Annals of the Institute of Statistical Mathematics (2021) 73:1249–1279
https://doi.org/10.1007/s10463-021-00784-5

Simultaneous confidence bands for nonparametric regression with missing covariate data

Li Cai[1] · Lijie Gu[2] · Qihua Wang[1] · Suojin Wang[3]

Received: 26 December 2019 / Revised: 11 December 2020 / Accepted: 16 December 2020 /
Published online: 8 June 2021

Abstract
We consider a weighted local linear estimator based on the inverse selection probability for nonparametric regression with missing covariates at random. The asymptotic distribution of the maximal deviation between the estimator and the true regression function is derived and an asymptotically accurate simultaneous confidence band is constructed. The estimator for the regression function is shown to be oracally efficient in the sense that it is uniformly indistinguishable from that when the selection probabilities are known. Finite sample performance is examined via simulation studies which support our asymptotic theory. The proposed method is demonstrated via an analysis of a data set from the Canada 2010/2011 Youth Student Survey.

Keywords Brownian motion · Maximal deviation · Simultaneous confidence band · Weighted local linear regression

1 Introduction

In nonparametric data analysis, one important problem is to detect the global shape of unknown curves or to test whether these curves follow some specific functional forms that describe the overall trend of the regression relationship. Many researchers have attempted to solve this problem by constructing nonparametric simultaneous confidence bands (SCBs) as a vital tool of global inference for unknown curves; see Johnston (1982), Zhou et al. (1998), Fan and Zhang (2000), Claeskens and Van Keilegom (2003), Zhao and Wu (2008), Cao et al. (2012), Cai et al. (2014), Cao et al. (2016), Cai et al. (2020) for the related theory and applications.

✉ Suojin Wang
sjwang@stat.tamu.edu

[1] School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China

[2] Soochow College and School of Mathematical Sciences, Soochow University, Suzhou 215006, China

[3] Department of Statistics, Texas A&M University, College Station, 77843 Texas, USA

Consider the common situation where observations $(X_i, Y_i, \varepsilon_i)_{i=1}^n$ are independent and identically distributed (i.i.d.) copies of (X, Y, ε) satisfying the following nonparametric regression model

$$Y = m(X) + \varepsilon, \tag{1}$$

where $\mathrm{E}(\varepsilon|X) = 0$, $\mathrm{var}(\varepsilon|X) = \sigma^2(X)$, and the mean function $m(\cdot)$ and the variance function $\sigma^2(\cdot)$ defined on a compact interval $[a, b]$ are unknown. In order to construct an asymptotically accurate SCB for the mean function $m(x)$, one requires to find a bound $L_{n,\alpha}$ such that $\lim_{n\to\infty} P\left(\sup_{x\in[a,b]} \left|\hat{m}(x) \quad -m(x)\right| \le L_{n,\alpha}\right) = 1-\alpha$, where $\hat{m}(x)$ is an estimator of $m(x)$ and $\alpha \in (0, 1)$ is a pre-specified error probability.

One classical approach to construct simultaneous confidence intervals is to first obtain the asymptotic distribution of $\left[\hat{m}(x)-m(x)\right]/\sqrt{\mathrm{var}\{\hat{m}(x)\}}$ which is often the standard normal distribution so that the pointwise confidence intervals for $m(x)$ are constructed. Then one can establish simultaneous confidence intervals for the values of the regression curve at the design points by Bonferroni's Inequality. One serious drawback of this approach is that the simultaneous confidence intervals are too conservative; see Eubank and Speckman (1993) for more details. Johnston (1982) and Härdle (1989) made a substantial improvement through studying the limiting distribution of the maximal deviation $\sup_{x\in[a,b]} \left|\hat{m}(x) - m(x)\right|$ for the kernel estimator and later Wang and Yang (2009) extended the results to the B spline regression. As formulated in the above works, Zheng et al. (2014) derived an SCB for the mean function of sparse functional data and Gu et al. (2014) considered an SCB for varying coefficient regression with sparse functional data, and Zheng et al. (2016) studied an SCB for generalized additive models. Furthermore, Gu and Yang (2015) proposed an SCB for the single-index link function, and Song and Yang (2009), Cai and Yang (2015) and Cai et al. (2019) studied SCBs for the variance function $\sigma^2(x)$. In addition, Härdle and Marron (1991) proposed an SCB for nonparametric regression based on the bootstrap where resampling is done from a suitably estimated residual distribution. Claeskens and Van Keilegom (2003) proposed bootstrap SCBs for $m(x)$ based on likelihood kernel regression. Chernozhukov et al. (2014) derived a Gaussian multiplier bootstrap procedure for constructing honest uniform confidence bands for a nonparametric function. Eubank and Speckman (1993), Hall and Titterington (1998), Wang (2012), Cai et al. (2014), and Cai et al. (2019) investigated SCBs for $m(x)$ in nonparametric regression with an equally spaced design.

All the above and other related works on SCBs for nonparametric regression are for fully observed data. To the best of our knowledge, there are no related works on SCBs for the data with partially missing observations which is a common situation in applications; see Little and Rubin (2019) for an introduction on missing data and many examples. When the data are not missing completely at random, using the complete case analysis by simply discarding the missing data can lead to a loss in efficiency and yield inconsistent estimates since the conditional distribution of the response given the observed covariates is in general not equal to the underlying true conditional distribution of the response given all the covariates.

A series of efforts have been made to deal with missing data. The main approaches include likelihood method, inverse selection probability weighted approach, imputation and EM algorithm. For example, Qin et al. (2009) considered likelihood approach, while Wang et al. (1997; 1998), Lipsitz et al. (1999) and Liang et al. (2004) studied an inverse selection probability weighting method. Hsu et al. (2014) proposed a nearest neighbor-based nonparametric multiple imputation approach to recover missing covariate information. Chen and Little (1999) applied the EM algorithm. See also Ibrahim et al. (2005), Kim and Shao (2013) and Little and Rubin (2019) for comprehensive overviews of statistical methods handling missing data. However, most of these existing works mainly study the consistency and asymptotic properties at any fixed point of the proposed estimator.

In this paper, we study the global inference for the mean function $m(x)$ by constructing an asymptotically accurate SCB when covariates are missing at random (MAR) meaning that the missingness mechanism depends only on variables that are fully observable. We employ a weighted estimator for $m(x)$ based on the inverse selection probability weights, which is shown to be oracally efficient in the sense that the estimator with estimated selection probabilities under a correctly specified model is uniformly as efficient as that with true selection probabilities. The asymptotic distribution of the maximal deviation of the estimator from the true mean function is provided and hence an asymptotically accurate SCB for $m(x)$ is constructed.

As an illustration, our proposed SCB is applied to the data collected from the Canada 2010/2011 Youth Student Survey to study the relationship between self-esteem and BMI. Figure 4 depicts the weighted local linear estimator and the SCB for the data. The null hypothesis of the mean function $m(x) = c_0 + c_1 x$ for some constants c_0 and c_1 is tested by our SCB with the minimum confidence level covering the null curve being 67.7%. Hence, with the p-value of 0.323 one cannot reject the null hypothesis; see Section 6 for more details.

The rest of the paper is organized as follows. Section 2 presents the main theoretical results and the detailed procedure to implement the proposed method. Finite sample simulation results and real data analyses are reported in Sections 3 and 4, respectively. Section 5 concludes the paper. Technical proofs of the main results are provided in the Appendix and the online Supplementary Material.

2 Main results

2.1 A new SCB for the mean function

When samples (X_i, Y_i) are fully observable, Fan and Gijbels (1996) proposed the local linear regression method to estimate $m(x)$ by solving

$$\operatorname{argmin}_{\beta_0,\beta_1\in\mathbb{R}} n^{-1}\sum_{i=1}^{n}\left\{Y_i-\beta_0-\beta_1\left(X_i-x\right)\right\}^2 K_h\left(X_i-x\right), \tag{2}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ is a rescaled kernel function with bandwidth h. However, when covariates are MAR, the complete case analysis in (2) by using only fully observed (X_i, Y_i) can result in a biased estimator for $m(x)$. Assume that the observed data are $(\delta_i, \delta_i X_i, Y_i)$, $i = 1, \ldots, n$, where $\delta_i = 1$ if X_i is observed and $\delta_i = 0$ otherwise, and $\pi_i = P(\delta_i = 1|Y_i, X_i) = P(\delta_i = 1|Y_i) = \pi(Y_i)$ is the selection probability by our MAR assumption. To accommodate the missingness, we apply the Horvitz and Thompson (1952)-type inverse selection weighted method by minimizing the following quantity with respect to (β_0, β_1),

$$n^{-1} \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \left\{Y_i - \beta_0 - \beta_1 (X_i - x)\right\}^2 K_h (X_i - x). \tag{3}$$

By least squares, one obtains the estimator $\hat{m}(x, \pi)$ for $m(x)$ with

$$\hat{m}(x, \pi) = e_0^T \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}, \tag{4}$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & \cdots & 1 \\ X_1 - x & \cdots & X_n - x \end{pmatrix}^T,$$

$\mathbf{W} = \frac{1}{n} \operatorname{diag}\left(\frac{\delta_1}{\pi_1} K_h (X_1 - x), \ldots, \frac{\delta_n}{\pi_n} K_h (X_n - x)\right)$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, and $e_0 = (1, 0)^T$. Here $\hat{m}(x, \pi)$ is used to emphasize its dependence on the selection probability function $\pi(y)$.

Note that the selection probability function $\pi(y)$ is generally unknown. Here we assume that $\pi(y)$ follows a parametric binary model $\pi(y, \boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ is some unknown parameter vector. For example, assuming a logistic regression model, $\pi_i = \pi(Y_i, \boldsymbol{\alpha}) = P(\delta_i = 1|Y_i) = \{1 + \exp(-\alpha_0 - \alpha_1 Y_i)\}^{-1}$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^T$. By applying the maximum likelihood approach, one easily obtains a root-n consistent estimate $\hat{\boldsymbol{\alpha}}$; see Robins et al. (1994) and Wang et al. (1998) for related studies and Hosmer and Lemeshow (2005) for a global statistic test for examining the pre-assumed binary regression model. Denote the resulting selection probability function estimator as $\hat{\pi}(y) = \hat{\pi}(y, \hat{\boldsymbol{\alpha}})$ and let $\hat{\pi}_i = \hat{\pi}\left(Y_i, \hat{\boldsymbol{\alpha}}\right)$, $i = 1, \ldots, n$. Thus, replacing π_i in (3) with $\hat{\pi}_i$, the feasible weighted estimator $\hat{m}\left(x, \hat{\pi}\right)$ of $m(x)$ is derived with

$$\hat{m}\left(x, \hat{\pi}\right) = e_0^T \left(\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X}\right)^{-1} \mathbf{X}^T \hat{\mathbf{W}} \mathbf{Y}, \tag{5}$$

where the symbols with a hat on the right side of the equation above are the same as those in equation (4) but with π_i replaced by $\hat{\pi}_i$.

For any function $\phi(x)$, we use $\phi^{(s)}(x)$ to represent its s-th order derivative, and for any integer $p \geq 0$ and use $C^{(p)}[c, d]$ to indicate the space of functions that have continuous p-th derivative on the interval $[c, d]$ with letting $C[c, d] = C^{(0)}[c, d]$. For any real positive sequences l_n and d_n, $l_n \ll d_n$ means $l_n/d_n \to 0$ as $n \to \infty$.

To construct an accurate SCB for the mean function $m(x)$, we need the following general assumptions:

(A1) *The mean function* $m(x) \in C^{(2)}[a,b]$ *and the density function* $f_X(x)$ *of* X *is positive in the open interval* (a,b) *with* $f_X(x) \in C^{(1)}[a,b]$. *Moreover, the joint density function* $f_{X,\varepsilon}(x,\varepsilon)$ *of* (X,ε) *has continuous first order partial derivative with respect to* x.

(A2) *The variance function* $\sigma^2(x)$ *is bounded on* $[a,b]$ *and* $\int \varepsilon^2 f_{X,\varepsilon|\delta=1}(x,\varepsilon)\,d\varepsilon$ *has a positive lower bound for all* $x \in [a,b]$, *where* $f_{X,\varepsilon|\delta=1}(x,\varepsilon)$ *is the joint density function of* (X,ε) *given* $\delta = 1$. *In addition, there exist constants* $\eta > 4$ *and* $M_\eta > 0$ *such that* $\mathrm{E}(|\varepsilon|^{2+\eta}\,|X) \le M_\eta$ *a.s.*

(A3) *The kernel function* $K(\cdot) \in C^{(1)}[-1,1]$ *is a symmetric probability density function.*

(A4) *The selection probability function* $\pi(y)$ *follows a parametric binary model and has a positive lower bound* c_π. *Moreover, it has bounded second order partial derivative with respect to* y *and has bounded first order partial derivative with respect to* $\boldsymbol{\alpha}$.

(A5) *The bandwidth* $h = h_n$ *satisfies* $n^{-1/3}\log n \ll h \ll n^{-1/5}\log^{-1/5} n$.

Assumptions (A1)–(A3) are elementary conditions in nonparametric kernel regression adapted from Johnston (1982), Härdle (1989), Wang and Yang (2009), and Cai et al. (2019). Assumption (A1) implies that the density function $f_X(x)$ defined in any compact subinterval of (a,b) is bounded away from zero. The condition $\eta > 4$ in Assumption (A2) can be relaxed to $\eta > 3$ but then the lower order restriction of the bandwidth is more complicated. For simplicity here we use $\eta > 4$. Assumption (A3) entails that $\mu_0(K) = 1$ and $\mu_1(K) = 0$ where $\mu_l(K) = \int_{-1}^{1} u^l K(u)\,du, l = 0, 1$. Assumption (A4) is typical in missing data analysis. The same condition appears in Wang et al. (1997) and Liang et al. (2004). Assumption (A5) is about the choice of bandwidth h. Technically, it keeps the bias at a lower rate than the variance and entails some negligible nonlinear remainder terms.

For any functions $\varphi_n(x)$ and $\phi_n(x), x \in \mathcal{D}$, we use $\varphi_n(x) = O\left(\phi_n(x)\right)$ and $\varphi_n(x) = o\left(\phi_n(x)\right)$ to mean "$\varphi_n(x)/\phi_n(x)$ is bounded and $\varphi_n(x)/\phi_n(x)$ tends to 0 as $n \to \infty$ for any fixed $x \in \mathcal{D}$", while use $\varphi_n(x) = U\left(\phi_n(x)\right)$ and $\varphi_n(x) = u\left(\phi_n(x)\right)$ to mean "$\varphi_n(x)/\phi_n(x)$ is bounded and $\varphi_n(x)/\phi_n(x)$ tends to 0 as $n \to \infty$ for all $x \in \mathcal{D}$ uniformly". We use O_p, o_p, U_p and u_p to denote the corresponding order symbols in probability.

Moreover, we let $[a_0, b_0]$ be any given closed subinterval of (a,b) so that it excludes the end points of a and b. The asymptotic uniform convergence properties of $\hat{m}\left(x, \hat{\pi}\right) - m(x)$ will be investigated on this compact subinterval to avoid the boundary effects of the kernel estimator. Since $[a_0, b_0]$ can be arbitrarily close to $[a,b]$, little is lost in exchange of technical convenience. The same strategy was employed in Härdle (1989), Gu and Yang (2015), Cai et al. (2019), etc.

Theorem 1 *Under Assumptions* (A1)–(A5), *as* $n \to \infty$, *uniformly for all* $x \in [a_0, b_0]$, *one has*

$$\hat{m}(x,\pi) - m(x) = V_n(x) + 2^{-1}h^2\mu_2(K)\,m^{(2)}(x) + u_p\left(h^2\right),$$

where $V_n(x) = n^{-1} f_X^{-1}(x) \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_h(X_i - x)\,\varepsilon_i$.

The proof of Theorem 1 is given in the Appendix. By the Central Limit Theorem, it is easy to see that the pointwise distribution of $\sqrt{nh}V_n(x)$ is approximately normally distributed with mean zero and a positive constant standard deviation. This together with Theorem 1 and $\sqrt{nh}h^2 \ll 1$ resulted from $h \ll n^{-1/5}\log^{-1/5} n$ in Assumption (A5) implies that $V_n(x)$ dominates the second and remaining terms of $\hat{m}(x,\pi) - m(x)$ uniformly.

Let $\Delta_n = \sum_{i=1}^n \delta_i$ be the number of complete cases and denote the ratio by $r_n = \Delta_n/n$. Since $\delta_1, \ldots, \delta_n$ are i.i.d., it is readily seen that

$$r_n = P(\delta_1 = 1) + O_p(n^{-1/2}). \tag{6}$$

We now give the following theorem which describes the limiting distribution of the maximal deviation between $\hat{m}(x,\pi)$ and $m(x)$. Its proof is given in the Appendix.

Theorem 2 *Under Assumptions* (A1)–(A5)*, as $n \to \infty$, for any $t \in \mathbb{R}$,*

$$P\left\{a_h\left[\sup_{x\in[a_0,b_0]}\left|\frac{(nh)^{1/2} r_n^{-1/2}\left\{\hat{m}(x,\pi) - m(x)\right\}}{d^{1/2}(x)}\right| - b_h\right] \le t\right\} \to \exp\{-2\exp(-t)\}, \tag{7}$$

where

$$a_h = \sqrt{-2\log(h/(b_0 - a_0))}, b_h = a_h + 2^{-1}a_h^{-1}\log\left(4^{-1}\pi^{-2}C(K)\right),$$

$$d(x) = \lambda(K)s(x)f_X^{-2}(x), s(x) = \int \frac{\varepsilon^2}{\pi^2(m(x)+\varepsilon)} f_{X,\varepsilon|\delta=1}(x,\varepsilon)\,d\varepsilon,$$

$$\lambda(K) = \textstyle\int_{-1}^1 K^2(u)\,du, C(K) = \lambda^{-1}(K)\int_{-1}^1 \left\{K^{(1)}(u)\right\}^2 du.$$

Note that the $\pi = 3.14\cdots$ in the definition of b_h above is a mathematical constant to be distinguished from the selection probability function $\pi(y)$. Note also that when data are fully observed, i.e., $\pi(y) \equiv 1$, $r_n \equiv 1$, $s(x)$ becomes $\sigma^2(x)f_X(x)$. In such a case, the result degenerates to that for the local linear estimator for fully observed data, which extends the result of Johnston (1982) for the Nadaraya-Watson kernel estimator to the local linear estimator under more general conditions.

The proof of Theorem 2 is quite involved. It uses the total probability formula that the probability of $\sup_{x\in[a_0,b_0]}|(nh)^{1/2}r_n^{-1/2}V_n(x)/d^{1/2}(x)|$ is the weighted average of its conditional probability given $\Delta_n = n_0$ with weights $P(\Delta_n = n_0), n_0 = 0, 1, 2, ..., n$; see the detailed argument in the Appendix. In the remainder of the theoretical development, we assume that the parametric model for π is correctly specified so that the estimator $\hat{\boldsymbol{\alpha}}$ satisfies $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = O_p(n^{-1/2})$. Theorem 3 below compares the difference between the estimator based on the true selection probability function π and that based on the estimated selection probability function $\hat{\pi}$. Its proof is given in the Appendix.

Theorem 3 *Under Assumptions* (A1)–(A5)*, as* $n \to \infty$*,*

$$\sup_{x \in [a_0, b_0]} \left| \hat{m}\left(x, \hat{\pi}\right) - \hat{m}\left(x, \pi\right) \right| = O_p\left(n^{-1/2}\right).$$

Combining Theorems 2 and 3 and Slutsky's Theorem, one obtains the following result:

Theorem 4 *Under Assumptions* (A1)–(A5)*, as* $n \to \infty$*, for any* $t \in \mathbb{R}$*,*

$$P\left\{a_h\left[\sup_{x \in [a_0, b_0]} \left| \frac{(nh)^{1/2} r_n^{-1/2} \left\{\hat{m}\left(x, \hat{\pi}\right) - m\left(x\right)\right\}}{d^{1/2}\left(x\right)} \right| - b_h\right] \leq t\right\}$$
$$\to \exp\left\{-2\exp\left(-t\right)\right\}.$$

Theorem 4 above can be used to construct a theoretical SCB for $m\left(x\right)$ which depends on unknown quantity $d\left(x\right)$. To obtain a feasible SCB, we estimate $d\left(x\right)$ by

$$\hat{d}_n\left(x\right) = \Delta_n^{-1} h \hat{f}_X^{-2}\left(x\right) \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i^2} K_h^2\left(X_i - x\right) \hat{\varepsilon}_i^2,$$

where $\hat{\varepsilon}_i = Y_i - \hat{m}\left(X_i, \hat{\pi}_i\right)$ and $\hat{f}_X\left(x\right)$ is the weighted kernel density pilot estimator of $f_X\left(x\right)$ with

$$\hat{f}_X\left(x\right) = n^{-1} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} K_{h_f}\left(X_i - x\right), \tag{8}$$

in which we recommend to use the Silverman's rule-of-thumb bandwidth (Silverman (1986), p.48) computed with complete data for h_f which has the order of $n^{-1/5}$.

Theorem 5 *Under Assumptions* (A1)–(A5)*, as* $n \to \infty$ *, one has*

$$\sup_{x \in [a_0, b_0]} \left| \hat{d}_n\left(x\right) - d\left(x\right) \right| = O_p\left(n^{-1/2} h^{-3/2} \log^{1/2} n\right).$$

Note that $n^{-1/2} h^{-3/2} \log^{1/2} n \ll \log^{-1} n$ by Assumption (A5). Therefore, we have the following corollary.

Corollary 1 *Under Assumptions* (A1)–(A5)*, for any* $\alpha \in (0, 1)$*, an asymptotic* $100\left(1 - \alpha\right)\%$ *simultaneous confidence band for* $m\left(x\right)$ *over any given* $[a_0, b_0] \subset (a, b)$ *is*

$$\hat{m}\left(x, \hat{\pi}\right) \pm (nh)^{-1/2} r_n^{1/2} \hat{d}_n^{1/2}\left(x\right) \left(b_h + a_h^{-1} q_\alpha\right), \tag{9}$$

where $q_\alpha = -\log\left\{-\frac{1}{2}\log\left(1 - \alpha\right)\right\}$ *and* a_h*,* b_h *are given in Theorem* 2*.*

2.2 Implementation

In this subsection, we describe the detailed procedure to implement the asymptotic SCB in (9). They will be used throughout Sections 3 and 4 for simulation studies and real data analysis.

The range of the covariate variable is taken as $[\hat{a}, \hat{b}]$ with $\hat{a} = \min_{\{1 \le i \le n, \delta_i = 1\}} X_i$ and $\hat{b} = \max_{\{1 \le i \le n, \delta_i = 1\}} X_i$, while the compact subinterval $[\hat{a}_0, \hat{b}_0]$ with $\hat{a}_0 = 0.9\hat{a} + 0.1\hat{b}$ and $\hat{b}_0 = 0.9\hat{b} + 0.1\hat{a}$ is regarded as the interval over which the SCBs are constructed. The quartic kernel, $K(u) = 15\left(1 - u^2\right)^2 I(|u| \le 1)/16$, is used for the weighted local linear estimator in (5) and the weighted kernel density estimator in (8), satisfying Assumption (A3).

Regarding the bandwidth selection for $\hat{m}$ in (4), we adopt $h = h_{\text{rot}} \log^{-\rho} n$ for some $\rho > 1/5$, where h_{rot} is the rule-of-thumb bandwidth in Fan and Gijbels (1996, Equation (4.3)) computed with the complete data. Note that the order of h_{rot} is $n^{-1/5}$ and hence the order of h is $n^{-1/5} \log^{-\rho} n$ which satisfies Assumption (A5). We have found in extensive simulations that $h = h_{\text{rot}} \log^{-1/4} n$ (i.e., $\rho = 1/4$) works quite well and that is what we recommend.

3 Simulation studies

In this section, we investigate the finite sample behaviors of the proposed SCB and the finite sample effect due to estimating the selection probabilities. For comparison, we also list the results of the complete case SCB for local linear regression by directly ignoring the missing covariates, denoted by SCB-CC.

The following four cases were examined:

$$\begin{aligned}
&\text{Case 1: } m(X) = \sin(\pi X), \sigma(X) = 1;\\
&\text{Case 2: } m(X) = \sin(\pi X), \sigma(X) = 2\exp(X)\{\exp(X) + 1\}^{-1};\\
&\text{Case 3: } m(X) = \exp(-6X^3/5), \sigma(X) = 1;\\
&\text{Case 4: } m(X) = \exp(-6X^3/5), \sigma(X) = 2\exp(X)\{\exp(X) + 1\}^{-1},
\end{aligned}$$

where $X \sim U[-1, 1]$, and the error $\varepsilon \sim N\left(0, \sigma^2(x)\right)$. Clearly, these scenarios include both homoscedastic errors (Case 1, Case 3) and heteroscedastic errors (Case 2, Case 4). Two models for the selection probability function were considered: (i) logistic model $\pi(Y) = P(\delta = 1|Y) = \{1 + \exp(-\alpha_0 - \alpha_1 Y)\}^{-1}$, and (ii) probit model $\pi(Y) = P(\delta = 1|Y) = \Phi\left(\alpha_0^* + \alpha_1^* Y\right)$, where Φ is the standard normal cumulative distribution function. We took (α_0, α_1) and $\left(\alpha_0^*, \alpha_1^*\right)$ as (1.8, 1) and (1, 0.5), respectively, leading to approximately 8% to 20% of the data missing (low proportion of missing). We also took (α_0, α_1) and $\left(\alpha_0^*, \alpha_1^*\right)$ as (0.2, 0.6) and (0.1, 0.3), respectively, leading to approximately 31% to 46% of the data missing (high proportion of missing). The sample sizes were $n = 200, 400, 600, 800$ and the confidence levels were $1 - \alpha = 0.95, 0.99$.

Table 1 Empirical coverage frequencies of the SCB in (9) and the SCB in the complete case (SCB-CC) with 1000 replications and their corresponding average widths (inside parentheses) under the selection probability model (i) with parameters $(\alpha_0, \alpha_1) = (1.8, 1)$

n	$1-\alpha$	Case 1		Case 2	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.911(1.448)	0.617(1.212)	0.911(1.412)	0.708(1.216)
	0.99	0.994(1.888)	0.904(1.581)	0.991(1.840)	0.936(1.584)
400	0.95	0.938(1.102)	0.422(0.910)	0.953(1.070)	0.573(0.910)
	0.99	0.993(1.422)	0.832(1.175)	0.994(1.380)	0.901(1.174)
600	0.95	0.954(0.934)	0.284(0.774)	0.955(0.908)	0.443(0.771)
	0.99	0.999(1.197)	0.705(0.992)	0.994(1.164)	0.850(0.988)
800	0.95	0.949(0.833)	0.180(0.690)	0.949(0.811)	0.349(0.689)
	0.99	0.998(1.063)	0.596(0.881)	0.996(1.035)	0.779(0.879)
n	$1-\alpha$	Case 3		Case 4	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.910(1.247)	0.851(1.150)	0.923(1.304)	0.832(1.172)
	0.99	0.991(1.622)	0.979(1.497)	0.992(1.692)	0.970(1.520)
400	0.95	0.938(0.945)	0.816(0.870)	0.942(0.985)	0.789(0.882)
	0.99	0.991(1.216)	0.976(1.120)	0.995(1.265)	0.970(1.133)
600	0.95	0.932(0.806)	0.800(0.744)	0.940(0.844)	0.757(0.756)
	0.99	0.994(1.031)	0.982(0.951)	0.997(1.077)	0.963(0.965)
800	0.95	0.934(0.720)	0.768(0.666)	0.937(0.760)	0.699(0.680)
	0.99	0.995(0.916)	0.962(0.847)	0.997(0.964)	0.940(0.863)

We first look at the performance of the proposed SCB in the cases where the selection probability models are correctly specified. Tables 1–4 give the coverage frequencies with 1000 replications that the true mean function was covered by the SCB in (9) and the SCB-CC at the equally spaced points $\hat{a}_0 + (\hat{b}_0 - \hat{a}_0)k/400$, $k = 0, \ldots, 400$. One can see that in all scenarios, the coverage frequencies of the proposed SCB in (9) are close to the nominal confidence levels 0.95 and 0.99 while the coverage frequencies of SCB-CC are far lower than the nominal levels, and the average widths of SCB-CC are systematically narrower than that of the proposed SCB. Meanwhile, the average widths of the SCBs decrease as the sample size n increases, as expected. All in all, it can be seen that the proposed SCB in (9) performs much better than the SCB-CC. This is because the local linear estimation in the complete case is generally biased for the underlying true function. These findings support our theoretical results.

We next investigate the sensitivity of the SCB to the selection probability model misspecification. Firstly, similar to Wang et al. (1997) we carried out a simulation study which has the same setting as that in Table 2 except that the selection probability is truncated above by 0.75. As a result, about 46% of the cases had missing covariates. Using the logistic regression model to fit $\pi(y)$ is not completely correct in this setting. Table 5 summarizes the simulation results under this misspecification. One can see that in all the scenarios, the coverage frequencies are quite close to those under the

Table 2 Empirical coverage frequencies of the SCB in (9) and the SCB in the complete case (SCB-CC) with 1000 replications and their corresponding average widths (inside parentheses) under the selection probability model (i) with parameters $(\alpha_0, \alpha_1) = (0.2, 0.6)$

n	$1-\alpha$	Case 1		Case 2	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.888(1.726)	0.568(1.501)	0.887(1.697)	0.564(0.496)
	0.99	0.977(2.260)	0.872(1.964)	0.976(2.220)	0.892(1.956)
400	0.95	0.916(1.303)	0.365(1.118)	0.922(1.267)	0.402(1.103)
	0.99	0.994(1.687)	0.774(1.447)	0.992(1.641)	0.816(1.428)
600	0.95	0.938(1.105)	0.219(0.945)	0.953(1.076)	0.247(0.932)
	0.99	0.993(1.422)	0.696(1.215)	0.997(1.385)	0.741(1.199)
800	0.95	0.942(0.979)	0.125(0.837)	0.950(0.957)	0.138(0.829)
	0.99	0.993(1.256)	0.562(1.074)	0.998(1.227)	0.607(1.062)
n	$1-\alpha$	Case 3		Case 4	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.907(1.453)	0.650(1.308)	0.911(1.536)	0.606(1.344)
	0.99	0.993(1.900)	0.924(1.710)	0.986(1.998)	0.921(1.749)
400	0.95	0.918(1.103)	0.525(0.988)	0.935(1.160)	0.495(1.016)
	0.99	0.991(1.425)	0.873(1.277)	0.992(1.492)	0.857(1.308)
600	0.95	0.935(0.948)	0.417(0.846)	0.948(0.999)	0.351(0.870)
	0.99	0.996(1.215)	0.836(1.085)	0.997(1.277)	0.826(1.112)
800	0.95	0.939(0.845)	0.318(0.758)	0.936(0.889)	0.248(0.777)
	0.99	0.992(1.078)	0.766(0.967)	1.000(1.131)	0.715(0.989)

correct specification of $\pi(y)$. Secondly, we also conducted simulations when the data were generated in the four cases above with the selection probability $\pi(Y) = \{1 + \exp(-1.8 - Y - 0.2Y^2)\}^{-1}$ having a quadratic term. Thus, using the logistic regression model to fit the selection probability is still not correct. Table 6 describes the simulation results under this misspecification. Likewise, one sees that the behaviors of the SCBs are similar to those under the correct specification of the selection probability. All this above suggests that the proposed SCB is not very sensitive to misspecification of the selection probability function.

To visualize the SCB for the mean function, Figures 1 and 2 were created based on two samples of size 400 and 800 for Case 1 and Case 4 under the logit missing mechanism with $(\alpha_0, \alpha_1) = (0.2, 0.6)$. One can see that the SCB for $n = 800$ is narrower and fits the true mean function better than those for $n = 400$, which corroborates our asymptotically theoretical results. For comparisons, the complete case estimates and the SCBs by ignoring the cases with missing covariates were also provided in Figures 1 and 2. These figures show that the SCBs do not contain the true curve completely and suggest that the estimated curves may be biased. Other settings yield similar results and hence they are omitted.

Table 3 Empirical coverage frequencies of the SCB in (9) and the SCB in the complete case (SCB-CC) with 1000 replications and their corresponding average widths (inside parentheses) under the selection probability model (ii) with parameters $(\alpha_0^*, \alpha_1^*) = (1, 0.5)$

n	$1-\alpha$	Case 1		Case 2	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.912(1.395)	0.671(1.226)	0.919(1.389)	0.759(1.238)
	0.99	0.996(1.820)	0.929(1.598)	0.989(1.808)	0.950(1.612)
400	0.95	0.944(1.062)	0.527(0.926)	0.946(1.043)	0.643(0.923)
	0.99	0.994(1.370)	0.888(1.194)	0.996(1.344)	0.940(1.191)
600	0.95	0.950(0.902)	0.409(0.787)	0.946(0.887)	0.540(0.784)
	0.99	0.996(1.156)	0.819(1.009)	1.000(1.136)	0.900(1.005)
800	0.95	0.952(0.808)	0.330(0.705)	0.948(0.791)	0.458(0.702)
	0.99	0.996(1.030)	0.745(0.899)	0.998(1.009)	0.856(0.894)
n	$1-\alpha$	Case 3		Case 4	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.914(1.246)	0.839(1.158)	0.918(1.298)	0.835(1.184)
	0.99	0.993(1.622)	0.979(1.507)	0.991(1.685)	0.968(1.536)
400	0.95	0.931(0.939)	0.815(0.872)	0.941(0.974)	0.789(0.887)
	0.99	0.992(1.209)	0.976(1.123)	0.995(1.252)	0.973(1.140)
600	0.95	0.928(0.803)	0.786(0.748)	0.943(0.836)	0.760(0.761)
	0.99	0.996(1.027)	0.979(0.957)	0.998(1.066)	0.968(0.972)
800	0.95	0.938(0.718)	0.752(0.670)	0.934(0.746)	0.702(0.681)
	0.99	0.997(0.914)	0.961(0.853)	0.997(0.948)	0.945(0.865)

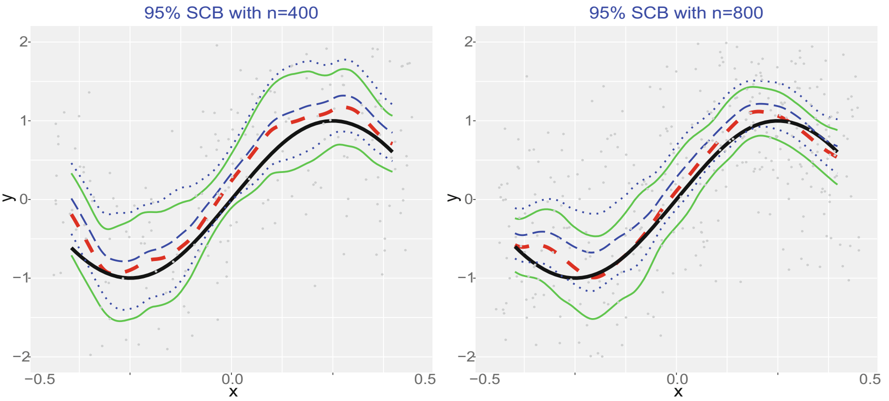

Fig. 1 Plots of the true mean function $m(x)$ (thick solid), the weighted local linear estimate $\hat{m}(x, \hat{\pi})$ (thick dashed) and the 95% SCB (solid line) for Case 1 under the selection probability model (i) with parameters $(\alpha_0, \alpha_1) = (0.2, 0.6)$ (about 45% missing). The complete case estimate (dashed) and the SCB (dotted) by ignoring the cases with missing covariates are also shown

Table 4 Empirical coverage frequencies of the SCB in (9) and the SCB in the complete case (SCB-CC) with 1000 replications and their corresponding average widths (inside parentheses) under the selection probability model (ii) with parameters $(\alpha_0^*, \alpha_1^*) = (0.1, 0.3)$

n	$1-\alpha$	Case 1		Case 2	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.900(1.682)	0.660(1.525)	0.882(1.666)	0.657(1.526)
	0.99	0.978(2.200)	0.907(1.995)	0.975(2.177)	0.911(1.993)
400	0.95	0.931(1.265)	0.514(1.137)	0.923(1.226)	0.548(1.114)
	0.99	0.995(1.636)	0.872(1.471)	0.998(1.588)	0.896(1.444)
600	0.95	0.944(1.067)	0.383(0.958)	0.946(1.047)	0.422(0.948)
	0.99	0.994(1.373)	0.821(1.232)	0.996(1.347)	0.845(1.219)
800	0.95	0.940(0.941)	0.296(0.845)	0.944(0.931)	0.296(0.843)
	0.99	0.997(1.207)	0.752(1.084)	0.999(1.193)	0.776(1.080)
n	$1-\alpha$	Case 3		Case 4	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.901(1.449)	0.689(1.342)	0.926(1.515)	0.673(1.374)
	0.99	0.992(1.896)	0.942(1.755)	0.989(1.976)	0.940(1.791)
400	0.95	0.922(1.103)	0.582(1.019)	0.933(1.150)	0.595(1.045)
	0.99	0.993(1.425)	0.899(1.316)	0.995(1.481)	0.905(1.346)
600	0.95	0.944(0.947)	0.511(0.873)	0.946(0.987)	0.473(0.894)
	0.99	0.997(1.214)	0.881(1.119)	0.997(1.262)	0.881(1.143)
800	0.95	0.944(0.838)	0.405(0.776)	0.943(0.876)	0.357(0.795)
	0.99	0.993(1.071)	0.846(0.991)	0.999(1.116)	0.805(1.013)

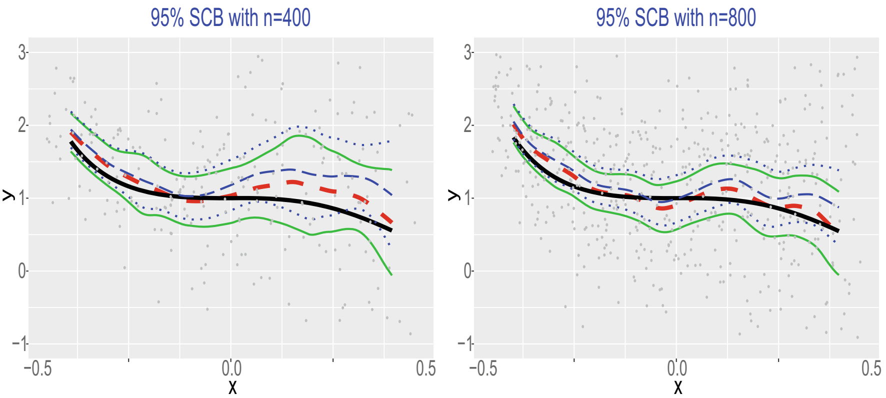

Fig. 2 Plots of the true mean function $m(x)$ (thick solid), the weighted local linear estimate $\hat{m}(x, \hat{\pi})$ (thick dashed) and the 95% SCB (solid) for Case 4 under the selection probability model (i) with parameters $(\alpha_0, \alpha_1) = (0.2, 0.6)$ (about 31% missing). The complete case estimate (dashed) and the SCB (dotted) by ignoring the cases with missing covariates are also shown

Table 5 Empirical coverage frequencies of the proposed SCB in (9) and the SCB in the complete case (SCB-CC) with 1000 replications and their corresponding average widths (inside parentheses) under using logistic regression to fit the underlying truncated logistic selection probability with parameters $(\alpha_0, \alpha_1) = (0.2, 0.6)$

n	$1-\alpha$	Case 1		Case 2	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.879(1.680)	0.582(1.490)	0.874(1.637)	0.622(1.487)
	0.99	0.978(2.200)	0.879(1.951)	0.970(2.142)	0.908(1.944)
400	0.95	0.901(1.262)	0.395(1.107)	0.909(1.222)	0.473(1.102)
	0.99	0.991(1.634)	0.782(1.434)	0.992(1.582)	0.859(1.426)
600	0.95	0.924(1.073)	0.237(0.939)	0.924(1.030)	0.329(0.925)
	0.99	0.993(1.381)	0.713(1.208)	0.994(1.327)	0.803(1.191)
800	0.95	0.923(0.948)	0.160(0.831)	0.933(0.917)	0.220(0.825)
	0.99	0.995(1.216)	0.599(1.066)	0.996(1.175)	0.707(1.057)
n	$1-\alpha$	Case 3		Case 4	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.907(1.369)	0.743(1.301)	0.917(1.434)	0.716(1.334)
	0.99	0.984(1.790)	0.950(1.702)	0.986(1.867)	0.951(1.736)
400	0.95	0.921(1.037)	0.654(0.986)	0.927(1.080)	0.644(1.010)
	0.99	0.994(1.340)	0.935(1.274)	0.993(1.391)	0.920(1.300)
600	0.95	0.937(0.885)	0.591(0.841)	0.944(0.926)	0.556(0.864)
	0.99	0.995(1.136)	0.930(1.079)	0.998(1.184)	0.904(1.105)
800	0.95	0.932(0.786)	0.512(0.748)	0.942(0.823)	0.435(0.770)
	0.99	0.994(1.004)	0.884(0.956)	0.997(1.048)	0.858(0.980)

Following a reviewer's suggestion, we also assessed the empirical performance of the statistical significance and power of the SCB test in the following setting:

$$\begin{cases} Y = m(X) + \sigma(X)\varepsilon, \sigma\ (X) = 2\exp(X)\left\{\exp(X)+1\right\}^{-1}, \\ \varepsilon \sim N(0,1), X \sim U\left[-1,1\right], \pi\ (Y) = \left\{1+\exp(-1.8-Y)\right\}^{-1}, \\ H_0 : m(X) = c_0 + c_1 X, \\ H_1 : m(X) = c_0 + c_1 X + \gamma\sin(\pi X), \end{cases} \tag{10}$$

where $c_0 = 1, c_1 = 6$ and $\gamma = 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5$. The nominal significance level of $\alpha = 0.05$ was used. Figure 3 shows the empirical power of the SCB test with 1000 replications. Note that when $\gamma = 0$, it is under the null hypothesis H_0 and the power degenerates to the type I error rate. One sees that the type I error rate is quite close to the significance level 0.05. When $\gamma \neq 0$, it is under the alternative hypothesis H_1. It is seen that the power of the SCB test increases as n and/or γ increase. All this supports our theoretical findings.

Table 6 Empirical coverage frequencies of the proposed SCB in (9) and the SCB in the complete case (SCB-CC) with 1000 replications and their corresponding average widths (inside parentheses) under using logistic regression to fit the underlying $\pi(Y) = \{1 + \exp(-1.8 - Y - 0.2Y^2)\}^{-1}$

n	$1-\alpha$	Case 1		Case 2	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.941(1.400)	0.840(1.248)	0.936(1.397)	0.838(1.251)
	0.99	0.995(1.822)	0.972(1.624)	0.992(1.817)	0.971(1.627)
400	0.95	0.955(1.057)	0.805(0.941)	0.951(1.052)	0.818(0.936)
	0.99	0.994(1.360)	0.972(1.211)	0.997(1.355)	0.978(1.205)
600	0.95	0.950(0.902)	0.728(0.801)	0.959(0.892)	0.738(0.793)
	0.99	0.996(1.154)	0.959(1.025)	0.998(1.141)	0.969(1.015)
800	0.95	0.957(0.802)	0.696(0.713)	0.964(0.797)	0.700(0.710)
	0.99	1.000(1.021)	0.943(0.908)	1.000(1.015)	0.952(0.904)
n	$1-\alpha$	Case 3		Case 4	
		SCB	SCB-CC	SCB	SCB-CC
200	0.95	0.913(1.263)	0.853(1.151)	0.916(1.324)	0.851(1.182)
	0.99	0.988(1.643)	0.982(1.498)	0.993(1.718)	0.981(1.534)
400	0.95	0.934(0.958)	0.835(0.870)	0.929(1.016)	0.834(0.888)
	0.99	0.991(1.233)	0.980(1.120)	0.995(1.305)	0.982(1.141)
600	0.95	0.932(0.819)	0.813(0.744)	0.938(0.872)	0.819(0.763)
	0.99	0.998(1.048)	0.983(0.951)	0.998(1.113)	0.979(0.973)
800	0.95	0.934(0.732)	0.798(0.668)	0.937(0.777)	0.785(0.683)
	0.99	0.997(0.932)	0.971(0.850)	0.998(0.987)	0.974(0.868)

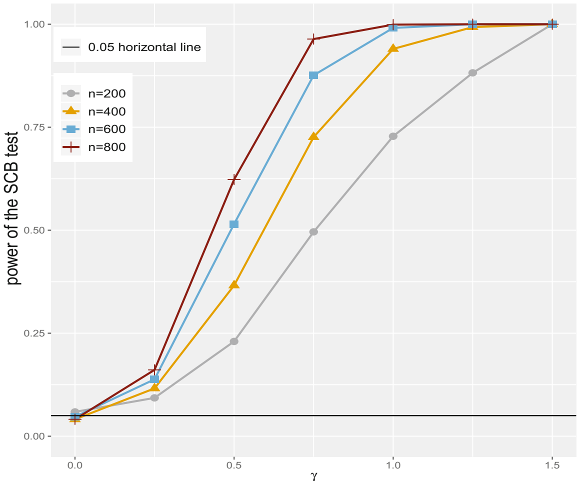

Fig. 3 Plot of the empirical power function of the SCB test in model (10) with 1000 replications. The nominal significance level is $\alpha = 0.05$

4 Real data analysis

In this section, we illustrate an application to the data from the Canada 2010/2011 Youth Student Survey. The 2010/2011 Youth Student Survey sponsored by Health Canada is a pan-Canadian, classroom-based survey on a representative youth students in grades 6–12 between October 2010 and June 2011. It aims to provide Health Canada, provinces, schools, communities, and parents with timely and reliable data on tobacco, alcohol and drug use in addition to other related issues about Canadian students; see more details in *2010-2011 YSS Student Survey Data Codebook* or from https://uwaterloo.ca/canadian-student-tobacco-alcohol-drugs-survey.

We focused on a subset of the data collected from white female youth students in grades 6–12 to study the relationship between self-esteem and Body Mass Index (BMI); see the interesting related discussions and further references in Habib et al. (2015) and ALAhmari et al. (2017). In this data set, the self-esteem was measured by using a score ranging from 0 to 12, and the BMI was computed by the weight over height in meter squared, ranging from 10.04 to 49.78. There were a total of 5343 students with having complete observations on self-esteem, while only 3565 students provided BMI (33.2% missing rate).

For the data missingness mechanism, we used the logistic regression to estimate $\pi(y)$. The fitted estimates are $\hat{\boldsymbol{\alpha}} = (0.82585, -0.015)^T$. To further judge how well the model fits the missingness pattern in the data, the Hosmer-Lemeshow goodness of fit test (Hosmer and Lemeshow (2005)) was employed with the p-value $= 0.17$. Thus one cannot reject the null hypothesis that the logistic model is correct. Figure 4 shows the inverse selection probability weighted local linear estimate $\hat{m}(x, \hat{\pi})$ (thick solid line) and the 95% and 65.6% SCBs (solid lines). The SCB was applied to test the null hypothesis H_0: $m(x) = c_0 + c_1 x$ where the coefficients $(c_0, c_1)^T$ were computed by the inverse selection probability weighted least square method; see the null curve (dashed line) in Figure 4. One can see that the null curve is completely covered by the 95% SCB. Thus the null hypothesis of the mean function being a linear function cannot be rejected at the significant level $= 0.05$. Applying Theorem 4, we obtained that the minimum confidence level containing the null curve is 67.7%; see the right panel of Figure 4. Therefore, the null hypothesis cannot be rejected with p-value $= 0.323$.

Moreover, one can also see that the mean curve has a general decreasing trend, i.e., there is a negative association between self-esteem and BMI among white female youth students in grades 6–12. This result agrees with that discovered by Habib et al. (2015). Meanwhile, according to Habib et al. (2015), a BMI between 20 and 25 is considered normal, a BMI between 25 and 30 is considered overweight, and a BMI $>$ 30 is considered obese. Therefore, even if female students were within normal weight range, their self-esteem was still decreased as BMI increased. This may be because female students in general are more likely to see themselves as obese or overweight and show dissatisfaction with their body image even if they have a healthy weight.

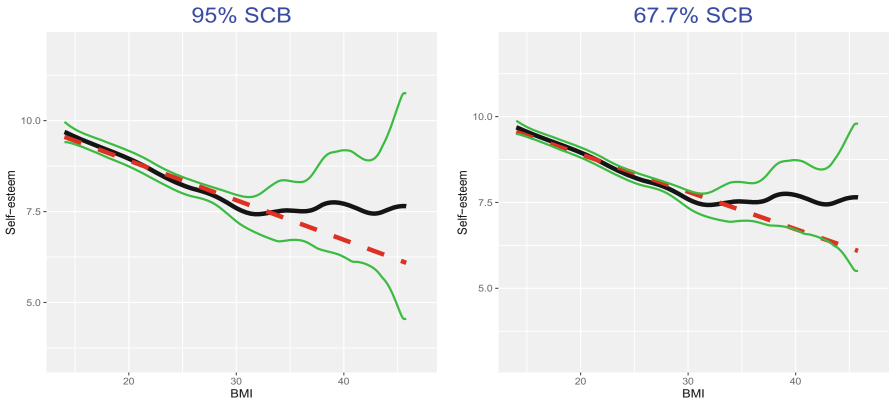

Fig. 4 Plots of the weighted local linear estimate $\hat{m}(x, \hat{\pi})$ (thick solid), the 95% and 67.7% SCBs (solid), and the null hypothesis weighted linear regression curve (dashed) for the youth student survey data collected from white female students

5 Concluding remarks

In this paper, asymptotically accurate SCBs were constructed for the nonparametric mean function with covariates missing at random by employing the weighted estimator based on inverse selection probabilities. The limiting distribution of the global estimation error (also known as maximal deviation) was derived, overcoming the main technical challenge on formulating such a confidence band. The proposed estimator for the mean function was shown to be oracally efficient in the sense that using root-n consistent selection probability estimates is as efficient as that when the selection probabilities were known as a prior. Simulation studies support our theoretical findings and the analysis of the Canada 2010/2011 Youth Student Survey data illustrates the versatility of the SCB. The methodology should also be suitable to partial linear models for missing covariate data (Wang, 2009). Further investigations may lead to similar constructions of SCBs for generalized nonparametric models, partial linear single-index models, varying coefficient models, and functional data with missing covariate data.

The traditional approach of using the asymptotic quantiles of the Gumbel extreme value distribution for the construction of the bands leads to a decay of logarithmic order in their coverage error. A bootstrap approximation can provide a substantial improvement; see, e.g., Hall (1991). Furthermore, one could potentially use recent results on anti-concentration of Gaussian processes (Chernozhukov et al., 2014), together with the multiplier bootstrap, in order to construct confidence bands whose coverage error decays polynomially fast. It would be desirable to explore how these ideas could be implemented successfully in our current setting of missing covariate data. All these are interesting problems for future research.

6 Supplementary information

The online Supplementary Material contains the proofs of the lemmas given in Appendices A.1 and A.2.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-021-00784-5.

Acknowledgements We would like to thank the Associate Editor and two referees for their helpful comments and suggestions that substantially improved an earlier version of this manuscript. This research was supported in part by the National Natural Science Foundation of China Award NSFC #11901521, #11701403, First Class Discipline of Zhejiang–A (Zhejiang Gongshang University–Statistics), Zhejiang Province Statistical Research Program #20TJQN04, the 2017 Jiangsu Overseas Visiting Scholar Program for University Prominent Young & Middle-aged Teachers and Presidents, Jiangsu Province Key-Discipline Program (Statistics) GD10700118, the National Natural Science Foundation of China (General Program 11871460, Key Program 11331011 and Program for Creative Research Group in China 61621003), a grant from the Key Lab of Random Complex Structure and Data Science, CAS, China, and the Simons Foundation Mathematics and Physical Sciences Program Award #499650.

A. Appendix

We use $a_n \sim b_n$ to represent $\lim_{n\to\infty} a_n/b_n = c$, where c is some nonzero constant. For any function $\varphi(u)$ defined on $[a, b]$, let $\|\varphi(u)\|_\infty = \|\varphi\|_\infty = \sup_{u\in[a,b]} |\varphi(u)|$.

A.1 Preliminaries

This section gives some lemmas that are needed in our theoretical development. Their proofs are given in the Supplementary Material.

Lemma 1 *(Theorem 1.2 of Bosq (1998)) Let $\xi_1, \ldots, \xi_n$ be independent random variables with mean* 0*. If there exists $c > 0$ such that (Cramér's Conditions)*

$$\mathrm{E}\left|\xi_i\right|^k \le c^{k-2} k!\, \mathrm{E}\,\xi_i^2 < +\infty \text{ for } 1 \le i \le n, k \ge 3,$$

then for any $t > 0$,

$$P\left\{\left|\sum_{i=1}^n \xi_i\right| > t\right\} \le 2\exp\left\{-\frac{t^2}{4\sum_{i=1}^n \mathrm{E}\,\xi_i^2 + 2ct}\right\}.$$

Lemma 2 *Under Assumptions* (A1)–(A5)*, for any integer $l \ge 0$, as $n \to \infty$, one has*

$$\sup_{x\in[a_0,b_0]} \left| n^{-1} \sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h(X_i - x)(X_i - x)^l \varepsilon_i \right| = O_p\left(n^{-1/2} h^{l-1/2} \log^{1/2} n\right).$$

In the following, we discuss the representations of the weighted estimators $\hat{m}(x, \pi)$ and $\hat{m}\left(x, \hat{\pi}\right)$, and break the errors $\hat{m}(x, \pi) - m(x)$ and $\hat{m}\left(x, \hat{\pi}\right) - m(x)$ into simpler parts to prove Theorems 1 and 3. Let

$$L_{n,l}(x) = n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i - x)(X_i - x)^l, l = 0, 1, 2,$$

and

$$M_{n,l}(x) = n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i - x)(X_i - x)^l\left\{Y_i - m(x) - m^{(1)}(x)(X_i - x)\right\}.$$

Then

$$\mathbf{X}^T\mathbf{W}\mathbf{X}=\begin{pmatrix}L_{n,0}(x) & L_{n,1}(x)\\ L_{n,1}(x) & L_{n,2}(x)\end{pmatrix},$$
$$\mathbf{X}^T\mathbf{W}\Big(\mathbf{Y}-m(x)\mathbf{X}e_0-m^{(1)}(x)\mathbf{X}e_1\Big)=\begin{pmatrix}M_{n,0}(x)\\ M_{n,1}(x)\end{pmatrix},$$

where $e_1 = (0, 1)^T$. By (4), one then has

$$\begin{aligned}\hat{m}(x,\pi) - m(x) &= e_0^T\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\left(\mathbf{Y}-m(x)\mathbf{X}e_0 - m^{(1)}(x)\mathbf{X}e_1\right)\\ &= e_0^T\begin{pmatrix}L_{n,0}(x) & L_{n,1}(x)\\ L_{n,1}(x) & L_{n,2}(x)\end{pmatrix}^{-1}\begin{pmatrix}M_{n,0}(x)\\ M_{n,1}(x)\end{pmatrix}.\end{aligned} \quad (11)$$

To further study $\hat{m}\left(x, \hat{\pi}\right)$, let

$$\hat{L}_{n,l}(x) = n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}K_h(X_i - x)(X_i - x)^l, l = 0, 1, 2,$$

and

$$\hat{M}_{n,l}(x) = n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}K_h(X_i - x)(X_i - x)^l\left\{Y_i - m(x) - m^{(1)}(x)(X_i - x)\right\}.$$

By (5), one then obtains that

$$\begin{aligned}\hat{m}\left(x,\hat{\pi}\right)-m(x) &= e_0^T\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{W}}\left(\mathbf{Y}-m(x)\mathbf{X}e_0 - m^{(1)}(x)\mathbf{X}e_1\right)\\ &= e_0^T\begin{pmatrix}\hat{L}_{n,0}(x) & \hat{L}_{n,1}(x)\\ \hat{L}_{n,1}(x) & \hat{L}_{n,2}(x)\end{pmatrix}^{-1}\begin{pmatrix}\hat{M}_{n,0}(x)\\ \hat{M}_{n,1}(x)\end{pmatrix}.\end{aligned} \quad (12)$$

Lemma 3 *Under Assumptions* (A1) *and* (A3)–(A5)*, as $n \to \infty$, uniformly for all $x \in [a_0, b_0]$, one has*

$$L_{n,l}(x) = h^l f_X(x)\mu_l(K) + u_p\left(h^{l+1}\right) + U_p\left(n^{-1/2}h^{l-1/2}\log^{1/2}n\right), l = 0, 1, 2.$$

Lemma 4 *Under Assumptions* (A1)–(A5)*, as $n \to \infty$, uniformly for all $x \in [a_0, b_0]$, one has*

$$M_{n,0}(x) = n^{-1} \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_h (X_i - x) \varepsilon_i + 2^{-1} m^{(2)}(x) f_X(x) \mu_2(K) h^2 + u_p \left(h^2\right)$$

and

$$M_{n,1}(x) = U_p \left(n^{-1/2} h^{1/2} \log^{1/2} n\right).$$

Lemma 5 *Under Assumptions* (A1)–(A5)*, as $n \to \infty$, one has*

$$\sup_{u \in [a_0, b_0]} \left|\hat{L}_{n,l}(x) - L_{n,l}(x)\right| = O_p\left(n^{-1/2}\right), l = 0, 1, 2,$$

and

$$\sup_{u \in [a_0, b_0]} \left|\hat{M}_{n,l}(x) - M_{n,l}(x)\right| = O_p\left(n^{-1/2}\right), l = 0, 1.$$

A.2 Conditional limiting extreme value distribution of $V_n(x)$

This section contains the main steps to obtain the conditional extreme value distribution of $V_n(x) = n^{-1} f_X^{-1}(x) \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_h (X_i - x) \varepsilon_i$ shown in Theorem 6 at the end of this section which will be used in the total probability formula in the proof of Theorem 2.

The Rosenblatt quantile transformation in Rosenblatt (1952) is adopted with

$$T(X, \varepsilon) = \left(X^*, \varepsilon^*\right) = \left(F_{X|\delta=1}(X), F_{\varepsilon|X,\delta=1}(\varepsilon|X)\right),$$

where $F_{X|\delta=1}(X)$ is the conditional distribution function of X given $\delta = 1$ and $F_{\varepsilon|X,\delta=1}(\varepsilon|X)$ is the conditional distribution function of ε given X and $\delta = 1$. This transformation produces mutually independent uniform random variables (X^*, ε^*) on $[0, 1]^2$. According to the strong approximation theorem in Tusnady (1977) (Theorem 1), there exists a sequence of two dimensional Brownian bridges B_n such that

$$\sup_{x,\varepsilon} |Z_n(x, \varepsilon) - B_n(T(x, \varepsilon))| = O_{a.s.}\left(n^{-1/2} \log^2 n\right), \tag{13}$$

where $Z_n(x, \varepsilon) = n^{1/2}\left\{F_n(x, \varepsilon) - F_{X,\varepsilon|\delta=1}(x, \varepsilon)\right\}$ with $F_n(x, \varepsilon)$ and $F_{X,\varepsilon|\delta=1}(x, \varepsilon)$ representing the empirical and the theoretical distribution of (X, ε) given $\delta = 1$. The transformation and the strong approximation results have been also used in Johnston (1982), Härdle (1989), and Wang and Yang (2009) for constructing SCBs for the nonparametric regression when data are fully observed.

To obtain the distribution of $\sup_{x \in [a_0, b_0]} |V_n(x)|$ conditional on $\Delta_n = n_0$, we will show the following Lemmas 6–8. Here $\{n_0\}$ is a sequence of numbers related to n with

$1 \le n_0 \le n$. By (6) it is clear that there exists a constant $r > 0$ such that $r \le \Delta_n/n \le 1$ in probability as $n \to \infty$. Thus we only need to consider $n_0 \ge r \times n$. That is, n_0 and n have the same order as $n \to \infty$. Therefore, to unify the notation in the following we will use n in the convergence rate.

Meanwhile, due to the i.i.d. assumption of the data, conditional on $\Delta_n = \sum_{i=1}^n \delta_i = n_0$ is equivalent to conditional on the event that there are n_0 elements in $\boldsymbol{\delta}_n = (\delta_1, ..., \delta_n)^T$ that are equal to 1 and the rest $(n - n_0)$ elements are equal to 0. Without loss of generality, let $\delta_i = 1$ for $i = 1, \ldots, n_0$ and $\delta_i = 0$ for $i = n_0 + 1, \ldots, n$.

Notice that, for $i = 1, \ldots, n$,

$$\begin{aligned} 0 &= \mathrm{E}\left\{ \frac{\delta_i}{\pi_i} K_h (X_i - x) \varepsilon_i \right\} = \mathrm{E}\left[\mathrm{E}\left\{ \frac{\delta_i}{\pi_i} K_h (X_i - x) \varepsilon_i \middle| \delta_i \right\} \right] \\ &= \mathrm{E}\left\{ \frac{1}{\pi_i} K_h (X_i - x) \varepsilon_i \middle| \delta_i = 1 \right\} P(\delta_i = 1). \end{aligned}$$

Thus, conditional on $\Delta_n = n_0$, $1 \le n_0 \le n$, by symmetry one has

$$\begin{aligned} &\mathrm{E}\left\{ \sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h (X_i - x) \varepsilon_i \middle| \Delta_n = n_0 \right\} \\ &= \mathrm{E}\left\{ \sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h (X_i - x) \varepsilon_i \middle| \delta_1 = \cdots = \delta_{n_0} = 1, \delta_{n_0+1} = \cdots = \delta_n = 0 \right\} \\ &= n_0 \, \mathrm{E}\left\{ \frac{1}{\pi_1} K_h (X_1 - x) \varepsilon_1 \middle| \delta_1 = 1 \right\} = 0 \end{aligned}$$

and

$$\begin{aligned} &\mathrm{var}\left\{ \sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h (X_i - x) \varepsilon_i \middle| \Delta_n = n_0 \right\} \\ &= \mathrm{E}\left[\{\sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h (X_i - x) \varepsilon_i\}^2 \middle| \delta_1 = \cdots = \delta_{n_0} = 1, \delta_{n_0+1} = \cdots = \delta_n = 0 \right] \\ &= n_0 \, \mathrm{E}\left(\frac{1}{\pi_1^2} K_h^2 (X_1 - x) \varepsilon_1^2 \middle| \delta_1 = 1 \right) \\ &= n_0 \int \frac{1}{\pi^2 (m(u) + \varepsilon)} K_h^2 (u - x) \varepsilon^2 f_{X,\varepsilon|\delta=1} (u, \varepsilon) \, du d\varepsilon \\ &= n_0 h^{-1} \int \frac{1}{\pi^2 (m(x + hv) + \varepsilon)} K^2 (v) \varepsilon^2 f_{X,\varepsilon|\delta=1} (x + hv, \varepsilon) \, dv d\varepsilon \\ &= n_0 h^{-1} \int K^2 (v) \, dv \int \frac{1}{\pi^2 (m(x) + \varepsilon)} \varepsilon^2 f_{X,\varepsilon|\delta=1} (x, \varepsilon) \, d\varepsilon \{1 + u(1)\} \\ &= n_0 h^{-1} \lambda(K) s(x) \{1 + u(1)\}. \end{aligned} \tag{14}$$

Moreover, as discussed above, conditional on $\Delta_n = n_0$ one can let $\delta_i = 1$ for $i = 1, \ldots, n_0$ and $\delta_i = 0$ for $i = n_0+1, \ldots, n$ without loss of generality. Then conditional on $\Delta_n = n_0$ one can write

$$\begin{aligned} V_n(x) &= n^{-1} f_X^{-1}(x) \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_h(X_i - x)\varepsilon_i \\ &= n^{-1} f_X^{-1}(x) \sum_{i=1}^{n_0} \frac{1}{\pi_i} K_h(X_i - x)\varepsilon_i. \end{aligned}$$

Conditional on $\Delta_n = n_0$ we now introduce the following standardized stochastic process:

$$\zeta_{1n_0}(x) = (n_0 h)^{1/2} s^{-1/2}(x)\, n_0^{-1} \sum_{i=1}^{n_0} \frac{1}{\pi_i} K_h(X_i - x)\varepsilon_i, \tag{15}$$

which can be rewritten as

$$\zeta_{1n_0}(x) = h^{1/2} s^{-1/2}(x) \int\int \frac{1}{\pi(m(u)+\varepsilon)} K_h(u - x)\,\varepsilon d Z_{n_0}(u, \varepsilon),$$

where $Z_{n_0}(u, \varepsilon)$ is the same as $Z_n(u, \varepsilon)$ in (13) but with n replaced by n_0.

Let $\kappa_n = n^{\theta}$ with $\frac{2}{3\eta} < \theta < \frac{1}{6}$ where $\eta > 4$ is given in Assumption (A2), which together with Assumption (A5) implies that

$$\kappa_n^{-\eta} h^{-2} \log n = O(1), \quad \kappa_n^2 n^{-1/2} h^{-1/2} (\log n)^{5/2} = o(1). \tag{16}$$

Then conditional on $\Delta_n = n_0$ one can define the following processes to approximate $\zeta_{1n_0}(x)$:

$$\begin{aligned} \zeta_{2n_0}(x) &= h^{1/2} s_n^{-1/2}(x) \int\int_{|\varepsilon| \le \kappa_n} \frac{1}{\pi(m(u)+\varepsilon)} K_h(u - x)\,\varepsilon d Z_{n_0}(u, \varepsilon), \\ \zeta_{3n_0}(x) &= h^{1/2} s_n^{-1/2}(x) \int\int_{|\varepsilon| \le \kappa_n} \frac{1}{\pi(m(u)+\varepsilon)} K_h(u - x)\,\varepsilon d B_{n_0}(T(u, \varepsilon)), \\ \zeta_{4n_0}(x) &= h^{1/2} s_n^{-1/2}(x) \int\int_{|\varepsilon| \le \kappa_n} \frac{1}{\pi(m(u)+\varepsilon)} K_h(u - x)\,\varepsilon d W_{n_0}(T(u, \varepsilon)), \end{aligned}$$

where $s_n(x) = \int_{|\varepsilon| \le \kappa_n} \frac{\varepsilon^2}{\pi^2(m(x)+\varepsilon)} f_{X,\varepsilon|\delta=1}(x, \varepsilon)\, d\varepsilon$, $B_{n_0}(T(u, \varepsilon))$ is the sequence of Brownian bridges in (13) and $W_{n_0}(T(u, \varepsilon))$ is the sequence of Wiener processes satisfying $B_{n_0}(u, s) = W_{n_0}(u, s) - us W_{n_0}(1, 1)$. Moreover, define

$$\zeta_{5n_0}(x) = h^{1/2} s_n^{-1/2}(x) \int s_n^{1/2}(u) K_h(u - x)\, dW(u),$$

and

$$\zeta_{6n_0}(x) = h^{1/2}\int K_h(u-x)\,dW(u),$$

where $W(u)$ is a two-sided Wiener process on $(-\infty,+\infty)$. Conditional on $\Delta_n = n_0$, according to Theorem 3.1 in Bickel and Rosenblatt (1952), one has

$$P\left[a_h\left\{\sup_{x\in[a_0,b_0]}\left|\zeta_{6n_0}(x)\right|/\lambda^{1/2}(K)-b_h\right\}\le t\,\middle|\,\Delta_n=n_0\right]\to\exp\{-2\exp(-t)\} \tag{17}$$

$\forall t\in\mathbb{R}$, as n_0 (and thus n) $\to\infty$. Here a_h, b_h, and $\lambda(K)$ are given in Theorem 2.

The proofs of the following Lemmas 6 and 7 are given in the Supplementary Material due to the space limitation.

Lemma 6 *Under Assumptions* (A1)–(A5)*, conditional on* $\Delta_n = n_0$*, for an increasing sequence* $\{n_0\}$*, as* $n_0\to\infty$*, one has*

$$\begin{aligned}
(a)\ &\sup_{x\in[a_0,b_0]}\left|\zeta_{2n_0}(x)\right.\\
&\left.-\zeta_{3n_0}(x)\right| = o_p\left(\log^{-1/2}n\right),\\
(b)\ &\sup_{x\in[a_0,b_0]}\left|\zeta_{3n_0}(x)\right.\\
&\left.-\zeta_{4n_0}(x)\right| = o_p\left(\log^{-1/2}n\right),\\
(c)\ &\sup_{x\in[a_0,b_0]}\left|\zeta_{5n_0}(x)\right.\\
&\left.-\zeta_{6n_0}(x)\right| = o_p\left(\log^{-1/2}n\right).
\end{aligned}$$

Lemma 7 *Conditional on* $\Delta_n = n_0$ *for an increasing sequence* $\{n_0\}$*, the stochastic processes* $\zeta_{4n_0}(x)$ *and* $\zeta_{5n_0}(x)$ *have the same asymptotic distribution as* $n_0\to\infty$*.*

Lemmas 6 and 7, expression (17), and Slutsky's Theorem imply that

$$P\left[a_h\left\{\sup_{x\in[a_0,b_0]}\left|\zeta_{2n_0}(x)\right|/\lambda^{1/2}(K)-b_h\right\}\le t\,\middle|\,\Delta_n=n_0\right]\to\exp\{-2\exp(-t)\} \tag{18}$$

$\forall t\in\mathbb{R}$, as $n_0\to\infty$.

Lemma 8 *Under Assumptions* (A1)–(A5)*, conditional on* $\Delta_n = n_0$ *for an increasing sequence* $\{n_0\}$*, one has*

$$\sup_{x\in[a_0,b_0]}\left|\zeta_{1n_0}(x)-\zeta_{2n_0}(x)\right| = o_p\left(\log^{-1/2}n\right),$$

as $n_0 \to \infty$.

Proof of Lemma 8. Define

$$\zeta^*_{1n_0}(x) = h^{1/2}s^{-1/2}(x)\int\int_{|\varepsilon|\le\kappa_n}\frac{1}{\pi(m(u)+\varepsilon)}K_h(u-x)\,\varepsilon dZ_{n_0}(u,\varepsilon).$$

To prove the lemma, it is sufficient to prove that conditional on $\Delta_n = n_0$

$$\sup_{x\in[a_0,b_0]}\left|\zeta_{1n_0}(x) - \zeta^*_{1n_0}(x)\right| = o_p\left(\log^{-1/2} n\right) \tag{19}$$

and

$$\sup_{x\in[a_0,b_0]}\left|\zeta_{2n_0}(x) - \zeta^*_{1n_0}(x)\right| = o_p\left(\log^{-1/2} n\right) \tag{20}$$

as $n_0 \to \infty$. In the following, we first show (20). By (18) and the fact that $b_h = O\left(\log^{1/2} n\right)$, one has $\sup_{x\in[a_0,b_0]}\left|\zeta_{2n_0}(x)\right| = O_p\left(\log^{1/2} n\right)$ which with (S.6) in the Supplementary Material implies that

$$\begin{aligned}\sup_{x\in[a_0,b_0]}\left|\zeta_{2n_0}(x)-\zeta^*_{1n_0}(x)\right| &= \sup_{x\in[a_0,b_0]}\Bigg|h^{1/2}\left\{s^{-1/2}(x) - s_n^{-1/2}(x)\right\}\\ &\quad\times\int\int_{|\varepsilon|\le\kappa_n}\frac{1}{\pi(m(u)+\varepsilon)}K_h(u-x)\,\varepsilon dZ_{n_0}(u,\varepsilon)\Bigg|\\ &= O_p\left(h^2\log^{-1/2} n\right) = o_p\left(\log^{-1/2} n\right).\end{aligned}$$

We next prove (19). Notice that

$$\begin{aligned}&\zeta_{1n_0}(x) - \zeta^*_{1n_0}(x)\\ &\quad= h^{1/2}s^{-1/2}(x)\int\int_{|\varepsilon|>\kappa_n}\frac{1}{\pi(m(u)+\varepsilon)}K_h(u-x)\,\varepsilon dZ_{n_0}(u,\varepsilon)\\ &\quad= s^{-1/2}(x)\sum_{i=1}^{n_0}\left(n_0^{-1}h\right)^{1/2}\left[\frac{1}{\pi(m(X_i)+\varepsilon_i)}K_h(X_i-x)\,\varepsilon_i I\{|\varepsilon_i|>\kappa_n\}\right.\\ &\left.-\mathrm{E}\left\{\frac{1}{\pi(m(X_i)+\varepsilon_i)}K_h(X_i-x)\,\varepsilon_i I\{|\varepsilon_i|>\kappa_n\}\middle|\,\delta_i=1\right\}\right].\end{aligned}$$

For convenience, we denote

$$\begin{aligned}\varsigma_{i,n}(x) = \left(n_0^{-1}h\right)^{1/2}\log^{1/2} n&\left[\frac{1}{\pi(m(X_i)+\varepsilon_i)}K_h(X_i-x)\,\varepsilon_i I\{|\varepsilon_i|>\kappa_n\}\right.\\ &\left.-\mathrm{E}\left\{\frac{1}{\pi(m(X_i)+\varepsilon_i)}K_h(X_i-x)\,\varepsilon_i I\{|\varepsilon_i|>\kappa_n\}\middle|\,\delta_i=1\right\}\right].\end{aligned}$$

To prove (19), it is sufficient to verify that

$$\sup_{x\in[a_0,b_0]}\left|\textstyle\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x\right)\right|=o_p\left(1\right).$$

By Theorem 15.6 in Billingsley (1968), it suffices to show: (i) conditional on $\Delta_n=n_0$, $\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x\right)\to 0$ in probability for any given $x\in[a_0,b_0]$ and (ii) the tightness of $\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x\right)$ conditional on $\Delta_n=n_0$, using the following moment condition:

$$\mathrm{E}\left\{\left|\left(\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x\right)-\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x_1\right)\right)\left(\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x_2\right)-\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x\right)\right)\right|\middle|\Delta_n=n_0\right\}\\ \leq C\left|x_2-x_1\right|^2$$

for any $x\in[x_1,x_2]$ and some constant $C>0$ that is independent of n_0.

Firstly, note that $\varsigma_{i,n}\left(x\right),\ 1\leq i\leq n$, are independent variables with $\mathrm{E}\{\varsigma_{i,n}\left(x\right)\mid\delta_i=1\}=0$ and

$$\begin{aligned}
&\mathrm{var}\{\varsigma_{i,n}\left(x\right)\left|\delta_i=1\right\}=\mathrm{E}\{\varsigma_{i,n}^2\left(x\right)\left|\delta_i=1\right\}\\
&\leq n_0^{-1}h\log n\,\mathrm{E}\left[\frac{1}{\pi^2\left(m\left(X_i\right)+\varepsilon_i\right)}K_h^2\left(X_i-x\right)\varepsilon_i^2 I\left\{\left|\varepsilon_i\right|>\kappa_n\right\}\middle|\delta_i=1\right]\\
&=n_0^{-1}h\log n\int\int_{|\varepsilon|>\kappa_n}\frac{1}{\pi^2\left(m\left(u\right)+\varepsilon\right)}K_h^2\left(u-x\right)\varepsilon^2 f_{X,\varepsilon|\delta=1}\left(u,\varepsilon\right)du d\varepsilon\\
&=n_0^{-1}\log n\int\int_{|\varepsilon|>\kappa_n}\frac{1}{\pi^2\left(m\left(x\right)+\varepsilon\right)}K^2(v)\,\varepsilon^2 f_{X,\varepsilon|\delta=1}(x,\varepsilon)\,dvd\varepsilon\left\{1+u\left(1\right)\right\}\\
&\leq n_0^{-1}\log n\int K^2(v)\,dv\int_{|\varepsilon|>\kappa_n}\frac{1}{\pi^2\left(m\left(x\right)+\varepsilon\right)}\varepsilon^2 f_{X,\varepsilon|\delta=1}(x,\varepsilon)\,d\varepsilon\left\{1+u\left(1\right)\right\}.
\end{aligned}$$

Thus, by (16), one has $\mathrm{var}\{\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x\right)\mid\Delta_n=n_0\}=n_0\,\mathrm{var}\{\varsigma_{i,n}\left(x\right)\mid\delta_i=1\}\to 0$ which together with Markov's inequality concludes that for any given $x\in[a_0,b_0]$,

$$\textstyle\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x\right)\to 0 \text{ in probability.}$$

Secondly, notice that

$$\begin{aligned}
&\mathrm{E}\left\{\left(\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x\right)-\sum_{i=1}^{n_0}\varsigma_{i,n}\left(x_1\right)\right)^2\middle|\Delta_n=n_0\right\}\\
&\quad=n_0^{-1}h\log n\sum_{i=1}^{n_0}\mathrm{E}\left[\left\{\frac{\left(K_h\left(X_i-x\right)-K_h\left(X_i-x_1\right)\right)}{\pi\left(m\left(X_i\right)+\varepsilon_i\right)}\varepsilon_i I\left\{\left|\varepsilon_i\right|>\kappa_n\right\}-\right.\right.\\
&\qquad\left.\left.\mathrm{E}\left(\frac{\left(K_h\left(X_i-x\right)-K_h\left(X_i-x_1\right)\right)}{\pi\left(m\left(X_i\right)+\varepsilon_i\right)}\varepsilon_i I\left(\left|\varepsilon_i\right|>\kappa_n\right)\middle|\delta_i=1\right)\right\}^2\middle|\Delta_n=n_0\right]\\
&\quad=h\log n\,\mathrm{E}\left[\left\{\frac{\left(K_h\left(X_1-x\right)-K_h\left(X_1-x_1\right)\right)}{\pi\left(m\left(X_1\right)+\varepsilon_1\right)}\varepsilon_1 I\left\{\left|\varepsilon_1\right|>\kappa_n\right\}-\right.\right.
\end{aligned}$$

$$\mathrm{E}\Bigg(\frac{(K_h(X_1 - x) - K_h(X_1 - x_1))}{\pi\,(m(X_1) + \varepsilon_1)} \varepsilon_1 I\,(|\varepsilon_1| > \kappa_n)\,) \Bigg| \delta_1 = 1 \Bigg) \Bigg\}^2 \Bigg| \delta_1 = 1 \Bigg].$$

Since $K(u) \in C^{(1)}[-1, 1]$ by Assumption (A3),

$$\begin{aligned} &\mathrm{E}\left\{ \left(\sum_{i=1}^{n_0} \varsigma_{i,n}(x) - \sum_{i=1}^{n_0} \varsigma_{i,n}(x_1) \right)^2 \middle| \Delta_n = n_0 \right\} \\ &\leq C_1 (x - x_1)^2 h^{-2} \log n \int_{|\varepsilon| > \kappa_n} \frac{1}{\pi^2 (m(x) + \varepsilon)} \varepsilon^2 f_{X,\varepsilon|\delta=1}(x, \varepsilon)\, d\varepsilon \end{aligned}$$

and

$$\begin{aligned} &\mathrm{E}\left\{ \left(\sum_{i=1}^{n_0} \varsigma_{i,n}(x_2) - \sum_{i=1}^{n_0} \varsigma_{i,n}(x) \right)^2 \middle| \Delta_n = n_0 \right\} \\ &\leq C_1 (x_2 - x)^2 h^{-2} \log n \int_{|\varepsilon| > \kappa_n} \frac{1}{\pi^2 (m(x) + \varepsilon)} \varepsilon^2 f_{X,\varepsilon|\delta=1}(x, \varepsilon)\, d\varepsilon \end{aligned}$$

for some constant $C_1 > 0$ that is independent of n_0. Therefore, by the Schwarz inequality, one has that

$$\begin{aligned} &\mathrm{E}\left\{ \left| \left(\sum_{i=1}^{n_0} \varsigma_{i,n}(x) - \sum_{i=1}^{n_0} \varsigma_{i,n}(x_1) \right) \left(\sum_{i=1}^{n_0} \varsigma_{i,n}(x_2) - \sum_{i=1}^{n_0} \varsigma_{i,n}(x) \right) \right| \middle| \Delta_n = n_0 \right\} \\ &\leq \left[\mathrm{E}\left\{ \left(\sum_{i=1}^{n_0} \varsigma_{i,n}(x) - \sum_{i=1}^{n_0} \varsigma_{i,n}(x_1) \right)^2 \middle| \Delta_n = n_0 \right\} \right]^{1/2} \times \\ &\qquad\qquad \left[\mathrm{E}\left\{ \left(\sum_{i=1}^{n_0} \varsigma_{i,n}(x_2) - \sum_{i=1}^{n_0} \varsigma_{i,n}(x) \right)^2 \middle| \Delta_n = n_0 \right\} \right]^{1/2} \\ &\leq C_1 |x - x_1| \, |x_2 - x| \, h^{-2} \log n \int_{|\varepsilon| > \kappa_n} \frac{1}{\pi^2 (m(x) + \varepsilon)} \varepsilon^2 f_{X,\varepsilon|\delta=1}(x, \varepsilon)\, d\varepsilon \end{aligned}$$

which together with (16) concludes that

$$\begin{aligned} &\mathrm{E}\left\{ \left| \left(\sum_{i=1}^{n_0} \varsigma_{i,n}(x) - \sum_{i=1}^{n_0} \varsigma_{i,n}(x_1) \right) \left(\sum_{i=1}^{n_0} \varsigma_{i,n}(x_2) - \sum_{i=1}^{n_0} \varsigma_{i,n}(x) \right) \right| \middle| \Delta_n = n_0 \right\} \\ &\qquad\qquad \leq C |x_2 - x_1|^2 \end{aligned}$$

for some $C > 0$ that is independent of n_0, verifying the tightness. The proof is completed. □

By the definitions of $V_n(x)$ in Theorem 1 and $\zeta_{1n_0}(x)$ in (15), one has $\zeta_{1n_0}(x) = (nh)^{1/2} r_n^{-1/2} s^{-1/2}(x) f_X(x) V_n(x)$ given $\Delta_n = n_0$. This together with Lemma 8, expression (18), and Slutsky's Theorem concludes the following result.

Theorem 6 *Under Assumptions* (A1)–(A5)*, one has that, for any* $t \in \mathbb{R}$*, as* $n_0 \to \infty$*,*

$$P\left[a_h\left\{\sup_{x\in[a_0,b_0]}\left|(nh)^{1/2}r_n^{-1/2}V_n(x)/d^{1/2}(x)\right|-b_h\right\}\le t\middle|\Delta_n=n_0\right] \to \exp\{-2\exp(-t)\}. \quad (21)$$

A.3 Proofs of the theorems in Section 2

Proof of Theorem 1. By Lemma 3 and Assumption (A5), one has

$$\mathbf{X}^T\mathbf{WX}=\begin{pmatrix}L_{n,0}(x) & L_{n,1}(x)\\ L_{n,1}(x) & L_{n,2}(x)\end{pmatrix}=f_X(x)\begin{pmatrix}1+u_p(h) & U_p(h^2)\\ U_p(h^2) & h^2\mu_2(K)+u_p(h^3)\end{pmatrix}$$

which implies that

$$\left(\mathbf{X}^T\mathbf{WX}\right)^{-1}=f_X^{-1}(x)\begin{pmatrix}1+u_p(h) & U_p(1)\\ U_p(1) & h^{-2}\mu_2^{-1}(K)+u_p\left(h^{-1}\right)\end{pmatrix}.$$

It together with (11) and Lemmas 2 and 4 concludes that uniformly for all $x\in[a_0,b_0]$,

$$\begin{aligned}&\hat{m}(x,\pi)-m(x)\\&=e_0^T\left\{f_X^{-1}(x)\begin{pmatrix}1+u_p(h) & U_p(1)\\ U_p(1) & h^{-2}\mu_2^{-1}(K)+u_p\left(h^{-1}\right)\end{pmatrix}\right\}\\&\quad\times\begin{pmatrix}n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i-x)\varepsilon_i+2^{-1}m^{(2)}(x)f_X(x)\mu_2(K)h^2+u_p\left(h^2\right)\\ U_p\left(n^{-1/2}h^{1/2}\log^{1/2}n\right)\end{pmatrix}\\&=V_n(x)+2^{-1}m^{(2)}(x)\mu_2(K)h^2+u_p\left(h^2\right)+U_p\left(n^{-1/2}h^{1/2}\log^{1/2}n\right)\\&=V_n(x)+2^{-1}m^{(2)}(x)\mu_2(K)h^2+u_p\left(h^2\right).\end{aligned}$$

The proof is completed. □

Proof of Theorem 2. According to Theorem 6, for any $t\in\mathbb{R}$, as $n_0\to\infty$,

$$P\left[a_h\left\{\sup_{x\in[a_0,b_0]}\left|(nh)^{1/2}r_n^{-1/2}V_n(x)/d^{1/2}(x)\right|-b_h\right\}\le t\middle|\Delta_n=n_0\right] \to \exp\{-2\exp(-t)\}.$$

Thus one has that for any given $\epsilon>0$ and $t\in\mathbb{R}$, there exists $N_0>0$ such that

$$\left|P\left[a_h\left\{\sup_{x\in[a_0,b_0]}\left|(nh)^{1/2}r_n^{-1/2}V_n(x)/d^{1/2}(x)\right|-b_h\right\}\le t\middle|\Delta_n=n_0\right]\right.$$

$$-\exp\left\{-2\exp\left(-t\right)\right\}\Bigg| < \frac{\epsilon}{2}$$

for all $n_0 \geq N_0$. On the other hand, since $\Delta_n/n \to P\left(\delta_1 = 1\right) > 0$ a.s., there exists $N > N_0$ such that when $n \geq N$, $P\left(\Delta_n \geq N_0\right) > 1 - \epsilon/2$. Therefore, unconditional on Δ_n, for $n \geq N$,

$$\begin{aligned}
&\left|P\left[a_h\left\{\sup_{x\in[a_0,b_0]}\left|(nh)^{1/2}r_n^{-1/2}V_n(x)/d^{1/2}(x)\right|-b_h\right\}\leq t\right]-\exp\left\{-2\exp\left(-t\right)\right\}\right|\\
&\leq \sum_{n_0=1}^{n}\left|P\left[a_h\left\{\sup_{x\in[a_0,b_0]}\left|(nh)^{1/2}r_n^{-1/2}V_n(x)/d^{1/2}(x)\right|-b_h\right\}\leq t\middle|\Delta_n = n_0\right]\right.\\
&\qquad\left.-\exp\left\{-2\exp\left(-t\right)\right\}\right|\times P(\Delta_n = n_0)+P(\Delta_n = 0)\\
&\leq \sum_{n_0=N_0}^{n}\left|P\left[a_h\left\{\sup_{x\in[a_0,b_0]}\left|(nh)^{1/2}r_n^{-1/2}V_n(x)/d^{1/2}(x)\right|-b_h\right\}\leq t\middle|\Delta_n = n_0\right]\right.\\
&\qquad\left.-\exp\left\{-2\exp\left(-t\right)\right\}\right|\times P(\Delta_n = n_0)+\frac{\epsilon}{2}<\epsilon.
\end{aligned}$$

This together with the fact that the dominating term of $\hat{m}\left(x, \pi\right) - m\left(x\right)$ is $V_n(x)$ as seen in Theorem 1 concludes Theorem 2. □

Proof of Theorem 3. By (11) and (12) one has

$$\begin{aligned}
\hat{m}\left(x,\pi\right)-\hat{m}\left(x,\hat{\pi}\right) &= e_0^T\begin{pmatrix}L_{n,0}(x) & L_{n,1}(x)\\ L_{n,1}(x) & L_{n,2}(x)\end{pmatrix}^{-1}\begin{pmatrix}M_{n,0}(x)\\ M_{n,1}(x)\end{pmatrix}\\
&\quad-e_0^T\begin{pmatrix}\hat{L}_{n,0}(x) & \hat{L}_{n,1}(x)\\ \hat{L}_{n,1}(x) & \hat{L}_{n,2}(x)\end{pmatrix}^{-1}\begin{pmatrix}\hat{M}_{n,0}(x)\\ \hat{M}_{n,1}(x)\end{pmatrix}.
\end{aligned}$$

By Lemma 5, it is easily seen that

$$\sup_{x\in[a_0,b_0]}\left|\hat{m}\left(x,\pi\right)-\hat{m}\left(x,\hat{\pi}\right)\right| = O_p\left(n^{-1/2}\right),$$

completing the proof. □

Proof of Theorem 5. By definition,

$$\hat{d}_n\left(x\right) = \frac{n}{\Delta_n}\hat{f}_X^{-2}\left(x\right)\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i^2}K_h^2\left(X_i-x\right)\hat{\varepsilon}_i^2. \tag{22}$$

Firstly, we study the uniform convergence property of $\frac{h}{n}\sum\limits_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i^2}K_h^2\left(X_i-x\right)\hat{\varepsilon}_i^2$. Notice that

$$
\begin{aligned}
&\sup_{x\in[a_0,b_0]}\left|\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i^2}K_h^2\left(X_i-x\right)\hat{\varepsilon}_i^2-\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i^2}K_h^2\left(X_i-x\right)\hat{\varepsilon}_i^2\right|\\
&=\sup_{x\in[a_0,b_0]}\left|\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i\left(\pi_i^2-\hat{\pi}_i^2\right)}{\hat{\pi}_i^2\pi_i^2}K_h^2\left(X_i-x\right)\left\{m\left(X_i\right)-\hat{m}\left(X_i,\hat{\pi}_i\right)+\varepsilon_i\right\}^2\right|\\
&=o_p\left(n^{-1/2}h^{-1}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
&\sup_{x\in[a_0,b_0]}\left|\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i^2}K_h^2\left(X_i-x\right)\hat{\varepsilon}_i^2-\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i^2}K_h^2\left(X_i-x\right)\varepsilon_i^2\right|\\
&=\sup_{x\in[a_0,b_0]}\left|\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i^2}K_h^2\left(X_i-x\right)\left(\hat{\varepsilon}_i^2-\varepsilon_i^2\right)\right|\\
&\le\sup_{x\in[a_0,b_0]}\left|\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i^2}K_h^2\left(X_i-x\right)\left\{m\left(X_i\right)-\hat{m}\left(X_i,\hat{\pi}_i\right)\right\}^2\right|\\
&\qquad+\sup_{x\in[a_0,b_0]}\left|2\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i^2}K_h^2\left(X_i-x\right)\left\{m\left(X_i\right)-\hat{m}\left(X_i,\hat{\pi}_i\right)\right\}\varepsilon_i\right|\\
&=O_p\left(n^{-1/2}h^{-3/2}\log^{1/2}n\right),
\end{aligned}
$$

which imply that

$$
\begin{aligned}
&\sup_{x\in[a_0,b_0]}\left|\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i^2}K_h^2\left(X_i-x\right)\hat{\varepsilon}_i^2-\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i^2}K_h^2\left(X_i-x\right)\varepsilon_i^2\right|\\
&\qquad=O_p\left(n^{-1/2}h^{-3/2}\log^{1/2}n\right).
\end{aligned}
\tag{23}
$$

Secondly, denote $\varepsilon_i^*=\frac{\delta_i\varepsilon_i^2}{\pi_i^2}-\mathrm{E}\left(\frac{\delta_i\varepsilon_i^2}{\pi_i^2}\middle|X_i\right)$. By applying the inequality in Lemma 1, the Borel-Cantelli Lemma, and the truncation and discretization method as in the proof of Lemma 2, one obtains that

$$
\sup_{x\in[a_0,b_0]}\left|\frac{h}{n}\sum_{i=1}^{n}K_h^2\left(X_i-x\right)\varepsilon_i^*\right|=O_p\left(n^{-1/2}h^{-1/2}\log^{1/2}n\right)\tag{24}
$$

as $n \to \infty$. Meanwhile, similar to the proof of Lemma 3, one can easily show that

$$\sup_{x\in[a_0,b_0]}\left|\frac{h}{n}\sum_{i=1}^{n}\mathrm{E}\left\{K_h^2(X_i-x)\frac{\delta_i\varepsilon_i^2}{\pi_i^2}\mid X_i\right\}-h\,\mathrm{E}\left\{K_h^2(X_1-x)\frac{\delta_1\varepsilon_1^2}{\pi_1^2}\right\}\right| = O_p\left(n^{-1/2}h^{-1/2}\log^{1/2}n\right). \tag{25}$$

Combining (23), (24), and (25), one has

$$\sup_{x\in[a_0,b_0]}\left|\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i^2}K_h^2(X_i-x)\hat{\varepsilon}_i^2-h\,\mathrm{E}\left\{\frac{\delta_1}{\pi_1^2}K_h^2(X_1-x)\varepsilon_1^2\right\}\right| = O_p\left(n^{-1/2}h^{-3/2}\log^{1/2}n\right).$$

Meanwhile, by Lemmas 3 and 5, and $h_f = O(n^{-1/5})$, one can easily obtain that

$$\sup_{x\in[a_0,b_0]}\left|\hat{f}_X(x)-f_X(x)\right| = o_p\left(h_f\right)+O_p\left(n^{-1/2}h_f^{-1/2}\log^{1/2}n\right)=o_p\left(n^{-1/5}\right).$$

Thus,

$$\sup_{x\in[a_0,b_0]}\left|\hat{f}_X^{-2}(x)\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i^2}K_h^2(X_i-x)\hat{\varepsilon}_i^2-f_X^{-2}(x)\,h\,\mathrm{E}\left\{\frac{\delta_1}{\pi_1^2}K_h^2(X_1-x)\varepsilon_1^2\right\}\right| = o_p\left(n^{-1/5}\right)+O_p\left(n^{-1/2}h^{-3/2}\log^{1/2}n\right)=O_p\left(n^{-1/2}h^{-3/2}\log^{1/2}n\right),$$

which together with the fact that

$$\begin{aligned} f_X^{-2}(x)\,h\,\mathrm{E}\left\{\frac{\delta_1}{\pi_1^2}K_h^2(X_1-x)\varepsilon_1^2\right\} &= f_X^{-2}(x)\,h\,\mathrm{E}\left\{\frac{1}{\pi_1^2}K_h^2(X_1-x)\varepsilon_1^2\middle|\delta_1=1\right\}P(\delta_1=1)\\ &= d(x)\,P(\delta_1=1)+u_p(h) \end{aligned}$$

implies

$$\sup_{x\in[a_0,b_0]}\left|\hat{f}_X^{-2}(x)\frac{h}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i^2}K_h^2(X_i-x)\hat{\varepsilon}_i^2-d(x)\,P(\delta_1=1)\right| = O_p\left(n^{-1/2}h^{-3/2}\log^{1/2}n\right). \tag{26}$$

It is easily seen from (22), (6), and (26) that

$$\sup_{x\in[a_0,b_0]}\left|\hat{d}_n\left(x\right)-d\left(x\right)\right|=O_p\left(n^{-1/2}h^{-3/2}\log^{1/2}n\right),$$

completing the proof. □

References

Al Ahmari, T., Alomar, A., Al Beeybe, J., Asiri, N., Al Ajaji, R., Al Masoud, R., Al-Hazzaa, M. (2017). Associations of self-esteem with body mass index and body image among Saudi college-age females. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, *1*, 1–9.

Bickel, P., Rosenblatt, M. (1973). On some global measures of deviations of density function estimates. *The Annals of Statistics*, *31*, 1852–1884.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes*. New York: Springer-Verlag.

Cai, L., Li, L., Huang, S., Ma, L., Yang, L. (2020). Oracally efficient estimation for dense functional data with holiday effects. *Test*, *29*(1), 282–306. https://doi.org/10.1007/s11749-019-00655-5.

Cai, L., Liu, R., Wang, S., Yang, L. (2019). Simultaneous confidence bands for mean and variance functions based on deterministic design. *Statistica Sinica*, *29*, 505–525.

Cai, T., Low, M., Ma, Z. (2014). Adaptive confidence bands for nonparametric regression functions. *Journal of the American Statistical Association*, *109*, 1054–1070.

Cai, L., Yang, L. (2015). A smooth simultaneous confidence band for conditional variance function. *Test*, *24*, 632–655.

Cao, G., Wang, L., Li, Y., Yang, L. (2016). Oracle efficient confidence envelopes for covariance functions in dense functional data. *Statistica Sinica*, *26*, 359–383.

Cao, G., Yang, L., Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics*, *24*, 359–377.

Chen, H., Little, R. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, *94*, 896–908.

Chernozhukov, V., Chetverikov, D., Kato, K. (2014). Anti-concentrition and honest, adaptive confidence bands. *The Annals of Statistics*, *42*, 1787–1818.

Claeskens, G., Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics*, *31*, 1852–1884.

Eubank, R., Speckman, P. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, *88*, 1287–1301.

Fan, J., Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. London: Chapman and Hall.

Fan, J., Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varyingcoefficient models. *Scandinavian Journal of Statistics*, *27*, 715–731.

Gu, L., Wang, L., Härdle, W., Yang, L. (2014). A simultaneous confidence corridor for varying coefficient regression with sparse functional data. *Test*, *23*, 806–843.

Gu, L., Yang, L. (2015). Oracally efficient estimation for single-index link function with simultaneous confidence band. *Electronic Journal of Statistics*, *9*, 1540–1561.

Habib, F., Al Fozan, H., Barnawi, N., Al Motairi, W. (2015). Relationship between body mass index, self-esteem and quality of life among adolescent saudi female. *Journal of Biology, Agriculture and Healthcare*, *5*, 2224–3208.

Hall, P. (1991). On convergence rates of suprema. *Probability Theory and Related Fields*, *89*, 447–455.

Hall, P., Titterington, D. (1988). On confidence bands in nonparametric density estimation and regression. *Journal of Multivariate Analysis*, *27*, 228–254.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis*, *29*, 163–179.

Härdle, W., Marron, J. (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, *19*, 778–796.

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

Hosmer, D., Lemeshow, S. (2005). *Applied Logistic Regression*2nd ed. New York: Wiley.

Hsu, C., Long, Q., Li, Y., Jacobs, E. (2014). A nonparametric multiple imputation approach for data with missing covariate values with application to colorectal adenoma data. *Journal of Biopharmaceutical Statistics*, *24*, 634–648.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*, 332–346.

Johnston, G. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *Journal of Multivariate Analysis*, *12*, 402–414.

Kim, J. K., Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. London: Chapman and Hall.

Liang, H., Wang, S., Robins, J., Carroll, R. (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association*, *99*, 357–367.

Lipsitz, S. R., Ibrahim, J. G., Zhao, L.-P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, *94*, 1147–1160.

Little, R., Rubin, D. (2019). *Statistical Analysis with Missing Data*3rd ed. New York: Wiley.

Qin, J., Zhang, B., Leung, D. (2009). Empirical likelihood in missing data problems. *Journal of the American Statistical Association*, *104*, 1492–1503.

Robins, J., Rotnitzky, A., Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*, 846–866.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of the Institute of Statistical Mathematics*, *23*, 470–472.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Song, Q., Yang, L. (2009). Spline confidence bands for variance function. *Journal of Nonparametric Statistics*, *21*, 589–609.

Tusnády, G. (1977). A remark on the approximation of the sample df in the multidimensional case. *Periodica Mathematica Hungarica*, *8*, 53–55.

Wang, Q. (2009). Statistical estimation in partial linear models with covariate data missing at random. *Annals of the Institute of Statistical Mathematics*, *61*, 47–84.

Wang, J. (2012). Modelling time trend via spline confidence band. *Annals of the Institute of Statistical Mathematics*, *64*, 275–301.

Wang, C., Wang, S., Carroll, R. (1998). Local linear regression for generalized linear models with missing data. *Annals of Statistics*, *26*, 1028–1050.

Wang, C., Wang, S., Zhao, L.-P., Ou, S.-T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, *92*, 512–525.

Wang, J., Yang, L. (2009). Polynomial spline confidence bands for regression curves. *Statistica Sinica*, *19*, 325–342.

Zhao, Z., Wu, W. (2008). Confidence bands in nonparametric time series regression. *Annals of Statistics*, *36*, 1854–1878.

Zheng, S., Liu, R., Yang, L., Härdle, W. (2016). Statistical inference for generalized additive models: simultaneous confidence corridors and variable selection. *Test*, *25*, 607–626.

Zheng, S., Yang, L., Hardle, W. (2014). A smooth simultaneous confidence corridor for the mean of sparse functional data. *Journal of the American Statistical Association*, *109*, 661–673.

Zhou, S., Shen, X., Wolfe, D. (1998). Local asymptotics of regression splines and confidence regions. *Annals of Statistics*, *26*, 1760–1782.