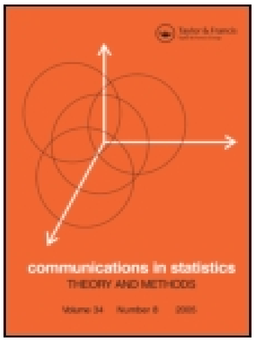

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: http://www.tandfonline.com/loi/lsta20

Estimating additive models with missing responses

Graciela Boente & Alejandra M. Martínez

To cite this article: Graciela Boente & Alejandra M. Martínez (2016) Estimating additive models with missing responses, Communications in Statistics - Theory and Methods, 45:2, 413-429, DOI: 10.1080/03610926.2013.815780

To link to this article: http://dx.doi.org/10.1080/03610926.2013.815780

Accepted author version posted online: 23 Jan 2015.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2016, VOL. 45, NO. 2, 413–429
http://dx.doi.org/10.1080/03610926.2013.815780

Taylor & Francis
Taylor & Francis Group

Estimating additive models with missing responses

Graciela Boente and Alejandra M. Martínez

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

ABSTRACT

For multivariate regressors, the Nadaraya–Watson regression estimator suffers from the well-known *curse of dimensionality*. Additive models overcome this drawback. To estimate the additive components, it is usually assumed 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 additive components estimation. 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 behavior of our procedures, under different scenarios.

ARTICLE HISTORY
Received 12 February 2013
Accepted 11 June 2013

KEYWORDS
Additive models; Kernel weights; Marginal integration; Missing Data; Non parametric regression

MATHEMATICS SUBJECT CLASSIFICATION
62G05; 62G08

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, non parametric and semiparametric methods have been developed for data analysis. In particular, non parametric regression models have gained importance when studying natural phenomenons with non linear complexity behavior. The non parametric regression model assumes that we have independent observations $(y_i, \mathbf{x}_i^{\mathrm{T}})$, $1 \le i \le n$, $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^d$ such that

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

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, neighborhoods of a point $\mathbf{x}$ become more sparse. This phenomenon is inherited by the convergence rate of the kernel regression estimators which 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, Hastie and Tibshirani (1990) introduced the so-called additive models which also provide the interpretation of univariate smoothers, since each component estimate can be plotted separately. In this sense, additive models combine the flexibility of

CONTACT Graciela Boente ✉ gboente@dm.uba.ar IMAS, CONICET and Departamento de Matemáticas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 1, Buenos Aires C1428EHA, Argentina.

the non parametric 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 have been extensively studied 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 data occur only on the response variables.

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 neighbor imputation (Chen and Shao, 2000), semiparametric estimation (Wang et al., 2004), non parametric multiple imputation (Aerts et al., 2002), empirical likelihood over the imputed values (Wang and Rao, 2002), among others. For a non parametric 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. In Sec. 2, the proposed estimators are introduced, besides, the problems arising when considering estimators using only the observations with no missing responses are described. Consistency for these estimators will be derived in Sec. 3 while the results of a simulation study are summarized in Sec. 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 independent of $\mathbf{x}_i$, $\mathbb{E}(\epsilon_i) = 0$ 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 MAR, that is, δ and Y are conditionally independent given $\mathbf{X}$, i.e.,

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

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

2.1. *Marginal integration estimators*

Among the methods to estimate the components under an additive model, we can mention the backfitting algorithm introduced by Buja et al. (1989) (see also Mammen et al., 1999) and the marginal integration procedure first introduced by Tjostheim and Auestad (1994) and Linton and Nielsen (1995). In this paper, we focus on the marginal integration method to estimate the additive components when missing responses arise. For the sake of completeness, we remind the definition of the estimators. For the sake of simplicity, from now on, $\mathbf{x}_{\underline{\alpha}}$ stands for the $(d-1)$-dimensional vector such that $\mathbf{x}_{\underline{\alpha}} = (x_1, \ldots, x_{\alpha-1}, x_{\alpha+1}, \ldots, x_d)^{\mathrm{T}}$. For any function $r : \mathbb{R}^d \to \mathbb{R}$ and for any $\mathbf{y}$, we allow the general notation $r(\mathbf{y}) = r(y_\alpha, \mathbf{y}_{\underline{\alpha}})$ to indicate the value of the function r calculated at the vector $\mathbf{y}$ with component α equal to y_α and the other ones equal to those of $\mathbf{y}$. This notation is used to point out with respect to which components we are adding or integrating.

Let $\mathbf{z}_i = (y_i, \mathbf{x}_i^{\mathrm{T}})$, $1 \leq i \leq n$, with the same distribution of $(Y, \mathbf{X}^{\mathrm{T}})$ such that $Y = m(\mathbf{X}) + \epsilon$ where the error ϵ is independent of $\mathbf{X}$, $\mathbb{E}(\epsilon) = 0$ and m satisfies the additive model (2), so that $\mu = \mathbb{E}(Y)$. Let $\widehat{\mu} = \sum_{i=1}^{n} y_i/n$ and consider the Nadaraya–Watson kernel estimator defined as $\widetilde{m}(\mathbf{x}) = \sum_{i=1}^{n} \mathcal{K}_h(\mathbf{x}_i - \mathbf{x})y_i / \sum_{j=1}^{n} \mathcal{K}_h(\mathbf{x}_i - \mathbf{x})$. The additive component g_α is then estimated through

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

One may wonder if, ignoring the vectors with missing responses, we will still obtain consistent estimators. That is, if the estimators obtained applying the previous procedure to the observations $\{\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_N}\} = \{(y_i, \mathbf{x}_i^{\mathrm{T}})\}_{\{i:\delta_i=1\}}$, where $N = \sum_{i=1}^{n} \delta_i$ lead to asymptotically unbiased estimators. This is one of the conditions needed to successfully apply the transfer principle described in Koul et al. (2012). Using the observations $\{\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_N}\}$ the obtained estimators are given by

$$\widehat{g}_{c,\alpha}(x_\alpha) = \frac{1}{N}\sum_{j=1}^{N} \widetilde{m}_c(x_\alpha, \mathbf{x}_{\underline{\alpha}i_j}) - \widehat{\mu}_c = \frac{1}{\sum_{i=1}^{n}\delta_i}\sum_{i=1}^{n} \delta_i \widetilde{m}_c(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - \widehat{\mu}_c\,,$$

where $\widehat{\mu}_c = \sum_{j=1}^{N} y_{i_j}/N = \sum_{i=1}^{n} \delta_i y_i / \sum_{i=1}^{n} \delta_i$ and $\widetilde{m}_c(\mathbf{x}) = \sum_{j=1}^{N} \mathcal{K}_h(\mathbf{x}_{i_j} - \mathbf{x}) y_{i_j} / \sum_{\ell=1}^{N} \mathcal{K}_h(\mathbf{x}_{i_\ell} - \mathbf{x})$. Note that $\widetilde{m}_c(\mathbf{x}) = \sum_{i=1}^{n} \mathcal{K}_h(\mathbf{x}_i - \mathbf{x}) \delta_i y_i / \sum_{j=1}^{n} \mathcal{K}_h(\mathbf{x}_i - \mathbf{x}) \delta_j$. We have used a subscript c to indicate that the estimator is computed using only the *complete* sample.

As is well known, under mild conditions, $\widetilde{m}_c(\mathbf{x}) \xrightarrow{a.s.} m(\mathbf{x})$, that is, the regression estimators based on the available observations are still consistent. Moreover, as shown in Theorem 3.2.1, the estimator $\widetilde{m}_c$ is uniformly consistent. Therefore, straightforward calculations and the fact that m satisfies the additive model (2) lead to

$$\frac{1}{N} \sum_{j=1}^{N} \widetilde{m}_c\big(x_\alpha, \mathbf{x}_{\underline{\alpha} i_j}\big) \xrightarrow{a.s.} \mu + g_\alpha(x_\alpha) + \frac{1}{p} \sum_{j=1, j \neq \alpha}^{d} \mathbb{E}[p(\mathbf{X}) g_j(X_j)] ,$$

where $(x_\alpha, \mathbf{X}_{\underline{\alpha}}) = (X_1, ..., X_{\alpha-1}, x_\alpha, X_{\alpha+1}, ..., X_d)$ and $p = \mathbb{E}p(\mathbf{X})$. On the other hand, since $\widehat{\mu}_c \xrightarrow{a.s.} p^{-1}\mathbb{E}[p(\mathbf{X})m(\mathbf{X})] = \mu + p^{-1} \sum_{j=1}^{d} \mathbb{E}[p(\mathbf{X}) g_j(X_j)]$, we get

$$\widehat{g}_{\alpha,c}(x_\alpha) \xrightarrow{a.s.} g_\alpha(x_\alpha) - \frac{1}{p}\mathbb{E}[p(\mathbf{X}) g_\alpha(X_\alpha)] ,$$

so that, the estimators are asymptotically biased. Hence, the transfer principle cannot be considered when estimating the marginal effects through the marginal integration procedure based on averaging over the covariates on the sample with $\delta_i = 1$, that is, using only the observations $\{\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_N}\}$ in the estimation procedure. The same conclusion is obtained if instead of averaging over the directions not of interest $\mathbf{X}_{\underline{\alpha}}$, one integrates using a fixed and known (d − 1)-dimensional measure.

It is worth noting that the asymptotic bias of $\widehat{g}_{\alpha,c}$ is not only due to the fact that the location estimator $\widehat{\mu}_c$ is not a consistent estimator of μ. Even by choosing a consistent estimator of μ, we do not obtain consistent estimators of g_α. Indeed, if one uses a consistent estimator of μ, we get that $\widehat{g}_{\alpha,c}(x_\alpha) \xrightarrow{a.s.} g_\alpha(x_\alpha) + p^{-1} \sum_{j=1, j\neq\alpha}^{d} \mathbb{E}[p(\mathbf{X}) g_j(X_j)]$, so, the bias is $p^{-1} \sum_{j=1, j\neq\alpha}^{d} \mathbb{E}[p(\mathbf{X}) g_j(X_j)]$. The key point for the loss of consistency is that the property $\mathbb{E}g_j(X_j) = 0$ is not inherited by the complete sample, that is, we cannot assume that $\mathbb{E}[p(\mathbf{X}) g_\alpha(X_\alpha)] = 0$.

For that reason, a modified procedure needs to be considered to estimate $g_\alpha(x_\alpha)$. In Sec. 2.2, we describe two procedures leading to consistent estimators of the marginal effects. The first one is based on the Nadaraya–Watson kernel estimator applied to the sample $\{\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_N}\}$, that is $\widetilde{m}_c(\mathbf{x})$. The second one is based on a modified internally normalized estimator. Both estimators turn out to be uniformly consistent. To obtain consistent estimators of $g_\alpha(x_\alpha)$, the estimators average over the direction $\mathbf{X}_{\underline{\alpha}}$ not only over the covariates with $\delta_i = 1$ but over all the covariates.

2.2. *Our proposal for data sets with missing responses*

Using the set of complete data $\{(y_i, \mathbf{x}_i^{\mathrm{T}})\}_{\{i:\delta_i=1\}}$ we can introduce two estimators of m. The first one, denoted $\widetilde{m}_{\mathrm{S}}^{(1)}$, equals $\widetilde{m}_c$, so it uses kernel weights modified multiplying by the indicator of the missing variables in order to adapt to the complete sample and avoid bias. However, for a better notation in the definitions to be given, we use $\widetilde{m}_{\mathrm{S}}^{(1)}$ instead of $\widetilde{m}_c$. 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

$$\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) = \sum_{i=1}^{n} w_{i,h}(\mathbf{x})\, y_i \left[\sum_{j=1}^{n} w_{j,h}(\mathbf{x})\right]^{-1}, \qquad \widetilde{m}_{\mathrm{S}}^{(2)}(\mathbf{x}) = \sum_{i=1}^{n} \frac{w_{i,h}(\mathbf{x})\, y_i}{\widehat{f}(\mathbf{x}_i)} \left[\sum_{k=1}^{n} \frac{w_{k,h}(\mathbf{x})}{\widehat{f}(\mathbf{x}_k)}\right]^{-1}, \tag{4}$$

where $w_{i,h}(\mathbf{x}) = \mathcal{K}_h(\mathbf{x} - \mathbf{x}_i)\delta_i$, $\widehat{f}(\mathbf{x}) = (1/n)\sum_{j=1}^{n} \mathcal{K}_h(\mathbf{x} - \mathbf{x}_j)$ is the kernel density estimator and $h = h_n$ is the bandwidth parameter. These regression estimators are based on the complete sample, i.e., discarding every incomplete pair of the original sample. For that reason, they are denoted with the subscript S as *simplified estimators*. This notation will be inherited by the marginal estimators, even though, they are obtained averaging over all the covariates and not only over those corresponding to $\delta_i = 1$.

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:

$$\widehat{\mu}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(\mathbf{x}_i), \qquad \widehat{\mu}^{(2)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i y_i}{\widehat{p}(\mathbf{x}_i)}, \tag{5}$$

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 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 $\widehat{\mu} = [\sum_{i=1}^{n} \delta_i y_i + (1 - \delta_i)\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}_i)]/n$ 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.

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 averages 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. The estimators are then defined as

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

where, as above, $\mathbf{x}_{\underline{\alpha}} = (x_1, \ldots, x_{\alpha-1}, x_{\alpha+1}, \ldots, x_d)^{\mathrm{T}}$.

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

$$\widehat{\widehat{g}}^{(1)}_{\alpha,\mathrm{S}}(x_\alpha) = \int \widetilde{m}^{(1)}_{\mathrm{S}}(x_\alpha, \mathbf{u}_{\underline{\alpha}}) q_{\underline{\alpha}}(\mathbf{u}_{\underline{\alpha}})\, d\mathbf{u}_{\underline{\alpha}} - \widehat{\mu}\,, \qquad \widehat{\widehat{g}}^{(2)}_{\alpha,\mathrm{S}}(x_\alpha) = \int \widetilde{m}^{(2)}_{\mathrm{S}}(x_\alpha, \mathbf{u}_{\underline{\alpha}}) q_{\underline{\alpha}}(\mathbf{u}_{\underline{\alpha}})\, d\mathbf{u}_{\underline{\alpha}} - \widehat{\mu}. \tag{6}$$

Hence, simplified estimators of the regression function that make use of the additive model assumption may be defined either as $\widehat{m}^{(j)}_{\mathrm{S}}(\mathbf{x}) = \sum_{\alpha=1}^{d} \widehat{g}^{(j)}_{\alpha,\mathrm{S}}(x_\alpha) + \widehat{\mu}$ or $\widehat{\widehat{m}}^{(j)}_{\mathrm{S}}(\mathbf{x}) = \sum_{\alpha=1}^{d} \widehat{\widehat{g}}^{(j)}_{\alpha,\mathrm{S}}(x_\alpha) + \widehat{\mu}$, for $j = 1, 2$, 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 i.i.d 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 $\mathbf{X}$.

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 $g : \mathbb{R} \to \mathbb{R}$, let $i_\alpha(g) = \inf_{x\in\mathcal{C}_\alpha} g(x)$ and $\|g\|_{\alpha,\infty} = \sup_{x\in\mathcal{C}_\alpha} |g(x)|$.

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,

$$\widehat{m}_Z(\mathbf{x}) = \sum_{i=1}^{n} \mathcal{K}_h\left(\mathbf{x} - \mathbf{x}_i\right) z_i \left[\sum_{i=1}^{n} \mathcal{K}_h\left(\mathbf{x} - \mathbf{x}_i\right)\right]^{-1}. \tag{7}$$

For the sake of completeness, 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 $|Y - m(\mathbf{x})| \le c$ a.s. for some $c < \infty$.*

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

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

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 the estimators introduced in Sec. 2, we will need the following set of assumptions:

- **D1.** $Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon$ with $\mathbb{E}(\epsilon) = 0$ and $\mathrm{VAR}(\epsilon) = 1$.
- **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 $i(p) > 0$.

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 non negative, bounded, and $\int \mathcal{K}(\mathbf{u})\, d\mathbf{u} = 1$.

K2. $\mathcal{K}(\mathbf{x}) = K(\|\mathbf{x}\|)$ for some non increasing 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 $\mathcal{C}_\alpha = \operatorname{sop} f_\alpha$.

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

Assumption **A1** sets that the considered model is an additive one, while **A2** ensures that the additive components g_j are identifiable. Separable regression models, as the one studied, 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.

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 $\widehat{\widehat{m}}$.

Assumption **K1** is a typical assumption for ordinary kernel smoothing while **K2** restricts the class of kernel functions to be chosen. Some relation between the bandwidth parameter

h_n and the sample size n is always necessary. To obtain the consistency of the proposals **H1** establishes conditions on the rate of convergence of the smoothing paramaters, which are standard in non parametric regression. Product kernels, as those considered in Hengartner and Sperlich (2005), can also be considered. In this case, $\mathcal{K}(\mathbf{x}) = \prod_{\ell=1}^{d} K_\ell(x_\ell)$ and $\mathcal{K}_h(\mathbf{x})$ is modified to $\mathcal{K}_{\mathbf{h}}(\mathbf{x}) = \prod_{\ell=1}^{d} K_\ell(x_\ell/h_\ell)(\prod_{\ell=1}^{d} h_\ell)^{-1}$ and assumption **H1** needs to be modified to $h_{\ell,n} \to 0$ and $n\prod_{\ell=1}^{d} h_{\ell,n}/\log n \to \infty$. For the sake of simplicity, we only state here the results when $h_{\ell,n} = h_n$ for all ℓ.

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).

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 Sec. 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 gives 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 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**, and **D1** to **D5**, **K1**, **K2**, and **H1**, we have that $\widehat{\mu}^{(1)} = (1/n)\sum_{i=1}^{n} \widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}_i) \xrightarrow{a.s.} \mu$.

The following result, whose proof is straightforward and can be found in Boente and Martínez (2012), states the strong consistency result for the estimators considered by Hirano et al. (2000):

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

Theorem 3.2.3. *Assume that **D2**, **A1**, **A2** (a), and **A3** hold. Let $\widehat{\mu}$ a consistent estimator of μ and $\widetilde{m}(\mathbf{x})$ such that $\sup_{\mathbf{x}\in\mathcal{C}} |\widetilde{m}(\mathbf{x}) - m(\mathbf{x})| \xrightarrow{a.s.} 0$. Define $\widehat{g}_\alpha(x_\alpha) = (1/n)\sum_{i=1}^{n} \widetilde{m}(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - \widehat{\mu}$. 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 *If **D1** to **D5**, **A1**, **A2** (a), **A3**, **K1**, **K2**, and **H1** hold, we have that*

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

(b) $\sup_{x\in\mathcal{C}_\alpha} |\widehat{g}_{\alpha,\mathrm{S}}^{(2)}(x) - g_\alpha(x)| \xrightarrow{a.s.} 0$, *$1 \le \alpha \le d$, 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. *Let $\widehat{\mu}$ a consistent estimator of μ and $\widetilde{m}(\mathbf{x})$ 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, under* ***D2***, ***D6***, ***D7***, ***A1***, ***A2*** *(b), and* ***A3***, *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 (6), which is stated below.

Corollary 3.2.2. *Assume* ***D1*** *to* ***D5***, ***D7***, ***A1***, *and* ***A2*** *(b) and* ***A3***, ***K1***, ***K2***, *and* ***H1*** *hold. Then, we have that*

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

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

Remark 3.2.1. As mentioned above, the strong uniform consistency results established also hold if product kernels are considered and different bandwidths are considered for each component. In this last situation, when no missing responses arise, under second derivative assumptions, Linton and Nielsen (1995) have shown that marginal integration methods attain the optimal univariate rate if we consider two components, that is, if $d = 2$. In particular, if $h_{1,n} = h_{2,n} = n^{-1/5}$ the optimal rate $n^{2/5}$ is attained. However, when $d > 2$ and the same bandwidth is used for all the directions, the integration method still suffers from the curse of dimensionality since it performs an average which reduces the order of the variance, but not of the bias. To reduce the bias on the nuisance directions when $d > 2$, Linton and Härdle (1996) consider a second order kernel on the direction of interest but higher order kernels on the other directions. A different approach to solve the challenging problem of reducing bias was given by Hengartner and Sperlich (2005) who introduced the internally normalized estimators. The internally normalized estimator of g_1 attains the univariate rate $n^{2/5}$ if the additive component g_1 is twice continuously differentiable, K_1 is a kernel of order 2, $h_{1,n} = n^{-1/5}$, $h_{\ell,n} = O(n^{-1/(3r_0)})$ with $r_0 = [[(d-1)/2]] + 1$ and K_ℓ are kernels of order r_0, for $\ell = 2, \ldots, d$. Even if asymptotic normality results are beyond the scope of this paper, it is expected that both proposals will attain the same convergence rate as the related estimators when no missing responses arise. In this sense, we expect that the internally corrected estimator will lead to a better order of convergence when $d > 2$. The asymptotic behavior of the preliminary estimators $\widetilde{m}^{(1)}$ and $\widetilde{m}^{(2)}$ can be derived using standard methods (see, for instance, Boente and Martínez, 2012). On the other hand, the study of the asymptotic behavior of $\widehat{\widehat{g}}^{(1)}_{\alpha,\mathrm{S}}$ and in particular, that of $\widehat{\widehat{g}}^{(2)}_{\alpha,\mathrm{S}}$ requires further study. This interesting topic may be the subject of future research.

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 $\widetilde{m}^{(1)}_{\mathrm{S}}$, $\widetilde{m}^{(2)}_{\mathrm{S}}$, $\widehat{m}^{(1)}_{\mathrm{S}}$, and $\widehat{m}^{(2)}_{\mathrm{S}}$, defined in Sec. 2. We perform $NR = 500$ 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 $z_i = m(\mathbf{x}_i) + u_i$, $1 \le i \le n$, 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:

$$m(x_1, x_2) = 4 + 24\,(x_1 - 0.5)^2 + 2\pi\,\sin(\pi x_2)\ . \tag{8}$$

To identify the marginal components and according to **A2**(a), their expectation is set equal to 0. Then, under (8), we have that $\mu = 10$ and the additive components are $g_1(x_1) = 24(x_1 - 0.5)^2 - 2$ and $g_2(x_2) = 2\pi \sin(\pi x_2) - 4$.

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, $p_2(\mathbf{x}) \equiv 0.8$, that is, MCAR responses are generated and $p_3(\mathbf{x}) = 0.4 + 0.5\cos^2(2x_1x_2 + 0.4)$. Besides, x_{i1}, x_{i2}, δ_i, and u_i are generated independently to each other.

For the smoothing procedure, we use the Epanechnikov multiplicative kernel $\mathcal{K}(\mathbf{x}) = K(x_1)K(x_2)$ where $K(u) = (3/4)(1 - u^2)\mathrm{I}_{[-1,1]}(u)$ and we choose $\mathcal{K}_{\mathbf{h}}(\mathbf{x}) = h^{-2}\prod_{\ell=1}^2 K_\ell(x_\ell/h)$.

The behavior of an estimator $\widehat{m}$ of m is measured using an approximation of the integrated squared error calculated at each replication as $\mathrm{ISE}(\widehat{m}) = (1/\ell^2)\sum_{s=1}^{\ell}\sum_{j=1}^{\ell}[m(\mathbf{u}_{js}) - \widehat{m}(\mathbf{u}_{js})]^2$, where $\mathbf{u}_{js} = (j/\ell, s/\ell)$, $1 \le j, s \le \ell$, $\ell = 50$. An approximation of the MISE is obtained averaging the ISE over replications.

In Boente and Martínez (2012), a preliminary study was performed to choose between the estimators $\widehat{\mu}^{(1)}$ and $\widehat{\mu}^{(2)}$ defined in (5). Based on the obtained results, we select $\widehat{\mu}_1$ as estimator of the marginal mean in the rest of our study. It is worth noticing that, when selecting the cross-validation bandwidth to estimate the additive components, the bandwidth for the estimator $\widehat{\mu}^{(1)}$ was kept fixed and equal to $h_n = 0.2$. Besides, when computing $\widetilde{m}_{\mathrm{S}}^{(2)}$, for the density estimator the bandwidth was chosen equal to 0.2.

Results with fixed bandwidths $h = 0.15, 0.2, 0.25$, and 0.3 are reported in Sec. 4.3 in Boente and Martínez (2012). We only report here the results corresponding to data-driven bandwidths.

4.2. *Data-driven bandwidths*

An important issue in any smoothing procedure is the choice of the smoothing parameter. Under a non parametric regression model, two commonly used approaches are cross-validation and plug-in. As is well known, plug-in methods require 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 (2000) 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. (2008) 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. Besides, since we have assumed that $\mathcal{K}_{\mathbf{h}}(\mathbf{x}) = h^{-2}\prod_{\ell=1}^2 K_\ell(x_\ell/h)$, we select the data-driven bandwidth as $\widehat{h} = \mathrm{argmin}_{h\in\mathbb{R}_+}\sum_{i=1}^n \delta_i\,[y_i - \widehat{m}_{-i,\mathrm{S}}(\mathbf{x}_i, h)]^2$, where $\widehat{m}_{-i,\mathrm{S}}(\cdot, h)$ represents the leave-one-out estimator corresponding to the simplified estimator $\widehat{m}_{\mathrm{S}}(\cdot, h)$ computed using the bandwidth h. As in cross-validation with complete data sets, the ith observation $(y_i, \mathbf{x}_i)$ is not used to predict y_i when $\delta_i = 1$, that is, when y_i is observed. In this way, we ensure that the observations used to calculate $\widehat{m}_{-i,\mathrm{S}}(\cdot, h)$ are independent of $\mathbf{x}_i$, the observation at which we evaluate $\widehat{m}_{-i,\mathrm{S}}(\cdot, h)$ to predict the i-response, when it is not missing.

4.2.1. *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, $\mathrm{MISE}(h) = \mathbb{E}\int[m(\mathbf{x}) - \widehat{m}(\mathbf{x}, h)]^2 d\mathbf{x}$, where $\widehat{m}(\cdot, h)$ denotes the estimator to be considered using as bandwidth the value 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 Sec. 4.1. For each value of the smoothing parameter, the value of the MISE was approximated by Monte Carlo as $\sum_{k=1}^{500} M(h, k)/500$, where for each replication k, $M(h, k) = \sum_{j=1}^{\ell}\sum_{s=1}^{\ell}[m(\mathbf{u}_{js}, h) - \widehat{m}(\mathbf{u}_{js}, h)]^2/\ell^2$, with $\mathbf{u}_{js} = (j/\ell, s/\ell)$, $1 \le j, s \le \ell$, and $\ell = 50$ as in the computation of the ISE. For each of the three missing probabilities, the optimal smoothing parameter h was selected over the grid $\mathcal{G}$ 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 1 reports the values obtained in each situation. We denote h_{OPT} the optimal bandwidth obtained.

4.2.2. *Cross-validation bandwidth*

We computed the data-driven bandwidths for each of the missing probabilities p_1, p_2, and p_3. As in Sec. 4.2.1, the data-driven smoothing parameter h was selected over the grid of points $\mathcal{G}$. Besides, when the minimization process leads to a value on the boundary, the search was carried on over the limits of the interval. Denote h_{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 Sec. 2. Table 3 summarizes the results obtained using as summary measure the $\mathrm{ISE}(\widehat{m})$. Besides, to evaluate the performance of the cross-validation bandwidths with respect to the optimal one, Table 2 reports 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 $\log(h_{\mathrm{CV}}/h_{\mathrm{OPT}})$. On the other hand, Fig. 1 shows the histograms of $\log(h_{\mathrm{CV}}/h_{\mathrm{OPT}})$ obtained for the estimators $\widehat{m}^{(1)}$ and $\widehat{m}^{(2)}$, under different missing schemes, respectively. Boxplots are given in Boente and Martínez (2012).

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 Table 3, the performance of the marginal and final estimators derived from the internally normalized regression estimator $\widetilde{m}^{(2)}$ is better than that obtained from

Table 1. Optimal smoothing parameters h_{opt} for each scenario and for each non parametric estimator.

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

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

	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 3. mise of the simplified estimators of m, g_1, and 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.

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

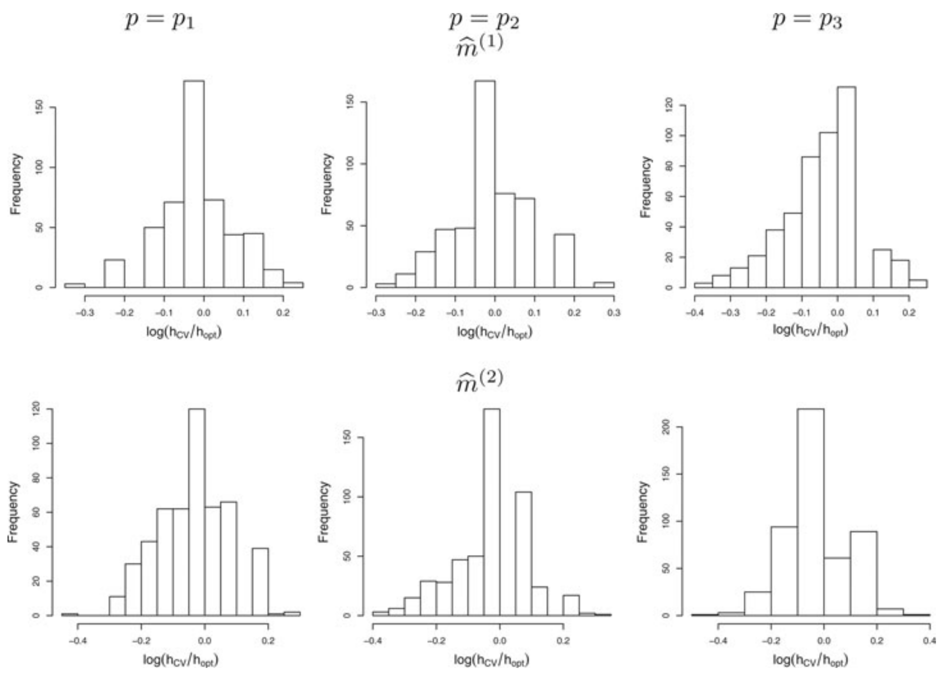

Figure 1. Histogram of $\log(h_{\text{CV}}/h_{\text{opt}})$ under different missing schemes $p_1(\mathbf{x}) \equiv 1$, $p_2(\mathbf{x}) \equiv 0.8$, and $p_3(\mathbf{x}) = 0.4 + 0.5(\cos(2x_1x_2 + 0.4))^2$. The upper and lower plots correspond to the optimal and data-driven selectors when using as estimates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

the Nadaraya–Watson estimator. The estimators of the additive components $\widehat{g}^{(2)}_{j,S}$, $j = 1, 2$, perform also better when a criterion which avoids border effects is considered (see Boente and Martínez, 2012). For these reasons we recommend the internally normalized regression estimators as a preliminary step to construct the marginal components.

Funding

This research was partially supported by Grants PIP112-200801-00216 from CONICET, PICT2011-397 from ANPCYT, and W276 from the Universidad de Buenos Aires at Buenos Aires, Argentina.

References

Aerts, M., Claeskens, G., Hens, N., Molenberghs, G. (2002). Local multiple imputation. *Biometrika* 89(2):375–388.

Boente, G., Martínez, A. (2012). Estimating Additive Models with Missing Responses. Available at: www.ic.fcen.uba.ar/preprints/Paper_aditivo_Boente_Martinez.pdf.

Buja, A., Hastie, T., Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Stat.* 17:453–555.

Chen, J. H., Shao, J. (2000). Nearest neighbor imputation for survey data. *J. Off. Stat.* 16:113–131.

Cheng, P. E. (1990). Applications of kernel regression estimation: a survey. *Commun. Stat. Ser. A, Theory Methods* 19:4103–4134.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Am. Stat. Assoc.* 89:81–87.

Cheng, P. E., Wei, L. J. (1986). Nonparametric inference under ignorable missing data process and treatment assignment. *Int. Stat. Symposium, Taipei, ROC* 1:97–112.

Chu, C. K. and Cheng, P. E. (1995). Nonparametric regression estimation with missing data. *J. Stat. Plan. Inference* 48:85–99.

Devroye, L. P. (1978). The uniform convergence of the Nadaraya-Watson regression function estimate. *Can. J. Stat.* 6:179–191.

González–Manteiga, W., Pérez–González, A. (2004). Nonparametric mean estimation with missing data. *Commun. Stat. Theory Methods* 33:277–303.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Berlin: Springer Series in Statistics, Springer.

Hastie, T. J., Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hengartner, N. W., Sperlich, S. (2005). Rate optimal estimation with the integration method in the presence of many covariates. *J. Multivar. Anal.* 95:246–272.

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

Koul, H. L., Muüller U. U., Schick A., (2012). The Transfer Principle: a tool for complete case analysis. *Ann. Stat.* 40:3031–3049.

Linton, O. B., Härdle, W. (1996). Estimation of additive regression models with known links. *Biometrika* 83:529–540.

Linton, O. B., Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82:93–101.

Mammen, E., Park, C. (2005). Bandwidth selection for smooth backfitting in additive models. *The Annals of Statistics* 33:1260–1294.

Mammen, E., Linton, O., Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Stat.* 27:1443–1490.

Martínez–Miranda, M. D., Raya–Miranda, R., González–Manteiga, W., González–Carmona, A. (2008). A bootstrap local bandwidth selector for additive models. *J. Comput. Graph. Stat.* 17:38–55.

Martínez–Miranda, M. D., Raya–Miranda, R. (2011). Data-driven local bandwidth selection for additive models with missing data. *Appl. Math. Comput.* 217:10328–10342.

Nadaraya, E. A. (1964). On estimating regression. *Theory Prob. Appl.* 9:141–142.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *J. Am. Stat. Assoc.* 33:101–116.

Newey, W. K. (1994). Kernel estimation of partial means. *Econ. Theory* 10:233–253.

Nielsen, J. P., Sperlich, S. (2005). Smooth backfitting in practise. *Journal of the Royal Statistical Society, Ser. B* 67:43–61.

Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *J. Multivar. Anal.* 73:166–179.

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

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* 14:590–606.

Tjostheim, D., Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: projections. *J. Am. Stat. Assoc.* 89:1398–1409.

Wang, Q., Linton, O., Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *J. Am. Stat. Assoc.* 99(466):334–345.

Wang, W., Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Ann. Stat.* 30:896–924.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā A* 26:359–372.

Appendix

Proposition A.1 due to Devroye (1978) will be used to derive the consistency of the estimators.

Proposition A.1. *Let* $(y_i, \mathbf{x}_i^{\mathrm{T}})_{i=1}^n$ *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 (7). 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 A.1. *Let* $\widehat{m}_{\delta Y}$ *and* $\widehat{m}_{\delta}$ *be defined as in (7). 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

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

Hence, it will be enough to show that

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta m}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| \xrightarrow{a.s.} 0, \tag{A.1}$$

$$\sup_{\mathbf{x}\in\mathcal{C}} |\widehat{m}_{\delta u}(\mathbf{x})| \xrightarrow{a.s.} 0. \tag{A.2}$$

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 and Proposition A.1, (A.1) follows.

It is easy to see that the sequence of independent and identically distributed variables $(\delta_i u_i)_{i=1}^n$ is also a uniformly generalized Gaussian sequence.

Effectively, using that the errors ϵ are independent of $(\delta, \mathbf{x})$ and that $\mathbb{E}(\epsilon) = 0$, we get $\mathbb{E}(\delta u|\mathbf{X} = \mathbf{x}) = p(\mathbf{x})\sigma(\mathbf{x})\mathbb{E}(\epsilon) = 0$. Then, for any $\lambda \in \mathbb{R}$, we get $\mathbb{E}(e^{\lambda\delta u}|\mathbf{X} = \mathbf{x}) = 1 - p(\mathbf{x}) + p(\mathbf{x})\mathbb{E}[e^{\lambda\sigma(\mathbf{x})\epsilon}]$. As $(\epsilon_i)_{i=1}^n$ is a sequence of independent, identically distributed, and uniformly generalized Gaussian variables, there exist $\tau \geq 0$ and $c \geq 0$ such that if $|\phi| < 1/c$, we get $\mathbb{E}(e^{\phi\epsilon}) \leq \exp\{\tau^2\phi^2/[2(1 - |\phi|c)]\}$. **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 $|\lambda| \leq 1/d$, $|\phi| = |\lambda|\sigma(\mathbf{x}) \leq 1/c$,

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

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

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

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 A.1, we obtain (A.2).

Finally, (b) can be obtained from (A.1) taking $Y \equiv 1$ or using Proposition A.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 A.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 $\sup_{\mathbf{x}\in\mathcal{A}} |\widehat{a}(\mathbf{x})/\widehat{f}(\mathbf{x}) - b(\mathbf{x})| \xrightarrow{a.s.} 0$

(b) $\sup_{\mathbf{x}\in\mathcal{A}} |\widehat{a}(\mathbf{x})/\widehat{f}(\mathbf{x}) - b(\mathbf{x})| \xrightarrow{a.s.} 0$, 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

$$\sup_{\mathbf{x}\in\mathcal{A}} |\widehat{a}(\mathbf{x}) - b(\mathbf{x})f(\mathbf{x})| \leq \sup_{\mathbf{x}\in\mathcal{A}} \left|\frac{\widehat{a}(\mathbf{x})}{\widehat{f}(\mathbf{x})} - b(\mathbf{x})\right| \left[\sup_{\mathbf{x}\in\mathcal{A}} |f(\mathbf{x})| + \sup_{\mathbf{x}\in\mathcal{A}} |\widehat{f}(\mathbf{x}) - f(\mathbf{x})|\right]$$
$$+ \sup_{\mathbf{x}\in\mathcal{A}} |b(\mathbf{x})| \sup_{\mathbf{x}\in\mathcal{A}} |\widehat{f}(\mathbf{x}) - f(\mathbf{x})|.$$

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

(b) Using that

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

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 Lemma A.1 since $\widetilde{m}_{\mathrm{S}}^{(1)}(\mathbf{x}) = \widehat{m}_{\delta Y}(\mathbf{x})/\widehat{m}_{\delta}(\mathbf{x})$. See Boente and Martínez (2012) for details.

(b) For the sake of simplicity denote $\widehat{f}(\mathbf{x}) = \widehat{f}_n(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{C}$. Recall that $w_{i,h}(\mathbf{x}) = \mathcal{K}_h\left(\mathbf{x}-\mathbf{x}_i\right)\delta_i$. 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 $B_0(\mathbf{x}) = (1/n)\sum_{i=1}^n w_{i,h}(\mathbf{x})/\widehat{f}(\mathbf{x}_i)$, $B_1(\mathbf{x}) = (1/n)\sum_{i=1}^n w_{i,h}(\mathbf{x})\, m(\mathbf{x}_i)/\widehat{f}(\mathbf{x}_i)$, and $B_2(\mathbf{x}) = (1/n)\sum_{i=1}^n w_{i,h}(\mathbf{x})\, u(\mathbf{x}_i)/\widehat{f}(\mathbf{x}_i)$. Hence, using that $i(p) > 0$ and Lemma A.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

$$B_{11}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n w_{i,h}(\mathbf{x})\frac{m(\mathbf{x}_i)}{f(\mathbf{x}_i)} \qquad \text{and}$$

$$B_{12}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n w_{i,h}(\mathbf{x})\, m(\mathbf{x}_i)\left[\frac{1}{\widehat{f}(\mathbf{x}_i)} - \frac{1}{f(\mathbf{x}_i)}\right].$$

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

$$\sup_{\mathbf{x}\in\mathcal{C}} |B_{11}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})| \xrightarrow{a.s.} 0, \tag{A.3}$$

$$\sup_{\mathbf{x}\in\mathcal{C}} |B_{12}(\mathbf{x})| \xrightarrow{a.s.} 0. \tag{A.4}$$

The fact that m and 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. 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 A.1 we get that $\sup_{\mathbf{x}\in\mathcal{C}}\left|B_{11}(\mathbf{x})/\widehat{f}(\mathbf{x}) - p(\mathbf{x})m(\mathbf{x})/f(\mathbf{x})\right| \xrightarrow{a.s.} 0$. On the other hand, **D2**, **K1**, **K2**, and **H1** imply that (see Prakasa Rao, 1983)

$$\sup_{\mathbf{x}\in\mathcal{C}}\left|\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right| \xrightarrow{a.s.} 0. \tag{A.5}$$

Thus, (A.3) follows from Lemma A.2.

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

$$|B_{12}(\mathbf{u})| \leq \|m\|_{0,\infty}\widehat{f}(\mathbf{u})\frac{\sup_{\mathbf{x}\in\mathcal{C}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|}{\inf_{\mathbf{x}\in\mathcal{C}}\widehat{f}(\mathbf{x})\inf_{\mathbf{x}\in\mathcal{C}}|f(\mathbf{x})|}$$

so, (A.4) follows easily from (A.5) 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

$$B_{21}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n w_{i,h}(\mathbf{x})\frac{u_i}{f(\mathbf{x}_i)} \qquad \text{and}$$

$$B_{22}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n w_{i,h}(\mathbf{x})\, u_i\left[\frac{1}{\widehat{f}(\mathbf{x}_i)} - \frac{1}{f(\mathbf{x}_i)}\right].$$

Effectively, using Proposition A.1 and Lemma A.2, we get that $\sup_{\mathbf{x}\in\mathcal{C}}|B_{21}(\mathbf{x})| \xrightarrow{a.s.} 0$. On the other hand, analogous arguments to those considered in the proof of (A.4), the Cauchy–Schwartz inequality and assumption **D6** lead to $\sup_{\mathbf{x}\in\mathcal{C}}|B_{22}(\mathbf{x})| \xrightarrow{a.s.} 0$, 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 Theorem 3.2.3. We begin by proving (a).

For any fixed $1 \leq \alpha \leq d$, we have that $\sup_{x_\alpha \in C_\alpha} |\widehat{g}_\alpha(x_\alpha) - g_\alpha(x_\alpha)| \leq B_1 + B_2 + B_3$ where $B_1 = \sup_{x_\alpha \in C_\alpha} \left|(1/n)\sum_{i=1}^n \widetilde{m}(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - m(x_\alpha, \mathbf{x}_{\underline{\alpha}i})\right|$, $B_3 = \sup_{x_\alpha \in C_\alpha} |(1/n)\sum_{i=1}^n m(x_\alpha, \mathbf{x}_{\underline{\alpha}i}) - \mu - g_\alpha(x_\alpha)|$, and $B_2 = |\widehat{\mu} - \mu|$. The consistency of $\widehat{\mu}$ entails that $B_2 \xrightarrow{a.s.} 0$. On the other hand, the uniform strongly convergence of $\widetilde{m}$ imply that $B_1 \xrightarrow{a.s.} 0$. Thus, in order to prove (a) it will be enough to show that $B_3 \xrightarrow{a.s.} 0$. Using that m satisfies **A1**, we get that $B_3 = \sup_{x_\alpha \in C_\alpha} |\sum_{\tau=1,\tau\neq\alpha}^d \sum_{i=1}^n g_\tau(x_{\tau i})/n|$. 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). □