de Poisson est largement utilisée pour étudier la relation entre un ensemble des covariables et une variable de comptage. Nous considérons la situation où la variable de comptage observée peut être censurée aléatoirement à droite (par exemple, dans une étude sur l'utilisation des soins de santé, les patients déclarant leur nombre de visites chez un médecin comme « 8 visites ou plus » fournissent un nombre censuré, c'est-à-dire, une borne inférieure sur le vrai nombre non observé). La littérature sur l'analyse des données de comptage censuré contient déjà plusieurs approches pour traiter de telles données, nous supposons en outre que les indicatrices de censure, qui indiquent si un comptage observé est censuré ou non, sont manquantes pour certains individus de l'échantillon. Pour cette deuxième contribution, nous proposons plusieurs méthodes d'estimation : regression calibration, imputation multiple et estimation AIPW (pondération par l'inverse de la probabilité de sélection augmentée). La consistance, la normalité asymptotique et l'estimation de la variance de nos estimateurs sont rigoureusement établies sous des conditions de régularité appropriées. Une étude de simulation comparant ces différentes méthodes est décrite.

Titre : Some estimation problems in nonparametric regression and Poisson regression models.

Key words : nonparametric regression, linear functional regression, count data analysis, missing data, asymptotic properties, simulations.

Abstract: This thesis has two main contributions. The first contribution deals with study of random projection based estimators for solving nonparametric regression or linear functional regression problems. More precisely, the projection kernels we consider in the construction of our estimators are given by the Gegenbauer polynomials Christoffel-Darboux kernel and the convolution Sinc- kernel. In particular, we provide error and convergence analyses of the proposed estimators under some regularity assumptions on the class of regression functions. Also, we study an orthogonal projection based estimator for the fast and stable solution of a linear functional regression (LFR) problem. This problem is solved under the usual assumption that the unknown slope function to be estimated is well approximated by its projection on a finite dimensional subspace of a Hilbert space. Finally, we conduct a simulation study to assess finite-sample properties of these estimators.
The second contribution of this work deals with censored count data regression with missing censoring information.

Note that Poisson regression is widely used to investigate the relationship between covariates and acountresponse. We consider the situation where the count of interest is randomly right-censored (for example,in a study of health-care utilization, patients reporting their number of visits to a doctor as“8 visits or more” provide a censored count, that is, a lower bound on the true unobserved count). The literature on censored count data analysis already contains several approaches for handling such data, we additionally suppose that the censoring indicator, which tells whether an observed count is censored or not, is missing for some subjects. For this second contribution, we propose several estimators based on the regression calibration, multiple imputation and augmented inverse probability weighting methods. Consistency, asymptotic normality and variance estimation of our estimators are rigorously established under appropriate regularity conditions. Simulation experiments are carried out to investigate the finite sample behaviour and relative performance of the proposed estimates.