On density and regression estimation with incomplete data

Mojirsheibani, Majid; Manley, Kevin; Pouliot, William

DOI:
10.1080/03610926.2016.1277751

On density and regression estimation with incomplete data

By

Majid Mojirsheibani[1] Kevin Manley[2]

Department of Mathematics, California State University, Northridge, CA 91330

and

William Pouliot[3]

Department of Economics, University of Birmingham, Birmingham, UK

Abstract

We consider the problem of estimation of a density function in the presence of incomplete data and study the Hellinger distance between our proposed estimator and the true density function. Here the presence of incomplete data is handled by utilizing a Horvitz-Thompson-type inverse weighting approach, where the weights are estimates of the unknown *selection probabilities*. We also address the problem of estimating a regression function with incomplete data.

Keywords: Convergence, incomplete data, empirical process, kernel, density.

1 Introduction

There has been a recent growing interest in the problem of density estimation in the presence of incomplete data. See, for example, Dubnicka (2009), Müller (2012), Tang et al (2012), Wang (2008), and Hazelton (2000) for kernel density estimation when auxiliary variables (covariates) are available, and also Zou et al (2015) for wavelet density estimators with incomplete data. Most of the work in the above cited references can be viewed, primarily, as the counterparts of the classical problem of density estimation with complete data as discussed, for example, by Rosenblatt (1956), Parzen (1962), Prakasa Rao (1983), and Devroye and Györfi (1985). Our interest in this paper is in the problem of density estimation with incomplete data, but with an approach that is closer in spirit to the work of van de Geer's (1993, 2000), where the author uses empirical process theory, based on fully observable data, to establish convergence results for the Hellinger distance between the true and the estimated densities. We will also extend our approach to deal with the problem of regression function estimation.

1.1 Background tools from empirical process

The justification and presentation of our main results will be facilitated with the aid of some empirical process theory results that are appropriate for situations where the data may be incomplete. Therefore, out of necessity, our main contributions start with some new results on empirical processes with incomplete data. To fix notation, let $\psi : \mathbb{R}^{d+p} \to \mathbb{R}$, $d > 0, p > 0$, and initially consider the estimation of the "mean" $\nu(\psi) := \mathbb{E}(\psi(\boldsymbol{Z}))$ based on the independently and identically distributed (iid) data $\mathbb{D}_n = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n\}$. However, here we are interested in the situation where the data $\boldsymbol{Z}_i = (\boldsymbol{X}_i', \boldsymbol{V}_i')' \in \mathbf{R}^{d+p}$, are not fully observed, more specifically $\boldsymbol{X}_i \in \mathbb{R}^d$ is always available, but $\boldsymbol{V}_i \in \mathbb{R}^p$ may be unobservable (unavailable) for various unknown reasons. To clarify our setup further, we also define the random variables $\xi_i = 1$ if $\boldsymbol{V}_i$ is observed, and $\xi_i = 0$ otherwise, and represent the data as

$$\mathbb{D}_n = \{(\boldsymbol{Z}_1, \xi_1), \ldots, (\boldsymbol{Z}_n, \xi_n)\} = \{(\boldsymbol{X}_1, \boldsymbol{V}_1, \xi_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{V}_n, \xi_n)\}.$$

Some important examples of the function ψ include $\psi(\boldsymbol{Z}) = \psi(\boldsymbol{X}, V) = V \in \mathbb{R}^1$, in which case the estimation of $\mathbb{E}(\psi(\boldsymbol{Z}))$ reduces to the usual mean estimation for $\mathbb{E}(V)$ when some of the V_i's are not

[1]Corresponding author. Email: majid.mojirsheibani@csun.edu .

[2]Email: kevin.manley.66@my.csun.edu

[3]Email: w.pouliot@bham.ac.uk

available. This case has been addressed and studied extensively in the literature; see, for example, Cheng (1994), Wang and Rao (2002), Hirano and Ridder (2003), Wang et al (2004), Rueda et al (2006), Müller (2009), and Kim and Yu (2011). Another example involves estimation of higher moments as well as mixed moments of the components of the random vector $\boldsymbol{Z}$. On the other hand, when the function ψ is of the form $\psi(\boldsymbol{Z}) \equiv \psi_{\boldsymbol{z}}(\boldsymbol{Z}) = \mathbb{I}\{\boldsymbol{Z} \leq \boldsymbol{z}\}$, $\boldsymbol{z} \in \mathbb{R}^{d+p}$, then $\mathbb{E}(\psi(\boldsymbol{Z}))$ is the cumulative distribution function (cdf) of $\boldsymbol{Z}$. The case of $\psi(\boldsymbol{Z}) \equiv \psi_{\boldsymbol{v}}(\boldsymbol{X}', \boldsymbol{V}') = \mathbb{I}\{\boldsymbol{V} \leq \boldsymbol{v}\}$ corresponds to the estimation of the marginal cdf of $\boldsymbol{V}$. These cases have been studied by, for example, Hu et al (2011), Liu et al (2011), Chenouri et al (2009), and Cheng and Chu (1996). When there are no missing $\boldsymbol{V}_i$'s in the data, $\mathbb{E}(\psi(\boldsymbol{Z}))$ can be estimated by the classical nonparametric empirical version

$$\nu_n(\psi) = n^{-1} \sum_{i=1}^{n} \psi(\boldsymbol{Z}_i)\,.$$

As for performance of this estimator, nonasymptotic exponential bounds are available on the uniform deviations on $\nu_n(\psi)$ from $\nu(\phi)$ (uniform in ψ) under various conditions. In fact for our future reference, we state one such result (see, for example, Pollard (1984; pp. 26-27):

Theorem 1 *Let Ψ be a class of functions $\psi : \mathbb{R}^{d+p} \to [-B, B]$, for some $0 < B < \infty$. Then, for every $\epsilon > 0$ and every $n \geq 1$, $\mathbb{P}\left\{\sup_{\psi\in\Psi} |\nu_n(\psi) - \nu(\psi)| > \epsilon\right\} \leq 8\,\mathbb{E}\left[\mathcal{N}_1\left(\epsilon/8\,,\Psi\,,\mathbb{D}_n\right)\right] e^{-n\epsilon^2/(128B^2)}$.*

Here, the term $\mathcal{N}_1(\varepsilon, \Psi, \mathbb{D}_n)$, called the *$\varepsilon$-covering number* of Ψ, is the cardinality of the smallest subclass of functions $\Psi_\epsilon = \{\psi_1, \cdots, \psi_{N(\epsilon)} : \mathbb{R}^{d+p} \to [-B, B]\}$, with the property that for fixed points $z_1, \ldots, z_n$ and for each $\psi \in \Psi$ there is a $\psi^\star \in \Psi_\varepsilon$ satisfying $\frac{1}{n}\sum_{i=1}^{n} |\psi(z_1) - \psi^\star(z_i)| < \epsilon$. Different extensions and variants of Theorem 1 are given by Alexander (1984), Massart (1990), Talagrand (1994), Giné (1996); also see the monograph by Vapnik (1998). Theorem 1 and its variants can provide tools to establish strongly consistency results for many important statistical estimation problems, including regression functions, density functions, and time series estimation; see, for example, the monograph by van de Geer (2000).

1.2 The difficulty with incomplete data

The situation can become quite different and challenging when not every $\boldsymbol{Z}_i = (\boldsymbol{X}_i, \boldsymbol{V}_i)'$ is fully observable; in particular, some of the $\boldsymbol{V}_i$'s may be missing. Of course, one may decide to estimate $\nu(\psi) = \mathbb{E}(\psi(\boldsymbol{Z}))$ based on the complete cases only, where a complete case refers to the fully observable $\boldsymbol{Z}_i$ (i.e. when $\xi_i = 1$). In this case the estimator can be expressed as $\tilde{\nu}(\psi) = \frac{1}{n'}\sum_{i=1}^{n} \xi_i \psi(\boldsymbol{Z}_i)$, where $n' = \sum_{i=1}^{n} \xi_i$. However, there are drawbacks with such estimators: (i) If a large proportion of the data (say 60 to 70 percent) have missing $\boldsymbol{V}_i$'s then, from a practical point of view, it makes sense to somehow revise $\tilde{\nu}_n$ to take into account the information which is available from $\boldsymbol{X}_i$'s. (ii) There are also theoretical reasons for not using the estimator $\tilde{\nu}_n$. For example, this estimator is not in general unbiased for $\nu(\psi)$ and therefore the corresponding empirical process, $\{\tilde{\nu}(\psi) - \nu(\psi) | \psi \in \Psi\}$ is not centered in general (not even asymptotically), and this plays a crucial role in establishing the theoretical properties of our propose density and regression estimators in this paper.

1.3 Summary of main results

Our main contributions may be summarized as follows.

1. We present revised version of $\tilde{\nu}(\psi)$ that take into account the missing covariates via an inverse weighting approach, where the weight functions are estimates of the selection probabilities. We propose a plug-in approach to replace these unknown selection probabilities with kernel regression and least-squares estimators

2. Under standard mild assumptions, we derive exponential bounds and inequalities similar to those of Theorem 1 that are suitable for our situation involving incomplete data.

3. We propose new density estimators in the presence of missing data. Furthermore, the new exponential bounds and inequalities in 2 above will be used to study the convergence properties of the proposed density estimator with respect to the Hellinger distance. Our methodology will also be applied to construct regression function estimators that are strongly optimal in the L_2 sense.

2 Main results

In this section we introduce the notion of missingness at random and suggest estimators of *selection probabilities* which will be used to construct our density and regression function estimators. We first establish a counterpart of Theorem 1, corresponding to the case of incomplete data: this result will help us to study our density and regression estimators later in this section.

2.1 Revised estimation of $\nu(\psi) := \mathbb{E}(\psi(\boldsymbol{Z}))$

The function $\mathbb{P}\{\xi = 1|\boldsymbol{Z}\}$, called the selection probability, plays an important role in estimation theory with incomplete data. In practice, this function is usually unknown and must be estimated. Under the commonly used assumption of data *Missing At Random* (MAR), it is assumed that the selection probability does not depend on $\boldsymbol{V}$ itself. In other words

$$\mathbb{P}\{\xi = 1|\boldsymbol{X}, \boldsymbol{V}\} = \mathbf{P}\{\xi = 1|\boldsymbol{X}\}(= \mathbf{E}(\xi|\boldsymbol{X})) \tag{1}$$

This assumption is essentially the baseline of analysis in the literature on incomplete data; see, for example, Cheng (1994), Cheng and Chu (1996), Wang and Rao (2002), Müller (2012), Tang et al (2012), and Bravo (2015). In what follows we shall focus on the case where the MAR assumption (1) holds. Define

$$\hat{\nu}(\psi) = n^{-1} \sum_{i=1}^{n} \xi_i \psi(\boldsymbol{Z}_i)/\hat{\pi}(\boldsymbol{X}_i) \tag{2}$$

The estimator of $\nu(\psi)$, where $\hat{\pi}(\boldsymbol{X}_i)$ is an estimator of the selection probability

$$\pi(\boldsymbol{X}_i) := \mathbb{E}(\xi_i|\boldsymbol{X}_i) = \mathbb{P}\{\xi_i = 1|\boldsymbol{X}_i\} \tag{3}$$

The estimator $\hat{\nu}(\psi)$ in (2) is in the spirit of the classical Horvitz-Thompson estimator (Horvitz and Thompson (1952)) in the sense that it works by weighting the complete cases by the inverse of the estimates of the selection probabilities, $\pi(\boldsymbol{X}_i)$. In fact, this approach has been used by many authors in the literature; see, for example, Robins et al. (1994) who propose a class of semiparametric estimators of regression coefficients based on inverse probability weighting estimating equations, with the imposed assumption that the missing probabilities are either known or can be modeled parametrically. Robins and Rotnitzky (1995) study efficient estimation in semiparametric multivariate regression models with missing response variables. Hirano et al. (2003) propose an estimator of the average treatment effect of a binary treatment using nonparametric methods for the missing probabilities. Wang et al. (2010) propose a class of kernel estimating equations to estimate a function $\theta(\mathbf{x}) = g\{E(Y|\boldsymbol{X} = \boldsymbol{x})\}$ whenever one has access to some auxiliary covariate vector $\boldsymbol{U}$. Here g is a known link function. Their method works by weighting the units with complete data by either the inverse of the true selection probability $\pi_i \equiv \pi(\boldsymbol{X}_i, \boldsymbol{U}_i) := P\{\zeta_i = 1|\boldsymbol{X}_i, \boldsymbol{U}_i)$, or an estimator of it, where π_i is assumed to have a known functional form. Unlike the above result, in this paper we do not assume the availability of any additional auxiliary covariates. Furthermore, we do not imposse assumptions that the functional form of the missing probabilities must always be known or be such that the functional form of the missing probabilities must always be known or be such that it can be modeled parametrically. Of course, if such assumptions hold then our proposed least squares methods will produce more accurate results, but our approach can also tackle the nonparametric case (via kernel methods) where missing probabilities are completely unknown. In this paper, we also derive exponential bounds on the performance of the proposed estimators. Unlike the above authors, here we carry out a Horvitz-Thompson type inversely weighted least squares

criterion to estimate the underlying regression function, where the weights are the inverses of the estimated selection probabilities of the complete cases. In the case of density estimation with incomplete data, our results are quite new and may be viewed as the counterparts of the classical work of van de Geer (1993, 200) who establishes strong Hellinger consistency of maximum likelihood density estimation.

As for the estimator of $\widehat{\pi}(\boldsymbol{X}_i)$ in (2), we consider two choices: (a) kernel regression and (b) the least squares.

(a) The kernel regression estimation of $\pi(\boldsymbol{X}_i) = \mathbb{E}(\xi_i|\boldsymbol{X}_i)$
Our first estimator is given by the kernel regression estimator

$$\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) = \sum_{k=1,\neq i}^{n} \xi_k \mathcal{K}\left(\frac{\boldsymbol{X}_i - \boldsymbol{X}_k}{h_n}\right) \Big/ \sum_{k=1,\neq i}^{n} \mathcal{K}\left(\frac{\boldsymbol{X}_i - \boldsymbol{X}_k}{h_n}\right), \tag{4}$$

with the convention $0/0 = 0$, where $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}_+$ is the kernel used, and h_n is the smoothing parameter of the kernel ($h_n \to 0$, as $n \to \infty$). Replacing $\widehat{\pi}(\cdot)$ in (2) by $\widehat{\pi}_{\mathsf{ker}}(\cdot)$, we find the following estimator of $\nu(\psi) = \mathbb{E}(\psi(\boldsymbol{Z}))$

$$\widehat{\nu}_n^{(\mathsf{ker})}(\psi) = n^{-1} \sum_{i=1}^{n} \xi_i \psi(\boldsymbol{Z}_i)/\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i)\,. \tag{5}$$

To assess the performance of $\widehat{\nu}_n^{(\mathsf{ker})}(\psi)$, we first state a number of assumptions.

(A1) $\pi_0 := \inf_{\boldsymbol{x}\in\mathbb{R}^d} \mathbb{P}\{\xi = 1|\boldsymbol{X} = \boldsymbol{x}\} > 0$.
(A2) The kernel $\mathcal{K}$ in (4) satisfies $\int_{\mathbb{R}^d} \mathcal{K}(\boldsymbol{x})d\boldsymbol{x} = 1$ and $\int_{\mathbb{R}^d} |x_i|\mathcal{K}(\boldsymbol{x})d\boldsymbol{x} < \infty$, for $i = 1, \ldots, d$, where $\boldsymbol{x} = (x_1, \ldots, x_d)'$. The smoothing parameter, h_n, of the kernel satisfies $h_n \to 0$ and $nh_n^d \to \infty$, as $n \to \infty$.
(A3) The random vector $\boldsymbol{X}$ has a compactly supported probability density function (pdf), $f(\boldsymbol{x})$, which is bounded away from zero on its compact support, i.e., $f_0 := \inf_{\boldsymbol{x}} f(\boldsymbol{x}) > 0$. Furthermore, f and its first-order partial derivatives are uniformly bounded.
(A4) The partial derivatives $\frac{\partial}{\partial x_i}\pi(\boldsymbol{x})$ exist for $i = 1, \ldots, d$ and are bounded uniformly, in $\boldsymbol{x}$, on the compact support of f, where $\pi(\boldsymbol{x}) = \mathbb{E}(\xi|\boldsymbol{X} = \boldsymbol{x})$.

The result below is a counterpart of Theorem 1 and is suitable for our incomplete data setup.

Theorem 2 *Let* Ψ *be a class of functions* $\psi : \mathbb{R}^{d+p} \to [-B, B],\ 0 < B < \infty$, *and let* $\widehat{\nu}_n^{(\mathsf{ker})}(\psi)$ *be as in (5). Then under assumptions (A1)–(A4), for every* $\epsilon > 0$ *there is a* $n_0 > 0$ *such that for all* $n > n_0$

$$\mathbb{P}\left\{\sup_{\psi\in\Psi} \left|\widehat{\nu}_n^{(\mathsf{ker})}(\psi) - \nu(\psi)\right| > \epsilon\right\} \;\leq\; 6n\, e^{-nh^d C_1\epsilon^2} + 8\mathbb{E}\left[\mathcal{N}_1\left(\frac{\pi_0\epsilon}{16}, \Psi, \mathbb{D}_n\right)\right] e^{-nC_2\epsilon^2},$$

where C_1 *and* C_2, *are positive constants not depending on* n *or* ϵ.

The proof of the above result appears in the Appendix. In passing we also note that Theorem 2 in conjunction with the Borel-Cantelli lemma imply that if $(nh_n^d)^{-1}\log n \to 0$ and $n^{-1}\log\mathbb{E}\left[\mathcal{N}_1\left(\pi_0\epsilon/16, \Psi, \mathbb{D}_n\right)\right] \to 0$, as $n \to \infty$ and $h_n \to 0$, then $\sup_{\psi\in\Psi} |\widehat{\nu}_n^{(\mathsf{ker})}(\psi) - \nu(\psi)| \overset{\text{a.s.}}{\to} 0$.

Next, we consider the least squares estimator.

(b) The least-squares estimator of $\pi(\boldsymbol{X}_i) = \mathbb{E}(\xi_i|\boldsymbol{X}_i)$

Let $\mathcal{P}$ be a known class of functions of the form $\tilde{\pi} : \mathbb{R}^d \to [\pi_0, 1]$, where π_0 is as in assumption (A1). The least-squares estimator of π is

$$\widehat{\pi}_{\mathsf{LS}} = \underset{\tilde{\pi}\in\mathcal{P}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \tilde{\pi}(\boldsymbol{X}_i))^2. \tag{6}$$

Therefore, upon taking $\widehat{\pi}$ to be $\widehat{\pi}_{\mathsf{LS}}$ in (2), we find the following estimator of $\nu(\psi) = \mathbb{E}(\psi(\boldsymbol{Z}))$

$$\widehat{\nu}^{(\mathsf{LS})}(\psi) = \frac{1}{n}\sum_{i=1}^{n}\frac{\xi_i\psi(\boldsymbol{Z}_i)}{\widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}_i)}\,. \tag{7}$$

The following result is a version of Theorem 2, corresponding to the least square estimtor.

Theorem 3 *Let Ψ be a class of functions $\psi : \mathbb{R}^{d+p} \to [-B, B],\ 0 < B < \infty$ and let $\widehat{\nu}_n^{(\mathsf{LS})}(\psi)$ be as in (7). Suppose that assumption (A1) holds and that $\pi \in \mathcal{P}$. Then for every $\epsilon > 0$, there is a $n_0 > 0$ such that for all $n > n_0$*

$$\begin{aligned}\mathbb{P}\left\{\sup_{\psi\in\Psi}\left|\widehat{\nu}_n^{(\mathsf{LS})}(\psi) - \nu(\psi)\right| > \epsilon\right\} \;\leq\; & 8\,\mathbb{E}\left[\mathcal{N}_1(C_{31}\epsilon, \Psi, \mathbb{D}_n)\right]\, e^{-nC_{32}\epsilon^2} \\ & + 8\,\mathbb{E}\left[\mathcal{N}_1(C_{33}\epsilon, \mathcal{P}, (\boldsymbol{X}_i)_{i=1}^n)\right]\, e^{-nC_{34}\epsilon^2} \\ & + 8\,\mathbb{E}\left[\mathcal{N}_1(C_{35}\epsilon^2, \mathcal{P}, (\boldsymbol{X}_i)_{i=1}^n)\right]\, e^{-nC_{36}\epsilon^4}\,,\end{aligned}$$

where C_{31} to C_{36} are positive constants not depending on n or ϵ.

The proof of this theorem appears in the Appendix.

In the next two sections we introduce our density and regression function estimators based on the approach and results of this section.

2.2 Density estimation

Once again let $\boldsymbol{Z} = (\boldsymbol{X}', \boldsymbol{V}')'$, where $\boldsymbol{X} \in \mathbb{R}^d$ is always observable but $\boldsymbol{V} \in \mathbb{R}^p$ may be missing at random in the sense that $\mathbb{P}\{\xi = 1|\boldsymbol{Z}\} \overset{\mathsf{MAR}}{=} \mathbb{P}\{\xi = 1|\boldsymbol{X}\} =: \pi(\boldsymbol{X})$. Here, as before, $\xi = 0$ if $\boldsymbol{V}$ is missing (and $\xi = 1$, otherwise). We are interested in estimating the probability density function of $\boldsymbol{Z}$, based on the data $\mathbb{D}_n = \{(\boldsymbol{Z}_1, \xi_1), \ldots, (\boldsymbol{Z}_n, \xi_n)\} = \{(\boldsymbol{X}_1, \boldsymbol{V}_1, \xi_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{V}_n, \xi_n)\}$. When $(X, V) \in \mathbb{R}^2$, Hazelton (2000) constructcs a kernel density estimator of the marginal distribution of V, using the X's as the auxiliary variables. This estimator, which is shown to be strongly uniformly consistent, is based on the distribution function estimator of Cheng and Chu (1996). A more recent result along these lines is the Horvitz-Thomson inverse weighting type density estimator of Dubnicka (2009) for the random variable V, based on the availability of an auxiliary random variable X. Our approach here, which does not assume the availability of any auxiliary variables, works as follows.

Suppose that the true pdf, g_0, of $\boldsymbol{Z}$ belongs to a class of densities $\mathcal{G}$. Clearly, when there are no missing data, the classical maximum likelihood estimator (MLE) of g_0 is $\widehat{g}_n = \operatorname{argmax}_{g\in\mathcal{G}}\sum_{i=1}^n \log g(\boldsymbol{Z}_i)$. Now, to tackle the presence of missing data define

$$\widehat{L}_n(g) = n^{-1}\sum_{i=1}^{n}\frac{\xi_i \log g(\boldsymbol{Z}_i)}{\widehat{\pi}(\boldsymbol{X}_i)}\,,\quad g \in \mathcal{G}\,, \tag{8}$$

where $\widehat{\pi}$ can be either $\widehat{\pi}_{\mathsf{ker}}$ as defined in (4) or $\widehat{\pi}_{\mathsf{LS}}$ as defined in (6). We consider the following estimator

$$\widehat{g}_n = \underset{g\in\mathcal{G}}{\operatorname{argmax}}\ \widehat{L}_n(g)\,.$$

How good is this MLE-type density estimator? To answer this question we first recall that the Hellinger distance ρ_{H} between two densities $g_1, g_2 \in \mathcal{G}$ is given by

$$\rho_{\mathsf{H}}(g_1, g_2) = \sqrt{\frac{1}{2}\int \left[\sqrt{g_1(\boldsymbol{z})} - \sqrt{g_2(\boldsymbol{z})}\right]^2 d\boldsymbol{z}}\,.$$

Hellinger-consistency of certain nonparametric density estimators have been studied by van de Geer (1993). For more on ρ_{H} and some of its properties see, for example, van de Geer (2000, Ch.4). The following result gives exponential performance bounds on the distance (Hellinger) between $\widehat{g}_n$ and the target density g_0.

Theorem 4 *Let $\rho_H(\widehat{g}_n, g_0)$ be the Hellinger distance between $\widehat{g}_n$ and g_0, and suppose that assumptions (A1)-(A4) hold. Then for every $\epsilon > 0$ there is an $n_0 > 0$ such that for all $n > n_0$*

$$\mathbb{P}\left\{\rho_H^2(\widehat{g}_n, g_0) > \epsilon\right\} \;\leq\; 8\mathbb{E}\left[\mathcal{N}_1\left(C_{16}\epsilon, \frac{\mathcal{G}}{g_0}, \mathbb{D}_n\right)\right] e^{-nC_{17}\epsilon^2} + \tau_n(\epsilon),$$

in which

$$\tau_n(\epsilon) = \begin{cases} 6n\, e^{-nh^d C_{18}\epsilon^2} & \text{if using } \widehat{\pi}_{\mathsf{ker}} \text{ in (8)},\\ 8\,\mathbb{E}\left[\mathcal{N}_1\left(C_{19}\epsilon, \mathcal{Q}, (\boldsymbol{X}_i)_{i=1}^n\right)\right] e^{-nC_{20}\epsilon^2} & \\ \quad + 8\,\mathbb{E}\left[\mathcal{N}_1\left(C_{21}\epsilon^2, \mathcal{Q}, (\boldsymbol{X}_i)_{i=1}^n\right)\right] e^{-nC_{22}\epsilon^4} & \text{if using } \widehat{\pi}_{\mathsf{LS}} \text{ in (8)}, \end{cases}$$

where $C_{16} - C_{22}$ are positive constants not depending on n or ϵ and

$$\frac{\mathcal{G}}{g_0} = \left\{\frac{g}{g_0}\,\mathbb{I}\{g_0 > 0\}\middle| g \in \mathcal{G}\right\}.$$

The above result can be used to establish almost-sure convergence results for $\widehat{g}_n$ with respect to the Hellinger distance. For example, with $\widehat{\pi} = \widehat{\pi}_{\mathsf{ker}}$, if $(nh_n^d)^{-1}\log n \to 0$ and $n^{-1}\log\mathbb{E}\left[\mathcal{N}_1\left(C_{16}\epsilon, \mathcal{G}/g_0, \mathbb{D}_n\right)\right] \to 0$, as $n \to \infty$ and $h_n \to 0$, then by the Borel-Cantelli lemma one has $\rho_{\mathsf{H}}(\widehat{g}_n, g_0) \overset{\text{a.s.}}{\to} 0$.

PROOF OF THEOREM 4

We prove the theorem for the case where $\widehat{\pi}{=}\widehat{\pi}_{\mathsf{ker}}$. The proof for the case where $\widehat{\pi}{=}\widehat{\pi}_{\mathsf{LS}}$ is almost identical and will not be given. Define the quantities

$$\begin{aligned} &\bar{g}_n = \frac{\widehat{g}_n + g_0}{2}\,, \quad \bar{g} = \frac{g + g_0}{2}\,, \quad g \in \mathcal{G}\\ &L_n\left(\frac{\bar{g}_n}{g_0}\,\mathbb{I}\{g_0 > 0\}\right) = \mathbb{E}\left[\log\frac{\bar{g}_n(\boldsymbol{Z})}{g_0(\boldsymbol{Z})}\,\mathbb{I}\{g_0(\boldsymbol{Z}) > 0\}\,\middle|\,\mathbb{D}_n\right]\\ &L\left(\frac{\bar{g}}{g_0}\,\mathbb{I}\{g_0 > 0\}\right) = \mathbb{E}\left[\log\frac{\bar{g}(\boldsymbol{Z})}{g_0(\boldsymbol{Z})}\,\mathbb{I}\{g_0(\boldsymbol{Z}) > 0\}\right]. \end{aligned}$$

We first show that

$$\rho_{\mathsf{H}}^2(\bar{g}_n, g_0) \leq \frac{1}{2}\left\{\widehat{L}_n\left(\frac{\bar{g}_n}{g_0}\,\mathbb{I}\{g_0 > 0\}\right) - L_n\left(\frac{\bar{g}_n}{g_0}\,\mathbb{I}\{g_0 > 0\}\right)\right\}, \tag{9}$$

where $\widehat{L}_n(\cdot)$ is as in (8). Here, (9) may be viewed as a version of Lemma 4.1 of van de Geer (2000) tailored to fit our current situation where $\boldsymbol{Z}_i$'s are allowed to have missing components; in fact, to prove (9), we borrow the arguments used in the proof of the cited result. First note that by the definition of $\widehat{g}_n$ we have

$$\forall\; 0 < g \in \mathcal{G}: \quad \widehat{L}_n\left(\frac{\widehat{g}_n}{g}\right) = \widehat{L}_n\left(\widehat{g}_n\right) - \widehat{L}_n(g) \geq 0$$

Also, by the concavity of the logarithmic function

$$\frac{1}{2}\,\log\frac{\widehat{g}_n(\boldsymbol{z})}{g_0(\boldsymbol{z})}\,\mathbb{I}\{g_0(\boldsymbol{z}) > 0\} \;\leq\; \log\frac{\bar{g}_n(\boldsymbol{z})}{g_0(\boldsymbol{z})}\,\mathbb{I}\{g_0(\boldsymbol{z}) > 0\}\,, \quad \text{where} \quad \bar{g}_n = \frac{\widehat{g}_n + g_0}{2}\,.$$

Thus

$$\begin{aligned} 0 &\leq \widehat{L}_n\left(\frac{\widehat{g}_n}{g_0}\,\mathbb{I}\{g_0>0\}\right) \leq 2\,\widehat{L}_n\left(\frac{\bar{g}_n}{g_0}\,\mathbb{I}\{g_0>0\}\right) \\ &= 2\left[\widehat{L}_n\left(\frac{\bar{g}_n}{g_0}\,\mathbb{I}\{g_0>0\}\right) - L_n\left(\frac{\bar{g}_n}{g_0}\,\mathbb{I}\{g_0>0\}\right)\right] + 2\,L_n\left(\frac{\bar{g}_n}{g_0}\,\mathbb{I}\{g_0>0\}\right). \end{aligned} \tag{10}$$

But $\bar{g}_n = (\widehat{g}_n + g_0)/2$ is also a density and therefore by Lemma 1.3 of van de Geer (2000)

$$L_n\left(\frac{\bar{g}_n}{g_0}\,\mathbb{I}\{g_0>0\}\right) \leq -2\,\rho_{\mathsf{H}}^2(\bar{g}_n, g_0)\,.$$

This last inequality together with (10) imply (9). Therefore

$$\begin{aligned} \rho_{\mathsf{H}}^2(\widehat{g}_n, g_0) &\leq 16\,\rho_{\mathsf{H}}^2(\bar{g}_n, g_0) \\ &\leq 8\left[\widehat{L}_n\left(\frac{\bar{g}_n}{g_0}\,\mathbb{I}\{g_0>0\}\right) - L_n\left(\frac{\bar{g}_n}{g_0}\,\mathbb{I}\{g_0>0\}\right)\right], \quad \text{(by (9))} \\ &\leq 8\sup_{g\in\mathcal{G}}\left|\widehat{L}_n\left(\frac{\bar{g}}{g_0}\,\mathbb{I}\{g_0>0\}\right) - L\left(\frac{\bar{g}}{g_0}\,\mathbb{I}\{g_0>0\}\right)\right|, \quad \text{(where } \bar{g} = \tfrac{g+g_0}{2}\text{)}. \end{aligned}$$

Next, define the class of functions

$$\mathcal{M} = \left\{ m : \mathbb{R}^{d+p} \to \mathbb{R} \,\middle|\, m(\boldsymbol{z}) = \log\frac{g(\boldsymbol{z}) + g_0(\boldsymbol{z})}{2g_0(\boldsymbol{z})}\,\mathbb{I}\{g_0(\boldsymbol{z})>0\},\ g\in\mathcal{G} \right\}$$

and note that, by Theorem 2, for every $\epsilon > 0$

$$\begin{aligned} \mathbb{P}\left\{\rho_{\mathsf{H}}^2(\widehat{g}_n, g_0) > \epsilon\right\} &\leq \mathbb{P}\left\{\sup_{g\in\mathcal{G}}\left|\widehat{L}_n\left(\frac{\bar{g}}{g_0}\,\mathbb{I}\{g_0>0\}\right) - L\left(\frac{\bar{g}}{g_0}\,\mathbb{I}\{g_0>0\}\right)\right| > \frac{\epsilon}{8}\right\} \\ &= \mathbb{P}\left\{\sup_{m\in\mathcal{M}}\left|\widehat{L}_n(m) - L(m)\right| > \frac{\epsilon}{8}\right\} \\ &\leq 6n\,e^{-nh_n^d C_{14}\epsilon^2} + 8\mathbb{E}\left[\mathcal{N}_1\left(\frac{\pi_0\epsilon}{128}, \mathcal{M}, \mathbb{D}_n\right)\right] e^{-nC_{15}\epsilon^2}, \end{aligned}$$

for n large enough, where C_{14} and C_{15} are positive constants not depending on n or ϵ. Now, let $m_1, m_2 \in \mathcal{M}$, where $m_k = \log\frac{g_k+g_0}{2g_0}\,\mathbb{I}\{g_0>0\}$, $g_k\in\mathcal{G}$, $k=1,2$, and observe that

$$\begin{aligned} |m_1(\boldsymbol{z}) - m_2(\boldsymbol{z})| &= \left|\log\frac{g_1(\boldsymbol{z}) + g_0(\boldsymbol{z})}{g_0(\boldsymbol{z})} - \log\frac{g_2(\boldsymbol{z}) + g_0(\boldsymbol{z})}{g_0(\boldsymbol{z})}\right| \mathbb{I}\{g_0(\boldsymbol{z})>0\} \\ &\leq \left|\frac{g_1(\boldsymbol{z})}{g_0(\boldsymbol{z})} - \frac{g_2(\boldsymbol{z})}{g_0(\boldsymbol{z})}\right| \mathbb{I}\{g_0(\boldsymbol{z})>0\}, \text{ (see van de Geer (2000, p.50))}. \end{aligned}$$

Therefore, $n^{-1}\sum_{i=1}^n \frac{\xi_i}{\pi(\boldsymbol{X}_i)}\,|m_1(\boldsymbol{Z}_i) - m_2(\boldsymbol{Z}_i)| \leq n^{-1}\sum_{i=1}^n \frac{\xi_i}{\pi(\boldsymbol{X}_i)}\left|\frac{g_1(\boldsymbol{Z}_i)}{g_0(\boldsymbol{Z}_i)} - \frac{g_2(\boldsymbol{Z}_i)}{g_0(\boldsymbol{Z}_i)}\right| \mathbb{I}\{g_0(\boldsymbol{Z}_i)>0\}$, which implies that if $\left\{\frac{g_1}{g_0}\,\mathbb{I}\{g_0>0\}, \ldots, \frac{g_N}{g_0}\,\mathbb{I}\{g_0>0\}\right\}$ is a minimal ϵ-cover of $\frac{\mathcal{G}}{g_0}$, then $\{m_1,\ldots,m_N\}$ is an ϵ-cover of $\mathcal{M}$, where $m_k = \log\frac{g_k+g_0}{2g_0}\,\mathbb{I}\{g_0>0\}$. Thus $\mathcal{N}_1(\epsilon, \mathcal{M}, \mathbb{D}_n) \leq \mathcal{N}_1(\epsilon, \frac{\mathcal{G}}{g_0}, \mathbb{D}_n)$, and this completes the proof of Theorem 4.

□

2.3 Regression function estimation

Let $(\boldsymbol{Z}, Y)$ be an $\mathbb{R}^{d+p}\times\mathbb{R}$-valued random vector with an unknown distribution. Here $\boldsymbol{Z} = (\boldsymbol{X}', \boldsymbol{V}')'$, where $\boldsymbol{X}\in\mathbb{R}^d$ is always observable but $\boldsymbol{V}\in\mathbb{R}^p$ may be missing at random in the sense that

$$\mathbb{P}\{\xi = 1|\boldsymbol{Z}, Y\} \overset{\mathsf{MAR}}{=} \mathbb{P}\{\xi = 1|\boldsymbol{X}, Y\} =: \pi(\boldsymbol{X}, Y)\,.$$

We note here that π is a map of the form $\mathbb{R}^{d+1} \to [0,1]$. Next, let $\mathbb{D}_n = \{(\boldsymbol{X}_1, \boldsymbol{V}_1, Y_1, \xi_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{V}_n, Y_n, \xi_n)\}$ represent the data (iid) and consider the following kernel regression estimator of $\pi(\boldsymbol{X}_i, Y_i) = \mathbb{P}\{\xi_i = 1 | \boldsymbol{X}_i, Y_i\}$

$$\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i, Y_i) = \sum_{j=1, \neq i}^{n} \xi_j \mathcal{H}\left((\boldsymbol{U}_i - \boldsymbol{U}_j)/\lambda_n\right) \Big/ \sum_{j=1, \neq i}^{n} \mathcal{H}\left((\boldsymbol{U}_i - \boldsymbol{U}_j)/\lambda_n\right), \text{ where } \boldsymbol{U}_j = (\boldsymbol{X}_j', Y_j)', \quad (11)$$

with the convention that $0/0 = 0$, where $\mathcal{H} : \mathbb{R}^{d+1} \to \mathbb{R}_+$ is the kernel used with the smoothing parameter λ_n satisfying $\lambda_n \to 0$ as $n \to \infty$. Alternatively, as a second choice, one may decide to consider the least-squares estimator of $\pi(\boldsymbol{X}_i, Y_i)$. More specifically, let $\mathcal{Q}$ be a known class of functions of the form $\tilde{\pi} : \mathbb{R}^{d+1} \to [\pi_{\mathsf{min}}, 1]$, where π_{min} is as in assumption (A1'), given below. Then the least-squares estimator is given by

$$\widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}_i, Y_i) = \underset{\tilde{\pi} \in \mathcal{Q}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \tilde{\pi}(\boldsymbol{X}_i, Y_i))^2. \quad (12)$$

Let Φ be a class of candidate regression functions of the form $\phi : \mathbb{R}^{d+p} \to \mathbb{R}$. Also, let $\phi^*(\boldsymbol{z}) = \mathbb{E}(Y | \boldsymbol{Z} = \boldsymbol{z})$ be the true regression function. Put

$$L(\phi) = \mathbb{E}[\phi(\boldsymbol{Z}) - Y]^2 \quad \text{and} \quad \widehat{L}_n(\phi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i |\phi(\boldsymbol{Z}_i) - Y_i|^2}{\widehat{\pi}(\boldsymbol{X}_i, Y_i)}, \quad \forall\, \phi \in \Phi, \quad (13)$$

where $\widehat{\pi}$ is either $\widehat{\pi}_{\mathsf{ker}}$ as defined in (11) or $\widehat{\pi}_{\mathsf{LS}}$ as defined in (12). Our proposed Horvitz-Thompson-based least squares estimator of the unknown regression function ϕ^* is given by

$$\phi_n = \underset{\phi \in \Phi}{\operatorname{argmin}} \widehat{L}_n(\phi). \quad (14)$$

To study the properties of the L_2 error of our estimator ϕ_n, we derive exponential performance bounds on the deviations of the L_2 error of ϕ_n from that of the best member of the class of candidate functions Φ, i.e., the quantity

$$\mathbb{E}([\phi_n(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2 \big| \mathbb{D}_n) - \inf_{\phi \in \Phi} \mathbb{E}[\phi(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2.$$

We first state the following counterparts of Assumptions *(A1)* - *(A4)*. Let $\boldsymbol{U} = (\boldsymbol{X}', Y)'$, and $\boldsymbol{u} = (\boldsymbol{x}', y)'$. Then the following assumption are exactly the same as those in *(A1)* - *(A4)*, but with d replaced by $d+1$, $\mathcal{K}$ by $\mathcal{H}$, h with λ_n, and π_0 with $\pi_{\min}$:
(A1') $\pi_{\min} := \inf_{\boldsymbol{u} \in \mathbb{R}^{d+1}} \mathbb{P}\{\xi = 1 | \boldsymbol{U} = \boldsymbol{u}\} > 0$.
(A2') The kernel $\mathcal{H}$ in (11) satisfies $\int_{\mathbb{R}^{d+1}} \mathcal{H}(\boldsymbol{u}) d\boldsymbol{u} = 1$ and $\int_{\mathbb{R}^{d+1}} |u_i| \mathcal{H}(\boldsymbol{u}) d\boldsymbol{u} < \infty$, for $i = 1, \ldots, (d+1)$. Also, the smoothing parameter λ_n satisfies $\lambda_n \to 0$ and $n\lambda_n^{d+1} \to \infty$, as $n \to \infty$.
(A3') The random vector $\boldsymbol{U}$ has a compactly supported pdf, which is bounded away from zero on its compact support. Furthermore, the pdf of $\boldsymbol{U}$ and its first-order partial derivatives are uniformly bounded.
(A4') The partial derivatives $\frac{\partial}{\partial u_i} \pi(\boldsymbol{u})$ exist for $i = 1, \ldots, (d+1)$ and are bounded uniformly, in $\boldsymbol{u}$, on the compact support of the pdf of $\boldsymbol{U}$.

Theorem 5 *Let Φ be a class of functions $\phi : \mathbb{R}^{d+p} \to [-B, B]$, for some $B < \infty$, and let ϕ_n be as in (14). Suppose that $|Y| \leq A < \infty$. Then, under assumptions (A1')-(A4'), for every $\epsilon > 0$ and n large enough*

$$\mathbb{P}\left\{ \left| \mathbb{E}\big([\phi_n(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2 \big| \mathbb{D}_n\big) - \inf_{\phi \in \Phi} \mathbb{E}[\phi(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2 \right| > \epsilon \right\}$$
$$\leq\ 8\, \mathbb{E}\left[\mathcal{N}_1\left(\frac{\pi_{\min} \epsilon}{64(A+B)}, \Phi, \mathbb{D}_n \right) \right] e^{-nC_{10}\epsilon^2} + \delta_n(\epsilon),$$

in which

$$\delta_n(\epsilon) = \begin{cases} 6ne^{-n\lambda_n^{d+1}C_{11}\epsilon^2} & \text{if using } \widehat{\pi}_{ker} \text{ in (13)}, \\ 8\,\mathbb{E}\left[\mathcal{N}_1\left(C_{12}\epsilon, \mathcal{Q}, (\boldsymbol{X}_i, Y_i)_{i=1}^n\right)\right] e^{-nC_{13}\epsilon^2} & \\ \quad +8\,\mathbb{E}\left[\mathcal{N}_1\left(C_{14}\epsilon^2, \mathcal{Q}, (\boldsymbol{X}_i, Y_i)_{i=1}^n\right)\right] e^{-nC_{15}\epsilon^4} & \text{if using } \widehat{\pi}_{LS} \text{ in (13)}, \end{cases}$$

where $C_{10} - C_{15}$ are positive constants not depending on n or ϵ.

We also note that the above result can be used to establish various almost-sure convergence results. For example, suppose that $\widehat{\pi}_{\text{ker}}$ is used in (13). Now, if $n^{-1}\log\{\mathbb{E}[\mathcal{N}_1(\pi_{\min}\epsilon/(64(A+B)), \Phi, \mathbb{D}_n)]\} \to 0$ and $(n\lambda_n^{d+1})^{-1}\log n \to 0$, as $n \to \infty$, then by an application of the Borel-Cantelli lemma $\mathbb{E}\big([\phi_n(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2\big|\mathbb{D}_n\big) \xrightarrow{\text{a.s.}} \inf_{\phi\in\Phi}\mathbb{E}[\phi(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2$. Clearly, if $\phi^* \in \Phi$ then $\inf_{\phi\in\Phi}\mathbb{E}[\phi(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2 = 0$, which yields $\mathbb{E}\big([\phi_n(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2\big|\mathbb{D}_n\big) \xrightarrow{\text{a.s.}} 0$, under the above conditions.

PROOF OF THEOREM 5.

We give a proof for the case where $\widehat{\pi}$ in (13) is the kernel estimator $\widehat{\pi}_{\text{ker}}$. The proof for the case where $\widehat{\pi}_{\text{LS}}$ is used is virtually the same and will not be given. First note that using the decomposition $\mathbb{E}\big([\phi_n(\boldsymbol{Z}) - Y]^2\big|\mathbb{D}_n\big) = \mathbb{E}\big([\phi_n(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2\big|\mathbb{D}_n\big) + \mathbb{E}[\phi^*(\boldsymbol{Z}) - Y]^2$ we can write

$$\begin{aligned} \mathbb{E}\big([\phi_n(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2\big|\mathbb{D}_n\big) &= \left\{\mathbb{E}\big([\phi_n(\boldsymbol{Z}) - Y]^2\big|\mathbb{D}_n\big) - \inf_{\phi\in\Phi}\mathbb{E}[\phi(\boldsymbol{Z}) - Y]^2\right\} \\ &\quad + \left\{\inf_{\phi\in\Phi}\mathbb{E}[\phi(\boldsymbol{Z}) - Y]^2 - \mathbb{E}[\phi^*(\boldsymbol{Z}) - Y]^2\right\}. \end{aligned} \tag{15}$$

But

$$\inf_{\phi\in\Phi}\mathbb{E}[\phi(\boldsymbol{Z}) - Y]^2 - \mathbb{E}[\phi^*(\boldsymbol{Z}) - Y]^2 = \inf_{\phi\Phi}\mathbb{E}[\phi(\boldsymbol{Z}) - \phi^*(\boldsymbol{Z})]^2. \tag{16}$$

Furthermore,

$$\begin{aligned} &\mathbb{E}\big([\phi_n(\boldsymbol{Z}) - Y]^2\big|\mathbb{D}_n\big) - \inf_{\phi\in\Phi}\mathbb{E}[\phi(\boldsymbol{Z}) - Y]^2 \\ &= \sup_{\phi\in\Phi}\left\{\mathbb{E}\big([\phi_n(\boldsymbol{Z}) - Y]^2\big|\mathbb{D}_n\big) - \widehat{L}_n(\phi_n) + \widehat{L}_n(\phi_n) - \widehat{L}_n(\phi) + \widehat{L}_n(\phi) - \mathbb{E}[\phi(\boldsymbol{Z}) - Y]^2\right\} \\ &\leq \sup_{\phi\in\Phi}\left\{\mathbb{E}\big([\phi_n(\boldsymbol{Z}) - Y]^2\big|\mathbb{D}_n\big) - \widehat{L}_n(\phi_n) + \widehat{L}_n(\phi) - \mathbb{E}[\phi(\boldsymbol{Z}) - Y]^2\right\} \\ &\qquad \text{(because } \widehat{L}_n(\phi_n) - \widehat{L}_n(\phi) \leq 0, \quad \forall\phi \in \Phi) \\ &\leq \left|\mathbb{E}\big([\phi_n(\boldsymbol{Z}) - Y]^2\big|\mathbb{D}_n\big) - \widehat{L}_n(\phi_n)\right| + \sup_{\phi\in\Phi}\left|\widehat{L}_n(\phi) - \mathbb{E}[\phi(\boldsymbol{Z}) - Y]^2\right| \\ &\leq 2\sup_{\phi\in\Phi}\left|\widehat{L}_n(\phi) - \mathbb{E}[\phi(\boldsymbol{Z}) - Y]^2\right|, \end{aligned} \tag{17}$$

where $\widehat{L}_n(\phi)$ is as in (13). Therefore in view of (15), (16), and (17), it is sufficient to show that the bound in Theorem 5 is also a bound on

$$\mathbb{P}\left\{\sup_{\phi\in\Phi}\left|\widehat{L}_n(\phi) - \mathbb{E}[\phi(\boldsymbol{Z}) - Y]^2\right| > \frac{\epsilon}{2}\right\}. \tag{18}$$

Now consider the class of functions of the form $\varphi(\boldsymbol{z}, y) = [y - \phi(\boldsymbol{z})]^2$ indexed by members of Φ, i.e., the class of functions

$$\mathcal{F} \equiv \mathcal{F}_\Phi = \left\{\varphi : \mathbb{R}^{d+p}\times\mathbb{R} \longrightarrow [-(A+B), A+B] \,\middle|\, \varphi(\boldsymbol{z}, y) = [y - \phi(\boldsymbol{z})]^2,\ \phi \in \Phi\right\}.$$

Then, by the definition of $\widehat{L}_n(\phi)$ in (13), we have

$$(18) = \mathbb{P}\left\{\sup_{\varphi\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \frac{\xi_i\varphi(\boldsymbol{Z}_i, Y_i)}{\widehat{\pi}_{\text{ker}}(\boldsymbol{X}_i, Y_i)} - \mathbb{E}[\varphi(\boldsymbol{Z}, Y)]\right| > \frac{\epsilon}{2}\right\}. \tag{19}$$

Furthermore, by an application of Theorem 2, for n large enough

$$(19) \quad \leq \quad 6n\, e^{-n\lambda_n^{d+1} C_{10}\epsilon^2} + 8\mathbb{E}\left[\mathcal{N}_1\left(\frac{\pi_{\min}\epsilon}{32}, \mathcal{F}, \mathbb{D}_n\right)\right] e^{-nC_{11}\epsilon^2},$$

where $\pi_{\min} = \inf_{\boldsymbol{x},y} \pi(\boldsymbol{x}, y) = \inf_{\boldsymbol{x},y} \mathbb{P}(\xi = 1 | \boldsymbol{X} = \boldsymbol{x}, Y = y)$, and where C_{10} and C_{11} are positive constants not depending on n or ϵ. To complete the proof, observe that for any $\varphi_1, \varphi_2 \in \mathcal{F}$

$$\begin{aligned} |\varphi_1(\boldsymbol{z}, y) - \varphi_2(\boldsymbol{z}, y)| &= \left|[y - \phi_1(\boldsymbol{z})]^2 - [y - \phi_2(\boldsymbol{z})]^2\right| \\ &\leq \left|[y - \phi_1(\boldsymbol{z})] + [y - \phi_2(\boldsymbol{z})]\right| \times \left|[y - \phi_1(\boldsymbol{z})] - [y - \phi_2(\boldsymbol{z})]\right| \\ &\leq 2(A + B)\left|\phi_1(\boldsymbol{z}) - \phi_2(\boldsymbol{z})\right|. \end{aligned}$$

Therefore $n^{-1}\sum_{i=1}^n |\varphi_1(\boldsymbol{Z}_i, Y_i) - \varphi_2(\boldsymbol{Z}_i, Y_i)| \leq 2(A + B)n^{-1}\sum_{i=1}^n |\phi_1(\boldsymbol{Z}_i) - \phi_2(\boldsymbol{Z}_i)|$, which implies that if $\{\phi_1, \ldots, \phi_N\}$ is a minimal $\epsilon/(2(A+B))$-cover of Φ with respect to the empirical L_1 norm, then $\{\varphi_1, \ldots, \varphi_N\}$ is an ϵ-cover of $\mathcal{F}$. Consequently, for every $t > 0$ we have $\mathcal{N}_1(t, \mathcal{F}, \mathbb{D}_n) \leq \mathcal{N}_1(t/(2(A + B)), \Phi, \mathbb{D}_n)$, and this completes the proof of Theorem 5.

□

Remarks.

Our proposed kernel estimators of the selection probabilities, as given by (4) or (14) play a crucial role in the development of our proposed density and regression function estimators. This can be noticed from the presence of the function $\widehat{\pi}$ in the denominator of the expressions in (8) and (13). The main issue with kernel type estimators is usually the choice of the bandwidth. In the case of kernel regression estimators, a popular choice of the bandwidth is the one that minimizes the Integrated Squared Error (ISE) of the corresponding kernel regression estimator. However, since ISE depends on the underlying unknown regression and density functions, Härdle and Marron (1985) replace them with "leave-one-out" estimators which are then used to de
ne their cross-validation bandwidth selection rule. A more recent approach is based on the cross-validation method of Racine and Li (2004), which is implemented in the 'R' package called "np" (see Racine and Hay
eld (2008)); in fact, we have used this method in our numerical studies of the next section. In the case of density estimation with missing data, one may also consider choosing the bandwidth as the minimizer of the mean Hellinger distance (MHD) or the mean weighted Hellinger distance (MWHD) proposed and studied by Ibrahim A. Ahmad and A. R. Mugdadi (2006). Unfortunately, the fact that our setup involves missing variables makes it very difficult to study MHD or MWHD analytically here. Alternatively, one may choose the bandwidth (from a grid of values) as the minimizer of an empirical version of MWHD, but we have not pursued that path in this paper.

3 Numerical examples

In what follows, we provide a number of numerical examples in order to assess the performance of our proposed estimators.

Example A. *[Density estimation.]* Here we consider the performance of the density estimator $\hat{g}_n$ as the maximizer of $\hat{L}(g)$ defined by (8), where $\hat{\pi}$ in (8) can be either $\hat{\pi}_{ker}$, defined in (4) or by $\hat{\pi}_{LS}$, defined in (6). We denote the corresponding density estimators by $\hat{g}_{\hat{\pi}_{ker}}$ and $\hat{g}_{\hat{pi}_{LS}}$, respectively. We have also considered the complete case (cc) density estimator, which uses the fully observed data only; this estimator will be denoted by $\hat{g}_{cc}$. Next, we carry our the numerical work, we generated n iid observations, $\boldsymbol{Z}_n$, $\imath = 1, \ldots, n$, from d-dimensional normal distributions; here we hav considered two different samples sizes, $n = 100$ and $n = 200$. Also, we use two different values of d. $d = 2$ and $d = 5$. For $d = 2$, we generated the actual data $\boldsymbol{Z}_n = (X_i, V_i,\ i = 1, \ldots, n$, from a bivariate Gaussian distribution with mean vector $(0, 1)'$ and covariance matrix $\Sigma = (\sigma_{if})$, where $\sigma_{1,1} = 1, \sigma_{2,2} = 2, \sigma_{1,2} = \sigma_{2,1} = 0$. Here X_i is always observable but V_i may be missing at random based on one of the following two missing

probability models for the function π defined in Here X_i is always observable but V_i may be missing at random based on one of the following two missing probability models for the function π defined in (3):

Model A. $\pi(x) := \mathbb{P}\{\zeta = 1|X = x\} = \exp(1 + 0.2x)/1 + \exp(1 + 0.2x)\}$,

Model B. $\pi(x) := \mathbb{P}\{\zeta = 1|X = x\} = 0.4[\exp(1 + 0.2x)/1 + \cos\{\exp(x) + \cos(x\sin(x)))\}]$.

For $d = 5$, the data were generated from a 5-dimensional Gaussian distribution with vector $(1, 1, 1, 1, 1)'$ and the covariance matrix $\Sigma = (\sigma_{ij})$, where $\sigma_{j,k} = 2^{-|j-k|}, 1 \leq j, k \leq 5$. Here, the last four components of $\boldsymbol{V}_i$ are allowed to be missing: more specifically, writing $\boldsymbol{Z}_i = (X_i), V_{i1}, V_{i2}, V_{i3}, V_{i4})'$, the component X_i is always observable, but $(V_{i1}, V_{i2}, V_{i3}, V_{i4})$ may be missing according to one of the following two missing probability models:

Model C. $\pi(x) := \mathbb{P}\{\zeta = 1|X = x\} = \exp(1 - 0.2x)/1 + \exp(1 - 0.2x)\}$,

Model D. $\pi(x) := \mathbb{P}\{\zeta = 1|X = x\} = 0.5[1 + \cos(2x\sin(x))\}]$.

The above choices of the missing probability mechanism result, roughly, in 30% missing data for Mode A and about 50% missing data for models B, C, and D. Let g be the true probability density function of $\boldsymbol{Z}$. In order to and the density estimators $\hat{g}_{\hat{\pi}_{ker}}$ and $\hat{g}_{\hat{\pi}_{LS}}$, we first used the cross-validation method of Racine and Li (2004) in the R package called "np" (see Racine and Hayfield 2008) to find the kernel estimator $\hat{\pi}_{ker}$, defined in (4). As for the parameters of the logistic missing probability mechanism $\hat{\pi}_{LS}$, defined in (6), we used nonlinear least squares regression (based on the R package "nls2"). To assess the performance of the estimators $\hat{g}_{\hat{\pi}_{ker}}$, $\hat{g}_{\hat{\pi}_{LS}}$, and the complete case estimator $\hat{g}_{cc}$, we computed the Hellinger distance between each estimator and the true density g based on two different sample sizes, $n = 100$ and $n = 200$. As a point of reference, we have also included the usual maximum likelihood estimator of g based on the full data of size n (i.e., when there are no missing data); this estimator is denoted by $\tilde{g}$. The entire above process was repeated a total of 500 times, each time using a sample of size n, and the average Hellinger distance were computed. The results for the case $d = 2$, which correspond to models A and B, appear in the first two rows of Table 1. The numbers appearing in brackets are the standard errors over 500 Monte Carlo runs. Table 1 shows both $\hat{g}_{\pi_{ker}}$ and $\hat{g}_{\pi_{LS}}$ tend to outperform the complete case estimator $\hat{g}_{cc}$ for models A and B. Under Model A, the estimator $\hat{g}_{\pi_{LS}}$ is slightly superior to the kernel based estimator $\hat{g}_{\pi_{ker}}$ which is not surprising because we are assuming that we know that the true underlying missing probability mechanism follows a logistic model. Similarly, under Model B, the kernel estimator does a better job in estimating the highly nonlinear trigonometric function $\pi(x)$ than the least squares method which is still assumes a logistic model. The estimator $\tilde{g}$, which is based on no missing data, is included only as a point of reference.

Table 1: Average Hellinger error, over 500 Monte Carlo runs, corresponding to the four models A, B, C, and D, for the proposed density estimators, $\hat{g}_{\hat{pi}_{ker}}$ and $\hat{g}_{\hat{\pi}_{LS}}$. Here $\hat{g}_{cc}$ is the complete case estimator (that discards all incomplete data points), and $\tilde{g}$ represents the estimator based on no missing data (it is included as a point of reference). The numbers in brackets are the standard errors.

	n=100				n=200			
	$\hat{g}_{\hat{pi}_{ker}}$	$\hat{g}_{\hat{\pi}_{LS}}$	$\hat{g}_{cc}$	$\tilde{g}$	$\hat{g}_{\hat{pi}_{ker}}$	$\hat{g}_{\hat{\pi}_{LS}}$	$\hat{g}_{cc}$	$\tilde{g}$
Model A	.788	.0762	.0810	.0622	.544	.0538	.0585	.0465
	(.00127)	(.00128)	(.00133)	(.00108)	(.00085)	(.00083)	(.00093)	(.00074)
Model B	.1124	.1159	.1227	.0662	.0832	.0904	.0981	.0465
	(.00183)	(.00204)	(.00220)	(.00108)	(.00131)	(.00150)	(.00164)	(.00074)
Model C	.1061	.1031	.1124	.0755	.0736	.0715	.0812	.0527
	(.00161)	(.00159)	(.00171)	(.00111)	(.00106)	(.00105)	(.00119)	(.00079)
Model D	.1508	.2388	.1854	.0755	.1208	.2082	.1729	.0527
	(.00191)	(.00719)	(.00148)	(.00111)	(.00145)	(.00664)	(.00106)	(.00079)

The restults for the case $d = 5$, which corresponds to models C and D, appear in the last two rows of

Table 2: Average L_2 error, over 500 Monte Carlo runs, corresponding to the four models A, B, C, and D, for the proposed density estimators, $\hat{g}_{\hat{pi}_{ker}}$ and $\hat{g}_{\hat{\pi}_{LS}}$. Here $\hat{g}_{cc}$ is the complete case estimator (that discards all incomplete data points), and $\tilde{g}$ represents the estimator based on no missing data (it is included as a point of reference). The numbers in brackets are the standard errors.

	n=100				n=200			
	$\hat{g}_{\hat{pi}_{ker}}$	$\hat{g}_{\hat{\pi}_{LS}}$	$\hat{g}_{cc}$	$\tilde{g}$	$\hat{g}_{\hat{pi}_{ker}}$	$\hat{g}_{\hat{\pi}_{LS}}$	$\hat{g}_{cc}$	$\tilde{g}$
Model A	.0349	.0338	.0364	.0295	.0241	.0237	.0261	.0206
	(.00061)	(.00060)	(.00063)	(.00051)	(.00039)	(.00037)	(.00043)	(.00034)
Model B	.0519	.0533	.0565	.0295	.0377	.0408	.0445	.0206
	(.00094)	(.00104)	(.00111)	(.00051)	(.00063)	(.00072)	(.00080)	(.00034)
Model C	.1076	.1060	.1107	.0909	.0898	.0885	.0942	.0759
	(.00083)	(.00083)	(.00085)	(.00069)	(.00066)	(.00066)	(.00072)	(.00058)
Model D	.1286	.1540	.1434	.0909	.1153	.1433	.1387	.0759
	(.00084)	(.00205)	(.00058)	(.00069)	(.00074)	(.00204)	(.00042)	(.00058)

Table 3: Average L_1 error, over 500 Monte Carlo runs, corresponding to the four models A, B, C, and D, for the proposed density estimators, $\hat{g}_{\hat{pi}_{ker}}$ and $\hat{g}_{\hat{\pi}_{LS}}$. Here $\hat{g}_{cc}$ is the complete case estimator (that discards all incomplete data points), and $\tilde{g}$ represents the estimator based on no missing data (it is included as a point of reference). The numbers in brackets are the standard errors.

	n=100				n=200			
	$\hat{g}_{\hat{pi}_{ker}}$	$\hat{g}_{\hat{\pi}_{LS}}$	$\hat{g}_{cc}$	$\tilde{g}$	$\hat{g}_{\hat{pi}_{ker}}$	$\hat{g}_{\hat{\pi}_{LS}}$	$\hat{g}_{cc}$	$\tilde{g}$
Model A	.1603	.1561	.1674	.1361	.1122	.1105	.1214	.0958
	(.00265)	(.00264)	(.00280)	(.00226)	(.00180)	(.00171)	(.00201)	(.00153)
Model B	.2290	.2336	.2472	.1361	.1686	.1809	.1960	.0958
	(.00374)	(.00410)	(.00438)	(.00226)	(.00265)	(.00298)	(.00325)	(.00153)
Model C	.2392	.2323	.2533	.1705	.1660	.1612	.1832	.1186
	(.00362)	(.00358)	(.00384)	(.00249)	(.00238)	(.00236)	(.00269)	(.00178)
Model D	.3396	.5313	.4170	.1705	.2723	.4657	.3892	.1186
	(.00428)	(.01555)	(.00332)	(.00249)	(.00333)	(.01482)	(.00238)	(.00178)

Table 1. As described earlier, the data were generated from a 5-dimensional Gaussian distribution with a unit mean vector and covariance $\Sigma = (\sigma_{jk})$ are known to be of the form $\sigma_{j,k} = 2^{-|j-k|}, 1 \leq j, k \leq 5$. Once again Table 1 shows that with the logistic missing probability mechanism of Model C, the estimator $\hat{g}_{\pi_{LS}}$ is superior to $\hat{g}_{\pi_{ker}}$ and that both of these estimators outperform the complete case estimator $\hat{g}_{cc}$. The situation is completely reversed for Model D (the fourth row of Table 1), where $\hat{g}_{\pi_{ker}}$ is by fare the best among the proposed estimators. In fact, we also note that here $\hat{g}_{\pi_{LS}}$ is a very poor estimator with a rather large standard deviation. Some of these facts are reflected in the boxplots of the 500 Hellinger errors (ie., distances) that appear in the first row of Figure 3 as well as the first row of Figure 3.
Although we have used the Hellinger distance to assess the performance of proposed density

estimators, we have also studied the L_2 and L_1 errors of each estimator. The results appear in Tables 2 and 3. As these tables show, one again $\hat{g}_{\hat{\pi}_{LS}}$ can be the best estimator when the assumptions of a logistic missing probability mechanism is indeed true. Otherwise, $\hat{g}_{\hat{\pi}_{ker}}$ is the best estimator.

We can draw the following conclusions form the results in tables 1, 2 and 3: if the missing probability mechanism, $\pi(x)$ has a known form (such as the logistic model), then $\hat{g}_{\hat{\pi}_{LS}}$ can be the best estimator. But in the more realistic case where one has no information about functional form $\pi(x)$, the density estimator $\hat{g}_{\hat{\pi}_{LS}}$ is, in general, the most appropriate one.

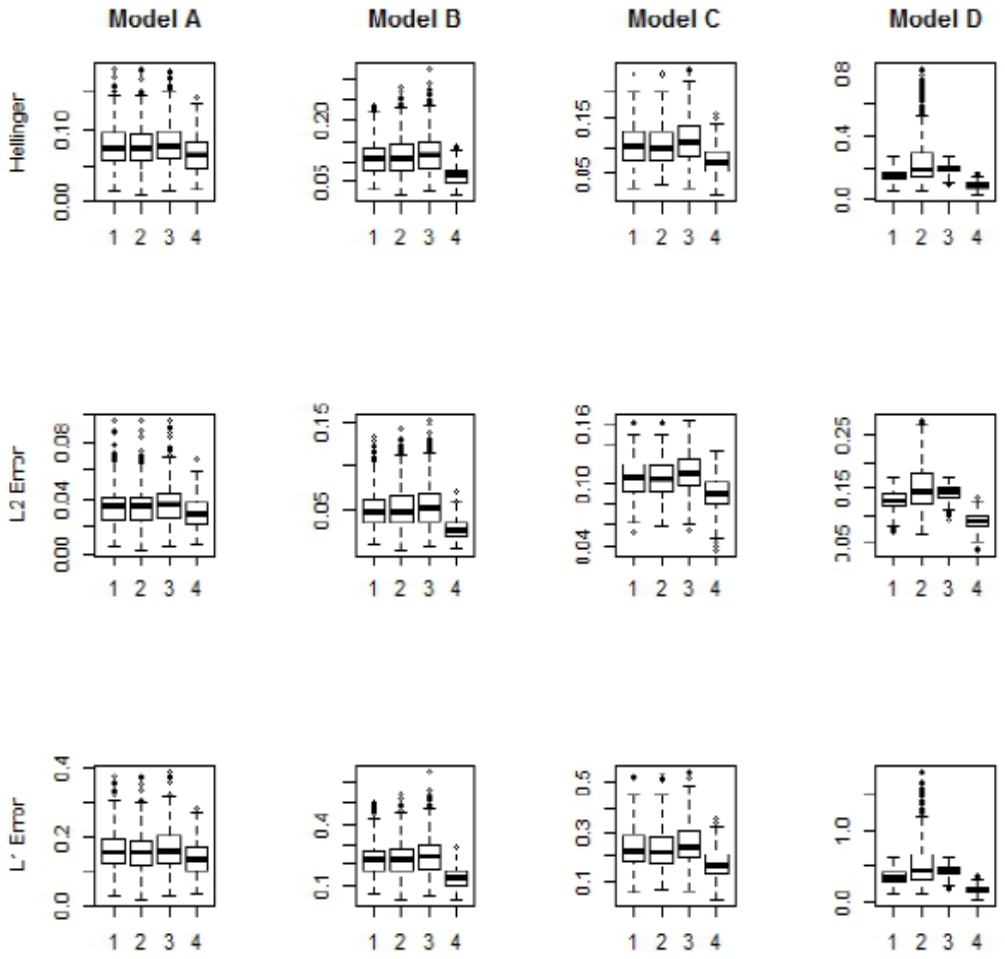


Figure 1: Boxplots of the errors of various estimator under different models, when $n = 100$. Within each of these 12 plots, boxplot 1 corresponds to $\hat{g}_{\hat{\pi}_{ker}}$, 2 corresponds to $\hat{g}_{\hat{\pi}_{LS}}$, 2 corresponds to $\hat{g}_{cc}$ and 4 corresponds to the case with no missing data (i.e. $\tilde{g}$).

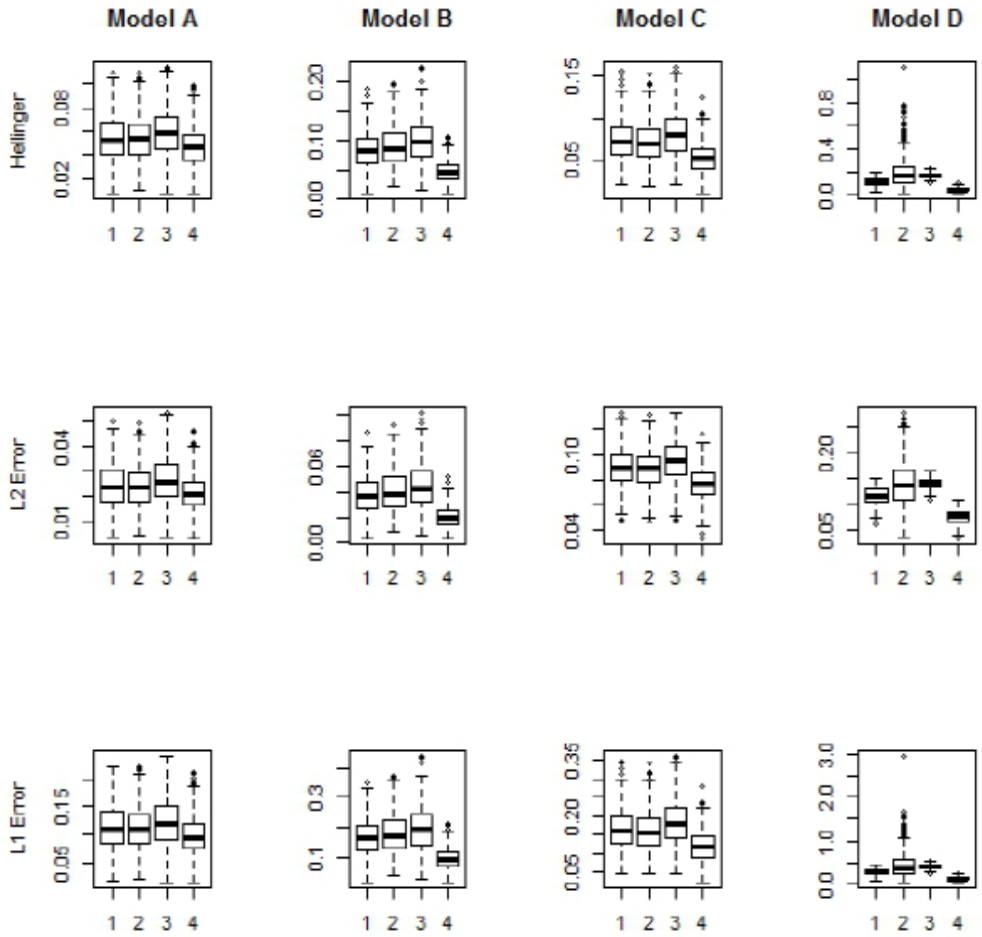


Figure 2: Boxplots of the errors of various estimator under different models, when $n = 200$. Within each of these 12 plots, boxplot 1 corresponds to $\hat{g}_{\hat{\pi}_{ker}}$, 2 corresponds to $\hat{g}_{\hat{\pi}_{LS}}$, 2 corresponds to $\hat{g}_{cc}$ and 4 corresponds to the case with no missing data (i.e. $\tilde{g}$).

Example B. *[Regression function estimation.]* Here we consider the performance of the following two versions of the regression function estimator ϕ_n defined via (13) and (14). The first estimator, denoted by $\phi_{n,ker}$ is the minimizer of (13) when $\hat{\pi}$ is taken to be the kernel estimator in (14). The second estimator, denoted by $\phi_{n,LS}$, is the minimizer of (14) when $\hat{\pi}$ is taken to be the least squares estimator in (14). We also consider the complete case estimator, denoted by $\phi_{n,cc}$, and the estimator based on the full data of size n (i.e., when there are no missing values). To perform our numerical studies, we generated samples $(Y_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$ of sizes $n = 100$ and $n = 200$ form

$$Y = sin(2Z_1) + Z_2^2 + Z_3 - \exp(-Z_4) + \varepsilon \text{ with } \varepsilon \approx N(0.0.5). \tag{20}$$

where ε is independent of $\boldsymbol{Z} = (Z_1, Z_2, Z_3, Z_4)$ and $\boldsymbol{Z}$ has a Gaussian distribution with mean $\mathbf{0}$ and the covariance matrix $\Sigma = (\sigma(ij)_{i,j>1}$, where $\sigma_{ij} = 2^{-|i-j|}$. In passing we also note that (20) is similar to the model used in Meier et al. (2009). Here, Z_1 and Z_2 are always observable, but Z_3 and Z_4 are allowed to be missing at random according to one of the following two missing probability mechanism:

Model (I). $\pi(z_1, z_2) := P\{\zeta = 1 | Z_1 = z_1, Z_2 = z_2\} = 0.4[1 + \cos(z_1 + z_2)]$.

Model (II). $\pi(z_1, z_2) := P\{\zeta = 1 | Z_1 = z_1, Z_2 = z_2\} = \exp(2 - 3z_1 - z_2)/[1 + \exp(2 + z_1 + z_2)].\}$

Model (I) results in apprximately 45% missing data where as model (II), which is logistic restults in aobut 75% missing data. Since the true underlying data generating model (20) is nerver known in practice, we decided to fit the partial second order model $D(Y|\boldsymbol{Z} = z) = z_1 + z_2 + z_3 + z_4 + z_4 + z_1^2 + z_2^2$. As in Example A, we used the cross-validation method of RAcine and Li (2004) in the R package "np" (Racine and Hayfield 2008) to find the kernel regression estimator $\hat{\pi}$ of π which is then sued to find $\phi_{n,ker}$ via (13) and (14). Simiarly, the parameters of the logistic missing probability mechanism were estimated using nonlinear least squares regression (based on the R package "nls2"),w hich are then used to find $\phi_{n,LS}$. To assess the performance of these two estimators we computed the empirical L_2 error of each estimator. This simulation process was repeated a total of 500 times (each time using samples of size $n = 100$ and $n = 200$ observations to find the least squares estimators $\phi_{n,ker}$, $\phi_{n,LS}$ and ϕ_{cc}) and the average L_2 errors were computed. The results appear in Table 4; the numbers appearing in brackets are the standard errors computed over 500 Monte Carlo runs. As a point of reference, we have also included the estimator corresponding to the case with no missing data this appears as $\tilde{\phi}_n$ in Table 4.

Table 4: Average L_2 error, over 500 Monte Carlo runs corresponding to models (I) and (II) for regression function estimators $\phi_{n,LS}$, $\phi_{n,LS}$, $\phi_{n,cc}$ and $\tilde{\phi}_n$. Here $\tilde{\phi}_n$ is based on the data with no missing values.

	n=100				n=200			
	$\hat{\phi}_{n,ker}$	$\hat{\phi}_{n,LS}$	$\hat{\phi}cc$	$\tilde{\phi}_n$	$\hat{\phi}_{n,ker}$	$\hat{\phi}_{n,LS}$	$\hat{\phi}cc$	$\tilde{\phi}_n$
Model (I)	.9733	1.0741	1.0537	.7324	.8008	.8814	.8690	.6704
	(.02463)	(.02284)	(.02849)	(.00598)	(.00758)	(.00730)	(.01039)	(.00337)
Model (II)	4.6476	4.4798	6.0594	.7324	2.8740	2.5824	5.2582	.6704
	(.27083)	(.27177)	(.30785)	(.00598)	(.14868)	(.14347)	(.20042)	(.00337)

Table 4 shows, $\phi_{n,ker}$ outperforms both $\phi_{n,LS}$ and ϕ_{ncc} under Model (I). On the other hand when Model (II) is correct, $\phi_{n,LS}$ is better. it is also important to ntice that, under Model (II), the error of $\phi_{n,cc}$ is substantially larger than every other estimator. In general, since the popular logistic model does not necessarily hold true, it would be safer to use $\phi_{n,ker}$ instead of $\phi_{n,LS}$ in practice. Figure 3 gives the boxplots of the 500 L_2 errors of various estimators. These boxplots show that the estimator are much more variable under Model (II).

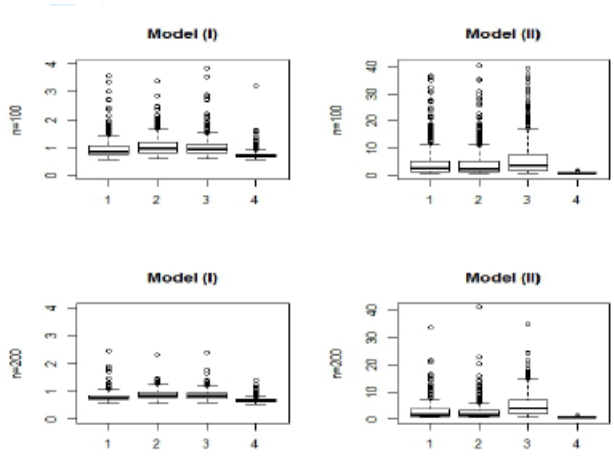


Figure 3: Boxplots of the L_2 errors of various regression estimators under Models (I) and (II). Within each of these 4 plots, boxplot 1 corresponds to $\phi_{n,ker}$, 2 corresponds to $\phi_{n,LS}$, 3 corresponds to $\phi_{n,cc}$ and 4 corresponds to the case with no missing data (i.e, $p\tilde{h}i_n$.

4 Appendix

To prove theorems 2 and 3 we first state two technical lemmas which may be of some independent theoretical interests as well.

Lemma 1 *Define* $\breve{\nu}_n(\psi) = n^{-1}\sum_{i=1}^n \xi_i\psi(\boldsymbol{Z}_i)/\pi(\boldsymbol{X}_i)$, *where the function* $\pi(\boldsymbol{X}_i)$ *is as in (3) and* ξ_i*'s are the Bernoulli random variable that appear in (1) . Suppose that Assumption (A1) holds. Then for every* $\epsilon > 0$ *and* $n \geq 1$

$$\mathbb{P}\left\{\sup_{\psi\in\Psi}\left|\breve{\nu}_n(\psi) - \nu(\psi)\right| > \epsilon\right\} \;\leq\; 8\,\mathbb{E}\left[\mathcal{N}_1\left(\frac{\pi_0\epsilon}{8}, \Psi, \mathbb{D}_n\right)\right] e^{-n\pi_0^2\epsilon^2/(128B^2)},$$

where B *is as in Theorem 2 and* $\pi_0 = \inf_{\boldsymbol{x}}\pi(\boldsymbol{x}) > 0$.

Lemma 2 *Let* f *be the pdf of the random vector* $\boldsymbol{X}$ *and put* $T(\boldsymbol{x}) = \pi(\boldsymbol{x})f(\boldsymbol{x})$, *where* $\pi(\boldsymbol{x}) = \mathbb{E}[\xi|\boldsymbol{X} = \boldsymbol{x}]$. *Let* $\widehat{T}(\boldsymbol{x}) = (nh_n^d)^{-1}\sum_{j=1}^n \xi_j\mathcal{K}((\boldsymbol{x} - \boldsymbol{X}_j)/h_n)$. *Then*
(i) *Under conditions (A2), (A3),and (A4),*

$$\left|\mathbb{E}\left[\widehat{T}(\boldsymbol{X})\middle|\boldsymbol{X}\right] - T(\boldsymbol{X})\right| \overset{a.s.}{\leq} ch_n,$$

where $c > 0$ *is a constant not depending on* n.
(ii) *For every constant* $\beta > 0$,

$$\mathbb{P}\left\{\left|\widehat{T}(\boldsymbol{X}) - \mathbb{E}\left[\widehat{T}(\boldsymbol{X})\middle|\boldsymbol{X}\right]\right| > \beta\middle|\boldsymbol{X} = \boldsymbol{x}\right\} \;\leq\; 2\exp\left\{\frac{-nh_n^d\beta^2}{2\|\mathcal{K}\|_\infty\left[2\|f\|_\infty + h_n^{2d}\beta/3\right]}\right\}.$$

PROOF OF THEOREM 2.
Let $\breve{\nu}_n(\psi) = n^{-1}\sum_{i=1}^n \xi_i\psi(\boldsymbol{Z}_i)/\pi(\boldsymbol{X}_i)$ and observe that

$$\begin{aligned}\mathbb{P}\left\{\sup_{\psi\in\Psi}\left|\widehat{\nu}_n^{(\mathsf{ker})}(\psi) - \nu(\psi)\right| > \epsilon\right\} \;&\leq\; \mathbb{P}\left\{\sup_{\psi\in\Psi}\left|\widehat{\nu}_n^{(\mathsf{ker})}(\psi) - \breve{\nu}_n(\psi)\right| > \frac{\epsilon}{2}\right\} + \mathbb{P}\left\{\sup_{\psi\in\Psi}\left|\breve{\nu}_n(\psi) - \nu(\psi)\right| > \frac{\epsilon}{2}\right\} \\ &:=\; \Delta_{n,1}(\epsilon) + \Delta_{n,2}(\epsilon)\,. \qquad (21)\end{aligned}$$

Now by Lemma 1, for every $\epsilon > 0$ and every $n \geq 1$,

$$\Delta_{n,2}(\epsilon) \leq 8\,\mathbb{E}\left[\mathcal{N}_1\left(\pi_0\epsilon/16, \Psi, \mathbb{D}_n\right)\right] e^{-n\pi_0^2\epsilon^2/(512B^2)}\,. \tag{22}$$

To deal with the term $\Delta_{n,1}(\epsilon)$ in (21) first note that

$$\left|\widehat{\nu}_n^{(\mathsf{ker})}(\psi) - \breve{\nu}_n(\psi)\right| \leq \frac{1}{n}\sum_{i=1}^n \xi_i\,|\psi(\boldsymbol{Z}_i)| \cdot \left|\frac{1}{\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i)} - \frac{1}{\pi(\boldsymbol{X}_i)}\right| \leq \frac{B}{n}\sum_{i=1}^n \left|\frac{1}{\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i)} - \frac{1}{\pi(\boldsymbol{X}_i)}\right|,$$

where $B = \|\psi\|_\infty$. Thus

$$\begin{aligned}
\Delta_{n,1}(\epsilon) &\leq \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n \left|\frac{1}{\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i)} - \frac{1}{\pi(\boldsymbol{X}_i)}\right| > \frac{\epsilon}{2B}\right\} &(23)\\
&\leq \mathbb{P}\left\{\left[\frac{1}{n}\sum_{i=1}^n \left|\frac{\pi(\boldsymbol{X}_i) - \widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i)}{\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i)\,\pi(\boldsymbol{X}_i)}\right| > \frac{\epsilon}{2B}\right] \cap \bigcap_{i=1}^n \left[\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) \geq 2^{-1}\pi_0\right]\right\}\\
&\quad + \mathbb{P}\left\{\bigcup_{i=1}^n \left[\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) < 2^{-1}\pi_0\right]\right\}\\
&\leq \mathbb{P}\left\{n^{-1}\sum_{i=1}^n |\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) - \pi(\boldsymbol{X}_i)| > (4B)^{-1}\pi_0^2\epsilon\right\} + \mathbb{P}\left\{\bigcup_{i=1}^n \left[\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) < 2^{-1}\pi_0\right]\right\}\\
&:= \Delta_{n,1}^{(i)}(\epsilon) + \Delta_{n,1}^{(ii)}(\epsilon)\,. &(24)
\end{aligned}$$

Let $T(\boldsymbol{X}) = \pi(\boldsymbol{X})f(\boldsymbol{X})$ and define $\widehat{T}(\boldsymbol{X}_i) = ((n-1)h_n^d)^{-1}\sum_{j=1,\neq i}^n \xi_i \mathcal{K}((\boldsymbol{X}_i - \boldsymbol{X}_j)/h_n)$ and $\widehat{f}(\boldsymbol{X}_i) = ((n-1)h_n^d)^{-1}\sum_{j=1,\neq i}^n \mathcal{K}((\boldsymbol{X}_i - \boldsymbol{X}_j)/h_n)$, and note that since $|\widehat{T}(\boldsymbol{X}_i)/\widehat{f}(\boldsymbol{X}_i)| \leq 1$, one finds $|\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) - \pi(\boldsymbol{X}_i)| \leq [|\widehat{T}(\boldsymbol{X}_i) - T(\boldsymbol{X}_i)| + |\widehat{f}(\boldsymbol{X}_i) - f(\boldsymbol{X}_i)|]/f(\boldsymbol{X}_i)$. Therefore

$$\begin{aligned}
\Delta_{n,1}^{(i)}(\epsilon) &\leq \sum_{i=1}^n \mathbb{P}\left\{\frac{1}{f(\boldsymbol{X}_i)}\left|\widehat{T}(\boldsymbol{X}_i) - T(\boldsymbol{X}_i)\right| > (8B)^{-1}\pi_0^2\epsilon\right\}\\
&\quad + \sum_{i=1}^n \mathbb{P}\left\{\frac{1}{f(\boldsymbol{X}_i)}\left|\widehat{f}(\boldsymbol{X}_i) - f(\boldsymbol{X}_i)\right| > (8B)^{-1}\pi_0^2\epsilon\right\}\\
&:= \sum_{i=1}^n p_{n,i}(\epsilon) + \sum_{i=1}^n q_{n,i}(\epsilon)\,. &(25)
\end{aligned}$$

But

$$\begin{aligned}
p_{n,i}(\epsilon) &\leq \mathbb{P}\left\{\left|\widehat{T}(\boldsymbol{X}_i) - \mathbb{E}(\widehat{T}(\boldsymbol{X}_i)|\boldsymbol{X}_i) + \mathbb{E}(\widehat{T}(\boldsymbol{X}_i)|\boldsymbol{X}_i) - T(\boldsymbol{X}_i)\right| > (8B)^{-1}f_0\pi_0^2\epsilon\right\}\\
&\quad \text{(because } f_0 := \inf f(\boldsymbol{x}) > 0 \text{ by Assumption (A3)),}\\
&\leq \mathbb{P}\left\{\left|\widehat{T}(\boldsymbol{X}_i) - \mathbb{E}(\widehat{T}(\boldsymbol{X}_i)|\boldsymbol{X}_i)\right| + (16B)^{-1}f_0\pi_0^2\epsilon > (8B)^{-1}f_0\pi_0^2\epsilon\right\}\\
&\quad \text{(by Part (i) of Lemma 2, for } n \text{ large enough)}\\
&= \mathbb{E}\left[\mathbb{P}\left\{\left|\widehat{T}(\boldsymbol{X}_i) - \mathbb{E}(\widehat{T}(\boldsymbol{X}_i)|\boldsymbol{X}_i)\right| > (16B)^{-1}f_0\pi_0^2\epsilon \,\middle|\, \boldsymbol{X}_i\right\}\right]\\
&\leq 2\exp\left\{\frac{-(n-1)h_n^d f_0^2\pi_0^4\epsilon^2}{(16B)^2(2\|\mathcal{K}\|_\infty)[2\|f\|_\infty + h_n^{2d}f_0\pi_0^2\epsilon/(48B)]}\right\}\\
&\quad \text{(by Part (ii) of Lemma 2)}\\
&\leq 2\exp\left\{\frac{-(n-1)h_n^d f_0^2\pi_0^4\epsilon^2}{2(16B)^2\|\mathcal{K}\|_\infty[2\|f\|_\infty + f_0/12]}\right\}, \quad \text{(for large } n\text{)}, &(26)
\end{aligned}$$

where we have used the fact that in bounding $\mathbb{P}\{|\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) - \pi(\boldsymbol{X}_i)| > (4B)^{-1}\pi_0^2\epsilon\}$, one only needs to consider $0 < \epsilon \leq 4B/\pi_0^2$ because $|\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) - \pi(\boldsymbol{X}_i)| \leq 1$. Similarly, since $\widehat{f}(\boldsymbol{X}_i)$ is the special case of

$\widehat{T}(\boldsymbol{X}_i)$ (take $\xi_j = 1$ in the definition of $\widehat{T}(\boldsymbol{X}_i)$, for all j), the above arguments leading to (26) give

$$q_{n,i}(\epsilon) \leq 2 \exp\left\{ \frac{-(n-1)h_n^d f_0^2 \pi_0^4 \epsilon^2}{2(16B)^2 \|\mathcal{K}\|_\infty [2\|f\|_\infty + f_0/12]} \right\},$$

for n large enough. Thus, in view of (25), for n large enough

$$\Delta_{n,1}^{(i)}(\epsilon) \leq 4n\, e^{-(n-1)h_n^d C_{15}\epsilon^2}, \tag{27}$$

where $C_{15} = \{2(16B)^2\|\mathcal{K}\|_\infty[2\|f\|_\infty + f_0/12]\}^{-1} f_0^2 \pi_0^4$. To complete the proof of Theorem 2, we also need to bound the term $\Delta_{n,1}^{(ii)}(\epsilon)$ in (24). Since $\mathbb{P}\{\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) < \pi_0/2\} \leq \mathbb{P}\{|\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) - \pi(\boldsymbol{X}_i)| > \pi_0/2\}$, the arguments that lead to the bound on $\Delta_{n,1}^{(i)}(\epsilon)$ (see (25) and (27)) yield

$$\begin{aligned} \Delta_{n,1}^{(ii)}(\epsilon) &\leq \sum_{i=1}^n \mathbb{P}\{\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) < \pi_0/2\} \leq \sum_{i=1}^n \mathbb{P}\{|\widehat{\pi}_{\mathsf{ker}}(\boldsymbol{X}_i) - \pi(\boldsymbol{X}_i)| > \pi_0/2\} \\ &\leq 2n\, e^{-(n-1)h_n^d C_{16}}, \end{aligned} \tag{28}$$

where $C_{16} = \{128\|\mathcal{K}\|_\infty(2\|f\|_\infty + f_0\pi_0/24)\}^{-1} f_0^2\pi_0^2$. Therefore, $\Delta_{n,1}(\epsilon) \leq \Delta_{n,1}^{(i)}(\epsilon) + \Delta_{n,1}^{(ii)}(\epsilon) \leq 6n\, e^{-nh^d C_{17}}$, with $C_{17} = \frac{1}{2}\min(C_{15}, C_{16})$. The theorem now follows from the bounds in (21), (22), (24), (27), and (28). □

PROOF OF THEOREM 3.
We start by writing,

$$\mathbb{P}\left\{ \sup_{\psi\in\Psi} \left|\widehat{\nu}_n^{(\mathsf{LS})}(\psi) - \nu(\psi)\right| > \epsilon \right\} \leq \mathbb{P}\left\{ \sup_{\psi\in\Psi} \left|\widehat{\nu}_n^{(\mathsf{LS})}(\psi) - \breve{\nu}_n(\psi)\right| > \frac{\epsilon}{2} \right\} + \mathbb{P}\left\{ \sup_{\psi\in\Psi} \left|\breve{\nu}_n(\psi) - \nu(\psi)\right| > \frac{\epsilon}{2} \right\},$$

where $\breve{\nu}_n(\psi) = n^{-1}\sum_{i=1}^n \xi_i\psi(\boldsymbol{Z}_i)/\pi(\boldsymbol{X}_i)$. But, by Lemma 1, for every $n \geq 1$,

$$\mathbb{P}\left\{ \sup_{\psi\in\Psi} \left|\breve{\nu}_n(\psi) - \nu(\psi)\right| > \frac{\epsilon}{2} \right\} \leq 8\,\mathbb{E}\left[\mathcal{N}_1\left(\pi_0\epsilon/16, \Psi, \mathbb{D}_n\right)\right] e^{-n\pi_0^2\epsilon^2/(512B^2)}.$$

Furthermore, since $|\widehat{\nu}_n^{(\mathsf{LS})}(\psi) - \breve{\nu}_n(\psi)| \leq Bn^{-1}\sum_{i=1}^n |(\widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}_i))^{-1} - (\pi(\boldsymbol{X}_i))^{-1}|$, one finds

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{\psi\in\Psi} \left|\widehat{\nu}_n^{(\mathsf{LS})}(\psi) - \breve{\nu}_n(\psi)\right| > \frac{\epsilon}{2} \right\} \\ &\leq \mathbb{P}\left\{ \frac{1}{n}\sum_{i=1}^n \left| \frac{1}{\widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}_i)} - \frac{1}{\pi(\boldsymbol{X}_i)} \right| > \frac{\epsilon}{2B} \right\} \leq \mathbb{P}\left\{ \frac{1}{n}\sum_{i=1}^n |\pi(\boldsymbol{X}_i) - \widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}_i)| > \frac{\pi_0^2\epsilon}{2B} \right\} \\ &\qquad \text{(the second inequality follows because } \widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}_i) \geq \pi_0 \text{ and } \pi(\boldsymbol{X}_i) \geq \pi_0) \\ &\leq \mathbb{P}\left\{ \sup_{\tilde{\pi}\in\mathcal{P}} \left| \frac{1}{n}\sum_{i=1}^n |\tilde{\pi}(\boldsymbol{X}_i) - \pi(\boldsymbol{X}_i)| - \mathbb{E}\,|\tilde{\pi}(\boldsymbol{X}_i) - \pi(\boldsymbol{X}_i)| \right| > \frac{\pi_0^2\epsilon}{4B} \right\} \\ &\qquad + \mathbb{P}\left\{ \mathbb{E}\left[|\widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}) - \pi(\boldsymbol{X})| \,\middle|\, \mathbb{D}_n \right] > \frac{\pi_0^2\epsilon}{4B} \right\} \\ &:= I_n(\epsilon) + II_n(\epsilon). \end{aligned} \tag{29}$$

Using Theorem 1, with Ψ replaced by the class $\mathcal{P}$, it is straightforward to see that

$$I_n(\epsilon) \leq 8\mathbb{E}\left[\mathcal{N}_1\left(\frac{\pi_0^2\epsilon}{32B}, \mathcal{P}, (\boldsymbol{X}_i)_{i=1}^n \right) \right] e^{-nC_{19}\epsilon^2},$$

where $C_{19} = \pi_0^4/(2048B^2)$. As for the term $II_n(\epsilon)$ in (29), let

$$\widehat{L}_n(\tilde{\pi}) = n^{-1}\sum_{i=1}^{n}(\xi_i - \tilde{\pi}(\boldsymbol{X}_i))^2\,, \forall\ \tilde{\pi} \in \mathcal{P}\,,$$

and note that by the Cauchy-Schwartz inequality

$$\begin{aligned} II_n(\epsilon) &\leq \mathbb{P}\left\{\mathbb{E}\left[(\widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}) - \pi(\boldsymbol{X}))^2\,\middle|\mathbb{D}_n\right] > \frac{\pi_0^4\epsilon^2}{16B^2}\right\} \\ &= \mathbb{P}\left\{\mathbb{E}\left[(\widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}) - \xi)^2\,\middle|\mathbb{D}_n\right] - \mathbb{E}\,(\pi(\boldsymbol{X}) - \xi)^2 > \frac{\pi_0^4\epsilon^2}{16B^2}\right\},\ \text{(since } \pi(\boldsymbol{X}) = \mathbb{E}(\xi|\boldsymbol{X})\text{)} \\ &\leq \mathbb{P}\left\{2\sup_{\tilde{\pi}\in\mathcal{P}}\left|\widehat{L}(\tilde{\pi}) - \mathbb{E}(\tilde{\pi}(\boldsymbol{X}) - \xi)^2\right| > \frac{\pi_0^4\epsilon^2}{16B^2}\right\}, \end{aligned} \tag{30}$$

where (30) follows from the fact that, since $\mathbb{E}(\pi(\boldsymbol{X}) - \xi)^2 = \inf_{\tilde{\pi}\in\mathcal{P}}\mathbb{E}(\tilde{\pi}(\boldsymbol{X}) - \xi)^2$, one has

$$\begin{aligned} &\mathbb{E}\left[(\widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}) - \xi)^2\,\middle|\mathbb{D}_n\right] - \mathbb{E}\,(\pi(\boldsymbol{X}) - \xi)^2 \\ &\quad\leq \sup_{\tilde{\pi}\in\mathcal{P}}\left\{\mathbb{E}\left[(\widehat{\pi}_{\mathsf{LS}}(\boldsymbol{X}) - \xi)^2\,\middle|\mathbb{D}_n\right] - \widehat{L}(\widehat{\pi}_{\mathsf{LS}}) + \widehat{L}(\widehat{\pi}_{\mathsf{LS}}) - \widehat{L}(\tilde{\pi}) + \widehat{L}(\tilde{\pi}) - \mathbb{E}(\tilde{\pi}(\boldsymbol{X}) - \xi)^2\right\} \\ &\quad\leq 2\sup_{\tilde{\pi}\in\mathcal{P}}\left|\widehat{L}(\tilde{\pi}) - \mathbb{E}(\tilde{\pi}(\boldsymbol{X}) - \xi)^2\right|, \quad \text{(because } \widehat{L}(\widehat{\pi}_{\mathsf{LS}}) - \widehat{L}(\tilde{\pi}) \leq 0,\ \forall\tilde{\pi}\in\mathcal{P}\text{)}. \end{aligned}$$

Finally, using Theorem 1, we find

$$(30) \leq 8\mathbb{E}\left[\mathcal{N}_1\left(\frac{\pi_0^4\epsilon^2}{(16)^2B}, \mathcal{P}, (\boldsymbol{X}_i)_{i=1}^n\right)\right]\ e^{-nC_{20}\epsilon^4}\,,$$

where $C_{20} = \pi_0^8/((128)(32)^2B^2)$. This completes the proof of Theorem 3. □

PROOF OF LEMMA 1.

The proof is based on the *symmetrization* arguments of Dudley and Pollard (Dudley (1978, P.925) and Pollard (1984, Sec. II.3)); also see van der Vaart and Wellner (1996, Sec. 2.3). Let $\mathbb{D}'_n = \{(\boldsymbol{X}'_1, \boldsymbol{V}'_1, \xi'_1), \ldots, (\boldsymbol{X}'_n, \boldsymbol{V}'_n, \xi'_n)\}$ be a hypothetical sample (a *ghost* sample) independent of the data $\mathbb{D}_n$, where $(\boldsymbol{X}'_i, \boldsymbol{V}'_i, \xi'_i) \overset{\text{iid}}{=} (\boldsymbol{X}_1, \boldsymbol{V}_1, \xi_1)$, $i = 1, \ldots, n$, and put

$$\breve{\nu}'_n(\psi) = n^{-1}\sum_{i=1}^{n}\xi'_i\psi(\boldsymbol{Z}'_i)/\pi(\boldsymbol{X}'_i)\,.$$

Next, fix the data $\mathbb{D}_n$ and observe that if $\sup_{\psi\in\Psi}|\breve{\nu}_n(\psi) - \nu(\psi)| > \epsilon$, then there is at least one $\psi_\epsilon \in \Psi$, which depends on $\mathbb{D}_n$ (but not $\mathbb{D}'_n$), such that $|\breve{\nu}_n(\psi_\epsilon) - \nu(\psi_\epsilon|\mathbb{D}_n)| > \epsilon$, where $\nu(\psi_\epsilon|\mathbb{D}_n) = \mathbb{E}[\psi_\epsilon(\boldsymbol{Z})|\mathbb{D}_n]$. Now observe that for $n\epsilon^2 \geq 8B^2/\pi_0^2$

$$\begin{aligned} \mathbb{P}\left\{\left|\breve{\nu}'_n(\psi_\epsilon) - \nu(\psi_\epsilon|\mathbb{D}_n)\right| < \frac{\epsilon}{2}\middle|\mathbb{D}_n\right\} &= 1 - \mathbb{P}\left\{\left|\breve{\nu}'_n(\psi_\epsilon) - \nu(\psi_\epsilon|\mathbb{D}_n)\right| \geq \frac{\epsilon}{2}\middle|\mathbb{D}_n\right\} \\ &\geq 1 - \sup_{\psi\in\Psi}\mathbb{P}\left\{\left|\breve{\nu}'_n(\psi) - \nu(\psi)\right| \geq \frac{\epsilon}{2}\right\} \\ &\geq 1 - \sup_{\psi\in\Psi}\mathbb{E}\big(\breve{\nu}'_n(\psi) - \nu(\psi)\big)^2/(\epsilon/2)^2, \text{(Markov's inequality)} \\ &\geq 1 - \frac{4}{n\epsilon^2}\sup_{\psi\in\Psi}\mathsf{Var}(\xi_1\psi(\boldsymbol{Z}_1)/\pi(\boldsymbol{X}_1)) \qquad (31) \\ &\geq 1 - \frac{4B^2}{n\pi_0^2\epsilon^2} \geq \frac{1}{2}, \quad \text{(because } n\epsilon^2 \geq 8B^2/\pi_0^2\text{)}, \end{aligned}$$

where (31) follows from the MAR assumption (**??**) and the fact that

$$\mathbb{E}[\xi\psi(\boldsymbol{Z})/\pi(\boldsymbol{X})] = \mathbb{E}[\mathbb{E}(\xi\psi(\boldsymbol{Z})/\pi(\boldsymbol{X})|\boldsymbol{Z})] \overset{\text{by (??)}}{=} \mathbb{E}(\psi(\boldsymbol{Z})) =: \nu(\psi)\,.$$

Therefore, for $n\epsilon^2 \geq 8B^2/\pi_0^2$, we have

$$\begin{aligned}
\frac{1}{2} &\leq \mathbb{P}\left\{\left|\breve{\nu}'_n(\psi_\epsilon) - \nu(\psi_\epsilon|\mathbb{D}_n)\right| < \frac{\epsilon}{2}\middle|\mathbb{D}_n\right\} \\
&\leq \mathbb{P}\left\{\left|\breve{\nu}_n(\psi_\epsilon) - \nu(\psi_\epsilon|\mathbb{D}_n)\right| - \left|\breve{\nu}'_n(\psi_\epsilon) - \breve{\nu}_n(\psi_\epsilon)\right| < \frac{\epsilon}{2}\middle|\mathbb{D}_n\right\} \\
&\leq \mathbb{P}\left\{\left|\breve{\nu}'_n(\psi_\epsilon) - \breve{\nu}_n(\psi_\epsilon)\right| > \frac{\epsilon}{2}\middle|\mathbb{D}_n\right\} \\
&\quad \text{(because by the definition of } \psi_\epsilon\text{, conditional on } \mathbb{D}_n\text{, } |\breve{\nu}_n(\psi_\epsilon) - \nu(\psi_\epsilon|\mathbb{D}_n)| > \epsilon\text{)} \\
&\leq \mathbb{P}\left\{\sup_{\psi\in\Psi}\left|\breve{\nu}'_n(\psi) - \breve{\nu}_n(\psi)\right| > \frac{\epsilon}{2}\middle|\mathbb{D}_n\right\}. \qquad (32)
\end{aligned}$$

Now observe that the far left and the far right sides of (32) do not depend on ψ_ϵ and that the chain of inequalities between them remain valid on the set $\{\sup_{\psi\in\Psi}|\breve{\nu}_n(\psi) - \nu(\psi)| > \epsilon\}$. Therefore, integrating the two far sides of (32) with respect to the distribution of $\mathbb{D}_n$, over this set, we find

$$\begin{aligned}
\mathbb{P}\{\sup_{\psi\in\Psi}|\breve{\nu}_n(\psi) - \nu(\psi)| > \epsilon\} &\leq 2\,\mathbb{P}\left\{\sup_{\psi\in\Psi}\left|\breve{\nu}'_n(\psi) - \breve{\nu}_n(\psi)\right| > \frac{\epsilon}{2}\right\} \\
&\leq 2\,\mathbb{P}\left\{\sup_{\psi\in\Psi}\frac{1}{n}\left|\sum_{i=1}^n\left[\frac{\xi'_i\psi(\boldsymbol{Z}'_i)}{\pi(\boldsymbol{X}'_i)} - \frac{\xi_i\psi(\boldsymbol{Z}_i)}{\pi(\boldsymbol{X}_i)}\right]\right| > \frac{\epsilon}{2}\right\}. \qquad (33)
\end{aligned}$$

Next, let $\sigma_1, \ldots, \sigma_n$ be iid random variables, independent of $\mathbb{D}_n$ and $\mathbb{D}'_n$,where $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$. Observing that σ_i's are random signs, we have

$$\begin{aligned}
(33) &= 2\,\mathbb{P}\left\{\sup_{\psi\in\Psi}\frac{1}{n}\left|\sum_{i=1}^n\sigma_i\left[\frac{\xi'_i\psi(\boldsymbol{Z}'_i)}{\pi(\boldsymbol{X}'_i)} - \frac{\xi_i\psi(\boldsymbol{Z}_i)}{\pi(\boldsymbol{X}_i)}\right]\right| > \frac{\epsilon}{2}\right\} \\
&\leq 4\,\mathbb{P}\left\{\sup_{\psi\in\Psi}\frac{1}{n}\left|\sum_{i=1}^n\frac{\sigma_i\xi_i\psi(\boldsymbol{Z}_i)}{\pi(\boldsymbol{X}_i)}\right| > \frac{\epsilon}{4}\right\} \\
&\leq 4\,\mathbb{E}\left[\mathbb{P}\left\{\sup_{\psi\in\Psi}\frac{1}{n}\left|\sum_{i=1}^n\frac{\sigma_i\xi_i\psi(\boldsymbol{Z}_i)}{\pi(\boldsymbol{X}_i)}\right| > \frac{\epsilon}{4}\middle|\mathbb{D}_n\right\}\right]. \qquad (34)
\end{aligned}$$

Now, put $\epsilon' = \epsilon/8$ and, for fixed $\mathbb{D}_n$, let $\Psi_{\epsilon'}$ be a weighted empirical L_1 ϵ'-cover of Ψ based on the weights $W_i = \xi_i/\pi(\boldsymbol{X}_i)$. That is, for each $\psi \in \Psi$ there is a $\psi^* \in \Psi_{\epsilon'}$ such that $n^{-1}\sum_{i=1}^n W_i\,|\psi(\boldsymbol{Z}_i) - \psi^*(\boldsymbol{Z}_i)| < \epsilon' = \epsilon/8$. Let $\Gamma_1(\epsilon', \Psi, \mathbb{D}_n)$ be the ϵ'-covering number of Ψ with respect to the weighted empirical L_1 norm. Then for some $\psi^* \in \Psi_{\epsilon'}$ we have

$$\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^n\frac{\sigma_i\xi_i\psi(\boldsymbol{Z}_i)}{\pi(\boldsymbol{X}_i)}\right| &\leq \frac{1}{n}\left|\sum_{i=1}^n\frac{\sigma_i\xi_i}{\pi(\boldsymbol{X}_i)}\left[\psi(\boldsymbol{Z}_i) - \psi^*(\boldsymbol{Z}_i)\right]\right| + \frac{1}{n}\left|\sum_{i=1}^n\frac{\sigma_i\xi_i\psi^*(\boldsymbol{Z}_i)}{\pi(\boldsymbol{X}_i)}\right| \\
&\leq \frac{\epsilon}{8} + \frac{1}{n}\left|\sum_{i=1}^n\frac{\sigma_i\xi_i\psi^*(\boldsymbol{Z}_i)}{\pi(\boldsymbol{X}_i)}\right|.
\end{aligned}$$

Consequently

$$\begin{aligned}
(34) &\leq 4\,\mathbb{E}\left[\mathbb{P}\left\{\sup_{\psi\in\Psi_{\epsilon'}}\frac{1}{n}\left|\sum_{i=1}^n\frac{\sigma_i\xi_i\psi(\boldsymbol{Z}_i)}{\pi(\boldsymbol{X}_i)}\right| > \frac{\epsilon}{8}\middle|\mathbb{D}_n\right\}\right] \\
&\leq 4\,\mathbb{E}\left[\Gamma_1(\epsilon', \Psi, \mathbb{D}_n)\cdot\max_{\psi\in\Psi_{\epsilon'}}\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^n\frac{\sigma_i\xi_i\psi(\boldsymbol{Z}_i)}{\pi(\boldsymbol{X}_i)}\right| > \frac{\epsilon}{8}\middle|\mathbb{D}_n\right\}\right] \\
&\leq 8\,\mathbb{E}\left[\Gamma_1(\epsilon/8, \Psi, \mathbb{D}_n)\right]\,e^{-n\pi_0^2\epsilon^2/(128B^2)}, \quad \text{(by Hoeffding's inequality).}
\end{aligned}$$

However, for all functions $\psi_1, \psi_2 : \mathbb{R}^{d+p} \to \mathbb{R}$ one has $\sum_{i=1}^n (\xi_i/\pi(\boldsymbol{X}_i))|\psi_1(\boldsymbol{Z}_i) - \psi_2(\boldsymbol{Z}_i)| \leq (1/\pi_0)\sum_{i=1}^n |\psi_1(\boldsymbol{Z}_i) - \psi_2(\boldsymbol{Z}_i)|$. Therefore, if $\{\phi_1, \ldots, \phi_N\}$ is a minimal $(\pi_0\epsilon)$-cover of Ψ with respect to the empirical L_1 norm, then it is an ϵ-cover of Ψ with respect to the weighted empirical L_1 norm. Thus, for every $\epsilon > 0$, we find $\Gamma_1(\epsilon, \Psi, \mathbb{D}_n) \leq \mathcal{N}_1(\epsilon\pi_0, \Psi, \mathbb{D}_n)$. Putting all the above together, we have, for $n\epsilon^2 \geq 8B^2/\pi_0^2$,

$$\mathbb{P}\{\sup_{\psi\in\Psi} |\breve{\nu}_n(\psi) - \nu(\psi)| > \epsilon\} \;\leq\; 8\,\mathbb{E}\left[\mathcal{N}_1\left(\frac{\pi_0\epsilon}{8}, \Psi, \mathbb{D}_n\right)\right] e^{-n\pi_0^2\epsilon^2/(128B^2)}\,.$$

When $n\epsilon^2 < 8B^2/\pi_0^2$ the lemma is trivially true (because the bound in the lemma will exceed 1).

□

PROOF OF LEMMA 2.

Part (i). The proof is similar to (and in fact easier than) that of Lemma 2.2 of Mojirsheibani and Montazeri (2007) and goes as follows. First note that

$$\mathbb{E}\left[\widehat{T}(\boldsymbol{X})\Big|\boldsymbol{X}\right] = h^{-d}\mathbb{E}\left[\xi_1\mathcal{K}((\boldsymbol{X}-\boldsymbol{X}_1)/h)\Big|\boldsymbol{X}\right] \stackrel{a.s.}{=} h_n^{-d}\mathbb{E}\left[\mathcal{K}((\boldsymbol{X}-\boldsymbol{X}_1)/h_n)\,\mathbb{E}[\xi_1|\boldsymbol{X},\boldsymbol{X}_1]\Big|\boldsymbol{X}\right].$$

Since $\mathbb{E}[\xi_1|\boldsymbol{X},\boldsymbol{X}_1] = \mathbb{E}[\xi_1|\boldsymbol{X}_1] = \pi(\boldsymbol{X}_1)$ (because $\boldsymbol{X}$ is independent of ξ_1 and $\boldsymbol{X}_1$), we find

$$\begin{aligned}\mathbb{E}\left[\widehat{T}(\boldsymbol{X})\big|\boldsymbol{X}\right] - T(\boldsymbol{X}) &= h_n^{-d}\mathbb{E}\left[(\pi(\boldsymbol{X}_1)-\pi(\boldsymbol{X}))\mathcal{K}((\boldsymbol{X}-\boldsymbol{X}_1)/h_n)\Big|\boldsymbol{X}\right] \\ &\quad + \mathbb{E}\left[\pi(\boldsymbol{X})\left\{h_n^{-d}\mathcal{K}((\boldsymbol{X}-\boldsymbol{X}_1)/h_n) - f(\boldsymbol{X})\right\}\Big|\boldsymbol{X}\right] \\ &:= R_{n,1}(\boldsymbol{X}) + R_{n,2}(\boldsymbol{X})\,.\end{aligned}$$

Now a one-term Taylor expansion gives

$$R_{n,1}(\boldsymbol{X}) = h_n^{-d}\mathbb{E}\left[\left(\sum_{i=1}^d (X_i - X_{1,i})\,\partial\pi(\boldsymbol{X}^*)/\partial X_i\right) \times \mathcal{K}((\boldsymbol{X}-\boldsymbol{X}_1)/h_n)\Big|\boldsymbol{X}\right],$$

where X_i and $X_{1,i}$ are the i^{th} components of $\boldsymbol{X}$ and $\boldsymbol{X}_1$, respectively, and $\boldsymbol{X}^*$ is a point on the interior of the line segment joining the points $\boldsymbol{X}$ and $\boldsymbol{X}_1$. Therefore,

$$\begin{aligned}|R_{n,1}(\boldsymbol{X})| &\leq C_{10}\sum_{i=1}^d \mathbb{E}\left[|X_{1,i} - X_i|\,h_n^{-d}\mathcal{K}((\boldsymbol{X}-\boldsymbol{X}_1)/h_n)\Big|\boldsymbol{X}\right] \\ &\qquad \text{(where } C_{10} = \max_{1\leq i\leq d}\sup_{\boldsymbol{x}}|\partial\pi(\boldsymbol{x})/\partial x_i| < \infty, \text{ by Assumption A4)} \\ &= C_{10}\sum_{i=1}^d \int |x_i - X_i|\,h_n^{-d}\mathcal{K}((\boldsymbol{X}-\boldsymbol{x})/h_n)f(\boldsymbol{x})d\boldsymbol{x} \\ &\leq C_{10}\|f\|_\infty \sum_{i=1}^d h_n \int |u_i|\mathcal{K}(\boldsymbol{u})d\boldsymbol{u} \;=\; C_{11}h_n, \quad \text{(by Assumptions A2 and A4)},\end{aligned}$$

where $0 < C_{11} < \infty$. As for the term $R_{n,2}(\boldsymbol{X})$, we have

$$\begin{aligned}|R_{n,2}(\boldsymbol{X})| &= \left|\pi(\boldsymbol{X})\int h^{-d}\mathcal{K}((\boldsymbol{X}-\boldsymbol{x})/h_n)[f(\boldsymbol{x}) - f(\boldsymbol{X})]d\boldsymbol{x}\right| \\ &= \left|\pi(\boldsymbol{X})\int [f(\boldsymbol{X} - h_n\boldsymbol{v}) - f(\boldsymbol{X})]\mathcal{K}(\boldsymbol{v})d\boldsymbol{v}\right| \\ &\leq \left(\sum_{i=1}^d \sup_{\boldsymbol{x}}|\partial f(\boldsymbol{x})/\partial x_i| \cdot \int |v_i|\mathcal{K}(\boldsymbol{v})d\boldsymbol{v}\right) h_n \;=\; C_{12}h_n\,,\end{aligned}$$

where $0 < C_{12} < \infty$, by Assumptions A2. This completes the proof of part (i).

Part (ii). For $j = 1, \ldots, n$, let

$$S_j(\boldsymbol{X}) \;=\; h_n^{-d}\left\{\xi_j \mathcal{K}((\boldsymbol{X}-\boldsymbol{X}_j)/h_n) - \mathbb{E}\big[\xi_j \mathcal{K}((\boldsymbol{X}-\boldsymbol{X}_j)/h_n)\big|\boldsymbol{X}\big]\right\}\,,$$

and observe that, conditional on $\boldsymbol{X}$, the terms $S_j(\boldsymbol{X})$ are independent, zero-mean random variables, bounded by $-h_n^{-d}\|\mathcal{K}\|_\infty$ and $+h_n^{-d}\|\mathcal{K}\|_\infty$. Furthermore, $\mathrm{Var}(S_j(\boldsymbol{X})|\boldsymbol{X}) = \mathbb{E}[S_j^2(\boldsymbol{X})|\boldsymbol{X}] \leq 2h_n^{-d}$ $\|\mathcal{K}\|_\infty\|f\|_\infty$. Therefore, by Bernstein's (1946) inequality,

$$\begin{aligned}
\mathbb{P}\left\{\left|\widehat{T}(\boldsymbol{X}) - \mathbb{E}\left[\widehat{T}(\boldsymbol{X})\middle|\boldsymbol{X}\right]\right| > \beta\middle|\boldsymbol{X}=\boldsymbol{x}\right\} &= \mathbb{P}\left\{\frac{1}{n}\left|\sum_{j=1}^n S_j(\boldsymbol{X})\right| > \beta\middle|\boldsymbol{X}=\boldsymbol{x}\right\} \\
&\leq 2\exp\left\{\frac{-n\beta^2}{2\left[2h_n^{-d}\|\mathcal{K}\|_\infty\|f\|_\infty + h_n^d\|\mathcal{K}\|_\infty\beta/3\right]}\right\},
\end{aligned}$$

which does not depend on $\boldsymbol{x}$. The lemma now follows upon integrating both sides with respect to the distribution of $\boldsymbol{X}$.

□

References

Ahmad, I.A. and Mugdadi, A.R. 2006. Weighted Hellinger distance as an error criterionn for bandwidth selection in kernel estimation. J. Nonpar. Statist. 18, 215-236.

Alexander, K. 1984. Probability inequalities for empirical processes and a law of the iterated logarithm. Ann. Probab. 12, 1041-1067.

Bernstein, S. 1946. The theory of probabilities. Gastehizdat Publishing House, Moscow.

Bravo, F. 2015. Semiparametric estimation with missing covariates. Journal of Multivariate Analysis, 139, 329-346.

Cheng P.E. 1994. Nonparametric estimation of mean functionals with data missing at random. Journal of the American Statistical Association, 87, 81-87.

Cheng, P.E. and Chu, C.K. 1996. Kernel estimation of distribution functions and quantiles with missing data. Statistica Sinica, 6, 63-78.

Chenouri, S., Mojirsheibani, M., and Montazeri, Z. 2009. Empirical measures for incomplete data with applications. Electron. J. Stat. 3, 1021-1038.

Devroye, L., Györfi, L., Lugosi, G. 1996. A probabilistic theory of pattern recognition. Springer-Verlag, New York.

Dubnicka, S. 2009. Kernel density estimation with missing data and auxiliary variables. Aust. N. Z. J. Stat. 51, 247-270.

Dudley, R.M. 1978. Central limit theorems for empirical measures. Ann. Probab. 6, 899-929.

Giné, E. 1996. Empirical processes and applications: an overview. Bernoulli 2, 1-8.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. 2002. A distribution-free theory of nonparametric regression. Springer-Verlag, New York.

Hädle, W., Marron, J. 1995 Optimal bandwidth seleciton in nonparametric regression function estimation. Ann Stat 13:1465-1481.

Hazelton, M.L. 2000. Marginal density estimation from incomplete bivariate data. Statistics & Probability Letters, 47, 75-84.

Hirano, K.I. and Ridder, G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71, 1161-1189.

Horvitz, D.G. and Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.

Hu, Z., Follmann, D., and Qin, J. 2011. Dimension reduced kernel estimation for distribution function with incomplete data. J. Statist. Plann. Inference 141, 3084-3093.

Kim, J.K. and Yu, C.L. 2011. A semiparametric estimation of mean functionals with nonignorable missing data. J. Amer. Statist. Assoc. 106, 157-165.

Little, R.J.A. and Rubin, D.B. 2002. Statistical Analysis With Missing Data, 2nd ed. Wiley, New York.

Liu, X., Liu, P., and Zhou, Y. 2011. Distribution estimation with auxiliary information for missing data. J. Statist. Plann. Inference 141, 711-724. Massart, P. 1990. The tight constant in the Dvoretzky-

Kiefer-Wolfowitz inequality. Ann. Probab. 18, 1269-1283.

Mojirsheibani, M. and Montazeri, Z. 2007. On nonparametric classification with missing covariates. J. Multivariate Anal. 98, 1051-1071.

Müller, U. 2012. Estimating the density of a possibly missing response variable in nonlinear regression. J. Statist Plann. Inference, 142, 1198-1214.

Müller, U. 2009. Estimating linear functionals in nonlinear regression with responses missing at random. Ann. Statist. 37, 2245-2277.

Parzen, E. 1962. On estimation of a probability density function and mode. Ann. Statist. 33, 1065-1076.

Pollard, D. 1984. Convergence of Stochastic Processes. Springer, New York.

Prakasa Rao, B.L.S. 1983. Nonparametric functional estimation. Academic Press, Orlando.

Racine, J. and Hay
eld, T. 2008. Nonparametric Econometrics: The np Package. J. Statistical Software, 27, 1-32.

Robins, J., Rotnitzky, A. 1995. Semiparametric efficiency in multivariate regression models with missing data. J. Amer. Statist. Assoc., 90, 122-129.

Robins, J., Rotnitzky, A., and Zhao, Lue, P. 1994. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 89, 846-866.

Rosenblatt, M. 1956. Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27, 832-837.

Rueda, M., Martínez, S., Martínez, H., and Arcos, A. (2006) Mean estimation with calibration techniques in presence of missing data. Comput. Statist. Data Anal. 50, 3263-3277.

Talagrand, M. 1994. Sharper bounds for Gaussian and empirical processes. Ann. Probab. 22, 280-76.

van de Geer, S. 2000. Applications of empirical process theory. Cambridge University Press, Cambridge.

van der Vaart, A. and Wellner, J. 1996. Weak convergence and empirical processes. Springer-Verlag, New York.

Vapnik, V. 1998. Statistical learning theory. John Wiley & Sons, Inc., New York.

Wang, Q. 2008. Probability density estimation with data missingat random when covariables are present. J. Statist. Plann. Inference, 138, 568-587.

Wang, Q. and Rao, J. 2002. Empirical likelihood-based inference under imputation for missing response data. Ann. Statist. 30, 890-924.

Wang, Q., Linton, O., Härdle, W. 2004. Semiparametric regression analysis with missing response at random. J. Amer. Statist. Assoc., 99, 334-345.

Wang, Q. and Qin, Y. 2010. Empirical likelihood confidence bands for distribution functions with missing responses. Journal of Statistical Planning and Inference, 140, 2778-2789.

Wang, L., Rotnitzky, A., Lin, X. 2010. Nonparametric regression with missing outcomes using weighted kernel estimating equations. J. Amer. Statist. Assoc., 105, 1135-1146.

Zou, Y., Liang, H.,Zhang, J. 2015. Nonlinear wavelet density estimation with data missing at random when covariates are present. Metrika, 78, 967-995

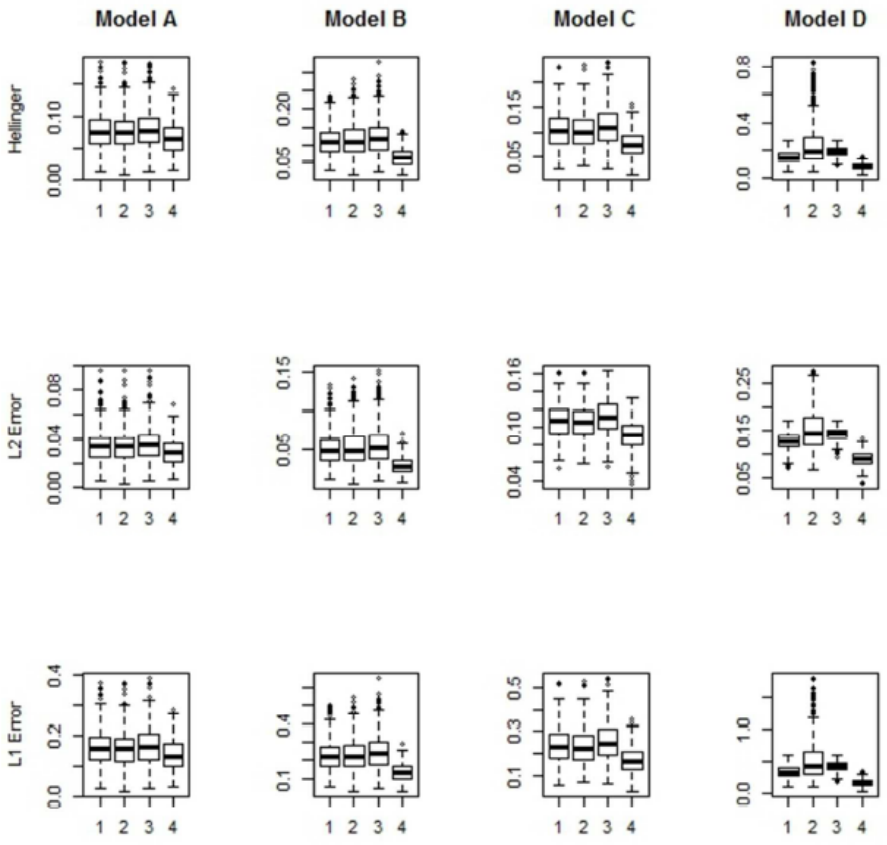
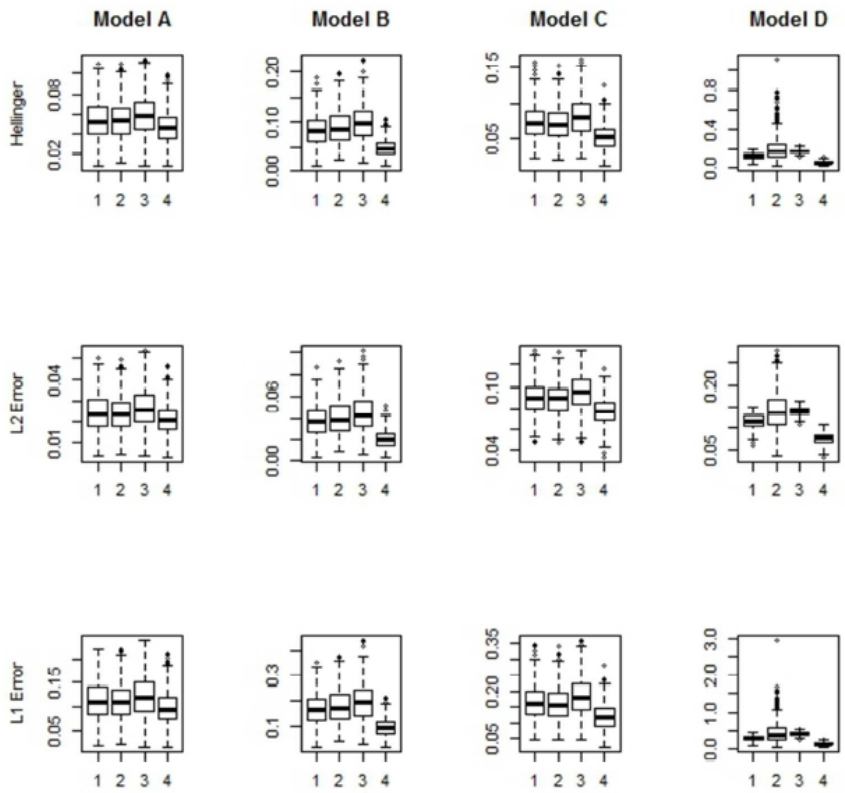
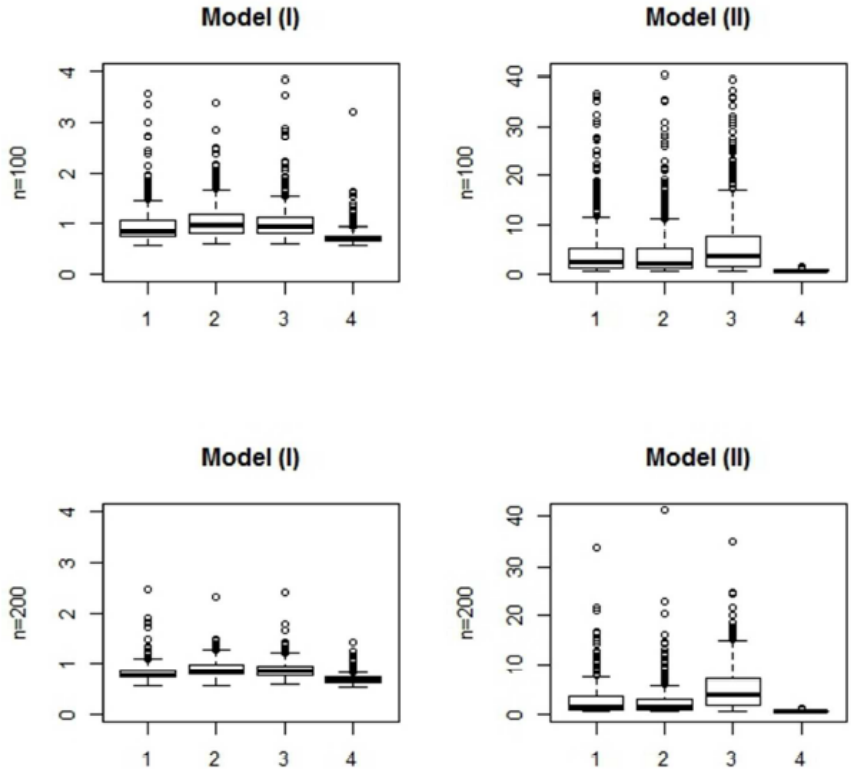


Figure 4:



Boxplots

Figure 5:



Boxplots for Example B

Figure 6: