Statistical Papers
https://doi.org/10.1007/s00362-022-01293-0

REGULAR ARTICLE

On the maximal deviation of kernel regression estimators with NMAR response variables

Majid Mojirsheibani[1]

Received: 27 April 2021 / Revised: 26 October 2021 / Accepted: 31 January 2022

Abstract
This article focuses on the problem of kernel regression estimation in the presence of nonignorable incomplete data with particular focus on the limiting distribution of the maximal deviation of the proposed estimators. From an applied point of view, such a limiting distribution enables one to construct asymptotically correct uniform bands, or perform tests of hypotheses, for a regression curve when the available data set suffers from missing (not necessarily at random) response values. Furthermore, such asymptotic results have always been of theoretical interest in mathematical statistics. We also present some numerical results that further confirm and complement the theoretical developments of this paper.

Keywords Kernel regression · Maximal deviation · Nonignorble missing · Uniform bands

1 Introduction

This paper deals with the problem of kernel regression estimation for incomplete data, with particular focus on the asymptotic distribution of the maximal deviation of the corresponding estimators under the *Not-Missing-At-Random* (NMAR) mechanism, also called the *nonignorable* missing mechanism. This is generally acknowledged to be a challenging problem in the missing data literature. Such limiting distributions have always been of both theoretical and applied interest in statistics and probability; see, for example, the classical results of Wandl (1980), Johnston (1982), Liero (1982), Konakov and Piterbarg (1984), Härdle (1989), and Muminov (2011, 2012) for the case of fully observable data. From an applied point of view, the limiting distribution of such statistics makes it possible to carry out various types of global inference

✉ Majid Mojirsheibani
majid.mojirsheibani@csun.edu

[1] Department of Mathematics, California State University, Northridge, CA 91330, USA

Published online: 10 February 2022

for an unknown regression curve over connected compact sets. These include tests of hypotheses as well as the construction of asymptotically correct uniform bands for an unknown regression curve. Uniform bands can be used to study particular characteristics such as the appearance, shape, and the overall variability of the unknown regression curve. In the case of fully observablle data, results along these lines appear in the work of Johnston (1982), Härdle (1989), Eubank and Speckman (1993), and Xia (1998), Dehuvels and Mason (2004), Massé and Meiniel (2014), Cai et al. (2014), and Proksch (2016). Other relevant results along these lines include the work of Sun and Loader (1994), Neumann and Polzehl (1998), Sun and Zhou (1998), Claeskens and Van Keilegom (2003), Song et al. (2012), Härdle and Song (2010), Nemouchi and Mohdeb (2010), Withers and Nadarajah (2012), Wang et al. (2013), Lütkepohl (2013), Sabbah (2014), Gu and Yang (2015), Lu and Kuriki (2017), Wojdyla and Szkutnik (2018), Yang and Barber (2019), Gardes (2020), Zhou et al. (2020), and Gu et al. (2021).

Virtually all of these papers address the situations where there are no missing data. In practice, some data values may be missing. Unobservable or missing data are widespread in survey data, public opinion polls, medical research data as well as the data collected in many other areas of scientific activities. Here, we assume that for various reasons some of the response values (i.e., Y) may be unavailable or unobservable. More formally, let (X, Y) be a random pair and consider the problem of estimating the regression function $m(x) = E(Y|X = x)$, based on the independent and identically distributed (iid) data (X_i, Y_i), $i = 1, \ldots, n$. Let $m_n(x)$ be the popular Nadaraya-Watson kernel estimator of the true regression function $m(x)$, i.e.,

$$m_n(x) = \frac{\sum_{i=1}^n Y_i \, \mathcal{K}((x - X_i)/h_n)}{\sum_{i=1}^n \mathcal{K}((x - X_i)/h_n)}, \tag{1}$$

where, in general, $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}_+$ is the kernel used with the *smoothing* parameter h_n. A particularly important measure of the global (uniform) accuracy of $m_n(\cdot)$, as an estimator of $m(\cdot)$, is the *maximal deviation* statistic, $\sup_{x \in \mathcal{C}} |m_n(x) - m(x)|$, where $\mathcal{C}$ is a connected and compact set. The limiting distribution of the properly normalized versions of this statistic has been well studied by many authors. The focus of this paper is to pursue such studies for the more realistic setup where the response variable Y may be unobservable or missing. More importantly, we assume a NMAR missing mechanism, also called the *nonignorable* missing pattern, where unlike the so-called *Missing-At-Random* scenarios, here the probability that Y is missing is allowed to depend on both Y and X (and not just X alone). The nonignorable missing pattern is generally acknowledged to be a far more challenging problem in the missing data literature; see, for example, Shao and Wang (2016).

In the case of predictive models such as regression, where the response variable Y has to be predicted based on the predictor X, Kim and Yu (2011) considered a rather flexible logistic missing probability model that works as follows. Define the indicator variable $\Delta = 1$ if Y is not missing (and $\Delta = 0$ otherwise). Then, Kim and Yu (2011) proposed and studied the flexible model

$$\pi_\gamma(x, y) := E\left[\Delta \,\middle|\, X = x, Y = y\right] = \frac{1}{1 + \exp\left\{g(x) + \gamma y\right\}}, \tag{2}$$

where g is a completely unknown function of the predictor X, and γ is an unknown parameter. The missing probability model (2) has been explored and exploited extensively in the literature; see, for example, Zhao and Shao (2015), Shao and Wang (2016), Morikawa et al. (2017), Uehara and Kim (2018), Morikawa and Kim (2018), Morikawa and Kano (2018), Fang et al. (2018), O'Brien et al. (2018), Maity et al. (2019), Sadinle and Reiter (2019), Zhao et al. (2019), Yuan et al. (2020), Chen et al. (2020), Liu and Yuan (2020), Mojirsheibani (2021), and Liu and Yau (2021). Of course, one may decide to consider more general nonparametric models instead of (2), but the estimation of such general models will become a difficult (if not impossible) issue. In fact, in view of the recent widespread use of the nonignorable model (2) in the literature, there appears to be the tacit consensus that (2) is versatile enough to be used in predictive models such as regression, and this will also be the direction of the current paper. In passing, we also note that if $\gamma = 0$, then (2) reduces to the simpler case of missing at random assumption.

We also note that under the unrealistic assumptions that $\pi_\gamma(x, y)$, the density $f(x)$ of X, or the conditional variance $\mathrm{Var}(Y|X = x)$ are known, one may still be able to study the limiting distribution of the maximal deviation of the estimator $m_n(x)$ with nonignorable missing Y's and construct asymptotically correct bands for $m(x)$. But such assumptions are not warranted in practice and will not be addressed here.

The paper is organized as follows. In Sect. 2.2 we develop a kernel regression estimator that takes into account the nonignorable missing mechanism that causes the absence of information in terms of the response variable Y. Our main result (Theorems 2 and 3) deal with the limiting distribution of the properly normalized version of the maximal deviation of the proposed kernel estimator, under standard conditions, without making any parametric assumptions about the underlying density of X or the the conditional variance $\mathrm{Var}(Y|X = x)$. The numerical results of Sect. 3 further confirm the good finite-sample performance of the proposed estimator.

2 Main results and methodology

2.1 The basic framework and preliminaries

We start by giving a brief overview of the standard setup where there are no missing data. More specifically, let $m_n(x)$ defined in (1) be the kernel regression estimator of the true regression function $m(x) = E[Y|X = x]$. Additionally, let

$$v_n^2(x) = \frac{\sum_{i=1}^n Y_i^2 \mathcal{K}((x - X_i)/h_n)}{\sum_{i=1}^n \mathcal{K}((x - X_i)/h_n)} - [m_n(x)]^2 \tag{3}$$

be the kernel estimator the conditional variance $v_0^2(x) = E[Y^2|X = x] - m^2(x)$. The asymptotic distribution of the statistic $\sup_{x\in[0,1]} \sqrt{f_n(x)/v_n^2(x)}\,\left|m_n(x) - m(x)\right|$, where $f_n(x)$ is the usual kernel density estimator of the density f of X, has been

extensively studied by a number of authors (see the references in the introduction section). Here, the interval $[0, 1]$ over which the supremum is taken, can be any connected compact subset of the interior of the support of f. To present our main results, we first state a number of assumptions, virtually all of which are as in Liero (1982). More specifically,

Assumption (A) The vector (X, Y) has a probability density function (pdf) with respect to the Lebesgue measure, denoted by $g(x, y)$. The response variable Y is bounded with probability one, i.e., $P\{B_1 \leq Y \leq B_2\} = 1$ for constants $-\infty < B_1 < B_2 < \infty$.

Assumption (B) The marginal pdf, f, of X is strictly positive on $(-\epsilon, 1+\epsilon)$, for some $\epsilon > 0$, and vanishes outside of a finite interval $[a, b]$, (thus $[0, 1] \in (a, b)$).

Assumption (C) The functions $v_0^2(x) = E[(Y - m(X))^2 | X = x]$, $m(x)$, and $f(x)$ are twice differentiable with bounded derivatives. Also, $v_0^2(x)$ is strictly positive on $[0, 1]$.

Next, put $Z = Y - m(X)$ and let $\widetilde{G}(x, z)$ and $\widetilde{g}(x, z)$ be, respectively, the cdf and the pdf of the vector (X, Z). Define $H(z|x)$ and $h(z|x)$ to be the conditional cdf and the conditional pdf of Z given X, respectively.

Assumption (D) The function $\widetilde{g}^{1/2}(x, z)$, where $\widetilde{g}$ is as defined above, is differentiable with respect to both x and z, and the partial derivatives are bounded. Also, the inverse functions H^{-1} and F^{-1} of H and F exist and $\frac{\partial}{\partial x} H^{-1}\big(z|F^{-1}(x)\big)$ and $\frac{\partial}{\partial z} H^{-1}\big(z|F^{-1}(x)\big)$ are bounded; here F is the cdf of the univariate random variable X.

Assumption (E) The kernel $\mathcal{K}$ is a density function with the bounded support $[-A, A]$ for some $A > 0$, it is continuously differentiable on $(-A, A)$, and satisfies $\int x\, \mathcal{K}(x)\, dx = 0$.

The following result can be found in Liero (1982).

Theorem 1 *Let $h_n = n^{-\delta}$, $\frac{1}{5} < \delta < \frac{1}{3}$, and suppose that assumptions* (A)–(E) *hold. Then*

$$\sqrt{2\delta \log n}\left\{\sqrt{\frac{nh_n}{c_K}} \sup_{x\in[0,1]} \sqrt{\frac{f_n(x)}{v_n^2(x)}} \Big|m_n(x) - m(x)\Big| - \varphi(n)\right\} \longrightarrow^d U, \quad \text{as } n \to \infty, \tag{4}$$

where $P\{U \leq u\} = \exp\big(-2\, e^{-u}\big)$, $c_K = \int K^2(t)\, dt$, and the term $\varphi(n)$ is given by (15).

The result in (4) can be used to construct confidence bands for $m(x)$. In fact, in view of (4),

$$m_n(x) \pm \left(\frac{c_K \cdot v_n^2(x)}{nh_n \cdot f_n(x)}\right)^{1/2} \left(\frac{x^{(\alpha)}}{\sqrt{2\log h_n^{-1}}} + \varphi(n)\right), \quad x \in [0, 1], \tag{5}$$

is an asymptotically correct $(1-\alpha)100\%$ confidence band for the regression function $m(x)$, where $x^{(\alpha)}$ is the solution of the equation $\exp\{-2\exp(-x)\} = 1-\alpha$. One can also carry out a test of hypothesis such as H_0: $m = m_0$, for a known function m_0, using the statistic $\tau_n := \sup_{x\in[0,1]} \sqrt{f_n(x)/v_n^2(x)}\,\big|m_n(x) - m_0(x)\big|$. Such a test rejects H_0 at level α, if $\tau_n > \sqrt{c_K/(nh_n)}\,\big\{\big(2\log h_n^{-1}\big)^{-1/2}\big[\log 2 - \log\log\big(\frac{1}{1-\alpha}\big)\big] + \varphi(n)\big\}$.

2.2 The proposed estimators and their maximal deviation

Suppose that the response variable is allowed to be missing nonignorably. Then, it is straightforward to see that the usual kernel regression estimator m_n in (1) is no longer available. At this stage, one may decide (by mistake) to consider an estimator based on the complete cases only, i.e., the estimator $m_n^{\mathrm{CC}}(x) := \sum_{i=1}^n \Delta_i Y_i\, \mathcal{K}((x - X_i)/h_n) \big/ \sum_{i=1}^n \Delta_i \mathcal{K}((x - X_i)/h_n)$. Unfortunately, this is the kernel estimator of the quantity $E(\Delta Y|X = x)\big/E(\Delta|X = x)$ which is in general different from the regression function $m(x) = E(Y|X = x)$ under the nonignorable missing mechanism (2).

To effectively deal with the presence of nonignorably missing response values, first consider the hypothetical situation where the function $\pi_\gamma(x, y)$, defined in (2), is completely known (unrealistic) and define

$$m_{\pi,n}(x) = \frac{\sum_{i=1}^n \frac{\Delta_i Y_i}{\pi_\gamma(X_i,Y_i)}\,\mathcal{K}\big((x - X_i)/h_n\big)}{\sum_{i=1}^n \mathcal{K}\big((x - X_i)/h_n\big)}. \tag{6}$$

Clearly, $m_{\pi,n}(x)$ is the kernel regression estimator of $E\big[\Delta Y\big/\pi_\gamma(X, Y)\big|X = x\big] = E[Y|X = x] = m(x)$, which follows from the simple argument that

$$E\left[\Delta Y/\pi_\gamma(X, Y)\big|X\right] = E\left(E\left[\Delta Y/\pi_\gamma(X, Y)\big|X, Y\right]\big|\, X\right) \overset{\text{via (2)}}{=} m(X). \tag{7}$$

Of course, in practice, $\pi_\gamma(x, y)$ is virtually always unknown and must be replaced by some estimator. Unfortunately, the problem is compounded by the fact that even if the selection probability $\pi_\gamma(x, y)$ is completely known, one cannot anticipate to be able to establish the conclusion of Theorem 1 for the maximal deviation of $m_{\pi,n}(x)$ in (6) from $m(x)$. The reason is that $m_{\pi,n}(x)$ is the kernel estimator of $E[Y^*|X = x]$, where $Y^* := \Delta Y/\pi_\gamma(X, Y)$ fails to have a density with respect to the Lebesgue measure (because $P\{Y^* = 0\} \neq 0$), and this in turn implies that Assumption (A) cannot hold for the response variable Y^*. To overcome this issue, we start by adding to $\Delta Y\big/\pi_\gamma(X, Y)$ a continuous random variable ε with finite support and $E(\varepsilon) = 0$; here, ε is independent of (X, Y, Δ). The use of ε (in estimation) as well as the choice of its probability distribution will be discussed in Remarks 1 and 2. The independence of ε and X implies that $E\big[\Delta Y\big/\pi_\gamma(X, Y) + \varepsilon\big|X = x\big] = m(x)$. Now, let $\varepsilon_1, \ldots, \varepsilon_n$ be independent copies of ε, independent of $(X_1, Y_1, \Delta_1), \ldots, (X_n, Y_n, \Delta_n)$, and define the revised version of (6) by

$$\widetilde{m}_{\pi,n}(x) = \sum_{i=1}^{n} \left\{ \left[\frac{\Delta_i Y_i}{\pi_\gamma(X_i, Y_i)} + \varepsilon_i \right] \mathcal{K}\left(\frac{x - X_i}{h_n} \right) \right\} \Big/ \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h_n} \right). \quad (8)$$

Since the function π_γ in (8) is completely unknown, it has to be replaced with an estimator. To this end, for each fixed γ, consider the following kernel-type estimator of $\pi_\gamma(x, y)$

$$\widetilde{\pi}_\gamma(x, y) = \left[1 + \frac{\sum_{j=1}^{n} \left[1 - (\Delta_j + \varepsilon_j) \right] \mathcal{K}((x - X_j)/\lambda_n)}{\sum_{j=1}^{n} \left[(\Delta_j + \varepsilon_j) \exp\{\gamma Y_j\} \right] \mathcal{K}((x - X_j)/\lambda_n)} \cdot \exp\{\gamma y\} \right]^{-1} \quad (9)$$

where ε_j's are as in (8) and λ_n is the smoothing parameter used here. We observe that the right side of (9) is justified, as a kernel estimator of $\pi_\gamma(x, y)$, by the fact that the term $\exp\{g(x)\}$ in (2) can alternatively be written as

$$\exp\{g(x)\} = \frac{E\big[1 - \Delta \big| X = x\big]}{E\big[\Delta \exp\{\gamma Y\} \big| X = x\big]} = \frac{E\big[1 - (\Delta + \varepsilon) \big| X = x\big]}{E\big[(\Delta + \varepsilon) \exp\{\gamma Y\} \big| X = x\big]}, \quad (10)$$

where the second equality in (10) follows because the zero-mean random variable ε is independent of (X, Y, Δ). Now, let $\widehat{\gamma}$ be any estimator of γ. This could be, for example, the estimator proposed by Kim and Yu (2011). In general, here we do not require $\widehat{\gamma}$ to be independent of the data $(X_1, Y_1, \Delta_1), \ldots, (X_n, Y_n, \Delta_n)$. In view of (9) and (8), we propose the following estimator of the true regression function $m(x)$

$$\widehat{m}_n(x) = \sum_{i=1}^{n} \left\{ \left[\frac{\Delta_i Y_i}{\widetilde{\pi}_{\widehat{\gamma}}(X_i, Y_i)} + \varepsilon_i \right] \mathcal{K}\left(\frac{x - X_i}{h_n} \right) \right\} \Big/ \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h_n} \right), \quad (11)$$

where $\widetilde{\pi}_{\widehat{\gamma}}(X_i, Y_i)$ is obtained from $\widetilde{\pi}_\gamma(X_i, Y_i)$ by substituting $\widehat{\gamma}$ for γ everywhere in (9). Next, to establish the limiting distribution of the maximal deviation of $\widehat{m}_n(x)$ from $m(x)$, let

$$Y^* = \Delta Y / \pi_\gamma(X, Y) + \varepsilon \quad \text{and} \quad Z = Y^* - E[Y^*|X] \left(\stackrel{\text{a.s.}}{=} Y^* - m(X) \right). \quad (12)$$

Now, define $\breve{G}(x, z)$ and $\breve{g}(x, z)$ to be the joint cdf and the joint pdf of the random pair (X, Z). Also, let $Q(z|x)$ and $q(z|x)$ be the conditional cdf and the conditional pdf of Z given X and consider the following revised version of Assumption (D):

Assumption (D′) The function $\breve{g}^{1/2}(x, z)$ has bounded partial derivatives with respect to both x and z. Additionally, Q^{-1} and F^{-1} (i.e., the inverse functions of Q and F) exist and both $\frac{\partial}{\partial x} Q^{-1}\big(z|F^{-1}(x)\big)$ and $\frac{\partial}{\partial z} Q^{-1}\big(z|F^{-1}(x)\big)$ are bounded. Here, F is the cdf of the univariate random variable X.

Assumption (E′) The kernel $\mathcal{K}$ satisfies Assumption (E). Also, $\mathcal{K}(x) = \mathcal{K}(-x)$.
To state the next assumption, let $\pi_\gamma(x, y) = E[\Delta|X = x, Y = y]$ be as in (2) and put $\psi(x) = E[Y^2 e^{\gamma Y}|X = x]$. Then

Assumption (F) The functions $\pi_\gamma(x, y)$ and $\psi(x)$ are twice differentiable in x, with bounded derivatives. Furthermore, $\inf_{x,y} \pi_\gamma(x, y) =: \pi_{\min} > 0$ for some $\pi_{\min} \in (0, 1]$. Throughout the paper, the response Y can be missing but X is always observable.

Assumption (G) The iid bounded random variables $\varepsilon, \varepsilon_1, \ldots, \varepsilon_n$ have mean zero and a density function that vanishes outside an interval (a_0, b_0), for some $-\infty < a_0 < b_0 < \infty$. Additionally, ε_i's are independent of the data (X_i, Y_i, Δ_i), $i = 1, \ldots, n$.

Here, the second part of assumption (F) is standard in missing data literature; see, for example, Kim and Yu (2011), Tang et al. (2014), or Shao and Wang (2016). Now, let $\widetilde{\pi}_{\widehat{\gamma}}(X_i, Y_i)$ be the quantity that appears in the definition of $\widehat{m}_n(x)$ in (11) and define

$$\widehat{v}^2_{\widetilde{\pi}}(x) = \left\{ \sum_{i=1}^{n} \left[\frac{\Delta_i Y_i}{\widetilde{\pi}_{\widehat{\gamma}}(X_i, Y_i)} + \varepsilon_i \right]^2 \mathcal{K}\left(\frac{x - X_i}{h_n} \right) \Big/ \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h_n} \right) \right\} - \left[\widehat{m}_n(x) \right]^2. \tag{13}$$

We note that, in view of (7) and Assumption (G), (13) is the kernel regression estimator of the conditional variance

$$v^2(x) = E\left[\left(\Delta Y / \pi_\gamma(X, Y) + \varepsilon \right)^2 \,\middle|\, X = x \right] - \left(E\left[\Delta Y / \pi_\gamma(X, Y) + \varepsilon \,\middle|\, X = x \right] \right)^2. \tag{14}$$

To state our main result for the estimator (11), define the quantity

$$\varphi(n) = \sqrt{2\delta \log n} + \begin{cases} \left[\log(C_1/\sqrt{\pi}) + \frac{1}{2} \log(\log n^\delta) \right] / \sqrt{2\delta \log n}, & \text{if } C_1 > 0, \\ \left[\frac{1}{2} \log(C_2/(2\pi^2)) \right] / \sqrt{2\delta \log n}, & \text{if } C_1 = 0, \end{cases} \tag{15}$$

where

$$C_1 = \frac{1}{2c_{\mathcal{K}}} \left[\mathcal{K}^2(A) + \mathcal{K}^2(-A) \right], \quad C_2 = \frac{1}{2c_{\mathcal{K}}} \int [\mathcal{K}'(t)]^2 \, dt, \quad \text{and} \quad c_{\mathcal{K}} = \int \mathcal{K}^2(t) \, dt, \tag{16}$$

with A as in assumption (E). Then, we have the following result for the proposed kernel regression estimator in the presence of nonignorable missing response variables.

Theorem 2 *Let $\widehat{m}_n(x)$ and $\widehat{v}^2_{\widetilde{\pi}}(x)$, be as in (11) and (13), respectively, and put $h_n = n^{-\delta}$ and $\lambda_n = (\log n)^{1/2} n^{-\beta}$ for any δ and β satisfying $1/5 < \beta < \delta < 1/3$. Let $\widehat{\gamma}$ be any estimator of γ in (9) satisfying*

$$\sqrt{n h_n \log n} \left| \widehat{\gamma} - \gamma \right| \to_p 0. \tag{17}$$

Then, under assumptions (A), (B), (C), (D′), (E′), (F), *and* (G), *one has*

$$P\left\{\sqrt{2\delta\log n}\left(\sqrt{\frac{nh_n}{c_{\mathcal{K}}}}\sup_{x\in[0,1]}\sqrt{\frac{f_n(x)}{\widehat{v}_{\widehat{\pi}}^2(x)}}\left|\widehat{m}_n(x)-m(x)\right|-\varphi(n)\right)\le z\right\}$$
$$\longrightarrow \exp\left(-2\,e^{-z}\right),$$

as $n\to\infty$, *where* $\varphi(n)$ *is as in* (15), $c_{\mathcal{K}}=\int\mathcal{K}^2(t)\,dt$, *and* $f_n(x)=(nh_n)^{-1}\sum_{i=1}^n\mathcal{K}((x-X_i)/h_n)$ *is the kernel estimator of* f.

Condition (17) on the rate of consistency of $\widehat{\gamma}$ is rather minimal as it is much weaker than the usual $\sqrt{n}$-consistency that many estimators enjoy in practice (this because $\sqrt{nh_n\log n}$ diverges much slower than $\sqrt{n}$, as $n\to\infty$). In fact, several $\sqrt{n}$-consistent estimators of γ have already been proposed and studied in the literature; for more on this see, for example, Kim and Yu (2011) and Shao and Wang (2016). In passing, we also note that Theorem 2 can be used to construct asymptotically correct $(1-\alpha)100\%$ uniform confidence bands for the regression function $m(x)$ in the presence of nonignorable missing response variables. These bands can be expressed as

$$\widehat{m}_n(x)\pm\left(\frac{c_{\mathcal{K}}\cdot\widehat{v}_{\widehat{\pi}}^2(x)}{nh_n\,f_n(x)}\right)^{1/2}\left(\frac{x^{(\alpha)}}{\sqrt{2\log h_n^{-1}}}+\varphi(n)\right),$$
$$\text{with } x^{(\alpha)}=-\left\{\log\log\left(\frac{1}{1-\alpha}\right)-\log 2\right\}. \tag{18}$$

Remark 1 The use of auxiliary random variables, $\varepsilon_1,\ldots,\varepsilon_n$, in our proposed approach is merely a technical device and is not necessary in practice. In fact, as our numerical results of Sect. 3 shows, in practice one can consider the more appealing choice of $\varepsilon=0$ and still expect virtually the same numerical results. The use of auxiliary or artificial random variables in estimation and inference is not new and has a long and successful history in statistical literature. Examples along these lines include the important problem of nearest-neighbor classification and pattern recognition for the cases where the p-dim covariate vectors do not have a density with respect to the Lebesgue measure. In such situations, the dimension is artificially increased to $p+1$ by including an additional random variable ε that has a pdf, and this makes it possible to establish the strong consistency of the nearest-neighbor classifier. Such auxiliary random variables are also used to perform tie-breaking in classification (see, for example, Devroye et al. 1996, pp 175–176). Perhaps, an even more important example of the use of auxiliary random variables is for the problem of weighted bootstrap approximation; see, for example, Praestgaard and Wellner (1993), Janssen and Pauls (2003), Janssen (2005), Horváth et al. (2000), Horváth (2000), Burke (1998, 2000), and Kojadinovic and Yan (2012).

Remark 2 The choice of the auxiliary random variables $\varepsilon_1,\ldots,\varepsilon_n$ in our estimation methodology is at the discretion of the practitioner. However, since the conditional

variance $v^2(x)$ in (14) may be represented as $E\big[Y^2/\pi_\gamma(X,Y)\big|X=x\big]+E(\varepsilon^2)-m^2(x)$, choosing ε to have a large variance yields an inflated estimator $\widehat{v}^2_{\widetilde{\pi}}(x)$ in (13). In other words, ε should be chosen to have a small variance. This is why we have taken ε to have a truncated $N\big(0,\ \sigma^2\big)$ distribution (truncated at $\pm 3\sigma$) with σ=0.001 in our numerical studies of Sect. 3. Fortunately, our numerical work also illustrates that even if ε is replaced by zero (which is more appealing in practice), one can still expect to obtain virtually the same numerical results. This further confirms the fact that the presence of a non-zero ε in our estimation methodology is only for theoretical considerations.

2.2.1 The estimator without noise terms

The presence of the noise terms $\varepsilon_1,\ldots,\varepsilon_n$ in the construction of our regression estimator is undesirable. However, as explained in Remark 1, and shown via simulations, the inclusion of such terms is only for technical purposes and not needed in practice. It is, nevertheless, still possible to eliminate these noise terms from the methodology if we adopt the approach of Konakov and Piterbarg (1984). This approach uses kernels that can be negative on some parts of $\mathbb{R}$, which may lead to negative density estimators. To present our alternative estimators, we start by defining the following noise-free counterparts of the estimators in (9) and (11):

$$\widetilde{\pi}_\gamma(x,y)=\left[1+\frac{\sum_{j=1}^n\big[1-\Delta_j\big]\mathcal{K}((x-X_j)/\lambda_n)}{\sum_{j=1}^n\big[\Delta_j\ \exp\{\gamma Y_j\}\big]\mathcal{K}((x-X_j)/\lambda_n)}\cdot\exp\{\gamma y\}\right]^{-1},\quad(19)$$

and

$$\widehat{m}_n(x)=\sum_{i=1}^n\left\{\left[\frac{\Delta_i Y_i}{\widetilde{\pi}_{\widehat{\gamma}}(X_i,Y_i)}\right]\mathcal{K}\left(\frac{x-X_i}{h_n}\right)\right\}\Big/\sum_{i=1}^n\mathcal{K}\left(\frac{x-X_i}{h_n}\right),\qquad(20)$$

where $\widetilde{\pi}_{\widehat{\gamma}}(X_i,Y_i)$ is obtained from $\widetilde{\pi}_\gamma(X_i,Y_i)$ by replacing γ in (19) by any estimator $\widehat{\gamma}$. Similarly, consider the following counterpart of (13)

$$\widehat{v}^2_{\widetilde{\pi}}(x)=\left\{\sum_{i=1}^n\left[\frac{\Delta_i Y_i}{\widetilde{\pi}_{\widehat{\gamma}}(X_i,Y_i)}\right]^2\mathcal{K}\left(\frac{x-X_i}{h_n}\right)\Big/\sum_{i=1}^n\mathcal{K}\left(\frac{x-X_i}{h_n}\right)\right\}-\big[\widehat{m}_n(x)\big]^2.\quad(21)$$

To study the limiting distribution of the estimator in (20), we also need the following revised versions of our earlier assumptions.

Assumption (A′) The response variable Y is bounded: $P\{B_1\le Y\ \le B_2\}=1$, where $-\infty<B_1<B_2<\infty$.

Assumption (B′) The function $\big(f(x)\,v_0^2(x)\big)^{1/2}$ is strictly positive on $[0,1]$ and satisfies a Lipschitz condition of order 1, where $v_0^2(x)=E[(Y-m(X))^2|X=x]$ and $f(x)$ is the marginal density of X.

To state the next assumption, let

$$Y^*=\Delta Y/\pi_\gamma(X,Y)\quad\text{and}\quad Z=Y^*-E[Y^*|X]\qquad(22)$$

and observe that in view of Assumption (A′) above, and Assumption (F), there are constants $-\infty < B_1' < B_2' < \infty$ such that $P(B_1' \le Z \le B_2') = 1$.

Assumption (C′) The density $f(x)$ is strictly positive on an open interval $(a, b) \ni [0, 1]$ and vanishes outside $[a, b]$. Furthermore, the distribution $\widetilde{G}(x, z)$ of (X, Z) has a density $\widetilde{g}(x, z)$ with respect to Lebesgue measure satisfying ess $\sup_{x,z} \widetilde{g}(x, z) < \infty$. Additionally, for the inverse function, T^{-1}, of the Rosenblatt (1952) transformation $T : \mathbb{R}^2 \to [0, 1]^2$, the function $p_{n,t}(T^{-1}(x', z'))$, where $(t', x') \in [0, 1]^2$, $t \in [0, 1]$, is twice continuously differentiable, where $p_{n,t}(x, z) = (\sqrt{c_{\mathcal{K}}}\, h_n)^{-1}\big(f(x)\, \nu_0^2(x)\big)^{-1/2} \cdot z\, \mathcal{K}\left((t - x)/h_n\right)$ and $c_{\mathcal{K}}$ is as in (16).

Assumption (D″) Both $f(x)$ and $m(x)$ have partial derivatives of order $\tau \ge 2$ that are bounded in (a, b), where (a, b) is as in Assumption (C′).

Assumption (E″). The kernel $\mathcal{K}$ is finite, $\int \mathcal{K}(x)dx = 1$, $\mathcal{K}'(x)$ is continuous, $\mathcal{K}(x) \to 0$ as $|x| \to \infty$, and $\mathcal{K}(-x) = \mathcal{K}(x)$. Furthermore, $\mathcal{K}$ satisfies the orthogonality condition $\int u^s \mathcal{K}(u)du = 0$, for $s = 1, \ldots, \tau$, where τ is as in Assumption (D″).

Now, put

$$\rho(n) = \sqrt{2\log(1/h_n) + \log\Big(\frac{1}{c_{\mathcal{K}}} \int (\mathcal{K}'(u))^2 du\Big) + 2\log(1/(2\pi))}, \tag{23}$$

where $c_{\mathcal{K}}$ is as in (16). Then we have the following counterpart of Theorem 2.

Theorem 3 *Let $\widehat{m}_n(x)$ and $\widehat{v}_{\widehat{\pi}}^2(x)$ be as in (20) and (21), respectively, and put $h_n = n^{-\delta}$ and $\lambda_n = (\log n)^{1/2} n^{-\beta}$ for any δ and β satisfying $1/5 < \beta < \delta < 1/3$. Let $\widehat{\gamma}$ be any estimator of γ in (9) satisfying (17). Then, under assumptions* (A′), (B′), (C′), (D″), (E″), *and* (F), *one has*

$$P\left\{\rho(n)\left(\sqrt{\frac{nh_n}{c_{\mathcal{K}}}} \sup_{x\in[0,1]} \sqrt{\frac{f_n(x)}{\widehat{v}_{\widehat{\pi}}^2(x)}} \left|\widehat{m}_n(x) - m(x)\right| - \rho(n)\right) \le z\right\} \longrightarrow \exp\left(-2\, e^{-z}\right),$$

as $n \to \infty$, where $\rho(n)$ is as in (23), and $c_{\mathcal{K}} = \int \mathcal{K}^2(t)\, dt$.

In passing, we note that the orthogonality condition $\int u^s \mathcal{K}(u)du = 0$, $s = 2$, appearing under Assumption (E″), implies that the kernel $\mathcal{K}$ can take negative values which is not desirable from a practical point of view as it might lead to negative density estimators.

3 Numerical examples

In this section we carry out some simulation studies to assess the finite-sample performance of the methods discussed in this paper. The results show that, in general, the proposed estimators perform well. We also take a close look at the performance of the complete-case estimator that is constructed based on the complete cases only. More specifically, in what follows we consider random samples (X_i, Y_i), $i = 1, \ldots, n$, of

sizes $n = 200, 400$, and 800 from the model

$$Y = 2 + \log(X+2)\,\cos(\pi\,(X+1)^{1.2}) + \exp\{-X^2\} + \nu_0(X)\cdot\mathcal{E},$$
$$\text{with}\ \ \mathcal{E} \sim \mathcal{N}\big(0,\ 0.1^2\big),$$

where $X \sim \text{Unif}(-1,\ 2)$ is independent of the random error $\mathcal{E}$, and $\nu_0^2(x) = 1 + \exp\big\{-(x+1)^2\big\}$. Here, the response variable Y_i is allowed to be missing according to the mechanism (2) with $\pi_\gamma(x,\,y) := \big[1 + \exp\big\{a_0 + a_1 x + \gamma y\big\}\big]^{-1}$. Two models are considered:

Model 1 $(a_0, a_1, \gamma) = (-1.5,\ 0.7,\ 0.2)$. This choice results in approximately 35% missing data.

Model 2 $(a_0, a_1, \gamma) = (0.3,\ -1.6,\ 0.7)$. This choice yields approximately 70% missing data.

For the estimators $\widetilde{\pi}_\gamma(x,\,y)$ and $\widehat{m}_n(x)$ (in (9) and (11)) and the bandwidth $h_n = n^{-\delta}$ we used the cross-validation approach of Racine and Li (2004) with the Epanechnikov kernel $K(u) = (0.75)(1-u^2)\cdot I\{|u| \le 1\}$, which is available in the R package "np" (Racine and Hayfield 2008). In the case of $\lambda_n = (\log n)^{1/2} n^{-\beta}$, a pilot study shows that larger values of λ_n perform better; therefore we took β to be very close to 0.2 which is in line with the requirement of Theorem 2. Regarding the choice of $\varepsilon_1, \ldots, \varepsilon_n$ used in (9) and (11), they are iid truncated $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.001$, where the truncation is at $\pm 3\sigma$. However, we also considered the more natural and appealing choice of $\varepsilon_i = 0$, $i = 1, \ldots, n$. To estimate γ in $\pi_\gamma(x, y)$, we employed the method of Kim and Yu (2011) based on a 15% follow-up sample of Y_i's; for details, see formula (23) in Sec. 4 of the cited reference. Next, for each of the three sample sizes, we computed the following quantity that appears in Theorem 2

$$W_n := \sqrt{2\delta\log n}\left(\sqrt{nh_n/c_K}\cdot\sup_{x\in[0,1]}\sqrt{\frac{\widehat{f}_n(x)}{\widehat{\nu}^2_{\widehat{p}_n}(x)}}\,\Big|\widehat{m}_n(x) - m(x)\Big| - \varphi(n)\right), \quad (24)$$

where the supremum functional in (24) was approximated by the maximum over a grid of 200 equally spaced values of x in the interval [0, 1]. Our initial pilot study shows that increasing the grid size to as large as 500 does not make any noticeable changes. Next, we observe that by Theorem 2, for large n, the quantity $W = \exp\{-2\exp(-W_n)\}$ is approximately a Uniform random variable on [0, 1]. Repeating the whole process above 400 times (each time based on a new data set of size n) yields $W_1, \ldots, W_{400}$. We also computed the above statistic based on the complete cases only, i.e., the statistic in Theorem 4, but with $m_n(x)$, $\nu_n^2(x)$, and $f_n(x)$ all computed based on the complete cases only. Repeating this process a total of 400 times, we obtain the counterparts of $W_1, \ldots, W_{400}$ for the complete cases; these will be denoted by $V_1, \ldots, V_{400}$. Figure 1 gives plots of the empirical distribution functions of $W_1, \ldots, W_{400}$ and $V_1, \ldots, V_{400}$ for each of the three sample sizes. The 45° line appearing in these plots represents the cumulative distribution function (CDF) of the Unif [0, 1] random variable. A comparison of the plots (a) and (b) clearly shows that the proposed estimator performs much

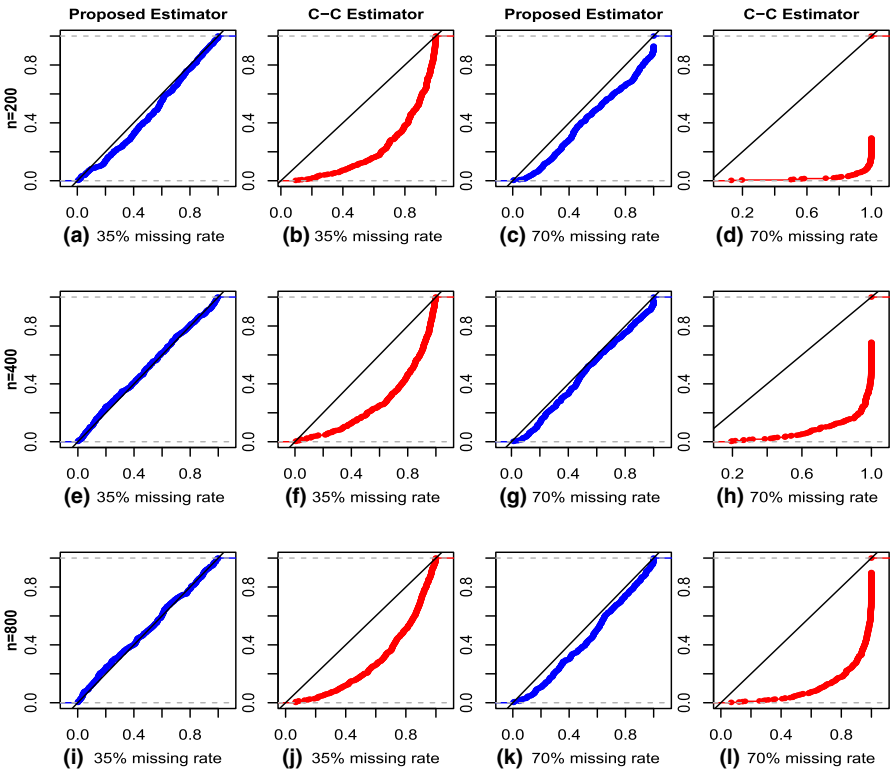

Fig. 1 Plots of empirical cdf's of $W_1, \ldots, W_{400}$ (based on the proposed estimator) and $V_1, \ldots, V_{400}$ (complete-case estimator). Here, ε_i' are iid $\mathcal{N}(0, \sigma^2)$ with σ=0.001 in (9) and (11)

better than the one based on complete cases at 35% missing rate; this follows from the fact that the empirical CDF of $W_1, \ldots, W_{400}$ is much better aligned with the 45° line. Similar conclusions hold at every sample size (200, 400, 600) and both missing rates (35% and 70%). Of course, in practice, the presence of the auxiliary random variables $\varepsilon_i, i = 1, \ldots, n$, in (9) and (11) is a nuisance. They are used only as a technical device in our work. The fact that ε_i's are chosen to have a very small variance ($\sigma^2 = 0.001^2$) suggests that, in practice, one should be able to replace them with the more realistic and appealing choice of $\varepsilon = 0$. To verify this, we also performed the same simulation study with $\varepsilon_i = 0$, $i = 1, \ldots, n$. The results, which appear in Fig. 2, are virtually indistinguishable from those in Fig. 1.

Next, we used our simulation results to construct confidence bands for the true regression function $m(x)$ over the unit interval $[0, 1]$. Three confidence levels are considered: 90%, 95%, and 99%. Table 1 presents these results for the 90% bands.

The coverage probabilities reported in this table represent the proportion of the 400 confidence bands that contain $m(x)$ for all $x \in [0, 1]$. This table also reports the average area of these confidence bands (averaged over 400 bands). There are a number of important facts to notice in Table 1, the most important of which is that the results are virtually identical regardless of whether ε_i's are $\mathcal{N}(0, \sigma^2)$, $\sigma = 0$, or the intuitively

Table 1 Coverage and the average area of confidence bands (averaged over 1000 bands) for the true regression function. Here, the nominal coverage probability is 90%

		$n = 200$		$n = 400$		$n = 600$	
	Missing =	35%	70%	35%	70%	35%	70%
Method							
As in (18) with	Coverage =	0.893	0.835	0.909	0.870	0.912	0.888
$\varepsilon_i \sim \mathcal{N}(0, 0.001^2)$	(Area) =	(2.8912)	(7.9887)	(2.3614)	(6.4364)	(2.1099)	(5.7355)
As in (18) with	Coverage =	0.893	0.835	0.909	0.870	0.912	0.888
$\varepsilon_i = 0,\ i = 1, \cdots, n$	(Area) =	(2.8911)	(7.9887)	(2.3614)	(6.4362)	(2.1099)	(5.7355)
Complete cases	Coverage =	0.549	0.043	0.670	0.188	0.703	0.327
used only	(Area) =	(0.2960)	(0.4126)	(0.2189)	(0.3765)	(0.1769)	(0.3204)

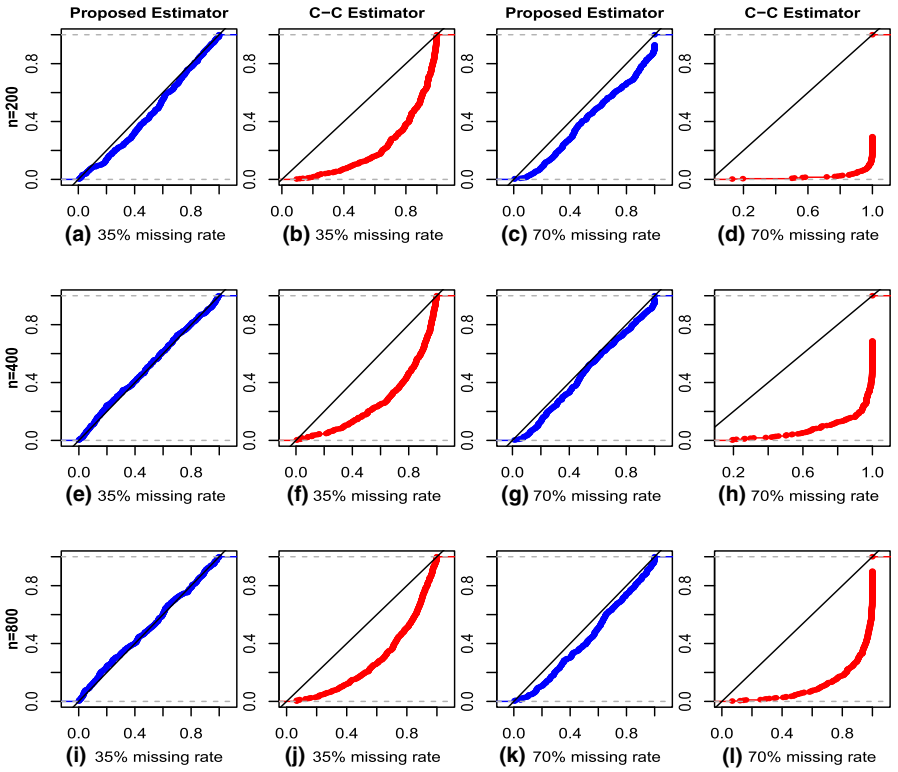

Fig. 2 Plots of empirical cdf's of $W_1, \ldots, W_{400}$ (based on the proposed estimator) and $V_1, \ldots, V_{400}$ (complete-case estimator). Here, $\varepsilon_1, \ldots, \varepsilon_n$=0 in (9) and (11)

more practical choice of ε_i=0. This is of course in line with the results in Figs. 1 and 2. The second important fact to notice in Table 1 is that the coverage of the complete-case-based bands deteriorates significantly as the missing rate changes from 35% to 70%, indicating that the proposed methods handles higher missing rates quite well. Table 1 also shows that the average areas of the confidence bands for the proposed method are higher than those of the complete-case bands. This does not imply that the proposed method's bands are unrealistically or unreasonably wider than what the theory prescribes. In fact, both Figs. 1 and 2, and the assertion of our main results in Theorem 2, confirm the fact that the proposed bands are precisely those that are backed up by the theory. It should also be mentioned that the bands based on the complete cases are actually not correct. This is because a kernel estimator of the regression function $m(x)$, based on the complete cases alone will replace $\sum_{i=1}^{n} Y_i \mathcal{K}((x - X_i)/h_n)$ by the quantity $\sum_{i=1}^{n} \Delta_i Y_i \mathcal{K}((x - X_i)/h_n)$ in the numerator of (1) in order to account for the missing Y_i's. But this mean that the new version of (1) will be the kernel regression estimator of $E[\Delta Y | X = x]$, instead of $E[Y | X = x]$, and the two will not be the same under a nonignorable missing mechanism.

Tables 2 and 3 present the corresponding results for 95% and 99% confidence bands, respectively. The results in these tables show that all conclusions are essentially the same as those in Table 1.

Acknowledgements This work is supported by the NSF Grant DMS-1916161 of Majid Mojirsheibani.

Funding The authors have not disclosed any funding.

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Conflict of interest The authors have not disclosed any competing interests.

Appendix: Proofs

To prove our main results, we first state a number of lemmas.

Lemma 1 *Let $\widetilde{\pi}_{\widehat{\gamma}}(x, y)$ be the estimator obtained from $\widetilde{\pi}_{\gamma}(x, y)$ upon replacing γ by any estimator $\widehat{\gamma}$ in (9). Then, under the conditions of Theorem 2, one has*

$$\sup_{x\in[0,1]} \max_{1\le i\le n} \left| \frac{1}{\widetilde{\pi}_{\widehat{\gamma}}(X_i, Y_i)} - \frac{1}{\widetilde{\pi}_{\gamma}(X_i, Y_i)} \right| \cdot \mathbf{I}\big\{x - Ah_n \le X_i \le x + Ah_n\big\}$$
$$= o_p\left(\frac{1}{\sqrt{nh_n \log n}}\right) \tag{25}$$

$$\sup_{x\in[0,1]} \max_{1\le i\le n} \left| \frac{1}{\widetilde{\pi}_{\gamma}(X_i, Y_i)} - \frac{1}{\pi_{\gamma}(X_i, Y_i)} \right| \cdot \mathbf{I}\big\{x - Ah_n \le X_i \le x + Ah_n\big\}$$
$$= \mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right) \tag{26}$$

Lemma 2 *Let $\widetilde{m}_{\pi,n}(x)$ and $\widehat{m}_n(x)$ be as in (8) and (11), respectively. Then,*

$$\sup_{x\in[0,1]} \left|\widehat{m}_n(x) - \widetilde{m}_{\pi,n}(x)\right| = o_p\left(\frac{1}{\sqrt{nh_n \log n}}\right) + \mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right). \tag{27}$$

To state our next lemma, we first need to define the following auxiliary quantities, which may be viewed as particular estimates of $v^2(x)$ defined in (14)

$$\widetilde{v}^2_{\pi}(x) = \sum_{i=1}^{n} \left\{ \left[\frac{\Delta_i Y_i}{\pi_{\gamma}(X_i, Y_i)} + \varepsilon_i \right]^2 \mathcal{K}\left(\frac{x - X_i}{h_n}\right) \right\} \Big/ \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h_n}\right)$$
$$- \big[\widetilde{m}_{\pi,n}(x)\big]^2 \tag{28}$$

Table 2 Coverage and the average area of confidence bands (averaged over 1000 bands) for the true regression function. Here, the nominal coverage probability is 95%

		$n = 200$		$n = 400$		$n = 600$	
	Missing =	35%	70%	35%	70%	35%	70%
Method							
As in (18) with	Coverage =	0.932	0.863	0.952	0.897	0.955	0.938
$\varepsilon_i \sim \mathcal{N}(0, 0.001^2)$	(Area) =	(3.2999)	(9.1864)	(2.6674)	(7.2706)	(2.3707)	(6.4443)
As in (18) with	Coverage =	0.932	0.863	0.952	0.897	0.955	0.938
$\varepsilon_i = 0,\ i = 1, \cdots, n$	(Area) =	(3.2999)	(9.1863)	(2.6674)	(7.2703)	(2.3707)	(6.4443)
Complete cases	Coverage =	0.686	0.071	0.798	0.286	0.839	0.367
used only	(Area) =	(0.3378)	(0.4709)	(0.2472)	(0.4253)	(0.2006)	(0.3602)

Table 3 Covrage and the average area of confidence bands (averaged over 1000 bands) for the true regression function. Here, the nominal coverage probability is 99%

		$n = 200$		$n = 400$		$n = 600$	
	Missing =	35%	70%	35%	70%	35%	70%
Method							
As in (18) with	Coverage =	0.973	0.886	0.988	0.941	0.992	0.958
$\varepsilon_i \sim \mathcal{N}(0, 0.001^2)$	(Area) =	(4.2253)	(11.7626)	(3.3605)	(9.1595)	(2.9612)	(8.0495)
As in (18) with	Coverage =	0.973	0.886	0.988	0.941	0.992	0.958
$\varepsilon_i = 0,\ i = 1, \cdots, n$	(Area) =	(4.2253)	(11.7621)	(3.3605)	(9.1593)	(2.9612)	(8.0494)
Complete cases	Coverage =	0.888	0.113	0.947	0.413	0.953	0.449
used only	(Area) =	(0.4328)	(0.6029)	(0.3115)	(0.5358)	(0.2483)	(0.4497)

$$\widetilde{\nu}^2_{\widetilde{\pi}}(x) = \sum_{i=1}^{n} \left\{ \left[\frac{\Delta_i Y_i}{\widetilde{\pi}_\gamma(X_i, Y_i)} + \varepsilon_i \right]^2 \mathcal{K}\left(\frac{x - X_i}{h_n} \right) \right\} \Big/ \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h_n} \right) - \left[\widetilde{m}_{\widetilde{\pi},n}(x) \right]^2, \tag{29}$$

where $\widetilde{\pi}_\gamma(x, y)$ is as in (9) and

$$\widetilde{m}_{\widetilde{\pi},n}(x) = \sum_{i=1}^{n} \left\{ \left[\frac{\Delta_i Y_i}{\widetilde{\pi}_\gamma(X_i, Y_i)} + \varepsilon_i \right] \mathcal{K}\left(\frac{x - X_i}{h_n} \right) \right\} \Big/ \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h_n} \right). \tag{30}$$

Lemma 3 *Let $\widehat{\nu}^2_{\widetilde{\pi}}(x)$, $\nu^2(x)$, $\widetilde{\nu}^2_{\widetilde{\pi}}(x)$, and $\widetilde{\nu}^2_{\pi}(x)$ be as in (13), (14), (29), and (28), respectively. Then*

$$\sup_{x\in[0,1]} \left| \widehat{\nu}^2_{\widetilde{\pi}}(x) - \widetilde{\nu}^2_{\widetilde{\pi}}(x) \right| = o_p\big(1/\sqrt{nh_n \log n}\,\big), \tag{31}$$

$$\sup_{x\in[0,1]} \left| \widetilde{\nu}^2_{\widetilde{\pi}}(x) - \widetilde{\nu}^2_{\pi}(x) \right| = \mathcal{O}_p\Big(\sqrt{(n\lambda_n)^{-1} \log n}\Big), \tag{32}$$

$$\sup_{x\in[0,1]} \left| \widetilde{\nu}^2_{\pi}(x) - \nu^2(x) \right| = \mathcal{O}_p\Big(\sqrt{(nh_n)^{-1} \log n}\Big). \tag{33}$$

Proof of Theorem 2 To prove Theorem 2, we first consider the following simple decomposition

$$\sup_{x\in[0,1]} \sqrt{\frac{f_n(x)}{\widehat{\nu}^2_{\widetilde{\pi}}(x)}} \left| \widehat{m}_n(x) - m(x) \right| = \sup_{x\in[0,1]} \sqrt{\frac{f_n(x)}{\widetilde{\nu}^2_{\pi}(x)}} \left| \widetilde{m}_{\pi,n}(x) - m(x) \right| + \mathcal{R}_n \tag{34}$$

where the remainder term, $\mathcal{R}_n$, is given by

$$\begin{aligned}
\mathcal{R}_n &= \sup_{x\in[0,1]} \sqrt{\frac{f_n(x)}{\widehat{\nu}^2_{\widetilde{\pi}}(x)}} \left| \widehat{m}_n(x) - m(x) \right| - \sup_{x\in[0,1]} \sqrt{\frac{f_n(x)}{\widetilde{\nu}^2_{\pi}(x)}} \left| \widetilde{m}_{\pi,n}(x) - m(x) \right| \\
&\le \sup_{x\in[0,1]} \sqrt{\frac{f_n(x)}{\widehat{\nu}^2_{\widetilde{\pi}}(x)}} \left| \widehat{m}_n(x) - \widetilde{m}_{\pi,n}(x) \right| \\
&\quad + \sup_{x\in[0,1]} \sqrt{\frac{\widetilde{\nu}^2_{\pi}(x)}{\widehat{\nu}^2_{\widetilde{\pi}}(x)}} \sqrt{\frac{f_n(x)}{\widetilde{\nu}^2_{\pi}(x)}} \left| \widetilde{m}_{\pi,n}(x) - m(x) \right| \\
&\quad - \sup_{x\in[0,1]} \sqrt{\frac{f_n(x)}{\widetilde{\nu}^2_{\pi}(x)}} \left| \widetilde{m}_{\pi,n}(x) - m(x) \right| \\
&\le \sup_{x\in[0,1]} \sqrt{\frac{f_n(x)}{\widehat{\nu}^2_{\widetilde{\pi}}(x)}} \left| \widehat{m}_n(x) - \widetilde{m}_{\pi,n}(x) \right|
\end{aligned}$$

$$+\left[\sup_{x\in[0,1]}\sqrt{\frac{\widetilde{\nu}_\pi^2(x)}{\widehat{\nu}_\pi^2(x)}}-1\right]\cdot\sup_{x\in[0,1]}\sqrt{\frac{f_n(x)}{\widetilde{\nu}_\pi^2(x)}}\left|\widetilde{m}_{\pi,n}(x)-m(x)\right|$$
$$=:\mathcal{R}_n(i)+\mathcal{R}_n(ii) \tag{35}$$

To deal with the first term on the r.h.s of (34), first observe that $\widetilde{m}_{\pi,n}(x)$ and $\widetilde{\nu}^2_{\pi,n}(x)$ that appear in this supremum term are, respectively, the kernel regression estimator of $E(Y^*|X=x)$ and the kernel estimator of the conditional variance of Y^* based on the iid "data" (X_i, Y_i^*), $i=1,\ldots,n$, where $Y^*=\Delta Y/\pi_\gamma(X,Y)+\varepsilon$; see (12). Furthermore, when assumptions (A), (F), and (G) hold, we have $P\{B_L\le Y^*\le B^U\}=1$ for finite constants B_L and B^U. In fact, one can take $B_L=\pi_{\min}^{-1}\min(0,B_1)+a_0$ and $B^U=\pi_{\min}^{-1}B_2+b_0$, where B_1 and B_2 are the constants in Assumption (A), the term $\pi_{\min}$ is as in assumption (F), and a_0 and b_0 are given in Assumption (G). Therefore, when Assumption (A) holds for the distribution of (X,Y) then, in view of assumptions (F) and (G), it also holds for the distribution of (X,Y^*) with B_1 and B_2 replaced by B_L and B^U. Additionally, it is not hard to show that, in view of Assumption (F), if $\nu_0^2(x):=E[(Y-m(X))^2|X=x]$ satisfies Assumption (C) then so does $\nu^2(x)$. Hence, in view of Theorem 1, and under assumptions (A), (B), (C), (D′), (E′), (F), and (G), the first term on the r.h.s of (34) satisfies

$$P\left\{\sqrt{2\delta\log n}\left(\sqrt{\frac{nh_n}{c_K}}\sup_{x\in[0,1]}\sqrt{\frac{f_n(x)}{\widetilde{\nu}_\pi^2(x)}}\left|\widetilde{m}_{\pi,n}(x)-m(x)\right|-\varphi(n)\right)\le u\right\}$$
$$\to\exp\left(-2e^{-u}\right) \tag{36}$$

where $c_K=\int K^2(t)\,dt$ and $\varphi(n)$ is as in (15). Now to finish the proof of Theorem 2, we have to show that $\sqrt{nh_n\log n}\,\mathcal{R}_n\to^p 0$, as $n\to\infty$. However, by (35), it is sufficient to show that $\sqrt{nh_n\log n}\,|\mathcal{R}_n(i)|\to^p 0$ and $\sqrt{nh_n\log n}\,|\mathcal{R}_n(ii)|\to^p 0$. To this end, first note that (36) yields

$$\sup_{x\in[0,1]}\sqrt{\frac{f_n(x)}{\widetilde{\nu}_\pi^2(x)}}\left|\widetilde{m}_{\pi,n}(x)-m(x)\right|=\mathcal{O}_p\left(\sqrt{\frac{\log n}{nh_n}}\right). \tag{37}$$

We also note that

$$\left|\sup_{x\in[0,1]}\sqrt{\frac{\widetilde{\nu}_\pi^2(x)}{\widehat{\nu}_\pi^2(x)}}-1\right|\le\sup_{x\in[0,1]}\left|\sqrt{\frac{\widetilde{\nu}_\pi^2(x)}{\widehat{\nu}_\pi^2(x)}}-1\right|\le\sup_{x\in[0,1]}\frac{\left|\widetilde{\nu}_\pi^2(x)-\widehat{\nu}_\pi^2(x)\right|}{\widehat{\nu}_\pi^2(x)}. \tag{38}$$

However, in view of (32) and (31),

$$\sup_{x\in[0,1]}\left|\widetilde{\nu}_\pi^2(x)-\widehat{\nu}_\pi^2(x)\right|=o_p\left(\frac{1}{\sqrt{nh_n\log n}}\right)+\mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right). \tag{39}$$

Also, observe that

$$\inf_{x\in[0,1]} \widehat{\nu}^2_{\widetilde{\pi}}(x) \geq -\sup_{x\in[0,1]} \left|\widehat{\nu}^2_{\widetilde{\pi}}(x) - \widetilde{\nu}^2_{\widetilde{\pi}}(x)\right| - \sup_{x\in[0,1]} \left|\widetilde{\nu}^2_{\widetilde{\pi}}(x) - \widetilde{\nu}^2_{\pi}(x)\right| - \sup_{x\in[0,1]} \left|\widetilde{\nu}^2_{\pi}(x) - \nu^2(x)\right| + \inf_{x\in[0,1]} \nu^2(x) \tag{40}$$

$$\inf_{x\in[0,1]} \widehat{\nu}^2_{\widetilde{\pi}}(x) \leq \sup_{x\in[0,1]} \left|\widehat{\nu}^2_{\widetilde{\pi}}(x) - \widetilde{\nu}^2_{\widetilde{\pi}}(x)\right| + \sup_{x\in[0,1]} \left|\widetilde{\nu}^2_{\widetilde{\pi}}(x) - \widetilde{\nu}^2_{\pi}(x)\right| + \sup_{x\in[0,1]} \left|\widetilde{\nu}^2_{\pi}(x) - \nu^2(x)\right| + \inf_{x\in[0,1]} \nu^2(x). \tag{41}$$

Now, taking the limit, as $n \to \infty$, of both sides of (40) and (41) and taking into account Lemma 3, we arrive at

$$0 < \lim_{n\to\infty} \inf_{x\in[0,1]} \widehat{\nu}^2_{\widetilde{\pi}}(x) < \infty. \tag{42}$$

This together with (39), (38), and (37) yields

$$|\mathcal{R}_n(ii)| \;=\; \mathcal{O}_p\left(\sqrt{\frac{\log n}{nh_n}}\right)\left[o_p\left(\frac{1}{\sqrt{nh_n \log n}}\right) + \mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right)\right],$$

from which we arrive at

$$\sqrt{nh_n \log n}\,|\mathcal{R}_n(ii)| \;=\; o_p\left(\sqrt{\frac{\log n}{nh_n}}\right) + \mathcal{O}_p\left(\frac{(\log n)^{3/2}}{\sqrt{n\lambda_n}}\right) \;=\; o_p(1).$$

To deal with the term $\mathcal{R}_n(i)$ in (35), first note that by Lemma 2

$$\begin{aligned}
&\sqrt{nh_n \log n} \sup_{x\in[0,1]} \left|\widehat{m}_n(x) - \widetilde{m}_{\pi,n}(x)\right| \\
&= \sqrt{nh_n \log n}\left[\mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right) + o_p\left(\frac{1}{\sqrt{nh_n \log n}}\right)\right] \\
&= \mathcal{O}_p\left(\sqrt{n^{\beta-\delta}\,(\log n)^{3/2}}\right) + o_p(1) = o_p(1),
\end{aligned}$$

where we have used the fact that $\beta < \delta$. Furthermore, since by (42), $\sup_{x\in[0,1]} \left|f_n(x)/\widehat{\nu}^2_{\widetilde{\pi}}(x)\right| \leq \left\{\sup_{x\in[0,1]} \left|f_n(x) - f(x)\right| + \sup_{x\in[0,1]} f(x)\right\} \big/ \inf_{x\in[0,1]} \widehat{\nu}^2_{\widetilde{\pi}}(x) \;=\; \mathcal{O}_p(1)$, one finds

$$\sqrt{nh_n \log n}\,|\mathcal{R}_n(i)| = o_p(1).$$

This completes the proof of Theorem 2. □

Proof of Theorem 3 The proof is similar to that of Theorem 2, but uses a result of Konakov and Piterbarg (1984, Theorem 1.1) instead of that of Liero (1982). □

Proof of Lemma 1 We start by defining the following quantities

$$\widehat{\phi}_1(x) = \sum_{j=1}^{n} \big[1 - (\Delta_j + \varepsilon_j)\big] \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right) \Big/ \sum_{j=1}^{n} \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right) \tag{43}$$

$$\widehat{\phi}_2(x) = \sum_{j=1}^{n} (\Delta_j + \varepsilon_j) \exp\{\widehat{\gamma} Y_j\} \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right) \Big/ \sum_{j=1}^{n} \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right) \tag{44}$$

$$\widetilde{\phi}_2(x) = \sum_{j=1}^{n} (\Delta_j + \varepsilon_j) \exp\{\gamma Y_j\} \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right) \Big/ \sum_{j=1}^{n} \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right). \tag{45}$$

$$\phi_2(x) = E\big[(\Delta + \varepsilon) \exp\{\gamma Y\} \big| X = x\big]. \tag{46}$$

$$\phi_1(x) = E\big[1 - (\Delta + \varepsilon) \big| X = x\big]. \tag{47}$$

Then it is straightforward to see

$$\begin{aligned}
&\left| \frac{1}{\widetilde{\pi}_{\widehat{\gamma}}(x, Y_i)} - \frac{1}{\widetilde{\pi}_{\gamma}(x, Y_i)} \right| \\
&= \left| \frac{-\exp\{\widehat{\gamma} Y_i\} \widehat{\phi}_1(x)}{\widehat{\phi}_2(x)} \cdot \frac{\widehat{\phi}_2(x) - \widetilde{\phi}_2(x)}{\widetilde{\phi}_2(x)} + \frac{\big[\exp\{\widehat{\gamma} Y_i\} - \exp\{\gamma Y_i\}\big] \widehat{\phi}_1(x)}{\widetilde{\phi}_2(x)} \right| \\
&\le \left| \frac{1}{\widetilde{\phi}_2(x)} \right| \left[\left| \frac{\exp\{\widehat{\gamma} Y_i\} \widehat{\phi}_1(x)}{\widehat{\phi}_2(x)} \right| \cdot \left| \widehat{\phi}_2(x) - \widetilde{\phi}_2(x) \right| \right. \\
&\quad \left. + \left| \big[\exp\{\widehat{\gamma} Y_i\} - \exp\{\gamma Y_i\}\big] \widehat{\phi}_1(x) \right| \right]
\end{aligned} \tag{48}$$

Now, put $c := \max(|B_1|, |B_2|)$, where B_1 and B_2 are as in Assumption (A), and observe that a one-term Taylor expansion gives

$$\begin{aligned}
\left| \widehat{\phi}_2(x) - \widetilde{\phi}_2(x) \right| &= \left| \frac{\sum_{j=1}^{n} (\Delta_j + \varepsilon_j) \big[\exp\{\widehat{\gamma} Y_j\} - \exp\{\gamma Y_j\}\big] \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right)}{\sum_{j=1}^{n} \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right)} \right| \\
&\le \left| \frac{c \sum_{j=1}^{n} (1 + |\varepsilon_j|) |\widehat{\gamma} - \gamma| \exp\big\{|\gamma^* - \gamma| c + \gamma c\big\} \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right)}{\sum_{j=1}^{n} \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right)} \right|, \\
&\quad (\gamma^* \text{ is a point on the interior of the line joining } \widehat{\gamma} \text{ and } \gamma) \\
&\le c \big(1 + |a_0| \vee b_0\big) |\widehat{\gamma} - \gamma| \exp\big\{|\gamma^* - \gamma| c + \gamma c\big\} \\
&\quad (\text{where } a_0 \text{ and } b_0 \text{ are as in Assumption (G)}) \\
&\le c_0 |\widehat{\gamma} - \gamma| \exp\big\{c\big[\gamma + |\widehat{\gamma} - \gamma| c\big\},
\end{aligned}$$

$$\text{where } c_0 = c\left(1 + |a_0| \vee b_0\right),$$
$$= o_p\left(\frac{1}{\sqrt{nh_n \log n}}\right) \cdot \mathcal{O}_p(1), \tag{49}$$

where the bound does not depend on x. Similarly, we note that

$$\begin{aligned}
&\left|\left[\exp\{\widehat{\gamma} Y_i\} - \exp\{\gamma Y_i\}\right] \widehat{\phi}_1(x)\right| \\
&\le \left|\exp\{\widehat{\gamma} Y_i\} - \exp\{\gamma Y_i\}\right| \frac{\sum_{j=1}^n \left|1 - (\Delta_j + \varepsilon_j)\right| \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right)}{\sum_{j=1}^n \mathcal{K}\left(\frac{x - X_j}{\lambda_n}\right)} \\
&\le \left(2 + |a_0| \vee b_0\right) c\,|\widehat{\gamma} - \gamma| \exp\left\{c\left[\gamma + |\widehat{\gamma} - \gamma|\right]\right\} \\
&= o_p\left(\frac{1}{\sqrt{nh_n \log n}}\right) \cdot \mathcal{O}_p(1),
\end{aligned} \tag{50}$$

where the bound in (50) does not depend on the particular x or Y_i. Now, observe that

$$\begin{aligned}
&\sup_{x \in [0,1]} \max_{1 \le i \le n} \left|\frac{1}{\widetilde{\pi}_{\widehat{\gamma}}(X_i, Y_i)} - \frac{1}{\widetilde{\pi}_{\gamma}(X_i, Y_i)}\right| \cdot \mathbf{I}\left\{x - Ah_n \le X_i \le x + Ah_n\right\} \\
&\le \max_{1 \le i \le n} \sup_{-h_n \le x \le 1 + Ah_n} \left\{ \left|\frac{1}{\widetilde{\phi}_2(x)}\right| \cdot \left[\left|\frac{\exp\{\widehat{\gamma} Y_i\} \widehat{\phi}_1(x)}{\widehat{\phi}_2(x)}\right| \cdot \left|\widehat{\phi}_2(x) - \widetilde{\phi}_2(x)\right| \right.\right. \\
&\quad \left.\left. + \left|\left[\exp\{\widehat{\gamma} Y_i\} - \exp\{\gamma Y_i\}\right] \widehat{\phi}_1(x)\right| \right] \right\}.
\end{aligned} \tag{51}$$

To deal with the right side of (51), first note that

$$\begin{aligned}
&\sup_{-h_n \le x \le 1 + Ah_n} \left|\frac{\exp\{\widehat{\gamma} Y_i\} \widehat{\phi}_1(x)}{\widehat{\phi}_2(x)}\right| \\
&\le \sup_{-h_n \le x \le 1 + Ah_n} \left\{ \left[\left|\left[\exp\{\widehat{\gamma} Y_i\} - \exp\{\gamma Y_i\}\right] \widehat{\phi}_1(x)\right| + \left|\left[\widehat{\phi}_1(x) - \phi_1(x)\right] \exp\{\gamma Y_i\}\right| \right.\right. \\
&\quad \left.\left. + \left|\phi_1(x) \exp\{\gamma Y_i\}\right| \right] \Big/ \left|\widehat{\phi}_2(x)\right| \right\}
\end{aligned} \tag{52}$$

Now, since the bound in (50) does not depend on any particular x or Y_i, one finds

$$\sup_{x \in [0,1]} \max_{1 \le i \le n} \left|\left[\exp\{\widehat{\gamma} Y_i\} - \exp\{\gamma Y_i\}\right] \widehat{\phi}_1(x)\right| = o_p\left(\frac{1}{\sqrt{nh_n \log n}}\right). \tag{53}$$

Next, let n be large enough so that $Ah_n < \epsilon$, where ϵ is as in assumption (B), and observe that by the results of Mack and Silverman (1982; Theorem B), one has

$$\sup_{x\in[0,1]} \max_{1\le i\le n} \left|\left[\widehat{\phi}_1(x) - \phi_1(x)\right] \exp\{\gamma Y_i\}\right| \le \mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right) \times \exp\{\gamma c\}$$
$$= \mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right), \tag{54}$$

where $c := \max(|B_1|, |B_2|)$ as before, and B_1 and B_2 are as in assumption (A). Furthermore,

$$\sup_{x\in[0,1]} \max_{1\le i\le n} |\phi_1(x) \exp\{\gamma Y_i\}| \le (1-\pi_{\min}) \exp\{\gamma c\} = \mathcal{O}(1). \tag{55}$$

We also need to deal with the infimum of the term $\left|\widehat{\phi}_2(x)\right|$ that appears in the denominator of (52). To this end, we first note that $\left|\widehat{\phi}_2(x)\right|$ can be upper- and lower-bounded as follows

$$|\phi_2(x)| - \left|\widetilde{\phi}_2(x) - \phi_2(x)\right| - \left|\widehat{\phi}_2(x) - \widetilde{\phi}_2(x)\right|$$
$$\le \left|\widehat{\phi}_2(x)\right| \le \left|\widehat{\phi}_2(x) - \widetilde{\phi}_2(x)\right| + \left|\widetilde{\phi}_2(x) - \phi_2(x)\right| + |\phi_2(x)|$$

Taking the infimum over $x \in [-h_n,\ 1 + Ah_n]$, we find $\inf_x |\phi_2(x)| - \sup_x \left|\widetilde{\phi}_2(x) - \phi_2(x)\right| - \sup_x \left|\widehat{\phi}_2(x) - \widetilde{\phi}_2(x)\right| \le \inf_x \left|\widehat{\phi}_2(x)\right| \le \sup_x \left|\widehat{\phi}_2(x) - \widetilde{\phi}_2(x)\right| + \sup_x \left|\widetilde{\phi}_2(x) - \phi_2(x)\right| + \sup_x \left|\phi_2(x)\right|$. Therefore, taking the limit as $n \to \infty$, one finds

$$0 < \varphi_0 \le \lim_{n\to\infty} \inf_{-h_n \le x \le 1+Ah_n} \left|\widehat{\phi}_2(x)\right| \le \exp\{\gamma c\}, \tag{56}$$

for a positive constant φ_0 not depending on n. Here, (56) follows from (49) in conjunction with Theorem B of Mack and Silverman (1982). Furthermore, similar (and in fact easier) arguments can also be used to show that

$$0 < \varphi_0 \le \lim_{n\to\infty} \inf_{-h_n \le x \le 1+Ah_n} \left|\widetilde{\phi}_2(x)\right| \le \exp\{\gamma c\}. \tag{57}$$

Now (25) follows from (57), (56), (55), (54), (53), (51), and (48). The proof of (26) is very similar to (and, in fact, easier than) that of (25) and therefore will not be given. □

Proof of Lemma 2 Let $\widetilde{m}_{\widetilde{\pi},n}(x)$ be as in (30), and note that

$$\left|\widetilde{m}_{\widetilde{\pi},n}(x) - \widetilde{m}_{\pi,n}(x)\right|$$
$$= \left|\frac{\sum_{j=1}^n (\Delta_j + \varepsilon_j) Y_j \left[\frac{1}{\widetilde{\pi}_\gamma(X_i,Y_i)} - \frac{1}{\pi_\gamma(X_i,Y_i)}\right] \mathcal{K}\left(\frac{x-X_j}{h_n}\right)}{\sum_{j=1}^n \mathcal{K}\left(\frac{x-X_j}{h_n}\right)}\right|$$

$$
\begin{aligned}
&\le \max_{1\le i\le n}\left\{\left|\frac{1}{\widetilde{\pi}_\gamma(X_i,Y_i)}-\frac{1}{\pi_\gamma(X_i,Y_i)}\right|\mathbf{I}\big\{x-Ah_n\le X_i\le x+Ah_n\big\}\right\}\\
&\quad\times\left[\sum_{j=1}^n\Big|(\Delta_j+\varepsilon_j)Y_j\Big|\mathcal{K}\left(\frac{x-X_j}{h_n}\right)\Big/\sum_{j=1}^n\mathcal{K}\left(\frac{x-X_j}{h_n}\right)\right]\\
&\le c_2\max_{1\le i\le n}\left\{\left|\frac{1}{\widetilde{\pi}_\gamma(X_i,Y_i)}-\frac{1}{\pi_\gamma(X_i,Y_i)}\right|\mathbf{I}\big\{x-Ah_n\le X_i\le x+Ah_n\big\}\right\},
\end{aligned}
$$

where c_2 is a positive constant not depending on n. Therefore, in view of (26),

$$
\sup_{x\in[0,1]}\left|\widetilde{m}_{\widetilde{\pi},n}(x)-\widetilde{m}_{\pi,n}(x)\right| \;=\; \mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right). \tag{58}
$$

Similarly, one has

$$
\begin{aligned}
&\left|\widehat{m}_n(x)-\widetilde{m}_{\widetilde{\pi},n}(x)\right|\\
&=\left|\frac{\sum_{j=1}^n(\Delta_j+\varepsilon_j)Y_j\left[\frac{1}{\widetilde{\pi}_{\widehat{\gamma}}(X_j,Y_j)}-\frac{1}{\widetilde{\pi}_\gamma(X_j,Y_j)}\right]\mathcal{K}\left(\frac{x-X_j}{h_n}\right)}{\sum_{j=1}^n\mathcal{K}\left(\frac{x-X_j}{h_n}\right)}\right|\\
&\le c_2\max_{1\le i\le n}\left\{\left|\frac{1}{\widetilde{\pi}_{\widehat{\gamma}}(X_i,Y_i)}-\frac{1}{\widetilde{\pi}_\gamma(X_i,Y_i)}\right|\mathbf{I}\big\{x-Ah_n\le X_i\le x+Ah_n\big\}\right\},
\end{aligned}
$$

which, together with (25), yields

$$
\sup_{x\in[0,1]}\left|\widehat{m}_n(x)-\widetilde{m}_{\widetilde{\pi},n}(x)\right| = o_p\left(\frac{1}{\sqrt{nh_n\log n}}\right). \tag{59}
$$

The proof of Lemma 2 now follows from (58) and (59) and the fact that $\left|\widehat{m}_n(x)-\widetilde{m}_{\pi,n}(x)\right|\le\left|\widehat{m}_n(x)-\widetilde{m}_{\widetilde{\pi},n}(x)\right|+\left|\widetilde{m}_{\widetilde{\pi},n}(x)-\widetilde{m}_{\pi,n}(x)\right|$. □

Proof of Lemma 3 We start with the proof of (31). First observe that

$$
\begin{aligned}
&\left|\widehat{v}^2_{\widetilde{\pi}}(x)-\widetilde{v}^2_{\widetilde{\pi}}(x)\right|\\
&\le\left|\left[\sum_{i=1}^n\Delta_i Y_i^2\left[\frac{1}{[\widetilde{\pi}_{\widehat{\gamma}}(X_i,Y_i)]^2}-\frac{1}{[\widetilde{\pi}_\gamma(X_i,Y_i)]^2}\right]\right.\right.\\
&\qquad\left.\left.\times\mathcal{K}\left(\frac{x-X_i}{h_n}\right)\right]\Big/\sum_{i=1}^n\mathcal{K}\left(\frac{x-X_i}{h_n}\right)\right|\\
&\qquad+2\left|\left[\sum_{i=1}^n\varepsilon_i\,\Delta_i Y_i\left[\frac{1}{\widetilde{\pi}_{\widehat{\gamma}}(X_i,Y_i)}-\frac{1}{\widetilde{\pi}_\gamma(X_i,Y_i)}\right]\right.\right.
\end{aligned}
$$

$$\times K\left(\frac{x-X_i}{h_n}\right)\Big]\Big/\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\Big|$$
$$+\Big|\big(\widetilde{m}_{\widetilde{\pi},n}(x)-\widehat{m}_n(x)\big)\big(\widetilde{m}_{\widetilde{\pi},n}(x)+\widehat{m}_n(x)\big)\Big|$$
$$=: \big|U_{n,1}(x)\big|+\big|U_{n,2}(x)\big|+\big|U_{n,3}(x)\big|. \tag{60}$$

However, we have

$$\begin{aligned}\big|U_{n,1}(x)\big| \le r_n(x)\cdot \max_{1\le i\le n}\Bigg[&\left|\frac{1}{\widetilde{\pi}_{\widehat{\gamma}}(X_i,Y_i)}-\frac{1}{\widetilde{\pi}_{\gamma}(X_i,Y_i)}\right|\\ &\times\Bigg\{\left|\frac{1}{\widetilde{\pi}_{\widehat{\gamma}}(X_i,Y_i)}-\frac{1}{\widetilde{\pi}_{\gamma}(X_i,Y_i)}\right|\\ &+2\left|\frac{1}{\widetilde{\pi}_{\gamma}(X_i,Y_i)}-\frac{1}{\pi_{\gamma}(X_i,Y_i)}\right|\\ &+2\left|\frac{1}{\pi_{\gamma}(X_i,Y_i)}\right|\Bigg\}\mathbf{I}\big\{x-Ah_n\le X_i\le x+Ah_n\big\}\Bigg],\end{aligned}$$

where $r_n(x)=\sum_{i=1}^{n}\Delta_i Y_i^2\,K((x-X_i)/h_n)/\sum_{i=1}^{n}K((x-X_i)/h_n)\le(|B_1|\vee|B_2|)^2$, where B_1 and B_2 are as in assumption (A). Therefore, in view of (25) and (26), we obtain

$$\begin{aligned}\sup_{x\in[0,1]}\big|U_{n,1}(x)\big| &= o_p\left(\frac{1}{\sqrt{nh_n\log n}}\right)\left\{o_p\left(\frac{1}{\sqrt{nh_n\log n}}\right)+\mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right)+\mathcal{O}_p(1)\right\}\\ &= o_p\left(\frac{1}{\sqrt{nh_n\log n}}\right).\end{aligned}$$

Similarly, we have

$$\sup_{x\in[0,1]}\big|U_{n,2}(x)\big| = o_p\big(1/\sqrt{nh_n\log n}\big).$$

Next, to deal with the term $\big|U_{n,3}(x)\big|$ in (60), we observe that $\big|U_{n,3}(x)\big|\le\big|\widetilde{m}_{\widetilde{\pi},n}(x)-\widehat{m}_n(x)\big|\times\big\{\big|\widetilde{m}_{\widetilde{\pi},n}(x)-\widehat{m}_n(x)\big|+2\big|\widetilde{m}_{\widetilde{\pi},n}(x)-\widetilde{m}_{\pi,n}(x)\big|+2\big|\widetilde{m}_{\pi,n}(x)-m(x)\big|+2|m(x)|\big\}$. Consequently, in view of (58) and (59) and the result of Mack and Silverman (1982, Theorem B), we get

$$\begin{aligned}&\sup_{x\in[0,1]}\big|U_{n,3}(x)\big|\\ &= o_p\left(\frac{1}{\sqrt{nh_n\log n}}\right)\left\{o_p\left(\frac{1}{\sqrt{nh_n\log n}}\right)+\mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right)+\mathcal{O}_p\left(\sqrt{\frac{\log n}{nh_n}}\right)+\mathcal{O}_p(1)\right\}\\ &= o_p\left(\frac{1}{\sqrt{nh_n\log n}}\right).\end{aligned}$$

Now, (31) follows from the above bounds together with (60). The proof of (32) is similar and goes as follows.

$$
\begin{aligned}
&\left|\widetilde{\nu}^2_{\widetilde{\pi}}(x) - \widetilde{\nu}^2_{\pi}(x)\right| \\
&\le \left|\left[\sum_{i=1}^{n} \Delta_i Y_i^2 \left[\frac{1}{[\widetilde{\pi}_\gamma(X_i, Y_i)]^2} - \frac{1}{[\pi_\gamma(X_i, Y_i)]^2}\right]\right.\right. \\
&\quad \left.\left. \times K\left(\frac{x - X_i}{h_n}\right)\right] \Big/ \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)\right| \\
&\quad + 2\left|\left[\sum_{i=1}^{n} \varepsilon_i \, \Delta_i Y_i \left[\frac{1}{\widetilde{\pi}_\gamma(X_i, Y_i)} - \frac{1}{\pi_\gamma(X_i, Y_i)}\right]\right.\right. \\
&\quad \left.\left. \times K\left(\frac{x - X_i}{h_n}\right)\right] \Big/ \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)\right| \\
&\quad + \left|\big(\widetilde{m}_{\widetilde{\pi},n}(x) - \widetilde{m}_{\pi,n}(x)\big)\big(\widetilde{m}_{\widetilde{\pi},n}(x) + \widetilde{m}_{\pi,n}(x)\big)\right| \\
&=: \left|T_{n,1}(x)\right| + \left|T_{n,2}(x)\right| + \left|T_{n,3}(x)\right|. \qquad (61)
\end{aligned}
$$

But

$$
\begin{aligned}
\left|T_{n,1}(x)\right| \le c_3 \max_{1\le i\le n} &\left[\left|\frac{1}{\widetilde{\pi}_\gamma(X_i, Y_i)} - \frac{1}{\pi_\gamma(X_i, Y_i)}\right| \cdot \left\{\left|\frac{1}{\widetilde{\pi}_\gamma(X_i, Y_i)} - \frac{1}{\pi_\gamma(X_i, Y_i)}\right|\right.\right. \\
&\left.\left. + 2\left|\frac{1}{\pi_\gamma(X_i, Y_i)}\right|\right\}\right] \mathbf{I}\big\{x - Ah_n \le X_i \le x + Ah_n\big\},
\end{aligned}
$$

where c_3 is a positive constant not depending on n. Therefore, by (26) and the second part of assumption (F), we have

$$
\begin{aligned}
\sup_{x\in[0,1]} \left|T_{n,1}(x)\right| &= \mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right)\mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right) + \mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right)\mathcal{O}(1) \\
&= \mathcal{O}_p\left(\sqrt{\frac{\log n}{n\lambda_n}}\right).
\end{aligned}
$$

Similarly, one has $\sup_{x\in[0,1]} \left|T_{n,2}(x)\right| = \mathcal{O}_p\left(\sqrt{\log n/(n\lambda_n)}\right)$. Furthermore, since

$$
\left|T_{n,2}(x)\right| \le \left|\widetilde{m}_{\widetilde{\pi},n}(x) - \widetilde{m}_{\pi,n}(x)\right|\Big[\left|\widetilde{m}_{\widetilde{\pi},n}(x) - \widetilde{m}_{\pi,n}(x)\right| + 2\left|\widetilde{m}_{\pi,n}(x)\right|\Big],
$$

one finds (in view of (58)) $\sup_{x\in[0,1]} \left|T_{n,2}(x)\right| = \mathcal{O}_p\left(\sqrt{\log n/(n\lambda_n)}\right)$. Now, (32) follows from (61) together with the above bounds. The proof of (33) is straightforward and, in fact, easier than those of (32) and (31), and hence will not be given. □

References

Al-Sharadqah A, Mojirsheibani M (2020) A simple approach to construct confidence bands for a regression function with incomplete data. AStA Adv Stat Anal 104:81–99

Burke M (1998) A Gaussian bootstrap approach to estimation and tests. In: Szyszkowicz EB (ed) Asymptotic methods in probability and statistics. North-Holland, Amsterdam, pp 697–706

Burke M (2000) Multivariate tests-of-fit and uniform confidence bands using a weighted bootstrap. Stat Probab Lett 46:13–20

Cai T, Low M, Zongming M (2014) Adaptive confidence bands for nonparametric regression functions. J Am Stat Assoc 109:1054–1070

Chen X, Diao G, Qin J (2020) Pseudo likelihood-based estimation and testing of missingness mechanism function in nonignorable missing data problems. Scand J Stat 47:1377–1400

Claeskens G, Van Keilegom I (2003) Bootstrap confidence bands for regression curves and their derivatives. Ann Stat 31:1852–1884

Dehuvels P, Mason D (2004) General asymptotic confidence bands based on kernel-type function estimators. Stat Inference Stoch Processes 7:225–277

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

Eubank R, Speckman P (2012) Confidence bands in nonparametric regression. J Am Stat Assoc 88:1287–1301

Fang F, Zhao J, Shao J (2018) Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values. Stat Sin 28:1677–1701

Gardes L (2020) Nonparametric confidence intervals for conditional quantiles with large-dimensional covariates. Electron J Stat 14:661–701

Gu L, Yang L (2015) Oracally efficient estimation for single-index link function with simultaneous confidence band. Electron J Stat 9:1540–1561

Gu L, Wang S, Yang L (2021) Smooth simultaneous confidence band for the error distribution function in nonparametric regression. Comput Stat Data Anal 155:107106

Härdle W (1989) Asymptotic maximal deviation of M-smoothers. J Multivar Anal 29:163–179

Härdle W, Song S (2010) Confidence bands in quantile regression. Econom Theory 26:1–22

Horváth L (2000) Approximations for hybrids of empirical and partial sums processes. J Stat Plan Inference 88:1–18

Horváth L, Kokoszka P, Steinebach J (2000) Approximations for weighted bootstrap processes with an application. Stat Probab Lett 48:59–70

Janssen A (2005) Resampling Student's t-type statistics. Ann Inst Stat Math 57:507–529

Janssen A, Pauls T (2003) How do bootstrap and permutation tests work? Ann Stat 31:768–806

Johnston G (1982) Probabilities of maximal deviations for nonparametric regression function estimates. J Multivar Anal 12:402–414

Kim JK, Yu C (2011) A semiparametric estimation of mean functionals with nonignorable missing data. J Am Stat Assoc 106:157–165

Kojadinovic I, Yan J (2012) Goodness-of-fit testing based on a weighted bootstrap: a fast large-sample alternative to the parametric bootstrap. Can J Stat 40:480–500

Konakov V, Piterbarg V (1984) On the convergence rate of maximal deviation distribution. J Multivar Anal 15:279–294

Liero H (1982) On the maximal deviation of the kernel regression function estimate. Ser Stat 13:171–182

Liu T, Yuan X (2020) Doubly robust augmented-estimating-equations estimation with nonignorable nonresponse data. Stat Pap 61:2241–2270

Liu Z, Yau CY (2021) Fitting time series models for longitudinal surveys with nonignorable missing data. J Stat Plan Inference 214:1–12

Lu X, Kuriki S (2017) Simultaneous confidence bands for contrasts between several nonlinear regression curves. J Multivar Anal 155:83–104

Lütkepohl H (2013) Reducing confidence bands for simulated impulse responses. Stat Pap 54:1131–1145

Mack Y, Silverman Z (1982) Weak and strong uniform consistency of kernel regression estimates. Z Wahrsch Verw Gebiete 61:405–415

Maity A, Pradhan V, Das U (2019) Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. Am Stat 73:340–349

Massé P, Meiniel W (2014) Adaptive confidence bands in the nonparametric fixed design regression model. J Nonparametr Stat 26:451–469

Mojirsheibani M (2021) On classification with nonignorable missing data. J Multivariate Anal 184:104775
Morikawa K, Kano Y (2018) Identification problem of transition models for repeated measurement data with nonignorable missing values. J Multivariate Anal 165:216–230
Morikawa K, Kim JK (2018) A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse. Stat Probab Lett 140:1–6
Morikawa K, Kim JK, Kano Y (2017) Semiparametric maximum likelihood estimation with data missing not at random. Can J Statist 45:393–409
Muminov M (2011) On the limit distribution of the maximum deviation of the empirical distribution density and the regression function. I. Theory Probab Appl 55:509–517
Muminov M (2012) On the limit distribution of the maximum deviation of the empirical distribution density and the regression function II. Theory Probab Appl 56:155–166
Nemouchi N, Mohdeb Z (2010) Asymptotic confidence bands for density and regression functions in the Gaussian case. J Afrika Statistika 5:279–287
Neumann M, Polzehl J (1998) Simultaneous bootstrap confidence bands in nonparametric regression. J Nonparametr Stat 9:307–333
O'Brien J, Gunawardena H, Paulo J, Chen X, Ibrahim J, Gygi S, Qaqish B (2018) The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. Ann Appl Statist 12:2075–2095
Praestgaard J, Wellner J (1993) Exchangeably weighted bootstraps of the general empirical process. Ann Probab 21:2053–2086
Proksch K (2016) On confidence bands for multivariate nonparametric regression. Ann Inst Stat Math 68:209–236
Racine J, Hayfield T (2008) Nonparametric econometrics: the np package. J Stat Softw 27:1–32
Racine J, Li Q (2004) Cross-validated local linear nonparametric regression. Stat Sin 14:485–512
Rosenblatt M (1952) Remarks on a multivariate transformation. Ann Math Stat 23:470–472
Sabbah C (2014) Uniform confidence bands for local polynomial quantile estimators. ESAIM: PS 18:265–276
Sadinle M, Reiter J (2019) Sequentially additive nonignorable missing data modelling using auxiliary marginal information. Biometrika 106:889–911
Shao J, Wang L (2016) Semiparametric inverse propensity weighting for nonignorable missing data. Biometrika 103:175–187
Song S, Ritov Y, Härdle W (2012) Bootstrap confidence bands and partial linear quantile regression. J Multivar Anal 107:244–262
Sun J, Loader C (1994) Simultaneous confidence bands for linear regression and smoothing. Ann Probab 22:1328–1345
Sun L, Zhou Y (1998) Sequential confidence bands for densities under truncated and censored data. Stat Probab Lett 40:31–41
Tang N, Zhao P, Zhu H (2014) Empirical likelihood for estimating equations with nonignorably missing data. Stat Sin 24:723–47
Uehara M, Kim JK (2018) Semiparametric response model with nonignorable nonresponse. Preprint. arXiv:1810.12519
Wandl H (1980) On kernel estimation of regression functions. Wissenschaftliche Sitzungen zur Stochastik (WSS-03), Berlin
Wang J, Cheng F, Yang L (2013) Smooth simultaneous confidence bands for cumulative distribution functions. J Nonparametr Stat 25:395–407
Withers C, Nadarajah S (2012) Maximum modulus confidence bands. Stat Pap 53:811–819
Wojdyla J, Szkutnik Z (2018) Nonparametric confidence bands in Wicksell's problem. Stat Sin 28:93–113
Xia Y (1998) Bias-corrected confidence bands in nonparametric regression. J R Stat Soc Ser B Stat Methodol 60:797–811
Yang F, Barber R (2019) Contraction and uniform convergence of isotonic regression. Electron J Stat 13:646–677
Yuan C, Hedeker D, Mermelstein R, Xie H (2020) A tractable method to account for high-dimensional nonignorable missing data in intensive longitudinal data. Stat Med 39:2589–2605
Zhao J, Shao J (2015) Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. J Am Stat Assoc 110:1577–1590
Zhao P, Wang L, Shao J (2019) Empirical likelihood and Wilks phenomenon for data with nonignorable missing values. Scand J Stat 46:1003–1024

Zhou S, Wang D, Zhu J (2020) Construction of simultaneous confidence bands for a percentile hyper-plane with predictor variables constrained in an ellipsoidal region. Stat Pap 61:1335–1346