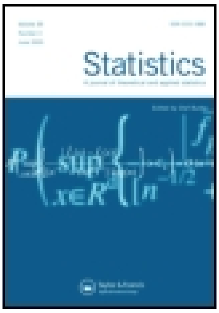

Statistics: A Journal of Theoretical and Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gsta20

Mean estimation with data missing at random for functional covariables

Frédéric Ferraty [a], Mariela Sued [b] & Philippe Vieu [a]

[a] Institut de Mathématiques, Université Paul Sabatier, Toulouse, France

[b] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Argentina

Available online: 24 Jan 2012

Statistics, iFirst, 2012, 1–19

Taylor & Francis
Taylor & Francis Group

Mean estimation with data missing at random for functional covariables

Frédéric Ferraty[a], Mariela Sued[b]* and Philippe Vieu[a]

[a] *Institut de Mathématiques, Université Paul Sabatier, Toulouse, France;* [b] *Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Argentina*

(*Received 3 October 2010; final version received 13 December 2011*)

In a missing-data setting, we want to estimate the mean of a scalar outcome, based on a sample in which an explanatory variable is observed for every subject while responses are missing by happenstance for some of them. We consider two kinds of estimates of the mean response when the explanatory variable is functional. One is based on the average of the predicted values and the second one is a functional adaptation of the Horvitz–Thompson estimator. We show that the infinite dimensionality of the problem does not affect the rates of convergence by stating that the estimates are root-*n* consistent, under missing at random (MAR) assumption. These asymptotic features are completed by simulated experiments illustrating the easiness of implementation and the good behaviour on finite sample sizes of the method. This is the first paper emphasizing that the insensitiveness of averaged estimates, well known in multivariate non-parametric statistics, remains true for an infinite-dimensional covariable. In this sense, this work opens the way for various other results of this kind in functional data analysis.

Keywords: averaged non-parametric estimates; missing at random; functional covariable; non-parametric functional kernel regression; root-*n* consistency

1. Introduction

Consider an observational study where, for every subject in the sample, we always observed a baseline variable $\mathbf{X}$, while a scalar response Y is missing by happenstance on some individuals. Our interest lies in estimating $\mu = \mathbb{E}[Y]$, the mean of the response variable. Without further assumptions, μ is not identified from the distribution of the observed data and so, no hope for estimating it from such a sample. In order to make the parameter identified, missing at random (MAR) is assumed. This hypothesis establishes that the value of the response does not provide additional information, on top of that given by the explanatory variables, to predict whether an individual will present a missing response (see [1]). To be more precise, one introduces a binary variable A such that $A = 1$ whenever a response is observed. The MAR assumption states that Y and A are conditional-independent, given $\mathbf{X}$.

*Corresponding author. Email: msued@dm.uba.ar

ISSN 0233-1888 print/ISSN 1029-4910 online

http://dx.doi.org/10.1080/02331888.2011.650172
http://www.tandfonline.com

There are different ways of estimating μ under MAR modelling. The first one is based on the fact that one can write

$$\mathbb{E}[Y|\mathbf{X}=\chi] = r(\chi) = \mathbb{E}[Y|(A=1,\mathbf{X}=\chi)]$$

and thus

$$\mu = \mathbb{E}[Y] = \mathbb{E}[r(\mathbf{X})]$$

may be expressed in terms of the distribution of the observed data. This leads to estimates of μ based on average of predicted values, which have been widely studied in the multivariate case (that is, when $\mathbf{X} \in \mathbb{R}^p$). Some authors postulate a parametric model for the regression function, while others prefer a non-parametric setting for estimating the regression function. For instance, Cheng [2] obtains the $\sqrt{n}$-asymptotic normality of such estimator using a non-parametric kernel regression estimate of r. An alternative way is based on the so-called propensity score

$$\pi(\chi) = \mathbb{E}[A|\mathbf{X}=\chi], \tag{1}$$

since under MAR assumption one can also write the parameter μ to be estimated as

$$\mu = \mathbb{E}\left[\frac{AY}{\pi(\mathbf{X})}\right].$$

In the standard multivariate problem, this representation has inspired the Horvitz–Thompson family of estimators. At this point, there exists a vast literature on the subject postulating parametric models for π, and also many authors have estimated the propensity score non-parametrically. For instance, Hirano *et al.* [3] proved $\sqrt{n}$-asymptotic normality when the propensity score is estimated by the Series Logit method.

Following the parametric approaches, it is necessary to assume in advance a working model either for the regression function or for the propensity score in order to get a consistent procedure for estimating μ. There is a third approach (doubly protected) that gives rise to a consistent estimator if at least one of the considered models is well specified, without knowing in advance which of them is correct. A recent survey and discussion on these three approaches can be found in Kan and Schafer [4] and Robins *et al.* [5].

The aim of this work is to look at what happens in situations in which the explanatory variable $\mathbf{X}$ takes values in some abstract space, being possibly of infinite dimension, in order to allow for applications involving functional data. Functional data analysis (FDA) is a field of research in statistics, which has been popularized by Ramsay and Silverman [6,7], is becoming more and more active. The recent advances can be found in special issues of various statistical journals (see, for instance [8–11]) or in the handbook by Ferraty and Romain [12]. The recent studies developed by Ferraty and Vieu [13–15] have opened the path to a wide literature on non-parametric modelling of functional statistical problems (see [16,17], for large sets of references).

This work takes part in the recent advances on non-parametric functional statistics by considering MAR modelling in a functional variable setting. More precisely, functional adaptations of the standard multivariate estimates are introduced in Section 2 and $\sqrt{n}$-consistency properties are stated in Section 3. Some simulations are reported in Section 4 discussing the interest of the method on finite samples. Section 5 is devoted to comments. Section 5.2 discusses how our hypotheses in the general functional framework are related with those usually assumed in the literature for multivariate data. Section 5.3 discusses some possible extensions of our results with special attention to asymptotic normality. It will also be discussed in Section 5.3 how our methodology extends directly to general mean functionals with obvious applications to moments and variance estimations.

Finally, it is worth stressing that this is the first paper showing that averaging non-parametric estimates is unsensitive to the dimensionality of the problem, since the parametric $\sqrt{n}$-rate is available even for infinite-dimensional variables. In this sense our guess is that this paper could open the way to several other results of this form in various problems involving functional data. The impact of this paper should therefore exceed the problem considered here (see discussion in Section 5.4).

2. Functional MAR modelling and estimates

2.1. *The model*

Consider a scalar response $Y \in \mathbb{R}$ and a baseline variable $\mathbf{X} \in \mathcal{F}$, where $\mathcal{F}$ is an abstract space dotted with a semi-metric $d(\cdot,\cdot)$ being possibly of infinite dimension. Let A be a binary variable, indicating that Y is observed when $A = 1$. While the randomness of the problem involves three variables $O = (\mathbf{X}, Y, A)$, one observes only

$$\tilde{O} = \begin{cases} \mathbf{X}, Y, A & \text{if } A = 1, \\ \mathbf{X}, A & \text{if } A = 0. \end{cases}$$

In the following, we denote by $O_i = (\mathbf{X}_i, Y_i, A_i)$, a sample of n independent variables distributed like O, but the purpose is to estimate $\mu = \mathbb{E}[Y]$ from the corresponding observed data $\{\tilde{O}_i; i = 1, \ldots, n\}$. This can be done by introducing the standard MAR hypothesis:

$$Y \text{ is independent of } A, \quad \text{given } \mathbf{X}. \tag{2}$$

Averaging directly observed responses would provide a consistent estimator of μ if A and Y were independent. In the present setting different approaches could be used. In Sections 2.2 and 2.3, we construct functional versions of two kinds of estimates which are classically used under MAR assumption with finite-dimensional covariables.

2.2. *Estimates based on averaged predicted values*

The main idea is the following: the MAR assumption implies that, given the explanatory variable, the conditional distribution of the response remains the same, regardless of the fact that the response is also observed:

$$Y|\mathbf{X} \sim Y|(\mathbf{X}, A = 1).$$

Then, one has

$$r(\chi) = \mathbb{E}[Y|\mathbf{X} = \chi] = \mathbb{E}[Y|(\mathbf{X} = \chi, A = 1)], \tag{3}$$

and the fact that $\mu = \mathbb{E}[Y] = \mathbb{E}[r(\mathbf{X})]$ with a fully observed variable $\mathbf{X}$ suggests us to estimate μ by averaging predicted values. Prediction will be done by considering a non-parametric estimator of the nonlinear regression operator r, in which a standard leave-one-out technique can be used to predict the response for the ith subject in the sample. Precisely, for each statistical unit i, a non-parametric estimate $\hat{r}_{-i}$ for the regression function based on observed data excluding the ith observation is considered to predict the response of the ith subject. Then we estimate μ with

$$\hat{\mu}_{\text{Reg}} = \frac{1}{n} \sum_{i=1}^{n} \hat{r}_{-i}(\mathbf{X}_i). \tag{4}$$

At this stage, one could think of using any kind of non-parametric functional estimate of r. Since the available theoretical background in this setting is mainly developed for Nadaraya–Watson kernel estimates (see [12], for wide discussion), we restrict our purpose to this kind of estimate. So, in the following, we will consider:

$$\hat{r}_{-i}(\chi) = \frac{\sum_{j\neq i} A_j Y_j K(h^{-1} d(\chi, \mathbf{X}_j))}{\sum_{j\neq i} A_j K(h^{-1} d(\chi, \mathbf{X}_j))}, \tag{5}$$

where $K(.)$ is a kernel function and $h = h_n$ is a sequence of positive real numbers which decreases to zero as n goes to infinity. Root-n consistency of the estimate (4) is stated in Theorem 1(a).

As pointed out by some reviewer of this paper and as discussed by Cheng [2] for finite-dimensional $\mathbf{X}$, instead of replacing all responses by their predicted values, we decide to use the observed responses and only predict the responses for those individuals whose responses are missing. This gives rise to the following estimator:

$$\hat{\mu}_{\text{Mix}} = \frac{1}{n} \sum_{i=1}^{n} (A_i Y_i + (1 - A_i) \hat{r}_{-i}(\mathbf{X}_i)). \tag{6}$$

Root-n consistency of the estimate (6) is also stated in Theorem 1(a).

2.3. *An alternative approach based on propensity score*

The Horvitz–Thompson family of estimators is inspired by the fact that, under MAR modelling, one can rewrite the problem in the following way

$$\mathbb{E}[Y] = \mathbb{E}\left[\frac{AY}{\pi(\mathbf{X})}\right], \tag{7}$$

and estimate the propensity score π using any non-parametric smoother of the new regression problem with response A and covariable $\mathbf{X}$. As mentioned in Section 2.2, we concentrate on kernel estimates

$$\hat{\pi}_{-i}(\chi) = \frac{\sum_{j\neq i} A_j K(h^{-1} d(\chi, \mathbf{X}_j))}{\sum_{j\neq i} K(h^{-1} d(\chi, \mathbf{X}_j))}, \tag{8}$$

leading to the following functional Horvitz–Thompson estimate:

$$\hat{\mu}_{\text{HT}} = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i Y_i}{\hat{\pi}_{-i}(\mathbf{X}_i)}. \tag{9}$$

Root-n consistency of the estimate (9) is stated in Theorem 1(b).

2.4. *Assumptions and notation*

Both estimates $\hat{\mu}_{\text{Reg}}$ and $\hat{\mu}_{\text{HT}}$, defined by Equations (4) and (9), are averaged regression estimators at random values $\mathbf{X}_i$. To control this extra randomness, uniform consistency of regression estimates will be used in this work to get the consistency and rate of convergence of the proposed estimators. Such results have already been obtained in the literature and they will be briefly recalled in the appendix (see Section A.2). They depend both on regularity conditions of the nonlinear functional operators to be estimated (i.e., r and π) and on topological considerations, related to the semi-metric d and to the subspace $\mathcal{S}_{\mathcal{F}}$ on which $\mathbf{X}$ takes its values. From these considerations, conditions H0–H6 to be specified below, arise. From now on, we consider a set $\mathcal{S}_{\mathcal{F}} \subseteq \mathcal{F}$ such that

H0 $\mathbb{P}(\mathbf{X} \in \mathcal{S}_{\mathcal{F}}) = 1$.

In order to understand the assumptions to be made, we introduce some definitions. The first one is related to the size of a set $\mathcal{S}$ in terms of its entropy, while the second one is the definition of a Lipschitz function related to some subset, which will be used to quantify regularity of some regression functions.

Definition 1 *Given a set $\mathcal{S} \subseteq \mathcal{F}$ and $\varepsilon > 0, N_\varepsilon(\mathcal{S})$ is the minimum number of open balls of radius ε needed to cover $\mathcal{S}$. The Kolmogorov's ε-entropy of the set $\mathcal{S}$ is given by $\psi_\mathcal{S}(\varepsilon) = \log(N_\varepsilon(\mathcal{S}))$.*

Definition 2 *We say that a function $m : \mathcal{F} \to \mathbb{R}$ is Lipschitz of order $b > 0$ on a subset $\mathcal{S}$ if there exists $C > 0$ such that for any $\chi_1, \chi_2 \in \mathcal{S}$, we have*

$$|m(\chi_1) - m(\chi_2)| \le Cd^b(\chi_1, \chi_2). \tag{10}$$

Remark 1 Note that any Lipschitz function on $\mathcal{S}$ with $\psi_\mathcal{S}(\varepsilon) < \infty$, for some $\varepsilon > 0$, is bounded.

Let $B(\chi, h)$ denote the ball of center χ and radius h for the topology associated with the semi-metric d:

$$B(\chi, h) = \{\chi' \in \mathcal{F} : d(\chi, \chi') < h\}.$$

Let C and or C' denote generic strictly positive real constants, which may change from line to line. We can now state the remaining assumptions:

H1 There exists a function ϕ such that $\forall \chi \in \mathcal{S}_\mathcal{F}$ and $\forall \epsilon > 0$:

$$0 < C\phi(\epsilon) \le \mathbb{P}(\mathbf{X} \in B(\chi, \epsilon)) \le C'\phi(\epsilon). \tag{11}$$

H2 The non-parametric model consists in the following assumptions on the regression function r and on the conditional moments $\sigma_m(\chi) = \mathbb{E}[|Y|^m|\mathbf{X} = \chi]$ which are assumed to exist for any integer $m > 0$:
(H2a) r is a bounded Lipschitz operator of order b_r on $\mathcal{S}_\mathcal{F}$ and $\mathbb{E}[Y^2] < \infty$,
(H2b) σ_m are continuous and $\exists C > 0, \forall \chi \in \mathcal{S}_\mathcal{F}, \sigma_m(\chi) \le C(m!)$.
H3 The propensity score π, defined in Equation (1), satisfies
(H3a) $\exists C > 0$ such that $C \le \pi(\chi)$, for all $\chi \in \mathcal{S}_\mathcal{F}$,
(H3b) π is a Lipschitz operator of order b_π on $\mathcal{S}_\mathcal{F}$.
H4 The kernel function K has to be a Lipschitz function on $[0, 1)$ which is
(H4a) non-negative, bounded and vanishing (at least) outside of $[0, 1]$,
(H4b) and, if $K(1) = 0$, also satisfying for all $t \in [0, 1)$, $-\infty < C < K'(t) < C' < 0$.
H5 The functions ϕ and $\psi_{\mathcal{S}_\mathcal{F}}$ are such that:
(H5a) $\exists C > 0, \exists \eta_0 > 0, \forall \eta < \eta_0, \phi'(\eta) < C$,
(H5b) $\exists C > 0, \exists n_0, \forall n \ge n_0, (\log n)^2/n\phi(h) < \psi_{\mathcal{S}_\mathcal{F}}(\log n/n) < C\sqrt{n}\phi(h)$,
(H5c) if $K(1) = 0$, it is required that $\exists C > 0, \exists \eta_0 > 0, \forall \eta < \eta_0, \int_0^\eta \phi(u)\,\mathrm{d}u > C\eta\phi(\eta)$,
(H5d) $\lim_{\epsilon\to 0} \psi_{\mathcal{S}_\mathcal{F}}(\epsilon) = \infty$.
H6 The Kolmogorov's ε-entropy of $\mathcal{S}_\mathcal{F}$ satisfies

$$\sum_{n=1}^{\infty} \exp\left\{(1-\beta)\psi_{\mathcal{F}_\mathcal{S}}\left(\frac{\log(n)}{n}\right)\right\} < \infty \quad \text{for some } \beta > 1.$$

Note that (H5b) implies the following facts that will be useful for us:

$$\psi_{\mathcal{S}_\mathcal{F}}\left(\frac{\log(n)}{n}\right) < \frac{n\phi(h)}{\log n} \quad \text{and} \quad \sqrt{n}\phi(h) \longrightarrow \infty.$$

A discussion about the low restriction of this set of assumptions will be provided later along Section 5.2.

3. Root-*n* consistency

Theorem 1 states the weak consistency of the estimates defined in Sections 2.2 and 2.3. The rate of convergence is the usual parametric one, that is, $\sqrt{n}$. The main point to be stressed is the fact that this rate does not depend on the dimensionality of the problem. While this fact has often been observed for multivariate variables **X** in various problems involving averaged estimators, this is the first time (to the best of our knowledge) that this phenomenon is observed for infinite-dimensional variables. Section 5 will pay specific attention to this appealing feature of our result.

THEOREM 1 *Suppose that the functional MAR assumption* (2) *holds.*

(a) *Assume that, for some* $C > 0$, $\sqrt{n}h^{b_r} \leq C$ *and* $\sqrt{n}h^{2b_\pi} \leq C$. *Then, under* (H0)–(H6), *we get that*

$$\sqrt{n}(\hat{\mu}_{\text{Reg}} - \mu) = O_p(1) \tag{12}$$

and

$$\sqrt{n}(\hat{\mu}_{\text{Mix}} - \mu) = O_p(1). \tag{13}$$

(b) *Assume that, for some* $C > 0$, $\sqrt{n}h^{b_\pi} \leq C$ *and* $\mathbb{E}[Y^2] < \infty$. *Then, under* (H0)–(H1) *and* (H3)–(H6) *we get that*

$$\sqrt{n}(\hat{\mu}_{\text{HT}} - \mu) = O_p(1). \tag{14}$$

The main steps of the proof are given below, while the technical details are presented in a number of lemmas, whose proofs are reported in the appendix.

Proof of Theorem 1 The general lines of the proof are the same for the three assertions (12)–(14). For these reasons, we have decomposed the proof of Theorem 1 into three parts. Firstly, we state some general considerations. Then, we give a detailed proof of the result (12), after which a much more sketched proof of Equation (14) is presented. To save space and because it is line by line similar to the one of (12), the proof of (13) is not presented.

(i) *Some general considerations and notations*. Both non-parametric regression estimates for $r(\chi)$ and $\pi(\chi)$ defined in Equations (5) and (8), respectively, can be written as a ratio

$$\hat{r}_{-i}(\chi) = \frac{\hat{r}_{i,2}(\chi)}{\hat{r}_{i,1}(\chi)} \quad \text{and} \quad \hat{\pi}_{-i}(\chi) = \frac{\hat{r}_{i,1}(\chi)}{\hat{r}_{i,3}(\chi)}, \tag{15}$$

where

$$\hat{r}_{i,1}(\chi) = \frac{1}{(n-1)} \sum_{j \neq i} A_j \frac{K(h^{-1}d(\chi, \mathbf{X}_j))}{f_h(\chi)}, \tag{16}$$

$$\hat{r}_{i,2}(\chi) = \frac{1}{(n-1)} \sum_{j \neq i} A_j Y_j \frac{K(h^{-1}d(\chi, \mathbf{X}_j))}{f_h(\chi)}, \tag{17}$$

$$\hat{r}_{i,3}(\chi) = \frac{1}{(n-1)} \sum_{j \neq i} \frac{K(h^{-1}d(\chi, \mathbf{X}_j))}{f_h(\chi)}, \tag{18}$$

$$f_h(\chi) = \mathbb{E}[K(h^{-1}d(\chi, \mathbf{X}_1))]. \tag{19}$$

Let us introduce the following notation:

$$r_1(\chi) = \pi(\chi), \quad r_2(\chi) = r(\chi)\pi(\chi), \quad r_3 = 1.$$

For any $o = (\chi, y, a)$, we will use the functions

$$F_h(o) = a\mathbb{E}_{\mathbf{X}_1}\left[\frac{1}{\pi(\mathbf{X}_1)}\frac{K(h^{-1}d(\chi,\mathbf{X}_1))}{f_h(\mathbf{X}_1)}(y - r(\mathbf{X}_1))\right]$$

and

$$\tilde{F}_h(o) = \mathbb{E}_{\mathbf{X}_1}\left[\frac{r(\mathbf{X}_1)(r_1(\mathbf{X}_1) - a)K(h^{-1}d(\chi,\mathbf{X}_1))}{r_1(\mathbf{X}_1)f_h(\mathbf{X}_1)}\right].$$

(ii) *Proof of Equation* (12). Straightforward calculations involving a Taylor expansion for $F(z) = 1/z$ around r_1, allow us to obtain the following expansion:

$$\hat{\mu}_{\text{Reg}} - \mu = V_n + U_n + R_n + S_n, \tag{20}$$

where

$$V_n = \frac{1}{n}\sum_{i=1}^{n}\{r(\mathbf{X}_i) - \mu\}, \quad U_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{r}_{i,2}(\mathbf{X}_i) - r(\mathbf{X}_i)\hat{r}_{i,1}(\mathbf{X}_i)}{r_1(\mathbf{X}_i)},$$

$$R_n = \frac{1}{n}\sum_{i=1}\frac{\hat{r}_{i,2}(\mathbf{X}_i)}{\tilde{r}_{i,1}^3(\mathbf{X}_i)}(\hat{r}_{i,1}(\mathbf{X}_i) - r_1(\mathbf{X}_i))^2,$$

$$S_n = \frac{1}{n}\sum_{i=1}\frac{r_2(\mathbf{X}_i) - \hat{r}_{i,2}(\mathbf{X}_i)}{r_1^2(\mathbf{X}_i)}(\hat{r}_{i,1}(\mathbf{X}_i) - r_1(\mathbf{X}_i)),$$

and $\tilde{r}_{i,1}(\mathbf{X}_i)$ is a middle point between $\hat{r}_{i,1}(\mathbf{X}_i)$ and $r_1(\mathbf{X}_1)$. The first term on the right-hand side of Equation (20) clearly satisfies $\sqrt{n}V_n = O_P(1)$. The second one, U_n, can be dealt with by means of asymptotic expansion for U-statistics (see Lemma 2 below). The last two terms, R_n and S_n, can be treated using uniform consistency results of the kernel estimates (see Lemma 3).

Lemma 2 *Assume that* (H0), (H1), (H2a) *and* (H3a) *hold. Additionally, assume that either* (H4) *or* (H5a) *is satisfied. Then, we have*:

$$\text{(a)} \quad \sqrt{n}U_n = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}F_h(O_i) + O_P\left(\sqrt{\frac{h^{2b_r}}{\phi(h)}}\right) + O_P\left(\frac{1}{\sqrt{n\phi(h)}}\right) + O(\sqrt{n}h^{b_r})$$

and

$$\text{(b)} \quad \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}F_h(O_i) = O_P(1) + O(\sqrt{n}h^{b_r}).$$

Lemma 3 *Under* (H0)–(H6), *we get that*

$$\text{(a)} \quad \sqrt{n}R_n = O_P(\sqrt{n}h^{2b_\pi}) + O_P\left(\frac{\psi_{\mathcal{S}_\mathcal{F}}(\log n/n)}{\sqrt{n}\phi(h)}\right), \tag{21}$$

and, for $b_2 = \min\{b_\pi, b_r\}$, we also get

$$\text{(b)} \quad \sqrt{n}S_n = O_P(\sqrt{n}h^{2b_r}) + O_P\left(\frac{\psi_{\mathcal{S}_\mathcal{F}}(\log n/n)}{\sqrt{n}\phi(h)}\right). \tag{22}$$

Finally, the proof of Equation (12) follows directly from Equation (20) together with Lemmas 2 and 3 that will be proved in the appendix.

(iii) *Proof of Equation* (14). This proof is very similar to the one above, and it will therefore be given with less details. As we did when dealing with $\hat{\mu}_{\text{Reg}}$, the representation (15) for $\hat{\pi}_{-i}$ allows us to obtain the following expansion:

$$\hat{\mu}_{\text{HT}} - \mu = \tilde{V}_n + \tilde{U}_n + \tilde{R}_n + \tilde{S}_n, \tag{23}$$

where

$$\tilde{V}_n = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_iY_i}{r_1(\mathbf{X}_i)} - \mu\right\}, \quad \tilde{U}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i}{r_1^2(\mathbf{X}_i)}\left\{r_1(\mathbf{X}_i)\hat{r}_{i,3}(\mathbf{X}_i) - \hat{r}_{i,1}(\mathbf{X}_i)\right\},$$

$$\tilde{R}_n = \frac{1}{n}\sum_{i=1}\frac{A_iY_i\hat{r}_{i,3}(\mathbf{X}_i)}{\tilde{r}_{i,1}^3(\mathbf{X}_i)}(\hat{r}_{i,1}(\mathbf{X}_i) - r_1(\mathbf{X}_i))^2,$$

$$\tilde{S}_n = \frac{1}{n}\sum_{i=1}\frac{A_iY_i(r_3 - \hat{r}_{i,3}(\mathbf{X}_i))}{r_1^2(\mathbf{X}_i)}(\hat{r}_{i,1}(\mathbf{X}_i) - r_1(\mathbf{X}_i)),$$

and $\tilde{r}_{i,1}(\mathbf{X}_i)$ is a middle point between $\hat{r}_{i,1}(\mathbf{X}_i)$ and $r_1(\mathbf{X}_1)$. As before in Lemma 2, one can treat $\tilde{U}_n$ by means of asymptotic expansions for U-statistics and get

$$\sqrt{n}\tilde{U}_n = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\tilde{F}_h(O_i) + O_P\left(\sqrt{\frac{h^{2b_\pi}}{\phi(h)}}\right) + O_P\left(\frac{1}{\sqrt{n\phi(h)}}\right) + O(\sqrt{n}h^{b_\pi}) \tag{24}$$

and then

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\tilde{F}_h(O_i) = O_P(1) + O(\sqrt{n}h^{b_\pi}). \tag{25}$$

Similarly, as done before, along the proof of Lemma 3, one can treat the terms $\tilde{R}_n$ and $\tilde{S}_n$ using uniform properties of kernel functional regressors and get

$$\sqrt{n}\tilde{R}_n = O_P(\sqrt{n}h^{2b_\pi}) + O_P\left(\frac{\psi_{\mathcal{S}_\mathcal{F}}(\log n/n)}{\sqrt{n}\phi(h)}\right), \tag{26}$$

and

$$\tilde{S}_n = O_P\left(\sqrt{n}h^{b_\pi}\sqrt{\frac{\psi_{\mathcal{S}_\mathcal{F}}(\log n/n)}{n\phi(h)}}\right) + O_P\left(\frac{\psi_{\mathcal{S}_\mathcal{F}}(\log n/n)}{\sqrt{n}\phi(h)}\right). \tag{27}$$

Finally, the order of convergence of $\hat{\mu}_{\text{HT}}$, as stated in Equation (14), follows directly from Equations (23)–(27) and from the obvious $\sqrt{n}$ consistency of the first term $\tilde{V}_n$.

The proof of Theorem 1 is now complete. ■

4. Some simulations

The aim of this section is to discuss the behaviour of the various mean estimates discussed before (namely $\hat{\mu}_{\text{Reg}}$, $\hat{\mu}_{\text{HT}}$ and $\hat{\mu}_{\text{Mix}}$) on finite sample situations. We also compute what we call the *naive estimate* which is the one obtained by averaging only the observed responses which are available and forgetting the possible information contained in the covariable $\mathbf{X}$, that is:

$$\hat{\mu}_{\text{Naive}} = \frac{\sum_{i=1}^{n} Y_iA_i}{\sum_{i=1}^{n} A_i}. \tag{28}$$

Of course, except in the very special case when the censoring process A and the covariate $\mathbf{X}$ are independent, the estimate $\hat{\mu}_{\text{Naive}}$ is not consistent and it has to be seen in our study only as a benchwark tool. Furthermore, the use of simulated processes allows us also to compute $\bar{Y}$. Of course this has to be seen as another benchmark for our analysis and not really as a competitor estimate since it is uncomputable with real data.

4.1. *Presentation of the model*

The general guidelines for building our simulated experiments were, in order to be close to what could happen in real-life situations, a two-populations sample (see Equation (29)) and a censoring scheme which affect more one population strongly than the other one (see Equation (30)).

4.1.1. *The simulated curves*

The functional variable $\mathbf{X}$ is chosen as a real-valued function with support $[0, \pi]$ whose observations are generated in the following way:

$$\mathbf{X}_i(t) = Z_i \cos(2t),$$

where t takes 100 equispaced values in $[0, \pi]$ and where Z_i are i.i.d. random variables distributed according to

$$Z_i \sim F_Z = 0.3\,F_1 + 0.7\,F_2, \quad i = 1, \dots, n, \tag{29}$$

where F_1 (respectively, F_2) is the cumulative distribution of a random variable distributed according to a $\mathcal{N}(10, 36)$ distribution (respectively, to a $\mathcal{N}(0, 36)$ distribution). To fix the ideas, Figure 1 presents a sample of $n = 200$ of such curves. Note that the choice of a high variance parameter (36 exactly) does not allow us to discover the heterogeneity in the set of curves by simple inspection.

4.1.2. *The responses and the censoring process*

The responses are obtained from the following regression model:

$$Y_i = r(\mathbf{X}_i) + \varepsilon_i,$$

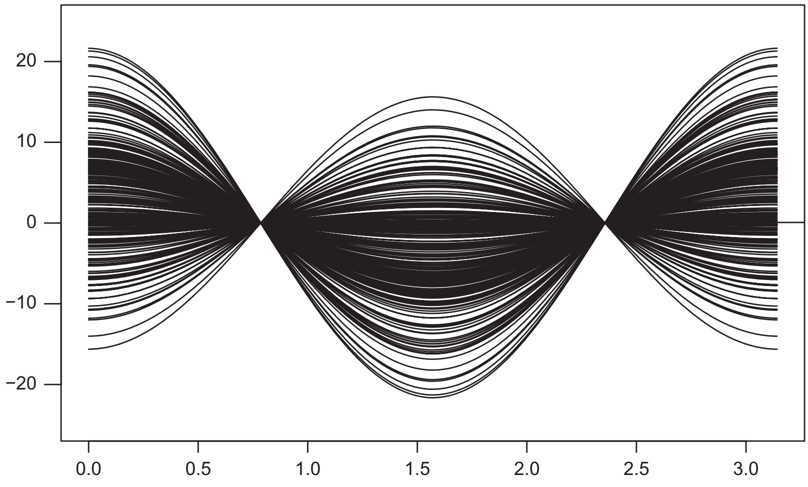

Figure 1. A sample of $n = 200$ simulated curves.

where $\varepsilon_i \sim \mathcal{N}(0, 0.05)$ and

$$r(\mathbf{X}) = \frac{2}{\pi}\int_0^{\pi} \mathbf{X}^2(t)\,\mathrm{d}t.$$

Finally, in order to simulate an MAR scheme, it remains to construct a procedure for deciding whether the response is observed or not. This is done through the following missing mechanism:

$$P(A = 1|\mathbf{X}) = \operatorname{expit}\left(\frac{2\alpha}{\pi}\int_0^{\pi} \mathbf{X}^2(t)\,\mathrm{d}t\right), \tag{30}$$

where $\operatorname{expit}(u) = \mathrm{e}^u/(1+\mathrm{e}^u)$, for $u \in \mathbb{R}$. Note that such a missing scheme has the tendency to censor more strongly the data from the first group (that is, those for which the corresponding Z_i follows a F_1 distribution) than those of the other group. The parameter α controls the degree of dependency between the functional covariable $\mathbf{X}$ and the censoring one A. Note finally that $\alpha = 0$ corresponds to the independence situation, while higher values of α will naturally make decrease in the degree of censoring in the sample. To keep control on this last quantity we have also computed

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i.$$

4.2. *Choosing the parameters of the estimates*

The estimates $\hat{\mu}_{\mathrm{Reg}}$ and $\hat{\mu}_{\mathrm{HT}}$ are based on the leave-one-out estimates

$$\hat{r}_{-i}(\chi) = \frac{\sum_{j\neq i} A_j Y_j K(h^{-1}d(\chi, \mathbf{X}_j))}{\sum_{j\neq i} A_j K(h^{-1}d(\chi, \mathbf{X}_j))} \quad \text{and} \quad \hat{\pi}_{-i}(\chi) = \frac{\sum_{j\neq i} A_j K(h^{-1}d(\chi, \mathbf{X}_j))}{\sum_{j\neq i} K(h^{-1}d(\chi, \mathbf{X}_j))},$$

which are both special cases of the functional Nadaraya–Watson regression-type estimator. Even though the method involves the estimation of complex mathematical objects (as, for instance, the nonlinear functional operators r and π), the simple form of Nadaraya–Watson kernel weights makes this implementation straightforward. Practically, one can directly use the R/S^+ function *funopare.knn.gcv*, presented in Ferraty and Vieu [15] and available on http://www.lsp.ups-tlse.fr/staph/npfda, which computes the leave-one-out estimates $\hat{r}_{-i}$ and $\hat{\pi}_{-i}$.

As usual in kernel non-parametric statistics the shape of the kernel function K is of minor importance and we choose a quadratic kernel, given by

$$K(u) = \tfrac{3}{4}(1 - u^2)I_{[0,1]}(u).$$

The choice of the bandwidth is more relevant. It is done (as a black box inside the function *funopare.knn.gcv*) by minimizing the cross-validation criterion over a set of bandwidths constructed by k-nearest ideas. The k-nearest methodology has the advantages of letting the bandwidth h being locally adapted and of making the minimization problem much more simpler. The reader wishing to know more on that point may have a look at the recent paper by Burba *et al.* [18]. Similarly, one may look at Benhenni *et al.* [19] for more information about cross-validated bandwidth choice in functional data situation. Finally, the measure of closeness between curves has been chosen to be the usual L_2-norm

$$d(\chi_1, \chi_2) = \sqrt{\int_0^{\pi} (\chi_1(t) - \chi_2(t))^2\,\mathrm{d}t}.$$

Table 1. Values Mean(MSE) for the various estimates over 100 replications.

α	Mean $\bar{A}$	$\hat{\mu}_{\text{Reg}}$	$\hat{\mu}_{\text{HT}}$	$\hat{\mu}_{\text{Naive}}$	$\hat{\mu}_{\text{Mix}}$
0	0.51	59.87 (73.78)	84.30 (934.13)	66.38 (72.20)	62.95 (51.35)
0.1	0.83	62.73 (49.21)	66.49 (43.09)	78.51 (211.57)	66.34 (40.08)
0.2	0.88	62.69 (49.39)	66.32 (40.03)	74.96 (129.85)	66.37 (41.74)
0.4	0.91	62.66 (49.53)	66.30 (39.97)	72.36 (87.33)	66.32 (40.00)
0.8	0.94	62.64 (49.63)	66.30 (39.97)	70.57 (65.08)	66.31 (39.98)
1.2	0.95	62.64 (49.65)	66.30 (39.97)	69.73 (57.80)	66.31 (39.97)

4.3. *The results*

In all the following, we have chosen $n = 200$ and have tried different values for the key parameter α. For each value of α we carried 100 independent replications of the experiment, and we have computed the mean squared error (MSE) (over these 100 replications) of the various estimates. The results are reported in Table 1. To keep in mind the effect of the censoring on our results, Table 1 reports also for each case the mean (still over the 100 replications) of the variable $\bar{A}$. Looking at Table 1, keep in mind that the true unknown value to be estimated is $\mu = \mathbb{E}Y = 66$. Recall also that the case $\alpha = 0$ plays a special role since it corresponds to the extreme situation when the censoring is independent from the functional covariate.

4.4. *First comments*

It appears from Table 1 that, except for the special case $\alpha = 0$ when the censoring is independent from the covariate, our three estimates $\hat{\mu}_{\text{Reg}}$, $\hat{\mu}_{\text{HT}}$ and $\hat{\mu}_{\text{Mix}}$ perform much better than the standard mean $\hat{\mu}_{\text{Naive}}$. The estimate based on propensity score $\hat{\mu}_{\text{HT}}$ performs better than the averaged estimate $\hat{\mu}_{\text{Reg}}$ when $\alpha \neq 0$. This is not surprising since the censoring process has the tendency to censor, in a stronger way, individuals from one group, leading to overestimation for $\hat{\mu}_{\text{Naive}}$ and underestimation for $\hat{\mu}_{\text{Reg}}$. In counterpart, when $\alpha = 0$, the estimate $\hat{\mu}_{\text{HT}}$ performs very poorly. This is also not surprising because, when there is independence between the censoring and the covariate, the estimate is based on an empirical idea that makes no sense (see formulas (1) and (7)). To conclude this comparative discussion between the various estimators, it is worth noting that, at least on this simulated experiment, the best estimator seems to be the mixed estimator $\hat{\mu}_{\text{Mix}}$ which works almost as well as the one based on propensity score when the censoring depends on the covariate (cases when $\alpha \neq 0$) without being too much affected in situations of independence (case when $\alpha = 0$).

4.5. *Complementary issues*

To fulfil our purpose, we present, for an arbitrarily selected value of α (namely $\alpha = 0.2$,) the histograms obtained for the various estimates over the 100 replications (see Figure 3) and also the corresponding box plots (see Figure 2). Plots for the whole set of values of α are not presented because they looked similar.

The results of Table 1 are confirmed in Figure 2. One can see the poor behaviour of the naive estimate compared with what can be obtained by any of the three other ones (recall that $\bar{Y}$ is just a benchmark tool which is unusable in practice). Also appearing is the superiority of the estimates $\hat{\mu}_{\text{HT}}$ and $\hat{\mu}_{\text{Mix}}$ over $\hat{\mu}_{\text{Reg}}$, as discussed in Section 4.4. Each $\hat{\mu}_{\text{HT}}$ and $\hat{\mu}_{\text{Mix}}$ performs as well as the uncomputable benchmark $\bar{Y}$.

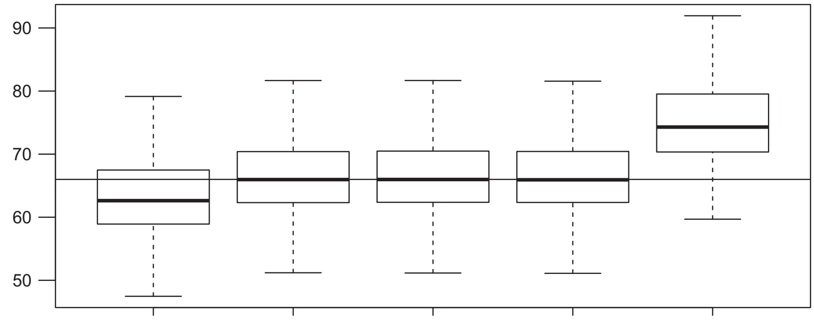

Figure 2. Box plots for $\hat{\mu}_{\text{Reg}}$, $\hat{\mu}_{\text{HT}}$, $\hat{\mu}_{\text{Mix}}$, $\bar{Y}$ and $\hat{\mu}_{\text{Naive}}$: $\alpha = 0.2$, $\mu = 66$.

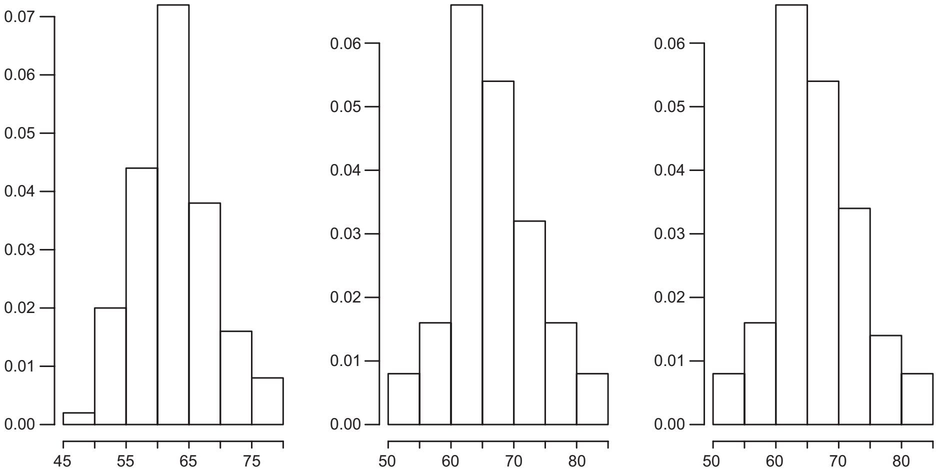

Figure 3. Observed frequencies of $\hat{\mu}_{\text{Reg}}$, $\hat{\mu}_{\text{HT}}$ and $\hat{\mu}_{\text{Mix}}$: $\alpha = 0.2$, $\mu = 66$.

Moreover another interesting feature of Figures 2 and 3 is to show that one may expect some positive response about possible asymptotic normal distribution for the estimates which is still an open theoretical question (see discussion in Section 5.3).

5. Discussion

5.1. *A few words on applications*

Even if we have decided to focus on the methodological and theoretical features of the Functional MAR model, it is worth devoting a few words on potential applications. While medicine and economy are the areas where estimation of mean responses under an MAR assumption finds most frequent applications, this issue may also become relevant in many other fields such as biology, social sciences, image analysis, etc. involving applied statistical problems. In particular, when doing causal inference, variables representing responses in hypothetical words under different treatments assignments are considered. Estimating their mean based on observed data can be treated as a missing data problem.

The methodological novelty in our work consists in letting the explanatory variable **X** to belong to some infinite-dimensional space: this is of great interest since more complex structures than those allowed by standard multivariate covariables can be used. For instance, in medical applications, the covariable may represent some pre-treatment assignment information, our approach allows the physician to collect information given by curves such as electrocardiogram or other complex clinical studies taking values in an infinite-dimensional space (see [6] or [20], for examples of curves data).

5.2. *Comments on the hypothesis*

It is beyond the scope of this paper to discuss in detail the various technical assumptions H1–H6 introduced earlier to ensure the good asymptotic behaviour of our estimate. They are widely commented in Ferraty *et al.* [21], through various examples of time continuous processes. We just wish to describe here shortly how they can be seen, in some sense, as being less restrictive than those appearing before in the usual multivariate literature. This is clear for conditions on the censoring, since they are reduced to (H3) which appears in most of the multivariate literature (see for instance [2]).

All other conditions are linked to the functional feature of the problem. Firstly note that, with the exception of the last part of (H5b) and of (H5d), all other assumptions are introduced with the sole purpose of ensuring uniform consistency of the kernel regression estimates as stated in Ferraty [21]. Therefore, they could be replaced by any other set of conditions implying the same uniform consistency property, without affecting the main results of our paper. To see their low degree of restrictivity, it suffices to look at the special case when the space $\mathcal{F}$ is equal to $\mathbb{R}^p$ (that is, the multivariate situation) and when d is the usual euclidian metric. In this case, if $\mathbf{X}$ has a density which is bounded below from 0 on a compact set $\mathcal{S}_{\mathcal{F}}$, then it is easy to see that we can chose as functions of ϕ and $\psi_{\mathcal{S}_{\mathcal{F}}}$ the following ones

$$\phi(\epsilon) \sim \epsilon^p \quad \text{and} \quad \psi_{\mathcal{S}_{\mathcal{F}}} \sim -c\log(\epsilon) \text{ as } \epsilon \sim 0,$$

and then all other conditions are true as long as the bandwidth satisfies the restrictions

$$\lim_{n\to\infty} h_n = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{\log n}{nh_n^p} = 0,$$

which are usual in multivariate non-parametric setting. Altogether, if we particularize our functional results to the multivariate setting when X has a density one finds back the same conditions as in the previous literature (see again for instance [2]), but our results work also without assuming a density for the explanatory variable (in this case the function ϕ is, of course, not the same). Moreover, even when the density exists, our methodology allows for other topology than the euclidian one, and this may have direct application for dimensionality reduction in multivariate regression problem. So there is real evidence of the fact that our set of conditions makes our methodology interesting not only for functional covariables (which is, however, its main goal) but also for multivariate covariates.

5.3. *On possible extensions of our results*

5.3.1. *General mean functional estimation*

In order to make the presentation more clear and because the main goal of this paper was to emphasize on the infinite dimensional feature of the covariable, we have only considered earlier the question of the estimation of the mean. However, general mean functionals could similarly be estimated. Clearly, assuming MAR, one has, for any known integrable function $g(Y)$,

$$\mathbb{E}[g(Y)] = \mathbb{E}[r_g(\mathbf{X})] = \mathbb{E}\left[\frac{Ag(Y)}{\pi(\mathbf{X})}\right]$$

with $r_g(\chi) = \mathbb{E}[g(Y)|(A = 1, \mathbf{X} = \chi)]$. If we were interested in estimating $\mathbb{E}[g(Y)]$, it would suffice to replace Y by $g(Y)$ into the estimates and similar asymptotic results would be obtained. In particular, this would allow us to estimate any moment of Y, and its variance too. The reader may have a look at the work by Cheng [2] in which this is done but with finite-dimensional explanatory variable $\mathbf{X}$.

5.3.2. *About asymptotic normality*

As pointed out by one reviewer of this paper, a natural question is to know what would be the asymptotic distribution of our estimates. In other words, one would like to get functional versions of the asymptotic normality results as stated in Cheng [2]. This has obvious appealing practical interest (for confidence band construction or hypothesis testing for instance). However, because of technical difficulties, which we will discuss here, this result is not presented here.

It should be noted that the asymptotic normality property in itself would not be the most difficult step, since one could naturally think in following our proofs (see, for instance, the proof of Lemma 2) for root-n consistency and using, as an additional probabilistic tool, a CLT theorem for U-statistics. This asymptotic normality would be in concordance with what has been observed in the simulated example (see Figure 3). For the asymptotic normality of the kernel regression operator itself, the key technical difficulty would be to give precise expressions for the constants involved in the asymptotic bias and variance of the estimates because they will depend on the smoothness of the regression operator r which is not an easy thing to control as r is a nonlinear functional operator. This fact has been highlighted by Masry [22], Ferraty *et al.* [23] and Delsol [24], but extension to the MAR problem and to the averaged estimates is still an open question. At this stage, it should also be mentioned that even such asymptotic normality theorem would not be directly of great practical interest (because of the high degree of complexity that one is expecting for the bias and variance) and should be completed by the construction of more trustable scheme like bootstrapping or resampling. Such ideas have been investigated recently by Ferraty *et al.* [25] for estimating the regression operator itself, but their extensions to MAR modelling is a second important open issue.

5.4. *About averaging non-parametric estimates*

In standard multivariate non-parametric statistics, there are many situations in which, by averaging or integrating initial estimates having non-parametric rates of convergence, one can construct new estimates having the parametric $\sqrt{n}$-rate. Examples of $\sqrt{n}$-consistent averaged non-parametric estimates involve, of course, the previous multivariate approaches of MAR modelling discussed in Section 1 (see, e.g. [2] or [3]), but also completely different fields like distribution function estimation for which one needs to integrate some non-parametric density estimate (see, e.g. [26], for the most recent advances) or bandwidth selection problems for which estimations of integrated functionals are necessary (see, e.g. [27], for a general discussion). The same idea may also be used, through some partial integration stage, to improve the rate of convergence of some multivariate non-parametric estimates, as for instance in the marginal integration procedure developed for additive non-parametric regression modelling (see, e.g. [28]).

While much work can be found in the literature regarding the multivariate setting, despite the wide range of potential applications, to the best of our knowledge there are still no precedents of this kind of results when functional covariables are used. In this sense, we guess that our paper could be not only of interest for functional MAR problems, but also as a starting point to many other problems for which averaged estimates involving functional variables have to be considered.

Acknowledgements

This work began while Mariela Sued was visiting the Institut de Mathématiques de Toulouse, supported by CONICET. The authors wish to express their gratitude to the members of the working group STAPH in the Institut de Mathématiques de Toulouse, for their continuous and fruitful comments and support. They also wish to thank profusely both Reviewers and Editors for their helpful analysis of the preliminary version of this work and whose comments and suggestions have greatly improved its quality. This research was partially supported by grants X–447 from University of Buenos Aires, PIP 0592 from CONICET and PICTs 00883 from ANPCyT.

References

[1] D. Rubins, *Inference and missing data*, Biometrika 63 (1976), pp. 581–592.

[2] P.E. Cheng, *Nonparametric estimation of mean functionals with data missing at random*, J. Amer. Statist. Assoc. 89 (1994), pp. 81–87.

[3] K. Hirano, G.W. Imbens, and G. Ridder, *Efficient estimation of average treatment effects using the estimated propensity score*, Econometrica 71(4) (2003), pp. 1161–1189.

[4] J.D.Y. Kang and J.L. Schafer, *Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data*, Statist. Sci. 22 (2007), pp. 523–539.

[5] M. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky, *Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable*, Statist. Sci. 22 (2007), pp. 544–559.

[6] J.O. Ramsay and B.W. Silverman, *Applied Functional Data Analysis*, Springer Series in Statistics, Springer, Berlin, 2002.

[7] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, 2nd ed., Springer, Berlin, 2005.

[8] M. Davidian, X. Lin, and J.L. Wang, *Introduction to the emerging issues in longitudinal and functional data analysis (with discussion)*, Statist. Sinica 14(3) (2004), pp. 613–629.

[9] W. González Manteiga and P. Vieu, *Introduction to the special issue on statistics for functional data*, Comput. Statist. Data Anal. 51(10) (2007), pp. 4788–4792.

[10] M. Valderrama, *An overview to modelling functional data*, Comput. Statist. 22 (2007), pp. 331–334.

[11] F. Ferraty, *High-dimensional data: A fascinating statistical challenge*, J. Multivariate Anal. 101 (2010), pp. 305–306.

[12] F. Ferraty and Y. Romain (eds), *Oxford Handbook on Functional Data Analysis*, Oxford University Press, Oxford, 2010.

[13] F. Ferraty and P. Vieu, *The functional nonparametric model and application to spectrometric data*, Comput. Statist. 17(4) (2002), pp. 545–564.

[14] F. Ferraty and P. Vieu, *Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination*, J. Nonparametr. Stat. 16 (2004), pp. 111–125.

[15] F. Ferraty and P. Vieu, *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Series in Statistics, Springer, New York, 2006.

[16] L. Delsol, *Advances on asymptotic normality in non-parametric functional time series analysis*, Statistics 43(1) (2009), pp. 13–33.

[17] F. Ferraty and P. Vieu, *Kernel regression estimation for functional data*, Oxford Handbook on Functional Data Analysis, Oxford University Press, Oxford, 2010.

[18] F. Burba, F. Ferraty, and P. Vieu, *k-nearest neighbour method in functional nonparametric regression*, J. Nonparametr. Stat. 21(4) (2009), pp. 453–469.

[19] K. Benhenni, F. Ferraty, M. Rachdi, and P. Vieu, *Local smoothing regression with functional data*, Comput. Statist. 22(3) (2007), pp. 353–369.

[20] F. Ferraty and P. Vieu, *Functional nonparametric statistics in action. The art of semiparametrics,* Contributions in Statistics, Physica-Verlag/Springer, Heidelberg, 2006, pp. 112–129.

[21] F. Ferraty, A. Laksaci, A. Tadj, and P. Vieu, *Rate of uniform consistency for nonparametric estimates with functional variables*, J. Statist. Plann. Inference 140 (2010), pp. 335–352.

[22] E. Masry, *Nonparametric regression estimation for dependent functional data: Asymptotic normality*, Stochast. Process. Appl. 115(1) (2005), pp. 155–177.

[23] F. Ferraty, A. Mas, and P. Vieu, *Nonparametric regression on functional data: Inference and practical aspects*, Aust. N. Z. J. Stat. 49(3) (2007), pp. 267–286.

[24] L. Delsol, *Nonparametric methods for α-mixing functional random variables*, Oxford Handbook on Functional Data Analysis, Oxford University Press, Oxford, 2010.

[25] F. Ferraty, I. Van Keilegom, and P. Vieu, *On the validity of the bootstrap in non-parametric functional regression*, Scand. J. Stat. 37(2) (2010), pp. 286–306.

[26] R. Liu and L. Yang, *Kernel estimation of multivariate cumulative distribution function*, J. Nonparametr. Stat. 20 (2010), pp. 661–677.

[27] P. Hall and J.S. Marron, *Estimation of integrated squared density derivatives*, Statist. Probab. Lett. 6 (1987), pp. 109–115.

[28] C. Camlong-Viot, P. Sarda, and P. Vieu, *Additive time series: The kernel integration method*, Math. Methods Statist. 9 (2000), pp. 358–375.

[29] A.J. Lee, *U-Statistics*, Marcel Dekker, New York, 1990.

Appendix: Proofs of technical lemmas

Let us start by recalling a result that will be widely used in the following. This result comes from Ferraty *et al.* [21] and it states that, under (H1) and (H4) or (H1) and (H5a), there exist constants $0 < C < C' < \infty$, such that

$$\forall \chi \in \mathcal{S}_{\mathcal{F}}, \quad C\phi(h) \le f_h(\chi) \le C'\phi(h). \tag{A1}$$

Keep in mind that all along these proofs C is a generic positive real constant, $o = (\chi, y, a)$ and $O_i = (\mathbf{X}_i, Y_i, A_i)$.

A.1. *Proof of Lemma* 2

Before proving the claimed results, we need to introduce some notation. Observe that

$$\begin{aligned} U_n &= \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{r}_{i,2}(\mathbf{X}_i) - r(\mathbf{X}_i)\hat{r}_{i,1}(\mathbf{X}_i)}{r_1(\mathbf{X}_i)} \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} \frac{A_j}{\pi(\mathbf{X}_i)}\frac{K(h^{-1}d(\mathbf{X}_i,\mathbf{X}_j))}{f_h(\mathbf{X}_i)}(Y_j - r(\mathbf{X}_i)). \end{aligned}$$

To get the asymptotic expansion presented in Equation (21) the main probabilistic tool will be the using of general results for U-statistics. So, in order to follow the traditional notation in this area, we write

$$U_n = \binom{n}{2}^{-1}\sum_{i<j} H_h(O_i, O_j),$$

where

$$H_h(O_i,O_j) = \frac{1}{2}\frac{A_j}{\pi(\mathbf{X}_i)}\frac{K(h^{-1}d(\mathbf{X}_i,\mathbf{X}_j))}{f_h(\mathbf{X}_i)}(Y_j - r(\mathbf{X}_i)) + \frac{1}{2}\frac{A_i}{\pi(\mathbf{X}_j)}\frac{K(h^{-1}d(\mathbf{X}_i,\mathbf{X}_j))}{f_h(\mathbf{X}_j)}(Y_i - r(\mathbf{X}_j)).$$

Let us also use the notation

$$H_{1,h}(o) = \mathbb{E}[H_h(o, O_1)].$$

(i) *Proof of Lemma* 2(a). We will show the following two intermediary results:

$$\sqrt{n}U_n = \sqrt{n}\frac{2}{n}\sum_{i=1}^{n} H_{1,h}(O_i) + O_P\left(\frac{1}{\sqrt{n\phi(h)}}\right) + O(\sqrt{n}h^{b_r}) \tag{A2}$$

and

$$\sqrt{n}\frac{2}{n}\sum_{i=1}^{n} H_{1,h}(O_i) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} F_h(O_i) + O_P\left(\sqrt{\frac{h^{2b_r}}{\phi(h)}}\right) + O(\sqrt{n}h^{b_r}). \tag{A3}$$

Let us consider

$$\theta_h = \mathbb{E}[H_{1,h}(O_1)] = \mathbb{E}\left[\frac{A_j}{\pi(\mathbf{X}_i)}\frac{K(h^{-1}d(\mathbf{X}_i,\mathbf{X}_j))}{f_h(\mathbf{X}_i)}(Y_j - r(\mathbf{X}_i))\right] \tag{A4}$$

and

$$U_n^* = \frac{2}{n}\sum_{i=1}^{n}(H_{1,h}(O_i) - \theta_h).$$

Then, we can write

$$\sqrt{n}\left(U_n - \frac{2}{n}\sum_{i=1}^{n} H_{1,h}(O_i)\right) = \sqrt{n}(U_n - \theta_h - U_n^*) - \sqrt{n}\theta_h.$$

Expansion (A2) can be deduced from the following two facts, to be proved:

$$n\mathrm{Var}(U_n^* - U_n + \theta_h) = O\left(\frac{1}{n\phi(h)}\right) \tag{A5}$$

and

$$\sqrt{n}\theta_h = O(\sqrt{n}h^{b_r}). \tag{A6}$$

To prove Equation (A5), write

$$n\mathrm{Var}(U_n^* - (U_n - \theta_h)) = n\mathrm{Var}(U_n^*) - 2n\mathrm{Cov}(U_n^*, U_n) + n\mathrm{Var}(U_n).$$

Let

$$\sigma_1^2(h) = \mathrm{Var}(H_{1,h}(O_1)), \quad \sigma_2^2(h) = \mathrm{Var}(H_h(O_1, O_2)).$$

Independence between observations guarantees that

$$n\mathrm{Var}(U_n^*) = 2^2\sigma_1^2(h).$$

On the other hand, the well-known formula for the variance of U-statistics (see [29, Theorem 3, Section 1.3,]) implies that

$$n\mathrm{Var}(U_n) = 2^2\frac{(n-2)}{(n-1)}\sigma_1^2(h) + \frac{1}{(n-1)}\sigma_2^2(h).$$

Finally, following the computations performed in the mentioned reference and observing that $\mathrm{Cov}(H_{1,h}(O_k), H_h(O_i, O_j)) = \mathrm{Cov}(H_h(O_k, O_{n+1}), H_h(O_i, O_j))$, we get that

$$n\mathrm{Cov}(U_n^*, U_n) = 2^2\sigma_1^2(h).$$

Now, from the definition of $H_h(O_1, O_2)$, we have that

$$\sigma_2^2(h) \le \mathbb{E}[H_h^2(O_1, O_2)] \le \mathbb{E}\left[\frac{A_j}{\pi^2(\mathbf{X}_i)}\frac{K^2(h^{-1}d(\mathbf{X}_i, \mathbf{X}_j))}{f_h^2(\mathbf{X}_i)}(Y_j - r(\mathbf{X}_i))^2\right] \simeq \frac{1}{\phi(h)}$$

and so, we conclude that Equation (A5) holds. The proof of Equation (A6) is much easier, since we get from Equation (A4) that

$$\begin{aligned}\theta_h &= \mathbb{E}\left[\frac{\pi(\mathbf{X}_j)}{\pi(\mathbf{X}_i)}\frac{K(h^{-1}d(\mathbf{X}_i, \mathbf{X}_j))}{f_h(\mathbf{X}_i)}(r(\mathbf{X}_j) - r(\mathbf{X}_i))\right] \\ &\simeq \sup_{d(\chi_1,\chi_2)\le h} |r(\chi_1) - r(\chi_2)|,\end{aligned}$$

and condition H2a allows us to conclude. At this stage, Equations (A5) and (A6) are proved, and therefore (A2) is also checked.

It remains just to prove Equation (A3). For that, note that we can write

$$2H_{1,h}(o_i) = G_h(o_i) + F_h(o_i),$$

where

$$G_h(o_i) = \mathbb{E}_{\mathbf{X}_j}\left[\frac{\pi(\mathbf{X}_j)}{\pi(\mathbf{X}_i)}\frac{K(h^{-1}d(\mathbf{X}_i, \mathbf{X}_j))}{f_h(\mathbf{X}_i)}(r(\mathbf{X}_j) - r(\mathbf{X}_i))\right].$$

Observing that

$$\mathbb{E}[G_h(O_i)] = \theta_h$$

and

$$\mathrm{Var}(G_h(O_i)) \le \mathbb{E}\left[\frac{\pi^2(\mathbf{X}_j)}{\pi^2(\mathbf{X}_i)}\frac{K^2(h^{-1}d(\mathbf{X}_i, \mathbf{X}_j))}{f_h^2(\mathbf{X}_i)}(r(\mathbf{X}_j) - r(\mathbf{X}_i))^2\right] \sim \frac{h^{2b_r}}{\phi(h)}$$

we arrive directly at

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} G_h(O_i) = O_P\left(\sqrt{\frac{h^{2b_r}}{\phi(h)}}\right) + O(\sqrt{n}h^{b_r}),$$

which is enough to prove Equation (A3). At this stage, Equations (A2) and (A3) have been checked and the proof of Lemma 2(a) is complete.

(ii) *Proof of Lemma* 2(b). Let us introduce the notation

$$M(\chi, h) = \mathbb{E}\left[\frac{1}{\pi(\mathbf{X}_1)}\frac{K(h^{-1}d(\chi,\mathbf{X}_1))}{f_h(\mathbf{X}_1)}\right]$$

and

$$N(\chi, h) = \mathbb{E}\left[\frac{r(\mathbf{X}_1)}{\pi(\mathbf{X}_1)}\frac{K(h^{-1}d(\chi,\mathbf{X}_1))}{f_h(\mathbf{X}_1)}\right].$$

and observe (see Remark 1) that M and N are uniformly bounded on $\chi \in \mathcal{S}_{\mathcal{F}}$ and on $h > 0$. Then,

$$F_h(O_i) = A_i Y_i M(\mathbf{X}_i, h) - A_i N(\mathbf{X}_i, h),$$

is a bounded random variable. Finally, by observing that

$$\mathbb{E}[F_h(O_i)] = \theta_h,$$

one gets directly that

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} F_h(O_i) = \sqrt{n}\theta_h + O_P(1).$$

This result together with Equation (A6) is enough to get the proof of Lemma 2-b.

A.2. *Proof of Lemma* 3

At this point, uniform rates of convergence of the pseudo (since they depend on the unknown quantity $f_h(\chi)$) estimators $\hat{r}_{i,1}$ and $\hat{r}_{i,2}$ are used. The following results are presented in Ferraty *et al.* [21]. They concern the kernel estimates $\hat{r}_1$ and $\hat{r}_2$, which are just differing from their leave-one-out versions defined in Equations (16) and (17) by the fact that they are constructed from the whole sample. Under the assumptions of Lemma 3, these authors proved that

$$\sup_{\chi\in\mathcal{S}_{\mathcal{F}}} |\hat{r}_1(\chi) - r_1(\chi)| = O(h^{b_\pi}) + O\left(\sqrt{\frac{\psi_{\mathcal{S}_{\mathcal{F}}}(\log n/n)}{n\phi(h)}}\right), \quad \text{a.s.} \tag{A7}$$

and

$$\sup_{\chi\in\mathcal{S}_{\mathcal{F}}} |\hat{r}_2(\chi) - r_2(\chi)| = O(h^{b_2}) + O\left(\sqrt{\frac{\psi_{\mathcal{S}_{\mathcal{F}}}(\log n/n)}{n\phi(h)}}\right), \quad \text{a.s.} \tag{A8}$$

The difference between $\hat{r}_{i,1}(\chi)$ and $\hat{r}_1(\chi)$ is given by

$$\hat{r}_{i,1}(\chi) - \hat{r}_1(\chi) = \frac{\hat{r}_1(\chi)}{n-1} - \frac{A_i K(h^{-1}d(\chi,\mathbf{X}_i))}{f_h(\mathbf{X}_i)(n-1)}.$$

So, using the notation

$$D_n = \sup_{\chi\in\mathcal{S}_{\mathcal{F}}} \max_{1\le i\le n} |\hat{r}_{i,1}(\chi) - r_1(\chi)|$$

and

$$W_n = \sup_{\chi\in\mathcal{S}_{\mathcal{F}}} \max_{1\le i\le n} \frac{1}{\tilde{r}_{i,1}(\chi)},$$

we have the following results which will be useful later:

$$D_n = O(1/n\phi(h)) + O(h^{b_\pi}) + O\left(\sqrt{\frac{\psi_{\mathcal{S}_{\mathcal{F}}}(\log n/n)}{n\phi(h)}}\right), \quad \text{a.s.}$$

and

$$\limsup_{n\to\infty} W_n \le C < \infty \quad \text{a.s.}$$

We are now in position to prove both assertions in Lemma 3.

(i) *Proof of Lemma* 3(a). Recall that

$$R_n = \frac{1}{n}\sum_{i=1} \frac{\hat{r}_{i,2}(\mathbf{X}_i)}{\tilde{r}^3_{i,1}(\mathbf{X}_i)}(\hat{r}_{i,1}(\mathbf{X}_i) - r_1(\mathbf{X}_i))^2,$$

and so one has directly

$$\sqrt{n}|R_n| \le W_n^3\sqrt{n}D_n^2\frac{1}{n}\sum_{i=1}^{n}|\hat{r}_{i,2}(\mathbf{X}_i)|.$$

Since

$$\mathbb{E}[|\widehat{r}_{i,2}(\mathbf{X}_i)|] \le \mathbb{E}\left[|Y_2|\frac{K(h^{-1}d(\mathbf{X}_1,\mathbf{X}_2))}{f_h(\mathbf{X}_1)}\right] \le C\mathbb{E}[|Y|],$$

it suffices to use the bounds obtained before for W_n and D_n to obtain Equation (21).

(ii) *Proof of Lemma* 3(b). Similarly, since

$$\hat{r}_{i,2}(\chi) - \hat{r}_2(\chi) = \frac{\hat{r}_2(\chi)}{n-1} - \frac{A_iY_iK(h^{-1}d(\chi,\mathbf{X}_i))}{f_h(\mathbf{X}_i)(n-1)},$$

one can write

$$\begin{aligned}\sqrt{n}|S_n| &= \sqrt{n}\left|\frac{1}{n}\sum_{i=1}\frac{r_2(\mathbf{X}_i) - \hat{r}_{i,2}(\mathbf{X}_i)}{r_1^2(\mathbf{X}_i)}(\hat{r}_{i,1}(\mathbf{X}_i) - r_1(\mathbf{X}_i))\right| \\ &\le C\sqrt{n}D_n\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{r}_2(\mathbf{X}_i)|}{n-1} + \frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_iK(0)}{f_h(\mathbf{X}_i)(n-1)} + \sup_{\chi\in\mathcal{S}_\mathcal{F}}|\hat{r}_2(\chi) - r_2(\chi)|\right\},\end{aligned}$$

and so Equation (22) holds by using the bound obtained before for D_n together with Equation (A8).