Estimating additive models with missing responses

Graciela Boente
Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET
Alejandra Martínez
Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET

Abstract

For multivariate regressors, the Nadaraya–Watson regression estimator suffers from the well–known *curse of dimensionality*. To overcome this drawback, additive regression models have been introduced. All the procedures developed, up to now, to estimate the components under an additive model, assume that we observe all the data. However, in many applied statistical analysis missing data occur. In this paper, we study the effect of missing responses on the estimation of the regression function, under an additive regression model. The estimators are based on marginal integration adapted to the missing situation. The proposed estimators turn out to be consistent under mild assumptions. A simulation study allows to compare the behaviour of the our procedures, under different scenarios.

Key Words: Additive models, Kernel weights, Nonparametric regression, Marginal integration, Missing Data

AMS Subject Classification:

1 Introduction

Most commonly used models in statistics are parametric and the assumption is that the observations in the sample belong to a known parametric family. In these cases, the problem consists in estimating or making inference on the unknown parameters. However, in many situations, this assumption may be relatively strong since the assumed parametric model may not be the correct one if there is some. On the other hand, as is well known, statistical methods developed for a particular parametric model can lead to wrong conclusions when they are applied to a slightly disturbed model. Due to these problems, nonparametric and semiparametric methods have been developed for data analysis. In particular, nonparametric regression models have gain importance when studying natural phenomenons with non linear complexity behaviour. Let us assume that we have independent observations $(y_i, \mathbf{x}_i^{\intercal})$, $1 \leq i \leq n$ such that $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^d$ and

$$y_i = m(\mathbf{x}_i) + \sigma(\mathbf{x}_i)\epsilon_i \qquad 1 \leq i \leq n. \tag{1}$$

where the errors ϵ_i are independent and independent of $\mathbf{x}_i$ with $\mathbb{E}(\epsilon_i) = 0$ and $\textsc{Var}(\epsilon_i) < \infty$. The estimation of m under model (1) needs multivariate smoothing techniques. Hence, it suffers from the well known *curse of the dimensionality* which is associated to the fact that as dimension increases, neighbourhoods of a point $\mathbf{x}$ become more sparse. This phenomenon is inherited by the convergence rate of the regression estimators that is not $\sqrt{n}$ as in the parametric case. Instead, when considering kernel estimators, the rate of convergence is $(nh_n^d)^{\frac{1}{2}}$ where h_n stands for the bandwidth or smoothing parameter used in the computation of the estimator. In order to solve this problem, several authors have considered the problem of reducing the dimension of the covariates in nonparametric models. Hastie y Tibshirani (1990) introduced additive models which solve the *curse of the dimensionality* and provide the easy interpretation of univariate smoothers since each component estimate can be plotted separately. In this sense, additive models combine the flexibility of the nonparametric models with the easy interpretation of the standard linear model. To be more precise, additive models assume that $m(\mathbf{x}) = \mu + \sum_{j=1}^{d} g_j(x_j)$ where the parameter $\mu \in \mathbb{R}$ and the smooth functions $g_j : \mathbb{R} \to \mathbb{R}$ are the quantities to be estimated. Estimators for additive models were studied by several authors and we refer to Hastie and Tibshirani (1990) or more recently, to Härdle *et al.* (2004).

Estimators for additive models are designed for complete data sets and problems arise when missing observations are present. In several situations, there might be a part of the design points on which the responses are missing. A fundamental issue of interest is to study the impact of the missing observations on the performance of the estimators that have been used. Even if there are many situations in which both the response and the explanatory variables are missing, we will focus our attention on those cases where missing data occur only in the responses. This situation arises in many biological experiments where the explanatory variables can be controlled. This pattern is common, for example, in the scheme of double sampling proposed by Neyman (1938), where first a complete sample is obtained and then some additional covariate values are computed since perhaps this is less expensive than to obtain more response values. Throughout this paper, we will assume that missing occurs only on the response variables.

The linear regression analysis of missing data was developed by Yates (1933) who proposed to impute the missing observations using least-square estimates. Along with the idea of imputing missing values through least-square predictions, Cochran (1968) used it to reduce bias in observational

studies, while Afifi and Elashoff (1969) gave asymptotic results of the proposals based on add-on process. In many situations, the incomplete observations are imputed via a preliminary estimator and then, one carries out the estimation of the conditional or unconditional mean of the response variable with the complete sample. The methods considered include kernel smoothing (Cheng, 1994; Chu and Cheng, 1995) nearest neighbour imputation (Chen and Shao, 2000), semiparametric estimation (Wang *et al.*, 2004), nonparametric multiple imputation (Aerts *et al.*, 2002), empirical likelihood over the imputed values (Wang and Rao, 2002), among others. For a nonparametric regression model, González–Manteiga and Pérez–Gonzalez (2004) considered an approach based on local polynomials to estimate the regression function when the response variable y is missing but the covariate $\mathbf{x}$ is totally observed. Wang *et al.* (2004) considered inference on the mean of y under regression imputation of missing responses based on a semiparametric regression model. In this paper, we will assume that the data are missing at random (MAR). Assuming MAR requires the existence of a random mechanism, such that the occurrence of a missing response is independent of the response given the covariates. On the other hand, the assumption of missing completely at random (MCAR) is more restrictive since it requires the missing happen stance is independent of both the response and the covariates. In practice, the assumption of MAR might be justified by the nature of the experiment when it is legitimate to assume that the missingness of the responses mainly depends on the covariates.

The aim of this paper is to describe methods of estimation under an additive model when responses are missing. The paper is organized as follows. The estimators to be considered are described in Section 2. Consistency for these estimators will be derived in Section 3 while the results of a simulation study are summarized in Section 4. Proofs are relegated to the Appendix.

2 The estimators

We will consider inference with an incomplete data set $(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i)$, $1 \leq i \leq n$ where $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ if y_i is missing and $(y_i, \mathbf{x}_i^{\mathrm{T}})$ satisfy model (1) where the errors ϵ_i are such that $\mathbb{E}(\epsilon_i) = 0$ and $m : \mathbb{R}^d \to \mathbb{R}$ is a regression function additive on each component of $\mathbf{x}$, i.e.,

$$m(\mathbf{x}) = \mu + \sum_{\alpha=1}^{d} g_\alpha(x_\alpha)\,, \tag{2}$$

where $g_\alpha : \mathbb{R} \to \mathbb{R}$ are unidimensional smooth functions such that $\mathbb{E}g_\alpha(x_\alpha) = 0$. The condition $\mathbb{E}g_\alpha(x_\alpha) = 0$ is set to identify the components in which case $\mu = \mathbb{E}(y_i)$.

Let $(Y, \mathbf{X}^{\mathrm{T}}, \delta)$ be a random vector with the same distribution as $(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i)$, with $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$. Our aim is to estimate, with the data set at hand, the regression components g_α. An ignorable missing mechanism will be imposed by assuming that Y is missing at random (MAR), that is, δ and Y are conditionally independent given $\mathbf{X}$, i.e.,

$$P\left(\delta = 1 | (Y, \mathbf{X})\right) = P\left(\delta = 1 | \mathbf{X}\right)) = p\left(\mathbf{X}\right)\,. \tag{3}$$

Our estimators will be based on the complete sample, i.e., discarding every incomplete pair of the original sample. For that reason, they will be denoted as *simplified estimators*.

Let $\mathcal{K}$ be a multivariate kernel function such that $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{K} \geq 0$, $\int \mathcal{K}(\mathbf{u})\, d\mathbf{u} = 1$, $\int \mathbf{u}\mathcal{K}(\mathbf{u})\, d\mathbf{u} = \mathbf{0}$, $\int \mathbf{u}\mathbf{u}^{\intercal}\mathcal{K}(\mathbf{u})\, d\mathbf{u} = \mu_2(\mathcal{K})\mathbf{I}_d$. On the other hand, we will denote by $\mathcal{K}_h(\mathbf{u}) = h^{-d}\mathcal{K}(\mathbf{u}/h)$.

Using the set of complete data $\{(y_i, \mathbf{x}_i^{\intercal})\}_{\{i:\delta_i=1\}}$ we can introduce two estimators of m. The first one denoted $\widetilde{m}_{\mathrm{S}}^{(1)}$ uses kernel weights modified multiplying by the indicator of the missing variables in order to adapt to the complete sample and avoid bias. On the other hand, the second one denoted $\widetilde{m}_{\mathrm{S}}^{(2)}$ is related to the internally normalized estimators considered in Hengartner and Sperlich (2005). To be more precise, $\widetilde{m}_{\mathrm{S}}^{(1)}$ and $\widetilde{m}_{\mathrm{S}}^{(2)}$ are defined as

$$\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) = \frac{\displaystyle\sum_{i=1}^{n} \mathcal{K}_h(\mathbf{x} - \mathbf{x}_i)\delta_i y_i}{\displaystyle\sum_{j=1}^{n} \mathcal{K}_h(\mathbf{x} - \mathbf{x}_j)\delta_j} \qquad\qquad \widetilde{m}_{\mathrm{S}}^{(2)}(\mathbf{x}) = \frac{\displaystyle\sum_{i=1}^{n} \frac{\mathcal{K}_h\left(\mathbf{x} - \mathbf{x}_i\right)\delta_i y_i}{\widehat{f}(\mathbf{x}_i)}}{\displaystyle\sum_{k=1}^{n} \frac{\mathcal{K}_h\left(\mathbf{x} - \mathbf{x}_k\right)\delta_k}{\widehat{f}(\mathbf{x}_k)}} \,. \tag{4}$$

where $\widehat{f}(\mathbf{x}) = (1/n)\sum_{j=1}^{n} \mathcal{K}_h\left(\mathbf{x} - \mathbf{x}_j\right)$ is the kernel density estimator and $h = h_n$ is the bandwidth parameter.

Let $\widehat{\mu}$ be an estimator of $\mu = \mathbb{E}(Y)$. Chen (1994) applied kernel regression imputation to estimate μ, see also Chu and Cheng (1995). Another possibility is to consider one of the following estimators

$$\widehat{\mu}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(\mathbf{x}_i) \qquad\qquad \widehat{\mu}^{(2)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i y_i}{\widehat{p}(\mathbf{x}_i)} \,,$$

where $\widehat{m}(\mathbf{x}_i)$ is an estimator of the regression function $m(\mathbf{x})$ such as $\widetilde{m}_{\mathrm{S}}^{(1)}$ or $\widetilde{m}_{\mathrm{S}}^{(1)}$. The estimator $\widehat{\mu}^{(2)}$ is the propensity score estimator and it assumes that the missingness probability p is estimated by $\widehat{p}$ when it is unknown. When $\widetilde{m}_{\mathrm{S}}^{(1)}$ is used as estimator of the regression function, the marginal estimator $\widehat{\mu}^{(1)}$ was previously considered by Cheng and Wei (1986) and Cheng (1990), while Chen (1994) obtained that the estimator $\widehat{\mu} = (1/n)\left(\sum_{i=1}^{n} \delta_i y_i + (1-\delta_i)\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}_i)\right)$ has the same asymptotic distribution as $\widehat{\mu}^{(1)}$. The main disadvantage of $\widehat{\mu}^{(1)}$ is that in practice, it inherits the *curse of dimensionality* problem of the kernel estimator even if its convergence rate will still be root$-n$. On the other hand $\widehat{\mu}^{(2)}$ needs a preliminary estimator of the missing probability. Usually, a parametric model is assumed for the missing probability so, only few parameters need to be estimated. Hirano *et al.* (2000) considered the estimator $\widehat{\mu}^{(2)}$ when a kernel estimator is used to estimate $p(\mathbf{x})$. See Wang *et al.* (2004) for a discussion on different estimators of the response mean.

For the sake of simplicity, from now on, the notation $m(x_\alpha, \mathbf{x}_{\underline{\alpha}i})$ indicates the value of the function m calculated at the vector $\mathbf{x}$ with component α equal to x_α and the other ones equal to those of $\mathbf{x}_i$.

Using the estimators defined in (4), four estimators of the marginal functions using marginal integration can be defined. Two of them are based on the Nadaraya–Watson estimator (Nadaraya, 1964, Watson, 1964) while the other ones are based on the internally normalized method introduced in Hengartner and Sperlich (2005). More precisely, the first procedure average over the observations which can be computationally expensive for large data sets while the second one proposes to marginally integrate the estimators defined through (4). Even if, in most situations, the integrals

cannot be computed analytically and numerical integration is needed, for large data sets, numerical integration over a grid of points may be less expensive than the former procedure which averages over all the data. The estimators are then defined as

$$\widehat{g}_{\alpha,\mathrm{S}}^{(1)}(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n}\widetilde{m}_{\mathrm{S}}^{(1)}(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - \widehat{\mu} \tag{5}$$

$$\widehat{g}_{\alpha,\mathrm{S}}^{(2)}(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n}\widetilde{m}_{\mathrm{S}}^{(2)}(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - \widehat{\mu}. \tag{6}$$

where $\mathbf{x}_{\underline{\alpha}}$ stands for the $(d-1)-$dimensional vector such that $\mathbf{x}_{\underline{\alpha}} = (x_1, \ldots, x_{\alpha-1}, x_{\alpha+1}, \ldots, x_d)^{\mathrm{T}}$ and for any $\mathbf{y}$, we allow the general notation $\widetilde{m}_{\mathrm{S}}^{(j)}(\mathbf{y}) = \widetilde{m}_{\mathrm{S}}^{(1)}(y_\alpha, \mathbf{y}_{\underline{\alpha}})$, $j = 1, 2$ to point out with respect to which components we are adding or integrating.

To introduce the second class of estimators, consider a product measure Q on $\mathbb{R}^d$ with $Q_{\underline{\alpha}}(\mathbf{x}_{\underline{\alpha}}) = Q(\mathbb{R}, \mathbf{x}_{\underline{\alpha}})d\mathbf{x}_{\underline{\alpha}}$ and set $qd\mathbf{x} = dQ$, $q_{\underline{\alpha}}d\mathbf{x}_{\underline{\alpha}} = dQ_{\underline{\alpha}}$. Then, the estimators are defined as

$$\widehat{\widetilde{g}}_{\alpha,\mathrm{S}}^{(1)}(x_\alpha) = \int \widetilde{m}_{\mathrm{S}}^{(1)}(x_\alpha, \mathbf{u}_{\underline{\alpha}}) q_{\underline{\alpha}}(\mathbf{u}_{\underline{\alpha}})\, d\mathbf{u}_{\underline{\alpha}} - \widehat{\mu} \tag{7}$$

$$\widehat{\widetilde{g}}_{\alpha,\mathrm{S}}^{(2)}(x_\alpha) = \int \widetilde{m}_{\mathrm{S}}^{(2)}(x_\alpha, \mathbf{u}_{\underline{\alpha}}) q_{\underline{\alpha}}(\mathbf{u}_{\underline{\alpha}})\, d\mathbf{u}_{\underline{\alpha}} - \widehat{\mu}. \tag{8}$$

Hence, simplified estimators of the regression function that make use of the additive model assumption may be defined as $\widehat{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) = \sum_{\alpha=1}^{d}\widehat{g}_{\alpha,\mathrm{S}}^{(1)}(x_\alpha) + \widehat{\mu}$, $\widehat{m}_{\mathrm{S}}^{(2)}(\mathbf{x}) = \sum_{\alpha=1}^{d}\widehat{g}_{\alpha,\mathrm{S}}^{(2)}(x_\alpha) + \widehat{\mu}$ or $\widehat{\widetilde{m}}_{\mathrm{S}}^{(1)}(\mathbf{x}) = \sum_{\alpha=1}^{d}\widehat{\widetilde{g}}_{\alpha,\mathrm{S}}^{(1)}(x_\alpha) + \widehat{\mu}$ and $\widehat{\widetilde{m}}_{\mathrm{S}}^{(2)}(\mathbf{x}) = \sum_{\alpha=1}^{d}\widehat{\widetilde{g}}_{\alpha,\mathrm{S}}^{(2)}(x_\alpha) + \widehat{\mu}$, respectively, depending if one uses the estimators that average or integrate the preliminary ones.

3 Consistency

3.1 Assumptions and notation

Let $(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i)_{i=1}^n$ be a sequence of independent and identically distributed vectors in $\mathbb{R}^{d+2}$ and $(Y, \mathbf{X}^{\mathrm{T}}, \delta)$ a vector with the same distribution as $(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta)$. Denote $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ and by μ the probability measure of $\mathbf{X}$. We remind some definitions that can be found, for instance, in Devroye (1978).

Definition 1. *The observations $(y_i)_{i=1}^n$ are uniformly bounded if $|Y - m(\mathbf{x})| \leq c$ a.s. for some $c < \infty$.*

Definition 2. *The random variables $(y_i)_{i=1}^n$ are uniformly generalized Gaussian if for some $\sigma \geq 0$ and $c \geq 0$*

$$\sup_{\mathbf{x}} \mathbb{E}\left[e^{\lambda(Y - m(\mathbf{x}))}|\mathbf{X} = \mathbf{x}\right] \leq e^{\frac{\sigma^2\lambda^2}{2(1-|\lambda|c)}}, \text{ for all } \quad |\lambda| \leq \frac{1}{c}\,.$$

Remark 3.1. It is clear that when the observations are uniformly bounded, they are uniformly generalized Gaussian. Besides, if $(y_i, \mathbf{x}_i)_{i=1}^n$ are such that $Y|\mathbf{X} = \mathbf{x} \sim N(m(\mathbf{x}), \sigma^2(\mathbf{x}))$ and $\sup_{\mathbf{x}\in\mathbb{R}^d}\sigma^2(\mathbf{x}) < \infty$, then $(y_i)_{i=1}^n$ are uniformly generalized Gaussian.

In order to derive consistency of $\widehat{m}_{\mathrm{S}}^{(1)}(\mathbf{x})$ and $\widehat{m}_{\mathrm{S}}^{(2)}(\mathbf{x})$, we will need the following set of assumptions

D1. $Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon$ with $\mathbb{E}(\epsilon) = 0$ and $\textsc{Var}(\epsilon) = 1$.

D2. The joint density of the covariates $f_{\mathbf{X}}$ is compactly supported, Lipschitz continuous and strictly bounded away from zero and infinity on the interior of its compact support denoted $\mathcal{C}$.

D3. $P(\delta = 1|\mathbf{X}, Y) = P(\delta = 1|\mathbf{X}) = p(\mathbf{X})$, with $p : \mathbb{R}^d \to \mathbb{R}$ continuous in $\mathcal{C}$ and such that $i(p) = \inf_{\mathbf{x}\in\mathcal{C}} p(\mathbf{x}) > 0$.

D4. $m : \mathbb{R}^d \to \mathbb{R}$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^+$ are continuous in $\mathcal{C}$.

D5. The errors ϵ are independent of $(\mathbf{X}, \delta)$. Furthermore, the sequence $(\epsilon_i)_{i=1}^n$ is uniformly generalized Gaussian.

D6. The sequence $(\epsilon_i^2)_{i=1}^n$ is uniformly generalized Gaussian.

D7. The product measure Q has a continuous density $q(\mathbf{x})$ (with respect to Lebesgue measure) bounded away from zero and infinity. Further, the support of Q is contained in the support of $f(\mathbf{x})$.

For the sake of simplicity, from now on, u and u_j stand for $u_j = \sigma(\mathbf{x}_j)\epsilon_j$ and $u = \sigma(\mathbf{X})\epsilon$, so $Y = m(\mathbf{X}) + u$.

Besides, we will need the following assumptions on the kernel $\mathcal{K}$ and the smoothing parameter h_n.

K1. $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}$ is nonnegative, bounded and $\int \mathcal{K}(\mathbf{u})\, d\mathbf{u} = 1$.

K2. $\mathcal{K}(\mathbf{x}) = K(\|\mathbf{x}\|)$ for some nonincreasing function $K : \mathbb{R}^+ \to \mathbb{R}^+$ such that

i) $u^d K(u) \to 0$ as $u \to \infty$,

ii) $K(u^*) > 0$ for some $u^* > 0$.

H1. $h_n \to 0$ and $nh_n^d / \log n \to \infty$.

The following assumptions will be used to derive the consistency of the marginal effects estimators under the additive model (2). It is worth noticing that, under **D2**, the density function of the component X_α, denoted by f_α, has a compact support denoted $\mathcal{C}_\alpha = \mathrm{sop} f_\alpha$.

A1. $m(\mathbf{x}) = \mu + \sum_{\alpha=1}^d g_\alpha(x_\alpha)$.

A2. a) $\mathbb{E}g_\alpha(X_\alpha) = 0$ for all $1 \le \alpha \le d$.
b) $\int g_\alpha(x_\alpha) q_\alpha(x_\alpha)\, dx_\alpha = 0$ where $q_\alpha(x)dx = dQ_\alpha(x)$ and Q_α is the $\alpha-$th marginal of the measure Q.

A3. g_α is a continuous function in $\mathcal{C}_\alpha$ for all $1 \le \alpha \le d$.

Remark 3.2. Separable regression models, as the one we are studying, are useful tools in analysing high-dimensional data sets because these models are not subject to the course of dimensionality, see, for instance, Stone (1986). Separable models are also of interest in econometric theory. Weak separable functions form a flexible class of functions which provides good approximations to continuous functions of several variables. Thus, even if the true underlying regression function is not separable, it may be well approximated by a separable one.

Remark 3.3. The assumptions mentioned above were considered by Buja *et al.* (1989), Hastie and Tibshirani (1990), Newey (1994), Tjostheim and Auestad (1994), Linton and Nielsen (1995), Hengartner and Sperlich (2005), Härdle *et al.* (2004) among others. These are rather typical assumptions for ordinary kernel smoothing.

A1 sets that the model under consideration is an additive one, while **A2** ensures that the additive components g_j are identifiable. Assumptions **D2** and **D4** state regularity conditions on the marginal density of $\mathbf{X}$ and on the conditional distribution function. Note that **D3** implies that some response variables are observed for all $\mathbf{x} \in \mathcal{C}$. This assumption ensures the uniform convergence all over the compact set $\mathcal{C}$. Condition **D5** is needed to obtain the almost surely uniform consistency of both preliminary estimators $\widetilde{m}_{\mathrm{S}}^{(1)}$ and $\widetilde{m}_{\mathrm{S}}^{(2)}$. To obtain asymptotic properties of the estimators based on the internally normalized method **D6** is also required. Condition **D7** allows us to interchange means with integrals to obtain the consistency of the estimators $\widehat{\widehat{g}}$ and $\widehat{\widehat{m}}$. Condition **K1** is a typical assumption for ordinary kernel smoothing. **K2** restricts the class of kernel functions to be chosen and establishes conditions on the rate of convergence of the smoothing paramaters, which are standard in nonparametric regression. Some relation between the bandwidth parameter h_n and the sample size n is always necessary. To obtain the consistency of the proposals **H1** is assumed.

Given a function $g : \mathbb{R}^d \to \mathbb{R}$, $i(g)$ and $\|g\|_{0,\infty}$ stand for $i(g) = \inf_{\mathbf{x}\in\mathcal{C}} g(\mathbf{x})$ and $\|g\|_{0,\infty} = \sup_{\mathbf{x}\in\mathcal{C}} |g(\mathbf{x})|$, respectively. Besides, for any function $g : \mathbb{R} \to \mathbb{R}$, we will denote by $i_\alpha(g) = \inf_{x\in\mathcal{C}_\alpha} g(x)$ and by $\|g\|_{\alpha,\infty} = \sup_{x\in\mathcal{C}_\alpha} |g(x)|$.

Finally, we will denote by $\widehat{m}_Z(\mathbf{x})$ the Nadaraya–Watson estimator of the regression function, $\mathbb{E}(Z|\mathbf{X})$, based on the observations $(z_i, \mathbf{x}_i^{\mathrm{T}})$ computed using with the kernel $\mathcal{K}$ and the bandwidth h_n, that is,

$$\widehat{m}_Z(\mathbf{x}) = \frac{\sum_{i=1}^n \mathcal{K}_{h_n}(\mathbf{x} - \mathbf{x}_i)\, z_i}{\sum_{i=1}^n \mathcal{K}_{h_n}(\mathbf{x} - \mathbf{x}_i)} \,. \tag{9}$$

3.2 Strong uniform convergence of the simplified estimators

We begin by proving strong consistency of the preliminary estimators $\widetilde{m}_{\mathrm{S}}^{(1)}$ and $\widetilde{m}_{\mathrm{S}}^{(2)}$ defined in (4) respectively.

Theorem 3.2.1. *Under **D1** to **D5**, **K1**, **K2** and **H1**, we have that*

a) $\sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0.$

b) $\sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}_{\mathrm{S}}^{(2)}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$ *if in addition, **D6** holds.*

As mentioned in Section 2, the estimators $\widehat{\mu}^{(1)}$ and $\widehat{\mu}^{(2)}$ have been previously considered in the literature, where, for instance, asymptotic normality was derived for different choices of the estimators $\widehat{m}(\mathbf{x})$ and $\widehat{p}(\mathbf{x})$. Proposition 3.2.1 below give a general consistency result, that will be useful in the sequel. Its proof is immediate so, it is omitted.

Proposition 3.2.1. *Let $\widetilde{m}$ be an estimator of the regression function such that $\sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$ and assume that **D1** and **D2** hold. Then, $\widehat{\mu} \xrightarrow{a.s.} \mu$ where $\widehat{\mu} = \sum_{i=1}^{n} \widetilde{m}(\mathbf{x}_i)/n$.*

A consequence of Theorem 3.2.1 and Proposition 3.2.1 is the strong consistency of the estimator $\sum_{i=1}^{n} \widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}_i)/n$ considered by Cheng and Wei (1986) and Cheng (1990). In particular, under **A1**, **A2**, **A3**, **D1** to **D5**, **K1**, **K2** and **H1**, we have that $\widehat{\mu}^{(1)} = (1/n)\sum_{i=1}^{n} \widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}_i) \xrightarrow{a.s.} \mu$.

We now state a strong consistency result for the estimators considered by Hirano *et al.* (2000).

Theorem 3.2.2. *Assume that **D1** to **D4** hold and let $\widehat{p}$ be an estimator of the missing probability such that $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{p}(\mathbf{x}) - p(\mathbf{x})| \xrightarrow{a.s.} 0$, then $\widehat{\mu}^{(2)} = (1/n)\sum_{i=1}^{n}(\delta_i y_i)/\widehat{p}(\mathbf{x}_i) \xrightarrow{a.s.} \mu$.*

Theorem 3.2.3. *Assume that **D2**, **A1**, **A2**a) and **A3** hold. Let $\widehat{\mu}$ a consistent estimator of μ and $\widetilde{m}(\mathbf{x})$ an estimator of the regression function such that $\sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$. Define*

$$\widehat{g}_\alpha(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}(x_\alpha, \mathbf{x}_{\underline{\alpha} i}) - \widehat{\mu}\,.$$

Then, we have that

a) $\sup_{x\in\mathcal{C}_\alpha} |\widehat{g}_\alpha(x_\alpha) - g_\alpha(x_\alpha)| \xrightarrow{a.s.} 0$

b) $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$, *where* $\widehat{m}(\mathbf{x}) = \sum_{\alpha=1}^{d} \widehat{g}_\alpha(x_\alpha) + \widehat{\mu}$.

Theorems 3.2.1 and 3.2.3 entail the consistency of the simplified estimators of the additive components which is stated in the following Corollary.

Corollary 3.2.1. *Assume **D1** to **D5**, **A1**, **A2**a) and **A3**. Let $\mathcal{K}$ a multivariate kernel satisfying **K1** and **K2** and $\mathbf{H} = h_n I_d$ where h_n satisfies **H1**. Then, we have that*

a) for $1 \le \alpha \le d$, $\sup_{x\in\mathcal{C}_\alpha} |\widehat{g}_{\alpha,\mathrm{S}}^{(1)}(x) - g_\alpha(x)| \xrightarrow{a.s.} 0$ and $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$.

*b) for $1 \le \alpha \le d$, $\sup_{x\in\mathcal{C}_\alpha} |\widehat{g}_{\alpha,\mathrm{S}}^{(2)}(x) - g_\alpha(x)| \xrightarrow{a.s.} 0$ and $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\mathrm{S}}^{(2)}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$, if in addition **D6** holds.*

Theorem 3.2.4. *Assume that **D2**, **D6**, **D7**, **A1**, **A2**b) and **A3** hold. Let $\widehat{\mu}$ a consistent estimator of μ and $\widetilde{m}(\mathbf{x})$ a regression function estimator such that $\sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$. Define*

$$\widehat{\widehat{g}}_{\alpha,\mathrm{S}}(x_\alpha) = \int \widetilde{m}_{\mathrm{S}}(x_\alpha, \mathbf{u}_{\underline{\alpha}}) q_{\underline{\alpha}}(\mathbf{u}_{\underline{\alpha}})\, d\mathbf{u}_{\underline{\alpha}} - \widehat{\mu}\,.$$

Then, we have that

a) $\sup_{x\in\mathcal{C}_\alpha} |\widehat{\widehat{g}}_\alpha(x_\alpha) - g_\alpha(x_\alpha)| \xrightarrow{a.s.} 0$

b) $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{\widehat{m}}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$, *where* $\widehat{\widehat{m}}(\mathbf{x}) = \sum_{\alpha=1}^{d} \widehat{\widehat{g}}_\alpha(x_\alpha) + \widehat{\mu}$.

From Theorems 3.2.1 and 3.2.4, we obtain the consistency of the estimators $\widehat{\widehat{g}}^{(1)}_{\alpha,\mathrm{S}}$ and $\widehat{\widehat{g}}^{(2)}_{\alpha,\mathrm{S}}$ defined through (7) and (8), which is stated below.

Corollary 3.2.2. *Assume* ***D1*** *to* ***D5***, ***D7***, ***A1***, ***A2****b) and* ***A3****. Let* $\mathcal{K}$ *a multivariate kernel satisfying* ***K1*** *and* ***K2*** *and* $\mathbf{H} = h_n I_d$ *where* h_n *satisfies* ***H1****. Then, we have that*

a) for $1 \leq \alpha \leq d$, $\sup_{x\in\mathcal{C}_\alpha} |\widehat{\widehat{g}}^{(1)}_{\alpha,\mathrm{S}}(x) - g_\alpha(x)| \xrightarrow{a.s.} 0$ *and* $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{\widehat{m}}^{(1)}_{\mathrm{S}}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$.

b) for $1 \leq \alpha \leq d$, $\sup_{x\in\mathcal{C}_\alpha} |\widehat{\widehat{g}}^{(2)}_{\alpha,\mathrm{S}}(x) - g_\alpha(x)| \xrightarrow{a.s.} 0$ *and* $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{\widehat{m}}^{(2)}_{\mathrm{S}}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$, *if in addition* ***D6*** *holds.*

4 Monte Carlo Study

4.1 General Description

This Section contains the results of a simulation study conducted with the aim of comparing the performance of the estimators $\widetilde{m}^{(1)}_{\mathrm{S}}$, $\widehat{m}^{(1)}_{\mathrm{S}}$, $\widehat{m}^{(2)}_{\mathrm{S}}$, defined in Section 2. We performed $NR = 500$ replications generating independent samples $\{(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i)\}_{i=1}^{n}$ of size $n = 500$. To this end, we first generate observations $(z_i, \mathbf{x}_i^{\mathrm{T}})$ such that

$$z_i = m(\mathbf{x}_i) + u_i, \qquad 1 \leq i \leq n\ ,$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}) \sim U([0,1] \times [0,1])$, $u = \sigma\epsilon$ with $\epsilon \sim N(0,1)$ and $\sigma = 0.5$, $m : \mathbb{R}^2 \to \mathbb{R}$ an additive function of the form

$$m(x_1, x_2) = 4 + 24\left(x_1 - \frac{1}{2}\right)^2 + 2\pi \sin(\pi x_2)\ . \tag{10}$$

Missing responses are defined using different missing schemes as $y_i = z_i$ if $\delta_i = 1$ and missing otherwise, where $\{\delta_i\}_{i=1}^{n}$ are generated under a MAR model with missing probability p equal to one of the following functions

- $p_1(\mathbf{x}) \equiv 1$ which corresponds to the situation of complete samples.
- $p_2(\mathbf{x}) \equiv 0.8$ which corresponds to missing completely at random responses.
- $p_3(\mathbf{x}) = 0.4 + 0.5(\cos(2x_1x_2 + 0.4))^2$.

Besides, x_{i1}, x_{i2}, δ_i and u_i are generated independently to each other.

To identify the marginal components and according to **A2**a), their expectation is set equal to 0. Then, for model (10), we have that $\mu = 10$ and the additive components are $g_1(x_1) =$

$24\,(x_1-0.5)^2-2$ and $g_2(x_2)=2\pi\sin(\pi x_2)-4$. For the smoothing procedure, we use the Epanechnikov multiplicative kernel $\mathcal{K}(\mathbf{x})=K(x_1)K(x_2)$ where $K(u)=(3/4)(1-u^2)\mathrm{I}_{[-1,1]}(u)$.

The behaviour of an estimator $\widehat{m}$ of m is measured using an approximation of the integrated squared error calculated at each replication as

$$\textsc{ise}(\widehat{m})=\frac{1}{\ell^2}\sum_{s=1}^{\ell}\sum_{j=1}^{\ell}\left(m\left(\mathbf{u}_{js}\right)-\widehat{m}\left(\mathbf{u}_{js}\right)\right)^2\ ,$$

where $\mathbf{u}_{js}=(j/\ell,s/\ell)$, $1\leq j,s\leq\ell$, $\ell=50$. An approximation of the MISE is obtained averaging the ISE over replications.

On the other hand, to avoid the high influence on the ISE of the estimation on the boundary, a weighted measure is introduced as

$$\textsc{wise}(\widehat{m})=\frac{1}{\ell^2}\sum_{s=1}^{\ell}\sum_{j=1}^{\ell}\left(m\left(\mathbf{u}_{js}\right)-\widehat{m}\left(\mathbf{u}_{js}\right)\right)^2\mathcal{W}(\mathbf{u}_{js})\ ,$$

where $\mathbf{u}_{js}=(j/\ell,s/\ell)$, $1\leq j,s\leq\ell$, $\ell=50$, $\mathcal{W}(\mathbf{x})=W(x_1)W(x_2)$ with $W(t)=I_{(\tau,1-\tau)}(t)$ with τ a parameter that may be taken as the bandwidth used in the computation of the estimator. The value WMISE refers to the average of WISE over replications.

Similar measures were used for the estimators of the additive components g_α. We first report the results obtained for the estimators of the location parameter μ.

4.2 Selecting the estimator of the expectation of Y

Recall that $Eg_\alpha(X_\alpha)=0$ and hence, $E(Y)=\mu=10$ under the additive model (10). In Section 2, we have introduced two estimators of μ defined as

$$\widehat{\mu}^{(1)}=\frac{1}{n}\sum_{i=1}^{n}\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}_i)\qquad\widehat{\mu}^{(2)}=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i y_i}{p(\mathbf{x}_i)}\ .$$

In order to choose one of them, we perform a preliminary study based on 1000 replications and samples of size $n=500$ to compare these two estimators, as well as the performance of the estimators of the additive components, $\widehat{g}_{\alpha,\mathrm{S}}^{(\ell)}$, $\ell=1,2$, $\alpha=1,2$, related to each one. Table 1 a) reports the mean, standard deviations (SD) and mean square errors (MSE) of the two estimators of μ, while Table 1b) gives the MISE for the two resulting estimators of g_α given in (5) and (6) when the missing probability equals $p_3(\mathbf{x})=0.4+0.5(\cos(2x_1x_2+0.4))^2$.

As observed in Table 1, by oversmoothing the kernel estimator, $\widetilde{m}_{\mathrm{S}}^{(1)}$, that is, selecting a bandwidth equal to $h_n=0.2$, the behaviour of $\widehat{\mu}^{(1)}$ in terms of MSE is much better than the behaviour of $\widehat{\mu}^{(2)}$. Note that the standard deviation of $\widehat{\mu}^{(2)}$ is three times larger that the deviation corresponding to $\widehat{\mu}^{(1)}$ and this is caused by the missing mechanism chosen which induces a missing average of observations close to 40%. It is worth noticing that when no missing responses arise, that is, when $p\equiv 1$, the estimator $\widehat{\mu}^{(2)}$ equals $\overline{y}$. In this case, the mean square error of $\widehat{\mu}^{(1)}$ is 0.0143, i.e., missing the observations only increase 27% times the MSE of the estimator. Similarly, the estimators of the

(a)

	$\widehat{\mu}^{(1)}$	$\widehat{\mu}^{(2)}$
Mean	10.00002	9.99558
SD	0.13520	0.41963
MSE	0.01828	0.17611

(b)

	ℓ =1		ℓ =2	
Estimator of μ	$\widehat{\mu}^{(1)}$	$\widehat{\mu}^{(2)}$	$\widehat{\mu}^{(1)}$	$\widehat{\mu}^{(2)}$
r =1	0.3046	2.1030	0.8131	2.8487
r =2	0.2891	0.3989	0.7800	1.1270

Table 1: (a) Summary measures for the estimators of the expectation of Y and (b) MISE of the estimators $\widehat{g}^{(\ell)}_{r,\mathrm{S}}$, $r, \ell = 1, 2$, according to the estimator of μ used under the missing mechanism p_3.

additive components have a better performance when using $\widehat{\mu}^{(1)}$, the advantage being larger for the first additive component estimator $\widehat{g}^{(\ell)}_{1,\mathrm{S}}$, for any of the two preliminary regression estimators $\widetilde{m}^{(1)}_{\mathrm{S}}$ or $\widetilde{m}^{(2)}_{\mathrm{S}}$. For these reasons, we have selected $\widehat{\mu}_1$ as estimator of the marginal mean in the rest of our study. It is worth noticing that, when selecting the bandwidth through a cross–validation method to estimate the additive components, the bandwidth for the estimator $\widehat{\mu}^{(1)}$ of the expectation of Y was kept fixed and equal to $h_n = 0.2$.

4.3 Results with fixed bandwidths

Before using a cross–validation method to select an automatic bandwidth for each sample, we have performed a simulation study to analyse the performance of the estimators over a fixed grid of bandwidths: $h = 0.15$, 0.2, 0.25 and 0.3. The density estimator was computed using a bandwidth equal to $h = 0.20$. Tables 2 and 3 report the obtained results. From these Tables, we observe that the best results are obtained for $h = 0.15$, in all cases. Besides taking into account the additive structure reduces the values of the MISE. Finally, the estimators based on the preliminary regression estimators internally normalized have a better performance. In particular, the advantage of the correction suggested by in Hengartner and Sperlich (2005) is observed when avoiding the border effects, that is, when considering the WMISE.

	$p = p_1$				$p = p_2$				$p = p_3$			
h	0.15	0.20	0.25	0.30	0.15	0.20	0.25	0.30	0.15	0.20	0.25	0.30
$\widetilde{m}^{(1)}_{\mathrm{S}}$	0.253	0.484	0.832	1.295	0.270	0.502	0.853	1.319	0.308	0.552	0.923	1.411
$\widehat{m}^{(1)}_{\mathrm{S}}$	0.217	0.456	0.811	1.279	0.227	0.471	0.831	1.305	0.252	0.511	0.893	1.391
$\widetilde{m}^{(2)}_{\mathrm{S}}$	0.183	0.363	0.654	1.066	0.197	0.376	0.668	1.083	0.224	0.401	0.702	1.134
$\widehat{m}^{(2)}_{\mathrm{S}}$	0.157	0.343	0.637	1.051	0.165	0.351	0.648	1.067	0.179	0.369	0.676	1.111
$\widehat{g}^{(1)}_{1,\mathrm{S}}$	0.119	0.237	0.406	0.624	0.124	0.244	0.416	0.637	0.143	0.270	0.453	0.685
$\widehat{g}^{(1)}_{2,\mathrm{S}}$	0.107	0.225	0.407	0.654	0.112	0.231	0.414	0.662	0.128	0.254	0.446	0.705
$\widehat{g}^{(2)}_{1,\mathrm{S}}$	0.091	0.184	0.327	0.522	0.095	0.189	0.333	0.531	0.107	0.203	0.352	0.556
$\widehat{g}^{(2)}_{2,\mathrm{S}}$	0.063	0.171	0.320	0.538	0.084	0.175	0.325	0.543	0.096	0.188	0.342	0.566

Table 2: MISE of the simplified estimators of m, g_1 y g_2 under different missingness probabilities $p_1(\mathbf{x}) \equiv 1$, $p_2(\mathbf{x}) \equiv 0.8$ and $p_3(\mathbf{x}) = 0.4 + 0.5(\cos(2x_1x_2 + 0.4))^2$.

	$p = p_1$				$p = p_2$				$p = p_3$			
h	0.15	0.20	0.25	0.30	0.15	0.20	0.25	0.30	0.15	0.20	0.25	0.30
$\widehat{g}^{(1)}_{1,\mathrm{S}}$	0.083	0.138	0.203	0.237	0.087	0.151	0.208	0.242	0.100	0.157	0.226	0.260
$\widehat{g}^{(1)}_{2,\mathrm{S}}$	0.026	0.053	0.104	0.161	0.027	0.054	0.104	0.161	0.019	0.040	0.085	0.134
$\widehat{g}^{(2)}_{1,\mathrm{S}}$	0.063	0.107	0.164	0.198	0.066	0.110	0.166	0.201	0.075	0.118	0.176	0.212
$\widehat{g}^{(2)}_{2,\mathrm{S}}$	0.017	0.030	0.061	0.099	0.018	0.030	0.061	0.098	0.013	0.020	0.045	0.081

Table 3: WMISE of the simplified estimators for the marginal functions under different missingness probabilities $p_1(\mathbf{x}) \equiv 1$, $p_2(\mathbf{x}) \equiv 0.8$ and $p_3(\mathbf{x}) = 0.4 + 0.5(\cos(2x_1x_2 + 0.4))^2$.

4.4 Data-driven bandwidths

An important issue in any smoothing procedure is the choice of the smoothing parameter. Under a nonparametric regression model, two commonly used approaches are cross–validation and plug–in. As is well known, plug-in methodologies require us to obtain theoretical expressions of the bias and the variance of regression estimators, which are not always available in practice. Among others, for additive models with no missing data, Opsomer (200) developed a plug-in bandwidth estimator for backfitting estimators, in the case of independence between the covariates while Mammen and Park (2005) introduced bandwidth selectors for smooth backfitting based on penalized sums of squared residuals. Finally, Nielsen and Sperlich (2005) developed a cross-validation method for the smooth backfitting estimator. Recently, a data-driven local bandwidth selector based on a Wild Bootstrap approximation of the mean squared error of the estimators was developed by Martínez–Miranda *et al.* and extended to the situation with missing responses by Martínez–Miranda and Raya–Miranda (2011). In our simulation study, we have selected as criterion the cross–validation method, performed over the observed observations, that is,

$$\widehat{\mathbf{h}} = \arg \min_{\mathbf{h}=(h_1,h_2)\in\mathbb{R}^2_+} \sum_{i=1}^{n} (y_i - \widehat{m}_{-i,\mathrm{S}}(\mathbf{x}_i, \mathbf{h}))^2 \delta_i \,. \tag{11}$$

where $\widehat{m}_{-i,\mathrm{S}}(\cdot, \mathbf{h})$ represents the leave-one-out estimator corresponding to $\widehat{m}_{\mathrm{S}}(\cdot, \mathbf{h})$. As in cross–validation with complete data sets, the $i-$th observation $(y_i, \mathbf{x}_i)$ is not used to predict y_i. In this way, we ensure that the observations used to calculate $\widehat{m}_{-i,\mathrm{S}}(\cdot, \mathbf{h})$ are independent of $\mathbf{x}_i$, the observation at which we evaluate $\widehat{m}_{-i,\mathrm{S}}(\cdot, \mathbf{h})$ to predict the $i-$response, when it is not missing.

4.4.1 Optimal Bandwidths

In order to have an asymptotic counterpart for the cross–validation bandwidth, an optimal deterministic smoothing parameter was selected for each of these estimators and for each missing probability using as goodness of fit criterion the mean integrated square error, MISE,

$$\mathrm{MISE}(\mathbf{h}) = E \int (m(\mathbf{x}) - \widehat{m}(\mathbf{x}, \mathbf{h}))^2 \, d\mathbf{x},$$

where $\widehat{m}(\cdot, \mathbf{h})$ denotes the estimator to be considered using as bandwidth the value $\mathbf{h}$. We performed 500 replications generating independent samples $\{(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i)\}_{i=1}^n$ of size $n = 500$ following the model described in Section 4.1. For each value of the smoothing parameter, the value of

the MISE was approximated by Monte Carlo as $\sum_{k=1}^{500} M(\mathbf{h},k)/500$, where for each replication k, $M(\mathbf{h},k) = \sum_{j=1}^{\ell}\sum_{s=1}^{\ell}\left(m\left(\mathbf{u}_{js},\mathbf{h}\right) - \widehat{m}\left(\mathbf{u}_{js},\mathbf{h}\right)\right)^2/\ell^2$, with $\mathbf{u}_{js} = (j/\ell, s/\ell)$, $1 \leq j,s \leq \ell$ and $\ell = 50$ as in the computation of the ISE. For each of the three missing probabilities, the optimal smoothing parameter $\mathbf{h}$ was selected on a diagonal grid of points in $\mathbb{R}^2$, that is, we assumed $h_1 = h_2 = h$, so that $\mathbf{h} = (h,h)$ with $h \in \mathcal{G}$ where $\mathcal{G} = \{0.03, 0.04\} \cup \mathcal{G}_0$ with $\mathcal{G}_0$ a grid of 14 equidistant points between 0.045 and 0.08. When the minimization process leads to a value on the boundary, the search was carried on over the limits of the interval. To be more precise, if in the first step the bandwidth selected equals 0.03, the minimization was carried on over the grid $\mathcal{G}_1 = \{0.015, 0.02, 0.025, 0.03, 0.035\}$. On the other hand, if the bandwidth selected was equal to 0.8, the minimization was done over the grid $\mathcal{G}_2 = \{0.0775, 0.08, 0.085, 0.09, 0.1\}$. Table 4 reports the values obtained in each situation. We denote $\mathbf{h}_{\text{OPT}} = (h_{\text{OPT}}, h_{\text{OPT}})$ the optimal bandwidth obtained.

	$p = p_1$	$p = p_2$	$p = p_3$
$\widehat{m}^{(1)}$	0.0550	0.0600	0.0675
$\widehat{m}^{(2)}$	0.0600	0.0650	0.0700

Table 4: Optimal smoothing parameters h_{OPT} for each scenario and for each nonparametric estimator.

4.4.2 Cross-validation bandwidth

A data–driven selector was discussed in Section 4.4. We have computed the data–driven bandwidths for each of the missing probabilities. As above, the data–driven smoothing parameter $\mathbf{h}$ was selected on a diagonal grid of points in $\mathbb{R}^2$, that is, $\mathbf{h} = (h,h)$ with $h \in \mathcal{G}$ and $\mathcal{G}$ as in Section 4.4.1. Besides, when the minimization process leads to a value on the boundary, the search was carried on over the limits of the interval.

We denote $\mathbf{h}_{\text{CV}} = (h_{\text{CV}}, h_{\text{CV}})$ the optimal bandwidth obtained. Due to the expensive computing time, we have performed $NR = 500$ replications. Once the optimal bandwidth (the asymptotic or the cross–validation one) is selected, the estimators are computed as described in Section 2. Tables 6 and 7 summarize the results obtained using the same measures defined in Section 4.1. Besides, to evaluate the performance of the cross-validation bandwidths with respect to the optimal one, Table 5 reports as summary measures, the minimum, the first quantile, the median, the third quantile and the maximum denoted respectively, Q^0, $Q^{0.25}$, $Q^{0.50}$, $Q^{0.75}$ and Q^1 as well as the mean of $\log\left(h_{\text{CV}}/h_{\text{OPT}}\right)$. On the other hand, Figures 1 and 2 show the histograms and boxplots of $\log\left(h_{\text{CV}}/h_{\text{OPT}}\right)$ obtained for the estimators $\widehat{m}^{(1)}$ and $\widehat{m}^{(2)}$, under different missing schemes, respectively.

When no missing responses arise, or under a completely at random missingness model, the cross–validation bandwidth for $\widehat{m}^{(1)}$ performs better than that obtained when using $\widehat{m}^{(2)}$. Even though, as shown in Tables 6 and 7, the performance of the marginal and final estimators derived from the internally normalized regression estimator $\widetilde{m}^{(2)}$ is better than that obtained from the Nadaraya-Watson estimator.

	Q^0	$Q^{0.25}$	$Q^{0.50}$	Mean	$Q^{0.75}$	Q^1
			$\widehat{m}^{(1)}$			
$p = p_1$	-0.31850	-0.09531	0.00000	-0.01530	0.04445	0.24120
$p = p_2$	-0.28770	-0.08701	0.00000	-0.01170	0.04082	0.28770
$p = p_3$	-0.35140	-0.11780	-0.03774	-0.04342	0.03637	0.23050
			$\widehat{m}^{(2)}$			
$p = p_1$	-0.40550	-0.13350	-0.04256	-0.03937	0.04082	0.28770
$p = p_2$	-0.36770	-0.12260	-0.03922	-0.03087	0.07411	0.32540
$p = p_3$	-0.44180	-0.07411	0.00000	-0.01569	0.06899	0.35670

Table 5: Summary measures of $\log(h_{\mathrm{CV}}/h_{\mathrm{OPT}})$ under the missing schemes $p_1(\mathbf{x}) \equiv 1$, $p_2(\mathbf{x}) \equiv 0.8$ and $p_3(\mathbf{x}) = 0.4 + 0.5(\cos(2x_1x_2 + 0.4))^2$.

	$p = p_1$	$p = p_2$	$p = p_3$
$\widetilde{m}_{\mathrm{S}}^{(1)}$	0.1574	0.1834	0.2253
$\widehat{m}_{\mathrm{S}}^{(1)}$	0.0361	0.0488	0.0773
$\widetilde{m}_{\mathrm{S}}^{(2)}$	0.1443	0.1692	0.2106
$\widehat{m}_{\mathrm{S}}^{(2)}$	0.0340	0.0458	0.0710
$\widehat{g}_{1,\mathrm{S}}^{(1)}$	0.0258	0.0325	0.0518
$\widehat{g}_{2,\mathrm{S}}^{(1)}$	0.0255	0.0298	0.0474
$\widehat{g}_{1,\mathrm{S}}^{(2)}$	0.0248	0.0311	0.0490
$\widehat{g}_{2,\mathrm{S}}^{(2)}$	0.0248	0.0287	0.0450

Table 6: MISE of the simplified estimators of m, g_1 y g_2 under different missing schemes, $p_1(\mathbf{x}) \equiv 1$, $p_2(\mathbf{x}) \equiv 0.8$ and $p_3(\mathbf{x}) = 0.4+0.5(\cos(2x_1x_2+0.4))^2$, when the bandwidth is selected using a cross–validation procedure.

	$p = p_1$	$p = p_2$	$p = p_3$
$\widehat{g}_{1,\mathrm{S}}^{(1)}$	0.0230	0.0285	0.0448
$\widehat{g}_{2,\mathrm{S}}^{(1)}$	0.0174	0.0185	0.0242
$\widehat{g}_{1,\mathrm{S}}^{(2)}$	0.0220	0.0270	0.0420
$\widehat{g}_{2,\mathrm{S}}^{(2)}$	0.0161	0.0175	0.0227

Table 7: WMISE of the simplified estimators for the marginal functions under under different missing schemes, $p_1(\mathbf{x}) \equiv 1$, $p_2(\mathbf{x}) \equiv 0.8$ and $p_3(\mathbf{x}) = 0.4 + 0.5(\cos(2x_1x_2 + 0.4))^2$, when the bandwidth is selected using a cross–validation procedure.

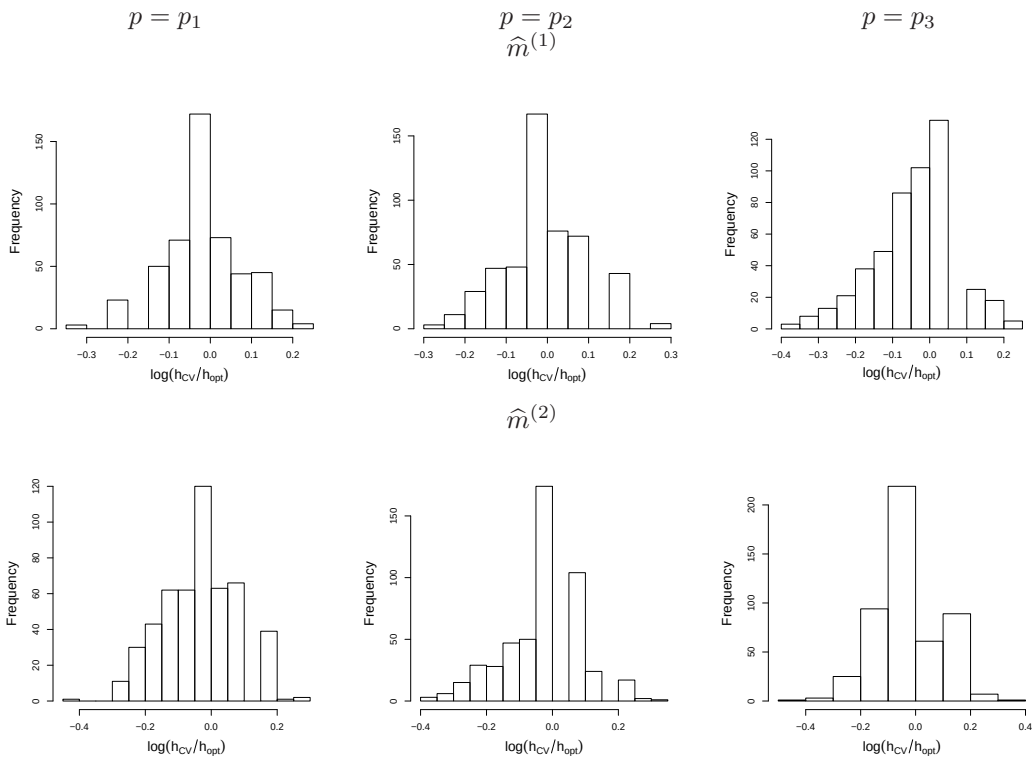

Figura 1: Histogram of $\log(h_{\mathrm{CV}}/h_{\mathrm{OPT}})$ under different missing schemes $p_1(\mathbf{x}) \equiv 1$, $p_2(\mathbf{x}) \equiv 0.8$ and $p_3(\mathbf{x}) = 0.4 + 0.5(\cos(2x_1x_2 + 0.4))^2$. The upper and lower plots correspond to the optimal and data driven selectors when using as estimates $\widehat{m}^{(1)}$ and $\widehat{m}^{(2)}$, respectively.

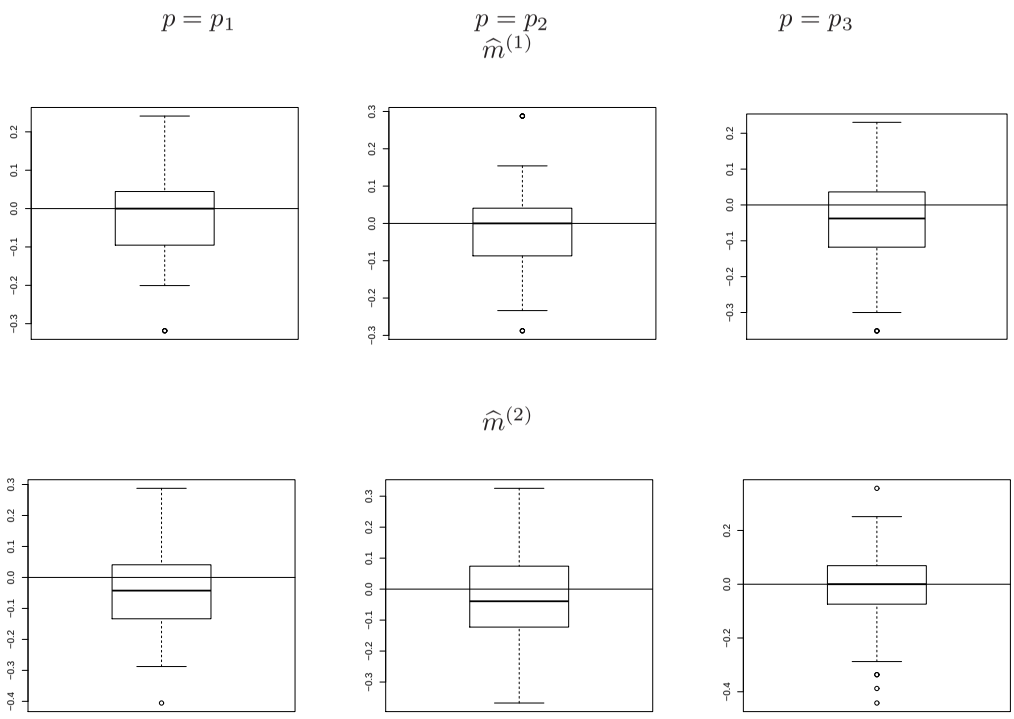

Figura 2: Boxplot of $\log(h_{\mathrm{CV}}/h_{\mathrm{OPT}})$ under different missing schemes, $p_1(\mathbf{x}) \equiv 1$, $p_2(\mathbf{x}) \equiv 0.8$ and $p_3(\mathbf{x}) = 0.4 + 0.5(\cos(2x_1x_2 + 0.4))^2$. The upper and lower plots correspond to the optimal and data driven selectors when using as estimates $\widehat{m}^{(1)}$ and $\widehat{m}^{(2)}$, respectively.

5 Appendix

In order to prove the consistency of the estimators, the following result due to Devroye (1978) will used. We state it for completeness.

Proposition 5.1. *Let $(y_i, \mathbf{x}_i^{\intercal})_{i=1}^n$ a sequence of independent and identically distributed variables and such that $(y_i)_{i=1}^n$ is a uniformly generalized Gaussian sequence. Denote $\widehat{m}_n(\mathbf{x}) = \widehat{m}_Y(\mathbf{x})$ the Nadaraya–Watson estimator defined in (9). Assume **K1**, **K2**, **H1**, m is bounded and continuous in the support of μ and that there exist $a, b > 0$ such that $\inf_{\mathbf{x}\in A} \mu(\mathcal{S}(\mathbf{x}, r)) \geq ar^d$, all $r \in [0, b]$, where $\mathcal{S}(\mathbf{x}, r)$ is the closed sphere with center $\mathbf{x}$ and radius r. Then, for any compact set A, we have that $\sup_{\mathbf{x}\in A} |\widehat{m}_n(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$.*

We first state some Lemmas that will be used in the sequel.

Lemma 5.1. *Let $\widehat{m}_{\delta Y}$ and $\widehat{m}_\delta$ be defined as in (9). Under **D1** to **D5**, **K1**, **K2** and **H1**, we have*

a) $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta Y}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| \xrightarrow{a.s.} 0$

b) $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta}(\mathbf{x}) - p(\mathbf{x})| \xrightarrow{a.s.} 0.$

Proof. We begin by proving a). Note that, as $\delta Y = \delta m(\mathbf{x}) + \delta u$, where $u = \sigma(\mathbf{x})\epsilon$, $\mathbb{E}(\delta Y|\mathbf{X} = \mathbf{x}) = p(\mathbf{x})m(\mathbf{x})$, so

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta Y} - p(\mathbf{x})m(\mathbf{x})| &= \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta m}(\mathbf{x}) + \widehat{m}_{\delta u}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| \\ &\leq \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta m}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta u}(\mathbf{x})|.\end{aligned}$$

Hence, it will be enough to show that

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta m}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| \xrightarrow{a.s.} 0 \tag{12}$$

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta u}(\mathbf{x})| \xrightarrow{a.s.} 0 \tag{13}$$

From **D4**, m is bounded in $\mathcal{C}$, then, the sequence of variables $(\delta_i m(\mathbf{x}_i))_{i=1}^n$ is a sequence of independent, identically distributed and uniformly bounded variables such that $\mathbb{E}[\delta m(\mathbf{X})|\mathbf{X} = \mathbf{x}] = m(\mathbf{x})\mathbb{E}[\delta|\mathbf{X} = \mathbf{x}] = m(\mathbf{x})p(\mathbf{x})$. Thus, using Remark 3.1, we obtain that $(\delta_i m(\mathbf{x}_i))_{i=1}^n$ is a uniformly generalized Gaussian sequence, hence (12) follows from Proposition 5.1.

Let now see that the sequence of independent and identically distributed variables $(\delta_i u_i)_{i=1}^n$ is also a uniformly generalized Gaussian sequence. Using that the errors ϵ are independent of $(\delta, \mathbf{x})$ and that $\mathbb{E}(\epsilon) = 0$, we get $\mathbb{E}(\delta u|\mathbf{X} = \mathbf{x}) = p(\mathbf{x})\sigma(\mathbf{x})\mathbb{E}(\epsilon) = 0$. For any $\lambda \in \mathbb{R}$, we have that

$$\begin{aligned}\mathbb{E}\left[e^{\lambda\delta u}|\mathbf{X} = \mathbf{x}\right] &= \mathbb{E}\left[e^{\lambda\delta\sigma(\mathbf{x})\epsilon}|\mathbf{X} = \mathbf{x}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{\lambda\delta\sigma(\mathbf{x})\epsilon}|\mathbf{X} = \mathbf{x}, Y\right]|\mathbf{X} = \mathbf{x}\right] = \\ &= \mathbb{E}\left[(1 - p(\mathbf{x})) + p(\mathbf{x})\, e^{\lambda\sigma(\mathbf{x})\epsilon}|\mathbf{X} = \mathbf{x}\right] = 1 - p(\mathbf{x}) + p(\mathbf{x})\mathbb{E}\left[e^{\lambda\sigma(\mathbf{x})\epsilon}\right]\end{aligned}$$

As $(\epsilon_i)_{i=1}^n$ is a sequence of independent, identically distributed and uniformly generalized Gaussian variables, there are $\tau \geq 0$ and $c \geq 0$ such that if $|\phi| < 1/c$, thus

$$\mathbb{E}\left(e^{\phi\epsilon}\right) \leq \exp\left\{\frac{\tau^2\phi^2}{2(1 - |\phi|c)}\right\}.$$

D4 entails that σ is bounded in $\mathcal{C}$, so taking $d = c\|\sigma\|_{0,\infty}^2$ and $\tilde{\tau} = \tau\|\sigma\|_{0,\infty}$ we obtain that, for all $|\lambda| \leq 1/d$, $|\phi| = |\lambda|\sigma(\mathbf{x}) \leq 1/c$, hence

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}} \mathbb{E}\left(\exp\left\{\lambda\sigma(\mathbf{x})\epsilon\right\}\right) &\leq \sup_{\mathbf{x}\in\mathcal{C}} \exp\left\{\frac{\tau^2\lambda^2\sigma^2(\mathbf{x})}{(1-|\lambda|\sigma(\mathbf{x})c)}\right\} \\ &\leq \sup_{\mathbf{x}\in\mathcal{C}} \exp\left\{\frac{\tau^2\lambda^2\|\sigma\|_{0,\infty}^2}{(1-|\lambda|c\|\sigma\|_{0,\infty}^2)}\right\} = \exp\left\{\frac{\tilde{\tau}^2\lambda^2}{(1-|\lambda|d)}\right\}.\end{aligned}$$

So if $|\lambda| \leq 1/d$, $1 \leq e^{\tilde{\tau}^2\lambda^2/(1-|\lambda|d)}$, we have that

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}} \mathbb{E}\left[e^{\lambda\delta u}|\mathbf{x}=\mathbf{x}\right] &\leq (1-p(\mathbf{x})) + p(\mathbf{x})\exp\left\{\frac{\tilde{\tau}^2\lambda^2}{(1-|\lambda|d)}\right\} \\ &\leq (1-p(\mathbf{x}))\exp\left\{\frac{\tilde{\tau}^2\lambda^2}{(1-|\lambda|d)}\right\} + p(\mathbf{x})\exp\left\{\frac{\tilde{\tau}^2\lambda^2}{(1-|\lambda|d)}\right\} = \exp\left\{\frac{\tilde{\tau}^2\lambda^2}{(1-|\lambda|d)}\right\},\end{aligned}$$

which entails that $(\delta_j u_j)_{j=1}^n$ is a uniformly generalized Gaussian sequence. As it is also an independent and identically distributed sequence of variables, from Proposition 5.1, we obtain (13).

Finally, b) can be obtained from (12) taking $Y \equiv 1$ or using Proposition 5.1 and the fact that the sequence of independent and identically distributed variables $(\delta_i)_{i=1}^n$ is a uniformly bounded sequence and so a uniformly generalized Gaussian sequence. □

Lemma 5.2. *Let $\mathcal{A}$ be a compact set, $b(\mathbf{x})$ and $f(\mathbf{x})$ two continuous functions in $\mathcal{A}$. Let $\widehat{f}(\mathbf{x}) = \widehat{f}_n(\mathbf{x})$ be such that $\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})| \xrightarrow{a.s.} 0$. Then we have that*

a) $\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{a}(\mathbf{x}) - b(\mathbf{x})f(\mathbf{x})| \xrightarrow{a.s.} 0$, for any $\widehat{a}(\mathbf{x}) = \widehat{a}_n$ such that $\sup_{\mathbf{x}\in\mathcal{A}}\left|\widehat{a}(\mathbf{x})/\widehat{f}(\mathbf{x}) - b(\mathbf{x})\right| \xrightarrow{a.s.} 0$.

b) $\sup_{\mathbf{x}\in\mathcal{A}}\left|\widehat{a}(\mathbf{x})/\widehat{f}(\mathbf{x}) - b(\mathbf{x})\right| \xrightarrow{a.s.} 0$, if $\inf_{\mathbf{x}\in\mathcal{A}} f(\mathbf{x}) > 0$ and $\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{a}(\mathbf{x}) - b(\mathbf{x})f(\mathbf{x})| \xrightarrow{a.s.} 0$.

Proof. a) Note that

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{a}(\mathbf{x}) - b(\mathbf{x})f(\mathbf{x})| &\leq \sup_{\mathbf{x}\in\mathcal{A}}|\widehat{a}(\mathbf{x}) - b(\mathbf{x})\widehat{f}(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{A}}|b(\mathbf{x})|\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})| \\ &\leq \sup_{\mathbf{x}\in\mathcal{A}}\left|\frac{\widehat{a}(\mathbf{x})}{\widehat{f}(\mathbf{x})} - b(\mathbf{x})\right|\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{A}}|b(\mathbf{x})|\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})| \\ &\leq \sup_{\mathbf{x}\in\mathcal{A}}\left|\frac{\widehat{a}(\mathbf{x})}{\widehat{f}(\mathbf{x})} - b(\mathbf{x})\right|\left[\sup_{\mathbf{x}\in\mathcal{A}}|f(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|\right] \\ &\quad + \sup_{\mathbf{x}\in\mathcal{A}}|b(\mathbf{x})|\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|\end{aligned}$$

Now a) follows from the fact that $b(\mathbf{x})$ and $f(\mathbf{x})$ are continuous functions and so, bounded over the compact set $\mathcal{A}$.

b) We have that

$$\sup_{\mathbf{x}\in\mathcal{A}}\left|\frac{\widehat{a}(\mathbf{x})}{\widehat{f}(\mathbf{x})} - b(\mathbf{x})\right| = \frac{\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{a}(\mathbf{x}) - b(\mathbf{x})\widehat{f}(\mathbf{x})|}{\inf_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x})|}$$

$$\leq \frac{\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{a}(\mathbf{x})-b(\mathbf{x})f(\mathbf{x})|+\sup_{\mathbf{x}\in\mathcal{A}}|b(\mathbf{x})|\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x})-f(\mathbf{x})|}{\inf_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x})|}$$
$$\leq \frac{\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{a}(\mathbf{x})-b(\mathbf{x})f(\mathbf{x})|+\sup_{\mathbf{x}\in\mathcal{A}}|b(\mathbf{x})|\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x})-f(\mathbf{x})|}{\inf_{\mathbf{x}\in\mathcal{A}}|f(\mathbf{x})|-\sup_{\mathbf{x}\in\mathcal{A}}|\widehat{f}(\mathbf{x})-f(\mathbf{x})|}$$

Now the result follows from the fact that $b(\mathbf{x})$ is bounded on $\mathcal{A}$, $\inf_{\mathbf{x}\in\mathcal{A}} f(\mathbf{x})>0$ and the uniform strong consistency of $\widehat{a}(\mathbf{x})$ and $\widehat{f}(\mathbf{x})$. □

Proof of Theorem 3.2.1. a) The result follows easily from the Lemma 5.1 since $\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) = \widehat{m}_{\delta Y}(\mathbf{x})/\widehat{m}_{\delta}(\mathbf{x})$. Effectively, let $\mathcal{N}$ the set of probability 0 such that $\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta Y}(\mathbf{x})-p(\mathbf{x})m(\mathbf{x})| \not\to 0$ or $\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x})-p(\mathbf{x})| \not\to 0$ and fix $\omega\notin\mathcal{N}$.

Using that $i(p)>0$, we have that, for $n\geq n_0$, $\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x})-p(\mathbf{x})|\leq i(p)/2$ so, $|\widehat{m}_{\delta}(\mathbf{x})|\geq i(p)/2$ for any $\mathbf{x}\in\mathcal{C}$. Hence, using that m is bounded in $\mathcal{C}$ and $\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) = \widehat{m}_{\delta Y}(\mathbf{x})/\widehat{m}_{\delta}(\mathbf{x})$, we get that

$$\sup_{\mathbf{x}\in\mathcal{C}}\left|\widetilde{m}_{\mathrm{S}}^{(1)}-m(\mathbf{x})\right| \leq \frac{\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta Y}(\mathbf{x})-m(\mathbf{x})\widehat{m}_{\delta}(\mathbf{x})|}{\inf_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x})|}$$
$$\leq \frac{\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta Y}(\mathbf{x})-m(\mathbf{x})p(\mathbf{x})|+\sup_{\mathbf{x}\in\mathcal{C}}|m(\mathbf{x})|\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x})-p(\mathbf{x})|}{\inf_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x})|}$$
$$\leq \frac{2}{i(p)}\left\{\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta Y}(\mathbf{x})-m(\mathbf{x})p(\mathbf{x})|+\|m\|_{0,\infty}\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{m}_{\delta}(\mathbf{x})-p(\mathbf{x})|\right\},$$

concluding the proof of a).

b) For the sake of simplicity denote $\widehat{f}(\mathbf{x}) = \widehat{f}_n(\mathbf{x})$ for all $\mathbf{x}\in\mathcal{C}$. As $y_i = m(\mathbf{x}_i)+u_i$, we have $\widetilde{m}_{\mathrm{S}}^{(2)}(\mathbf{x}) = (B_1(\mathbf{x})+B_2(\mathbf{x}))/B_0(\mathbf{x})$ where

$$B_0(\mathbf{x}) = \frac{1}{nh_n^d}\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{\delta_i}{\widehat{f}(\mathbf{x}_i)}$$
$$B_1(\mathbf{x}) = \frac{1}{nh_n^d}\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{\delta_i m(\mathbf{x}_i)}{\widehat{f}(\mathbf{x}_i)}$$
$$B_2(\mathbf{x}) = \frac{1}{nh_n^d}\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{\delta_i u(\mathbf{x}_i)}{\widehat{f}(\mathbf{x}_i)}$$

Hence, using that $i(p)>0$ and Lemma 5.2, it will be enough to show that

i) $\sup_{\mathbf{x}\in\mathcal{C}}|B_1(\mathbf{x})-p(\mathbf{x})m(\mathbf{x})| \xrightarrow{a.s.} 0$

ii) $\sup_{\mathbf{x}\in\mathcal{C}}|B_2(\mathbf{x})| \xrightarrow{a.s.} 0$

iii) $\sup_{\mathbf{x}\in\mathcal{C}}|B_0(\mathbf{x})-p(\mathbf{x})| \xrightarrow{a.s.} 0$

i) $B_1(\mathbf{x})$ can be written as $B_1(\mathbf{x}) = B_{11}(\mathbf{x})+B_{12}(\mathbf{x})$ where

$$B_{11}(\mathbf{x}) = \frac{1}{nh_n^d}\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{\delta_i m(\mathbf{x}_i)}{f(\mathbf{x}_i)}$$

$$B_{12}(\mathbf{x}) = \frac{1}{nh_n^d}\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\delta_i m(\mathbf{x}_i)\left[\frac{1}{\widehat{f}(\mathbf{x}_i)}-\frac{1}{f(\mathbf{x}_i)}\right].$$

Thus, the proof of i) will be completed if we show that

$$\sup_{\mathbf{x}\in\mathcal{C}}|B_{11}(\mathbf{x})-p(\mathbf{x})m(\mathbf{x})| \xrightarrow{a.s.} 0 \tag{14}$$

$$\sup_{\mathbf{x}\in\mathcal{C}}|B_{12}(\mathbf{x})| \xrightarrow{a.s.} 0 \tag{15}$$

The fact that m y f are bounded in $\mathcal{C}$ entails that the sequence of i.i.d. variables $(\delta_i m(\mathbf{x}_i)/f(\mathbf{x}_i))_{i=1}^n$ are uniformly bounded, so uniformly generalized Gaussian. Using that $\mathbb{E}(\delta m(\mathbf{X})/f(\mathbf{X})|\mathbf{X}=\mathbf{x}) = m(\mathbf{x})p(\mathbf{x})/f(\mathbf{x})$ and Proposition 5.1 we get that

$$\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{1}{\widehat{f}(\mathbf{x})}B_{11}(\mathbf{x})-\frac{p(\mathbf{x})m(\mathbf{x})}{f(\mathbf{x})}\right| \xrightarrow{a.s.} 0\,.$$

On the other hand, **D2**, **K1**, **K2** and **H1** imply that (see Prakasa Rao, 1983)

$$\sup_{\mathbf{x}\in\mathcal{C}}\left|\widehat{f}(\mathbf{x})-f(\mathbf{x})\right| \xrightarrow{a.s.} 0\,, \tag{16}$$

Thus, (14) follows from Lemma 5.2.

Using that $\mathbf{X}$ has compact support, m is bounded on the support of $\mathbf{X}$ and $\mathcal{K}\geq 0$, we obtain that

$$\begin{aligned}
|B_{12}(\mathbf{u})| &= \frac{1}{nh_n^d}\left|\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{u}-\mathbf{x}_i}{h_n}\right)\delta_i m(\mathbf{x}_i)\left[\frac{1}{\widehat{f}(\mathbf{x}_i)}-\frac{1}{f(\mathbf{x}_i)}\right]\right| \\
&\leq \|m\|_{0,\infty}\frac{1}{nh_n^d}\sum_{j=1}^{n}\mathcal{K}\left(\frac{\mathbf{u}-\mathbf{x}_j}{h_n}\right)\frac{\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{f}(\mathbf{x})-f(\mathbf{x})|}{\inf_{\mathbf{x}\in\mathcal{C}}\hat{f}(\mathbf{x})\inf_{\mathbf{x}\in\mathcal{C}}|f(\mathbf{x})|} \\
&\leq \|m\|_{0,\infty}\widehat{f}(\mathbf{u})\frac{\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{f}(\mathbf{x})-f(\mathbf{x})|}{\inf_{\mathbf{x}\in\mathcal{C}}\hat{f}(\mathbf{x})\inf_{\mathbf{x}\in\mathcal{C}}|f(\mathbf{x})|}
\end{aligned}$$

so, (15) follows easily from (16) and the fact that $i(f)>0$.

ii) The proof follows similar steps to those used in i) since $B_2(\mathbf{x}) = B_{21}(\mathbf{x}) + B_{22}(\mathbf{x})$ with

$$\begin{aligned}
B_{21}(\mathbf{x}) &= \frac{1}{nh_n^d}\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\frac{\delta_i u_i}{f(\mathbf{x}_i)} \\
B_{22}(\mathbf{x}) &= \frac{1}{nh_n^d}\sum_{i=1}^{n}\mathcal{K}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\delta_i u_i\left[\frac{1}{\widehat{f}(\mathbf{x}_i)}-\frac{1}{f(\mathbf{x}_i)}\right]
\end{aligned}$$

Recall that we have already shown that the sequence of variables $(\delta_i u_i)_{i=1}^n$ is an independent, identically distributed and uniformly generalized Gaussian sequence then, as $i(f)>0$, we obtain that the sequence of variables $(\delta_i u_i/f(\mathbf{x}_i))_{i=1}^n$ is also an independent, identically distributed and uniformly generalized Gaussian sequence such that $\mathbb{E}\left[\delta u/f(X)|X\right] = p(X)\sigma(X)/f(X)\mathbb{E}(\epsilon) = 0$. Hence, using Proposition 5.1 and Lemma 5.2, we get that $\sup_{\mathbf{x}\in\mathcal{C}}|B_{21}(\mathbf{x})| \xrightarrow{a.s.} 0$.

Using analogous arguments to those considered in the proof of (15) and using that the sequence $(\delta_i|u_i|)_{i=1}^n$ in an independent, identically distributed and uniformly generalized Gaussian sequence, we easily get that $\sup_{\mathbf{x}\in\mathcal{C}}|B_{22}(\mathbf{x})| \xrightarrow{a.s.} 0$, concluding the proof of ii).

iii) Note that $B_0(\mathbf{x})$ corresponds to $B_1(\mathbf{x})$ when $m \equiv 1$. Therefore, iii) follows from i). □

Proof of Theorem 3.2.2. Note that $\widehat{\mu}^{(2)} = \widetilde{\mu} + R_n$ where

$$\begin{aligned}\widetilde{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i y_i}{p(\mathbf{x}_i)} \\ R_n &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i y_i}{p(\mathbf{x}_i)\widehat{p}(\mathbf{x}_i)}\left(\widehat{p}(\mathbf{x}_i) - p(\mathbf{x}_i)\right) .\end{aligned}$$

Using that $B_n = \sup_{\mathbf{x}\in\mathcal{C}}|\widehat{p}(\mathbf{x}) - p(\mathbf{x})| \xrightarrow{a.s.} 0$ and that $i(p) > 0$ we have that on a set with probability 1, for n large enough the following bound for R_n holds

$$R_n \leq \frac{2\,B_n}{i(p)^2}\frac{1}{n}\sum_{i=1}^{n}|y_i|$$

and so, $R_n \xrightarrow{a.s.} 0$ since $\mathbb{E}|Y| < \infty$. It only remains to show that $\widetilde{\mu} \xrightarrow{a.s.} \mu$. For that purpose, define $z_i = \delta_i y_i / p(\mathbf{x}_i)$, $\{z_i\}_{i=1}^n$ is a sequence of independent and identically distributed variables such that $|z_1| \leq |y_1|/i(p)$, hence $\mathbb{E}(|z_i|) < \infty$. On the other hand, **D3** and **A2** entail that

$$\mathbb{E}\left(\mathbb{E}\left[\frac{\delta Y}{p(\mathbf{x})}|\mathbf{x}\right]\right) = \mathbb{E}\left(\frac{p(\mathbf{x})}{p(\mathbf{x})}\mathbb{E}[Y|\mathbf{x}=\mathbf{x}]\right) = \mathbb{E}(Y) = \mu .$$

Then, using strong law of large numbers, we get the result. □

Proof of Theorem Theorem 3.2.3. We begin by proving a). Let $1 \leq \alpha \leq d$. We have that

$$\begin{aligned}\sup_{x_\alpha\in C_\alpha}|\widehat{g}_\alpha(x_\alpha) - g_\alpha(x_\alpha)| \leq & \sup_{x_\alpha\in C_\alpha}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{m}(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - m(x_\alpha, \mathbf{x}_{\underline{\alpha}i})\right| + |\widehat{\mu} - \mu| \\ & + \sup_{x_\alpha\in C_\alpha}\left|\frac{1}{n}\sum_{i=1}^{n}m(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - \mu - g_\alpha(x_\alpha)\right| = B_1 + B_2 + B_3.\end{aligned}$$

The uniform strongly convergence of $\widetilde{m}$ and the fact that $B_1 \leq \sup_{\mathbf{x}\in\mathcal{C}}|\widetilde{m}(\mathbf{x}) - m(\mathbf{x})|$, imply that $B_1 \xrightarrow{a.s.} 0$. On the other hand, $B_2 \xrightarrow{a.s.} 0$ since $\widehat{\mu}$ is a consistent estimator of μ. Then, in order to prove a) it will be enough to show that $B_3 \xrightarrow{a.s.} 0$. Using that m satisfies the additive model given in **A1**, we get

$$\begin{aligned}B_3 &= \sup_{x_\alpha\in C_\alpha}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{\tau=1,\tau\neq\alpha}^{d}g_\tau(x_{\tau i}) + g_\alpha(x_\alpha) - g_\alpha(x_\alpha)\right\}\right| = \sup_{x_\alpha\in C_\alpha}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{\tau=1,\tau\neq\alpha}^{d}g_\tau(x_{\tau i})\right| \\ &= \left|\sum_{\tau=1,\tau\neq\alpha}^{d}\frac{1}{n}\sum_{i=1}^{n}g_\tau(x_{\tau i})\right| .\end{aligned}$$

Since $\mathbb{E}|g_\tau(X_\tau)| < \infty$ and **A2**a) holds, the result follows now from the strong law of large numbers.

b)The proof follows easily from a) and the consistency of $\widehat{\mu}$ using the bound $\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \leq |\widehat{\mu} - \mu| + \sum_{\alpha=1}^{d} \sup_{x_\alpha \in C_\alpha} |\widehat{g}_\alpha(x_\alpha) - g_\alpha(x_\alpha)|$. □

PROOF OF THEOREM 3.2.4. The proof follows using analogous arguments to those considered in the proof of Theorem 3.2.3 changing the averages to integrals and using **A2**b).

References

[1] Aerts, M.; Claeskens, G.; Hens, N. and Molenberghs, G. (2002). Local multiple imputation. *Biometrika* **89**, 2, 375–388.

[2] Afifi, A. and Elashoff, R. M. (1969). Missing observations in multivariate statistics III: Large sample analysis of simple linear regression. *J. Amer. Statist. Assoc.*, **64**, 337-358.

[3] Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion).*Ann. Statis.* **17**, 453-555.

[4] Chen, J. H. and Shao, J. (2000). Nearest neighbor imputation for survey data. *J. Official Statist.* **16**, 113-131.

[5] Cheng, P. E. (1990) Applications of kernel regression estimation: A survey. *Comm. Statist., Ser. A, Theory and Methods*, **19**, 4103-4134.

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

[7] Cheng, P. E. and Chu, C.K. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statist. Sinica* **6**, 63-78.

[8] Cheng, P. E. and Wei, L. J. (1986). Nonparametric inference under ignorable missing data process and treatment assignment. *Int. Statist. Symposium, Taipei, ROC.* **1**, 97-112.

[9] Chu, C. K.and Cheng, P. E. (1995). Nonparametric regression estimation with missing data. *J. Statist. Plann. Inference* **48**, 85-99.

[10] Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias from observational studies. *Biometrics*, **24**, 205- 213.

[11] Devroye, L. P. (1978). The Uniform Convergence of the Nadaraya-Watson Regression Function Estimate. *The Canadian Journal of Statistics* **6**, 179-191.

[12] González–Manteiga, W. and Pérez–González, A. (2004). Nonparametric mean estimation with missing data. *Comm. Statist. Theory Methods* **33**, 277-303.

[13] Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004) *Nonparametric and semiparametric models*, Springer Series in Statistics, Springer, Berlin.

[14] Hastie, T.J. and Tibshirani, R. J.(1990). *Generalized Additive Models*. Chapman and Hall.

[15] Hengartner, N. W. and Sperlich, S. (2005). Rate optimal estimation with the integration method in the presence of many covariates.*Journal of Multivariate Analysis* **95** 246-272.

[16] Hirano, K., G. Imbens, G. Ridder, (2000). Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score. NBER Technical Working Paper 251.

[17] Linton, O. B. and Nielsen, J. P. (1999). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93-101.

[18] Martínez–Miranda, M.D., Raya–Miranda, R., González–Manteiga, W. and González–Carmona, A. (2008). A bootstrap local bandwidth selector for additive models. *J. Comput. Graph. Statist.*, **17**, 38-55.

[19] Martínez–Miranda, M.D. and Raya–Miranda, R. (2011). Data-driven local bandwidth selection for additive models with missing data. *Applied Mathematics and Computation* **217** 10328-10342.

[20] Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9**, 141-142.

[21] Neyman, J. (1938). Contribution to the theory of sampling human populations. *J. Amer. Statist. Assoc.* **33** 101-116.

[22] Newey, W. K. (1994). Kernel estimation of partial means. *Econom. Theory* **10** 233-253.

[23] Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. *J. Multivariate Anal.* **73** 166179.

[24] Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation.* Academic Press, London.

[25] Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statis.* **14**, 590-606.

[26] Tjostheim, D. and Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.* **89** 1398-1409.

[27] Wang, Q.; Linton, O. and Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *J. Amer. Statist. Assoc.* **99**, 466, 334–345.

[28] Wang, W. and Rao, J. N. K., 2002. Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.* **30**, 896-924.

[29] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā A* **26**, 359-372.

[30] Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emporium J. Exp. Agriculture* **1**, 129-142.