Statistica Sinica **33** (2023), 281-302
doi:https://doi.org/10.5705/ss.202021.0073

ORACLE-EFFICIENT GLOBAL INFERENCE FOR VARIANCE FUNCTION IN NONPARAMETRIC REGRESSION WITH MISSING COVARIATES

Li Cai and Suojin Wang

Zhejiang Gongshang University and Texas A&M University

Abstract: We propose a new bias-corrected spline-kernel estimator and a smooth simultaneous confidence band (SCB) as a global inference tool for the conditional variance function in a nonparametric regression when the covariates are missing at random. To adapt to the possible missingness of the covariates, we employ a Horvitz–Thompson-type weighted spline smoothing to fit the nonparametric regression function. Based on the squared residuals, the weighted kernel method is then applied to estimate the variance function. Synthesizing the spline smoothing and kernel regression in one estimator takes advantage of the fast computing speed of the spline regression, and of the flexible local estimation and easy SCB constructions of the kernel smoothing. The proposed estimator is shown to be oracle-efficient in the sense that it is as efficient as the ideal one when the mean function and the selection probabilities are known by the "oracle", which we use to establish an asymptotically correct SCB for the variance function. The findings of our empirical finite-sample studies support our asymptotic theory. An application to a data set from the Canada 2010/2011 Youth Student Survey illustrates the usefulness of the proposed techniques.

Key words and phrases: B-spline regression, local linear regression, missing at random, oracle efficiency, simultaneous confidence band.

1. Introduction

Variance function estimation is an important procedure in many statistical analyses, such as stochastic control, risk analysis, the construction of confidence intervals for a regression function, and the estimation of smoothing parameters. Research on inferences for the variance function includes the works of Hall and Carroll (1989) and Wang et al. (2008), who studied the effect of the unknown mean on the estimation of the variance function, and Müller and Stadtmüller (1987, 1993), Brown and Levine (2007), and Cai and Wang (2008), who considered difference-based adaptive nonparametric estimators of the variance function.

Corresponding author: Suojin Wang, Department of Statistics, Texas A&M University, College Station, TX 77843, USA. E-mail: sjwang@stat.tamu.edu.

In addition, Ruppert et al. (1997), Fan and Yao (1998), Song and Yang (2009), and Cai and Yang (2015) estimated the variance function by applying residual-based nonparametric methods, and Ziegelmann (2002) and Yu and Jones (2004) derived likelihood-based local linear estimators of the conditional variance function. Moreover, Zhang (2013) studied a residual-based estimation for the variance function of functional data.

Most existing works have focused on complete data sets, in which both the response and the covariate variables are fully observed, without missing values. In practice, however, problems arise when missing observations are present, which is common in research areas such as psychology, biomedicine, environmental science, and socioeconomy. See, for instance, Sun and Wang (2019) and Cai et al. (2021) for examples of missing covariates in the field of psychological sciences, Särndal and Lundström (2005) and Liang, Wang and Carroll (2007) for missing response examples in the field of social and clinical studies, and Meng (2000) for an example of missing responses in genetics. If such missing values are not completely missing at random (MAR), any statistical results based on the complete case, ignoring the missing data, are generally biased. It is thus important to handle missing data properly.

We consider the following heteroscedastic nonparametric model:

$$Y = g(X) + \varepsilon, \tag{1.1}$$

with the observations of covariate X partially missing, where $\mathrm{E}(\varepsilon|X = x) = 0, \mathrm{E}\left(\varepsilon^2|X = x\right) = \sigma^2(x)$, and $g(x)$ and $\sigma^2(x)$ are unknown conditional mean and variance functions, respectively, defined on a compact interval $[a, b]$. Let the observations $\{X_i, Y_i\}_{i=1}^n$ and the unobserved errors $\{\varepsilon_i\}_{i=1}^n$ be independent and identically distributed (i.i.d.) copies of (X, Y, ε) from (1.1), and denote δ_i, for $1 \le i \le n$, as binary indicator variables, with $\delta_i = 1$ if X_i is observed, and $\delta_i = 0$ otherwise. Furthermore, let

$$\pi_i = P(\delta_i = 1|X_i, Y_i) = P(\delta_i = 1|Y_i) = \pi(Y_i)$$

be the selection probability, which, conditional on Y_i, does not depend on X_i, that is, X_i is MAR. This MAR assumption is common in missing data analyses; see Liang et al. (2004) and Pérez-González, Vilar-Fernández and González-Manteiga (2010), among others.

In many applications, we are interested in examining the overall shape of the noise variance function or testing whether certain functional forms are adequate in describing its global trend. This is also our main focus in this study.

Specifically, we aim to provide an accurate global inference tool–a simultaneous confidence band (SCB)–for $\sigma^2(x)$ when the covariates are MAR. Existing SCB studies tend to focus on fully observed data. For example, Härdle (1989) and Xia (1998) constructed SCBs for the univariate kernel regression. Wang and Yang (2009) proposed SCBs for the polynomial spline regression, and Cai, Low and Ma (2014) derived adaptive SCBs for wavelet smoothing. Later, Gu and Yang (2015) extended these SCBs to the link function for a single-index model, Zheng et al. (2016) to generalized additive models, and Song and Yang (2009) and Cai and Yang (2015) to the variance function in nonparametric models. Moreover, Degras (2011), Cao, Yang and Todem (2012), Ma, Yang and Carroll (2012), and Wang et al. (2020) proposed SCBs for functional data. Zhao and Wu (2008) considered SCBs for nonparametric time series regression. For partially missing data, Al-Sharadqah and Mojirsheibani (2019) and Cai et al. (2021) studied SCBs for the kernel-type regression for the mean function.

However, to the best of our knowledge, there are no related global inference studies for the variance function when the data are not completely observed. In the existing literature, Chen and Shao (2001) proposed a jackknife estimator for the variance function when the response variable is MAR. Pérez-González, Vilar-Fernández and González-Manteiga (2010) proposed local polynomial estimators based on the decomposition $\sigma^2(x) = \mathrm{E}(Y^2|X = x) - g^2(x)$ for the conditional variance function in a fixed-design nonparametric regression in which the response variable is MAR. Nevertheless, both only studied their corresponding pointwise properties.

Here, we construct an oracle-efficient SCB as a global inference tool for the variance function for missing covariate data. In the estimation procedure, we first employ a Horvitz–Thompson-type weighted spline regression to estimate the regression function. Based on the squared residuals, we then apply the weighted kernel smoothing to fit the variance function. The proposed estimator is shown to be oracle-efficient in the sense that it is as efficient as the ideal one obtained when the selection probabilities and the mean function are known by the "oracle". Using this oracle efficiency, an asymptotic accurate SCB is constructed for the variance function. The proposed SCB is smooth because it comes from a kernel regression. It also has an excellent convergence rate for the kernel smoothing and good coverage probabilities, generally close to the nominal level when the sample sizes are moderately large.

The remainder of this paper is organized as follows. Our main estimation procedure and theoretical results are described in Section 2. Some insights of the proofs of our asymptotic theory are given in Section 3. Specific implementation

steps to apply the proposed method are provided in Section 4. Simulation results and a real-data analysis are discussed in Section 5. Section 6 concludes the paper. Technical proofs are provided in the Supplementary Material.

2. Main Results

In this section, we present our bias-corrected estimation procedure for the variance function, as well as its uniform theoretical properties. By convention, for any t-dimensional vector $\mathbf{v}$ with ith entry being v_i, we write $\mathbf{v} = (v_1, \ldots, v_t)^T$ as a column vector.

According to the nonparametric model (1.1), one has $\varepsilon = Y - g(X)$ and $\varepsilon_i = Y_i - g(X_i)$, for $1 \le i \le n$. Denote $R = \varepsilon^2$ and $R_i = \varepsilon_i^2$ as the squared errors. Clearly, if the mean function $g(x)$ were known by the oracle, one could compute R_i at the observable points X_i. Then, by the fact that $\mathrm{E}((\delta_i/\pi_i)R_i|X_i) = \sigma^2(X_i)$, to accommodate the missingness, one could employ the bias-corrected Horvitz and Thompson (1952) type Nadaraya–Watson kernel method to estimate the variance function $\sigma^2(x)$; that is, for each x, we estimate $\sigma^2(x)$ by minimizing the quantity

$$\underset{c_0 \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} (R_i - c_0)^2 K_h(X_i - x)$$

if $g(X_i)$ and π_i are known, where $h = h_n$ is a sequence of smoothing parameters called the bandwidth, and $K_h(u) = K(u/h)/h$ is a recaled kernel function by the bandwidth h. Denote the resulting estimator as $\tilde{\sigma}^2_{\mathrm{NW}}(x)$. A simple least squares calculation leads to

$$\tilde{\sigma}^2_{\mathrm{NW}}(x) = n^{-1} \tilde{f}_X^{-1}(x) \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_h(X_i - x) R_i, \tag{2.1}$$

where $\tilde{f}_X(x) = n^{-1} \sum_{i=1}^{n} (\delta_i/\pi_i) K_h(X_i - x)$. Obviously, this would-be estimator $\tilde{\sigma}^2_{\mathrm{NW}}(x)$ is infeasible, because it relies on the unavailable mean function $g(x)$ and the (usually) unknown selection probabilities π_i. However, it is useful as a benchmark against which feasible estimates can be compared, and as a pivotal medium by which an asymptotic SCB can be constructed for the variance function.

To mimic the infeasible estimator $\tilde{\sigma}^2_{\mathrm{NW}}(x)$, a new bias-corrected spline-kernel estimator for $\sigma^2(x)$ is proposed:

$$\hat{\sigma}^2_{\mathrm{SNW}}(x) = n^{-1} \hat{f}_X^{-1}(x) \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} K_h(X_i - x) \hat{R}_i, \tag{2.2}$$

where $\hat{f}_X(x) = n^{-1}\sum_{i=1}^{n}(\delta_i/\hat{\pi}_i)K_h(X_i - x)$, $\hat{\pi}_i = \hat{\pi}(Y_i)$ is an estimator of $\pi(Y_i)$, and $\hat{\mathbf{R}} = (\hat{R}_1, \ldots, \hat{R}_n)^T$, with $\hat{R}_i = \hat{\varepsilon}_i^2 = (Y_i - \hat{g}_p(X_i))^2$, for $1 \le i \le n$, being the squared residuals. Here, the inverse selection weighted spline estimator $\hat{g}_p(X_i)$ for $g(X_i)$ is given in (2.3). In particular, the selection probabilities are estimated by applying the usual maximum likelihood approach, with the assumption that the selection probabilities follow a parametric binary model, such as a logistic or probit binary regression model (see Hosmer and Lemeshow (2005) for a global test statistic for examining a pre-assumed binary regression model). Then, the resulting estimates $\hat{\pi}_i$ for π_i are known to be root-n consistent. Furthermore, $\hat{g}_p(\cdot)$ is the weighted spline estimator for the mean function $g(\cdot)$ with

$$\hat{g}_p(\cdot) = \underset{m \in G_N^{(p-2)}[a,b]}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} \{Y_i - m(X_i)\}^2, \tag{2.3}$$

where $G_N^{(p-2)}$ is the spline space of functions that are piecewise polynomials of degree $(p-1)$ for some positive integer $p > 0$, defined as follows.

Divide the interval $[a, b]$ into $(N+1)$ subintervals $\{S_j = [\chi_j, \chi_{j+1})\}_{j=0}^{N-1}$ and $S_N = [\chi_N, \chi_{N+1}]$ using a sequence of equally spaced points $\{\chi_j\}_{j=1}^{N}$, called interior knots, given as

$$a = \chi_0 < \chi_1 < \cdots < \chi_N < \chi_{N+1} = b.$$

The $G_N^{(p-2)} = G_N^{(p-2)}[a, b]$ is the space of functions that are polynomials of degree $(p-1)$ on each subinterval S_j and that have a continuous $(p-2)$th derivative on the interval $[a, b]$. For instance, $G^{(-1)}$ represents the space of constant functions on each S_j, and $G^{(0)}$ is the space of functions that are linear on each S_j and continuous on $[a, b]$. Following de Boor (2001), denote the spline basis of $G_N^{(p-2)}$ as $B_{J,p}(x), 1-p \le J \le N$. One has $G_N^{(p-2)} = \{\sum_{J=1-p}^{N} \alpha_{J,p} B_{J,p}(x), \alpha_{J,p} \in \mathbb{R}, x \in [a, b]\}$. It is straightforward that for any $x \in [a, b]$, at most $(p+1)$ of the numbers of $B_{1-p,p}(x), \ldots, B_{N,p}(x)$ are between zero and one, with the remainder being zero. At the same time, algebra shows that the spline space $G_N^{(p-2)}$ can also be spanned by the truncated power basis $\{1, x, \ldots, x^{p-1}, (x - \chi_j)_+^{p-1}, 1 \le j \le N\}$, which is often used in the practice; see de Boor (2001) for details.

The estimator $\hat{g}_p(x)$ in (2.3) is defined in terms of $\{B_{j,p}(x)\}_{j=1-p}^{N}$ as follows:

$$\hat{g}_p(x) = \sum_{J=1-p}^{N} \hat{\lambda}_{J,p} B_{J,p}(x),$$

where the coefficients $\hat{\lambda}_{1-p,p}, \ldots, \hat{\lambda}_{N,p}$ are the solution of the following least-squares problem:

$$\left(\hat{\lambda}_{1-p,p}, \ldots, \hat{\lambda}_{N,p}\right)^T = \underset{(\lambda_{1-p,p},\ldots,\lambda_{N,p})\in R^{N+p}}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} \left\{ Y_i - \sum\nolimits_{J=1-p}^{N} \lambda_{J,p} B_{J,p}\left(X_i\right) \right\}^2. \tag{2.4}$$

Simple calculations yield

$$\hat{g}_p\left(x\right) = \left(B_{1-p,p}(x), \ldots, B_{N,p}(x)\right) \left(\mathbf{B}^T \hat{\mathbf{\Delta}} \mathbf{B}\right)^{-1} \mathbf{B}^T \hat{\mathbf{\Delta}} \mathbf{Y}, \tag{2.5}$$

where

$$\mathbf{B} = \begin{pmatrix} B_{1-p,p}(X_1) \cdots B_{N,p}(X_1) \\ \vdots \quad \ddots \quad \vdots \\ B_{1-p,p}(X_n) \cdots B_{N,p}(X_n) \end{pmatrix}, \quad \hat{\mathbf{\Delta}} = \operatorname{diag}\left(\frac{\delta_1}{\hat{\pi}_1}, \ldots, \frac{\delta_n}{\hat{\pi}_n}\right)^T,$$

and $\mathbf{Y} = \left(Y_1, \ldots, Y_n\right)^T$.

Recall that $\tilde{\sigma}^2_{\mathrm{NW}}$ in (2.1) and $\hat{\sigma}^2_{\mathrm{SNW}}\left(x\right)$ in (2.2) are obtained by applying the weighted Nadaraya–Watson kernel method to smooth the samples $(X_i, R_i)_{i=1}^n$ and $(X_i, \hat{R}_i)_{i=1}^n$, respectively, which ensures that the resulting variance estimator is nonnegative. Of course, one can also employ the weighted spline local linear method to estimate $\sigma^2(x)$. Specifically, the infeasible local linear estimator is

$$\tilde{\sigma}^2_{\mathrm{LL}}\left(x\right) = e_1^T (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{R},$$

where $e_1 = (1, 0)^T$, $\mathbf{W} = \operatorname{diag}((\delta_1/\pi_1) K_h\left(X_1 - x\right), \ldots, (\delta_n/\pi_n) K_h\left(X_n - x\right))$, and

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{pmatrix}, \quad \mathbf{R} = \left(R_1, \ldots, R_n\right)^T.$$

The feasible weighted local linear estimator mimicking $\tilde{\sigma}^2_{\mathrm{LL}}\left(x\right)$ is defined as

$$\hat{\sigma}^2_{\mathrm{SLL}}\left(x\right) = e_1^T (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{W}} \hat{\mathbf{R}},$$

where $\hat{\mathbf{W}} = \operatorname{diag}((\delta_1/\hat{\pi}_1) K_h\left(X_1 - x\right), \ldots, (\delta_n/\hat{\pi}_n) K_h\left(X_n - x\right))$ and $\hat{\mathbf{R}} = (\hat{R}_1, \ldots, \hat{R}_n)^T$, with $\hat{R}_i = \left(Y_i - \hat{g}_p\left(X_i\right)\right)^2$, for $1 \le i \le n$. Note that if one employs the weighted local linear estimator instead of the weighted Nadaraya–Watson kernel estimator in the second estimation step, the obtained variance estimator would no longer be guaranteed to be nonnegative. However, it has advantages of automatic

boundary correction, design adaption, and higher asymptotic efficiency; see Fan and Gijbels (1996) for details. In the following, for brevity, we use $\tilde{\sigma}^2(x)$ as a generic term to denote both $\tilde{\sigma}^2_{\mathrm{NW}}(x)$ and $\tilde{\sigma}^2_{\mathrm{LL}}(x)$, and $\hat{\sigma}^2(x)$ to represent the corresponding $\hat{\sigma}^2_{\mathrm{SNW}}(x)$ and $\hat{\sigma}^2_{\mathrm{SLL}}(x)$, when there is no confusion.

Denote the error term $Z = R - \sigma^2(X)$, and let $f_{X,Z}(x,z)$ be the joint density function of (X,Z). For sequences of real numbers p_n and q_n, we write $p_n \ll q_n$ to mean $p_n/q_n \to 0$ as $n \to \infty$, and write $p_n \sim q_n$ to mean there exists a constant $d \neq 0$ such that $\lim_{n\to\infty} p_n/q_n = d$. Denote $C^{(p)}[a,b]$ as the space of functions that have a continuous pth derivative on the interval $[a,b]$. We use the following general conditions for our theoretical development.

(A1) *The mean function* $g(x) \in C^{(p)}[a,b], p \geq 2$. *The density function* $f_X(x)$ *of* X *is positive in the open interval* (a,b), *with* $f_X(x) \in C^{(1)}[a,b]$, *and* $f_{X,Z}(x,z)$ *has a continuous first-order partial derivative with respect to* x.

(A2) *The variance function* $\sigma^2(x) \in C^{(1)}[a,b]$ *and* $\mathrm{E}\left(Z^2|\delta = 1, X = x\right)$ *has a positive lower bound on* $[a,b]$. *In addition, there exists a constant* $\eta > 4$ *such that* $\mathrm{E}(|Z|^{2+\eta}\,|X)$ *is bounded.*

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

(A4) $\pi(y)$ *follows a parametric binary model and has a positive lower bound denoted by* c_π. *Moreover, it has a bounded second-order partial derivative with respect to* y, *and has a bounded first-order partial derivative with respect to the parameters.*

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

(A6) *The number of interior knots* N *satisfies* $\max\{n^{1/(4p)}, h^{-1/2(p-1)} \times \log^{1/2(p-1)} n\} \ll N \ll \min\{n^{1/2}\log^{-1} n, n^{1/3}h^{1/3}\log^{-2/3} n\}$.

Assumptions (A1)–(A3) are elementary conditions for nonparametric spline and kernel smoothing, and are adapted from Fan and Yao (1998), Song and Yang (2009), and Cai et al. (2021). Assumption (A4) ensures the root-n consistency that $\sup_y |\hat{\pi}(y) - \pi(y)| = O_p(n^{-1/2})$. Assumptions (A5) and (A6) concern the choice of the bandwidth h for the weighted kernel smoothing and the number of interior knots N for the weighted B-spline regression, respectively. Specifically, one can take any undersmoothing bandwidth $h \sim n^{-1/5}\log^{-t} n$ for $t > 1/5$ and any optimal interior knots $N \sim n^{1/(2p+1)}$ fulfilling Assumptions (A5) and (A6). The detailed data-driven procedure is given in Section 4.

Let $[a_0, b_0] \subset (a, b)$ be an arbitrary compact subinterval of interest. Applying the weighted local linear SCB theory in Theorem 2 of Cai et al. (2021) to the sample $(\delta_i, X_i, R_i)_{i=1}^n$, one easily obtains the following result.

Theorem 1. *Under Assumptions* (A1)–(A5), *for any* $t \in \mathbb{R}$,

$$P\left\{ a_h \left[\sup_{x\in[a_0,b_0]} \left| (nh)^{1/2} \kappa_n^{-1/2} v^{-1/2}(x) \left\{ \tilde{\sigma}^2(x) - \sigma^2(x) \right\} \right| - b_h \right] \le t \right\}$$
$$\to \exp(-2\exp(-t)), \tag{2.6}$$

as $n \to \infty$, *where* $\kappa_n = n^{-1} \sum_{i=1}^n \delta_i$,

$$a_h = \left\{ -2\log\left(\frac{h}{b_0 - a_0} \right) \right\}^{1/2}, \quad b_h = a_h + 2^{-1} a_h^{-1} \log(4^{-1}\pi^{-2} C_K),$$
$$C_K = \frac{\int (K^{(1)}(u))^2 du}{\int K^2(u) du}, \quad v(x) = d(x) f_X^{-2}(x) \int K^2(u) du,$$
$$d(x) = \int \frac{(\varepsilon^2 - \sigma^2(x))^2}{\pi^2 (m(x) + \varepsilon)} f_{X,\varepsilon|\delta=1}(x, \varepsilon) d\varepsilon.$$

In Theorem 1, $f_{X,\varepsilon|\delta=1}(x, \varepsilon)$ is the conditional joint density function of the observed (X, ε), and π in b_h is a mathematical constant $3.14\cdots$, to be distinguished from the selection probability function $\pi(y)$. Thus, an asymptotic $100(1-\alpha)\%$ infeasible SCB for $\sigma^2(x)$, $x \in [a_0, b_0]$ is

$$\tilde{\sigma}^2(x) \pm (nh)^{-1/2} \kappa_n^{1/2} v^{1/2}(x) \left(a_h^{-1} q_{1-\alpha} + b_h \right), \tag{2.7}$$

where $q_{1-\alpha} = -\log((-1/2)\log(1-\alpha))$.

Theorem 2. *Under Assumptions* (A1)–(A6), *as* $n \to \infty$,

$$\sup_{x\in[a,b]} \left| \hat{\sigma}^2(x) - \tilde{\sigma}^2(x) \right| = O_p(n^{-1/2}).$$

Theorem 2 shows that, regardless of whether the selection probabilities and the mean function are known by the oracle or are estimated by the proposed method, the estimator $\hat{\sigma}^2(x)$ for $\sigma^2(x)$ is asymptotically as efficient as the ideal one $\tilde{\sigma}^2(x)$.

Theorem 2 and Slutsky's Theorem, together with (2.6), imply the following result.

Theorem 3. *Under Assumptions* (A1)–(A6)*, for any* $t \in \mathbb{R}$*,*

$$P\left\{a_h\left[\sup_{x\in[a_0,b_0]}\left|(nh)^{1/2}\kappa_n^{-1/2}v^{-1/2}(x)\left\{\hat{\sigma}^2(x)-\sigma^2(x)\right\}\right|-b_h\right]\le t\right\}$$
$$\to \exp(-2\exp(-t)),$$

as $n\to\infty$*. Hence, for any* $\alpha\in(0,1)$*, an asymptotic* $100(1-\alpha)\%$ *oracle-efficient SCB for* $\sigma^2(x), x\in[a_0,b_0]$ *is*

$$\hat{\sigma}^2(x)\pm(nh)^{-1/2}\kappa_n^{1/2}v^{1/2}(x)\left(a_h^{-1}q_{1-\alpha}+b_h\right).$$

Note that the theoretical SCBs above for $\sigma^2(x)$ rely on the unknown quantity $v(x)$. Following Cai et al. (2021), we estimate $v(x)$ by

$$\hat{v}(x)=\kappa_n^{-1}h_0\hat{f}_X^{-2}(x)\,n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i^2}K_{h_0}^2(X_i-x)\hat{Z}_i^2,$$

where $\hat{Z}_i=\hat{R}_i-\hat{\sigma}^2(X_i)$. The bandwidth for $\hat{f}_X(x)$ here may be different from that used in (2.2); we recommend using Silverman's rule-of-thumb (ROT) bandwidth in Silverman (1998, p.48) with the order of $n^{-1/5}$ computed using the complete data. According to Theorem 5 of Cai et al. (2021), under the condition that $n^{-1/3}\log n\ll h_0\ll n^{-1/5}\log^{1/5}n$, we have

$$\sup_{x\in[a_0,b_0]}|\hat{v}(x)-v(x)|=O_p(n^{-1/2}h_0^{-3/2}\log^{1/2}n).$$

In the simulation studies, we found $h_0=2h$ (where h is given in the implementation section (Section 4)) works quite well, and is what we recommend. Because $n^{-1/2}h_0^{-3/2}\log^{1/2}n\ll\log^{-1}n$, the following main result is obtained by applying Slutsky's Theorem.

Theorem 4. *Under Assumptions* (A1)–(A6)*, for any* $\alpha\in(0,1)$*, an asymptotic* $100(1-\alpha)\%$ *SCB for* $\sigma^2(x)$ *over* $[a_0,b_0]$ *is*

$$\hat{\sigma}^2(x)\pm(nh)^{-1/2}\kappa_n^{1/2}\hat{v}^{1/2}(x)\left(a_h^{-1}q_{1-\alpha}+b_h\right). \tag{2.8}$$

3. Error Decomposition

In this section, we show in detail that the weighted two-step estimator $\hat{\sigma}^2(x)$ is asymptotically as efficient as the infeasible one $\tilde{\sigma}^2(x)$ in Theorem 2. For clarity, we focus on proving the part between $\tilde{\sigma}^2_{\mathrm{NW}}(x)$ and $\hat{\sigma}^2_{\mathrm{SNW}}(x)$. The proof of that for $\tilde{\sigma}^2_{\mathrm{SLL}}(x)$ and $\hat{\sigma}^2_{\mathrm{SLL}}(x)$ is similar, but with more tedious arguments; see the detailed

discussion below Proposition 1. To study the estimate error $\hat{\sigma}^2_{\mathrm{SNW}}(x) - \tilde{\sigma}^2_{\mathrm{NW}}(x)$, we first discuss the spline space $G_N^{(p-2)}$ and the representation of the weighted spline estimator $\hat{g}_p(x)$ for $g(x)$.

Write $\mathbf{Y}$ as the sum of a signal vector $\mathbf{g}$ and a noise vector $\mathbf{E}$:

$$\mathbf{Y} = \mathbf{g} + \mathbf{E}, \ \ \mathbf{g} = (g(X_1), \ldots, g(X_n))^T, \ \ \mathbf{E} = (\varepsilon_1, \ldots, \varepsilon_n)^T.$$

Projecting this relationship into the space $G_N^{(p-2)}$, one has

$$\hat{g}_p(x) = \tilde{g}_p(x) + \tilde{\varepsilon}_p(x), \tag{3.1}$$

where

$$\tilde{g}_p(x) = \sum_{J=1-p}^{N} \tilde{\alpha}_{J,p} B_{J,p}(x), \ \ \tilde{\varepsilon}_p(x) = \sum_{J=1-p}^{N} \tilde{\beta}_{J,p} B_{J,p}(x).$$

The vectors $(\tilde{\alpha}_{1-p,p}, \ldots, \tilde{\alpha}_{N,p})^T$ and $(\tilde{\beta}_{1-p,p}, \ldots, \tilde{\beta}_{N,p})^T$ are solutions to (2.4), but with Y_i replaced by $g(X_i)$ and ε_i, respectively.

For our theoretical development, we introduce another infeasible weighted estimator $\hat{\sigma}^{*2}_{\mathrm{SNW}}(x)$ based on the true selection probabilities, defined as

$$\hat{\sigma}^{*2}_{\mathrm{SNW}}(x) = n^{-1} \tilde{f}_X^{-1}(x) \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_h(X_i - x) \hat{R}_i^*,$$

where $\hat{R}_i^* = (\hat{R}_1^*, \ldots, \hat{R}_n^*)^T$ and $\hat{R}_i^* = \hat{\varepsilon}_i^{*2} = (Y_i - \hat{g}_p^*(X_i))^2$, for $1 \le i \le n$, which are the squared residuals from the following weighted spline estimator:

$$\hat{g}_p^*(\cdot) = \underset{m \in G_N^{(p-2)}[a,b]}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \{Y_i - m(X_i)\}^2. \tag{3.2}$$

Therefore,

$$\hat{g}_p^*(x) = (B_{1-p,p}(x), \ldots, B_{N,p}(x)) (\mathbf{B}^T \mathbf{\Delta} \mathbf{B})^{-1} \mathbf{B}^T \mathbf{\Delta} \mathbf{Y}, \tag{3.3}$$

where $\mathbf{\Delta} = \mathrm{diag}(\delta_1/\pi_1, \ldots, \delta_n/\pi_n)^T$, and $\mathbf{B}$ is given in (2.5). Similar to (3.1), one has the decomposition $\hat{g}_p^*(x) = \tilde{g}_p^*(x) + \tilde{\varepsilon}_p^*(x)$, where

$$\tilde{g}_p^*(x) = \sum_{J=1-p}^{N} \tilde{\alpha}_{J,p}^* B_{J,p}(x), \ \ \tilde{\varepsilon}_p^*(x) = \sum_{J=1-p}^{N} \tilde{\beta}_{J,p}^* B_{J,p}(x), \tag{3.4}$$

in which the vectors $(\tilde{\alpha}^*_{1-p,p}, \ldots, \tilde{\alpha}^*_{N,p})^T$ and $(\tilde{\beta}^*_{1-p,p}, \ldots, \tilde{\beta}^*_{N,p})^T$ are the solutions to (3.2), but with Y_i replaced by $g(X_i)$ and ε_i, respectively.

Taking the difference between $\hat{\sigma}^{*2}_{\mathrm{SNW}}(x)$ and $\tilde{\sigma}^{2}_{\mathrm{NW}}(x)$, one gets

$$\begin{aligned}
&\hat{\sigma}^{*2}_{\mathrm{SNW}}(x) - \tilde{\sigma}^{2}_{\mathrm{NW}}(x) \\
&= n^{-1}\tilde{f}_X^{-1}(x)\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i - x)\left(\hat{R}_i^* - R_i\right) \\
&= n^{-1}\tilde{f}_X^{-1}(x)\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i - x)\left\{\left(g(X_i) + \varepsilon_i - \tilde{g}_p^*(X_i) - \tilde{\varepsilon}_p^*(X_i)\right)^2 - \varepsilon_i^2\right\} \\
&= I_1(x) + I_2(x) + I_3(x), \qquad (3.5)
\end{aligned}$$

where

$$\begin{aligned}
I_1(x) &= \tilde{f}_X^{-1}(x)\, n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i - x)\left(g(X_i) - \tilde{g}_p^*(X_i)\right)^2 \\
&\quad + \tilde{f}_X^{-1}(x)\, n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i - x)\,\tilde{\varepsilon}_p^{*2}(X_i) \\
&\quad - 2\tilde{f}_X^{-1}(x)\, n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i - x)\left(g(X_i) - \tilde{g}_p^*(X_i)\right)\tilde{\varepsilon}_p^*(X_i), \\
I_2(x) &= 2\tilde{f}_X^{-1}(x)\, n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i - x)\left(g(X_i) - \tilde{g}_p^*(X_i)\right)\varepsilon_i, \\
I_3(x) &= -2\tilde{f}_X^{-1}(x)\, n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i - x)\,\varepsilon_i\tilde{\varepsilon}_p^*(X_i).
\end{aligned}$$

Likewise, one has

$$\begin{aligned}
&\hat{\sigma}^{2}_{\mathrm{SNW}}(x) - \hat{\sigma}^{*2}_{\mathrm{SNW}}(x) \\
&= n^{-1}\hat{f}_X^{-1}(x)\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}K_h(X_i - x)\,\hat{R}_i - n^{-1}\tilde{f}_X^{-1}(x)\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(X_i - x)\,\hat{R}_i^* \\
&= J_1(x) + J_2(x) + J_3(x), \qquad (3.6)
\end{aligned}$$

where

$$\begin{aligned}
J_1(x) &= \hat{f}_X^{-1}(x)n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}K_h(X_i - x)\left(\hat{R}_i - \hat{R}_i^*\right), \\
J_2(x) &= \hat{f}_X^{-1}(x)n^{-1}\sum_{i=1}^{n}\left(\frac{\delta_i}{\hat{\pi}_i} - \frac{\delta_i}{\pi_i}\right)K_h(X_i - x)\,\hat{R}_i^*,
\end{aligned}$$

$$J_3(x) = \left(\hat{f}_X^{-1}(x) - \tilde{f}_X^{-1}(x)\right) n^{-1} \sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_h\left(X_i - x\right) \hat{R}_i^*.$$

Therefore, the estimation error $\hat{\sigma}^2_{\mathrm{SNW}}(x) - \tilde{\sigma}^2_{\mathrm{NW}}(x)$ can be decomposed as

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{SNW}}(x) - \tilde{\sigma}^2_{\mathrm{NW}}(x) &= \hat{\sigma}^2_{\mathrm{SNW}}(x) - \hat{\sigma}^{*2}_{\mathrm{SNW}}(x) + \hat{\sigma}^{*2}_{\mathrm{SNW}}(x) - \tilde{\sigma}^2_{\mathrm{NW}}(x) \\ &= I_1(x) + I_2(x) + I_3(x) + J_1(x) + J_2(x) + J_3(x).\end{aligned}$$

Proposition 1. *Under Assumptions* (A1)–(A6)*, as* $n \to \infty$*, one has*

(a) $\sup_{x\in[a,b]}|I_1(x)| = O_p(N^{-2p} + n^{-1}N\log n)$, $\sup_{x\in[a,b]}|J_1(x)| = O_p(n^{-1/2})$;

(b) $\sup_{x\in[a,b]}|I_2(x)| = O_p(n^{-1/2}h^{-1/2}N^{1-p}\log^{1/2} n)$, $\sup_{x\in[a,b]}|J_2(x)| = O_p(n^{-1/2})$;

(c) $\sup_{x\in[a,b]}|I_3(x)| = O_p(n^{-1}h^{-1/2}N^{3/2}\log n)$, $\sup_{x\in[a,b]}|J_3(x)| = O_p(n^{-1/2})$.

The proof of the results for the weighted spline Nadaraya–Watson estimator in Theorem 2 relies on Proposition 1. In order to prove the corresponding part for the weighted spline local linear estimator, one can extend Proposition 1 by including the terms that contain one more element $(X_i - x)$ in each summation of $I_1(x), I_2(x), I_3(x), J_1(x), J_2(x)$, and $J_3(x)$. These do not add a great deal of difficulty, but are rather lengthy and tedious, and thus are omitted.

4. Implementation

In this section, we describe how to implement the proposed oracle-efficient estimator and the SCBs for the variance function. These will be used in Section 5 for both the simulation studies and the real-data analysis.

The quartic kernel function $K(u) = 15(1-u^2)^2/16$, for $-1 \le u \le 1$, is used for the weighted kernel regression, which satisfies Assumption (A3). We take the range of the covariate variable as $(\hat{a}, \hat{b})$, with $\hat{a} = \min\{X_i : 1 \le i \le n, \delta_i = 1\}$ and $\hat{b} = \max\{X_i : 1 \le i \le n, \delta_i = 1\}$, and the compact interval $[\hat{a}_0, \hat{b}_0] = [0.9\hat{a} + 0.1\hat{b}, 0.1\hat{a} + 0.9\hat{b}]$ is chosen as the subinterval of interest over which the SCBs are constructed for the variance function.

Using formulae for $\tilde{\varepsilon}_p(x)$ and $\tilde{g}_p(x)$ parallel to (S2.1) and (S2.2), respectively, in the Supplementary Material, it is easy to show that the optimal order of the number of interior knots N for the weighted spline regression is $n^{1/(2p+1)}$, satisfying Assumption (A6). In the following, we discuss an approach to select N using the BIC. Denote the predictor for the ith response Y_i by $\hat{Y}_i(N) = \hat{g}_p(X_i)$, for $1 \le i \le n, \delta_i = 1$, which depends on the knot number N. The optimal $\hat{N}^{\mathrm{opt}}$ is obtained by minimizing the following BIC value for N over the range of

$\left[0.05n^{1/(2p+1)}, \min\left\{10n^{1/(2p+1)}, n/4-p\right\}\right]$:

$$\mathrm{BIC} = \log(\mathrm{MSE}) + \frac{2(N+p)\log n}{n},$$

where $\mathrm{MSE} = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i(N))^2$ and $N+p$ is the number of parameters in (2.4). Here, the range of $\left[0.05n^{1/(2p+1)}, \min\left\{10n^{1/(2p+1)}, n/4-p\right\}\right]$ ensures that the chosen knot number $\hat{N}^{\mathrm{opt}}$ is of order $n^{1/(2p+1)}$ and the total number of the parameters in the least square estimation (2.4) is less than $n/4$. Although any spline order $p \geq 2$ may be applied to estimate the mean function $g(x)$, we used cubic splines ($p = 4$) in our empirical studies in Section 5.

To further use a kernel regression to compute $\hat{\sigma}^2_{\mathrm{SLL}}(x)$ and $\hat{\sigma}^2_{\mathrm{SNW}}(x)$, one can take the bandwidth $h = h_{\mathrm{rot}}\log^{-t} n$ for any $t > 1/5$, where h_{rot} is the ROT bandwidth with order of $n^{-1/5}$ computed for the complete case; see Fan and Gijbels (1996), Equation (4.3) for the explicit formula. Clearly, the chosen bandwidth $h \sim n^{-1/5}\log^{-t} n$ satisfies Assumption (A5). In our empirical studies, we applied local linear smoothing in the second estimation step for the variance function. We found in extensive simulations that $h = h_{\mathrm{rot}}\log^{-1/2} n$ works quite well, and hence is what we recommend.

5. Empirical Studies

5.1. Simulation studies

In this section, we investigate the finite-sample behavior of the proposed estimator and the asymptotic SCBs for the conditional variance function.

The following four cases are examined:

Case 1: $g(x) = x^3\exp(x) + 1, \sigma(x) = x^2 + 0.5$;
Case 2: $g(x) = x^3\exp(x) + 1, \sigma(x) = 3\exp(x)/\{2(\exp(2x)+1)\}$;
Case 3: $g(x) = \sin(\pi x) + x^2 + 1, \sigma(x) = x^2 + 0.5$;
Case 4: $g(x) = \sin(\pi x) + x^2 + 1, \sigma(x) = 3\exp(x)/\{2(\exp(2x)+1)\}$.

The covariate X was generated from the uniform distribution $U[0,1]$, and the error ε from the standard normal distribution $N(0,1)$. We considered the following selection probability functions: (1) $\pi_1(y) = \{1+\exp(-2y)\}^{-1}$, leading to $5\% - 12\%$ of X observations missing; and (2) $\pi_2(y) = \{1+\exp(0.9-y)\}^{-1}$, leading to $28\% - 38\%$ of X observations missing. The sample sizes were $n = 200, 400, 600, 800$, and the confidence levels were $1-\alpha = 0.95, 0.99$. We considered the local linear smoothing approach in the second estimation step for the

Table 1. Empirical coverage frequencies of the SCB in (2.8), the SCB in the complete case (SCB-CC), and the infeasible SCB in (2.7) (SCB*), as well as their corresponding average widths (inside parentheses), under missingness mechanism $\pi_1(y)$ with 1,000 replications.

n	$1-\alpha$	Case 1			Case 2		
		SCB	SCB-CC	SCB*	SCB	SCB-CC	SCB*
200	0.95	0.875(1.46)	0.803(1.18)	0.910(1.64)	0.873(0.93)	0.760(0.75)	0.888(1.02)
	0.99	0.962(1.82)	0.930(1.47)	0.971(2.04)	0.959(1.15)	0.916(0.93)	0.966(1.26)
400	0.95	0.932(1.26)	0.840(0.92)	0.944(1.33)	0.916(0.79)	0.789(0.60)	0.932(0.83)
	0.99	0.985(1.55)	0.955(1.14)	0.986(1.64)	0.985(0.97)	0.942(0.74)	0.991(1.01)
600	0.95	0.950(1.14)	0.818(0.79)	0.951(1.21)	0.928(0.72)	0.746(0.52)	0.927(0.74)
	0.99	0.988(1.40)	0.955(0.97)	0.988(1.49)	0.985(0.87)	0.925(0.64)	0.983(0.90)
800	0.95	0.965(1.02)	0.802(0.71)	0.965(1.06)	0.949(0.66)	0.746(0.47)	0.950(0.68)
	0.99	0.995(1.25)	0.961(0.87)	0.995(1.30)	0.994(0.80)	0.941(0.58)	0.994(0.82)
n	$1-\alpha$	Case 3			Case 4		
		SCB	SCB-CC	SCB*	SCB	SCB-CC	SCB*
200	0.95	0.866(1.31)	0.761(1.09)	0.888(1.45)	0.885(0.76)	0.826(0.67)	0.905(0.79)
	0.99	0.951(1.64)	0.891(1.37)	0.960(1.82)	0.971(0.95)	0.941(0.84)	0.979(0.99)
400	0.95	0.897(1.06)	0.758(0.82)	0.911(1.17)	0.930(0.59)	0.864(0.51)	0.932(0.61)
	0.99	0.969(1.32)	0.895(1.03)	0.976(1.45)	0.992(0.74)	0.973(0.64)	0.993(0.76)
600	0.95	0.922(0.93)	0.730(0.70)	0.929(1.02)	0.941(0.52)	0.841(0.44)	0.942(0.53)
	0.99	0.979(1.16)	0.897(0.87)	0.981(1.27)	0.992(0.64)	0.961(0.54)	0.991(0.65)
800	0.95	0.934(0.83)	0.718(0.62)	0.942(0.88)	0.956(0.47)	0.853(0.39)	0.956(0.47)
	0.99	0.983(1.03)	0.881(0.77)	0.987(1.09)	0.995(0.57)	0.973(0.49)	0.993(0.58)

variance function.

We first examine the performance of the proposed SCB for the variance function when the selection probability function is correctly specified. Tables 1 and 2 show the average widths of the SCB in (2.8) and the coverage frequencies with which the true variance function is totally covered by the SCB at the equally spaced points $\{\hat{a}_0 + k(\hat{b}_0 - \hat{a}_0)/400, k = 0, 1, \ldots, 400\}$, with 1,000 replications. For comparison, we have also listed the coverage frequencies of the SCB in the complete case by directly ignoring the missing data, denoted by SCB-CC, and those of the infeasible SCB in (2.7), denoted by SCB*. In the latter case, $v(x)$ is computed in the same way as $\hat{v}(x)$ in (2.8), but using the true mean function and the true selection probability function. Tables 1 and 2 show that in all cases, (i) the coverage frequencies of the proposed SCB approach the confidence levels 0.95 and 0.99 as the sample size n increases, and they are close to those of the infeasible ones, and (ii) the obtained SCBs perform better than those of the SCB-CC, and the widths of the SCBs become narrower as the sample size n increases.

Table 2. Empirical coverage frequencies of the SCB in (2.8), the SCB in the complete case (SCB-CC), and the infeasible SCB in (2.7) (SCB*), as well as their corresponding average widths (inside parentheses), under missingness mechanism $\pi_2(y)$ with 1,000 replications.

n	$1-\alpha$	Case 1			Case 2		
		SCB	SCB-CC	SCB*	SCB	SCB-CC	SCB*
200	0.95	0.846(1.66)	0.793(1.43)	0.902(1.87)	0.800(1.06)	0.793(0.92)	0.877(1.17)
	0.99	0.939(2.07)	0.907(1.78)	0.966(2.33)	0.931(1.31)	0.927(1.14)	0.956(1.45)
400	0.95	0.917(1.36)	0.834(1.09)	0.933(1.45)	0.908(0.90)	0.870(0.739)	0.924(0.95)
	0.99	0.970(1.68)	0.941(1.35)	0.978(1.80)	0.979(1.11)	0.973(0.91)	0.989(1.17)
600	0.95	0.930(1.23)	0.820(0.94)	0.940(1.30)	0.927(0.80)	0.873(0.64)	0.940(0.83)
	0.99	0.981(1.52)	0.940(1.16)	0.989(1.60)	0.981(0.98)	0.967(0.78)	0.988(1.01)
800	0.95	0.950(1.10)	0.832(0.84)	0.951(1.14)	0.946(0.73)	0.881(0.58)	0.956(0.74)
	0.99	0.992(1.36)	0.949(1.03)	0.993(1.40)	0.992(0.88)	0.977(0.71)	0.997(0.91)
n	$1-\alpha$	Case 3			Case 4		
		SCB	SCB-CC	SCB*	SCB	SCB-CC	SCB*
200	0.95	0.813(1.47)	0.752(1.22)	0.876(1.64)	0.838(0.86)	0.809(0.75)	0.890(0.92)
	0.99	0.926(1.85)	0.864(1.53)	0.957(2.06)	0.957(1.08)	0.939(0.94)	0.968(1.16)
400	0.95	0.905(1.22)	0.760(0.92)	0.926(1.32)	0.893(0.69)	0.857(0.58)	0.907(0.71)
	0.99	0.974(1.52)	0.906(1.15)	0.983(1.64)	0.979(0.85)	0.969(0.72)	0.981(0.88)
600	0.95	0.916(1.07)	0.732(0.77)	0.915(1.16)	0.931(0.59)	0.855(0.49)	0.931(0.61)
	0.99	0.970(1.33)	0.872(0.96)	0.979(1.44)	0.982(0.74)	0.968(0.61)	0.986(0.75)
800	0.95	0.944(0.96)	0.704(0.69)	0.943(1.00)	0.944(0.54)	0.856(0.44)	0.946(0.54)
	0.99	0.984(1.19)	0.877(0.85)	0.990(1.25)	0.993(0.66)	0.972(0.54)	0.996(0.67)

All of these results are consistent with our theoretical findings in Theorems 2 and 4.

We next investigate the sensitivity of the proposed SCB to a misspecification of the missingness mechanism. The selection probability functions $\pi_1^\dagger(y) = \{1 + \exp(-0.3 - 1.6y - 0.2y^2)\}^{-1}$ and $\pi_2^\dagger(y) = \{1 + \exp(0.2 - 1.5y - 0.2\sin(y))\}^{-1}$ were used for this purpose. It is thus not completely correct to employ a linear logistic regression to fit the selection probabilities. Tables 3 and 4 show the coverage frequencies and the average widths of the SCB, SCB-CC, and SCB* in these misspecified cases. In all cases, the proposed SCB performs similarly to the infeasible ones and to those under the correct specification of the selection probabilities. Furthermore, the proposed SCB performs better than that in the complete case. This suggests that the proposed method is not very sensitive to a misspecification of the missingness mechanism.

Figure 1 plots the variance estimate $\hat{\sigma}^2_{\mathrm{SLL}}(x)$ (dashed) and the 95% SCB (thick solid) for $\sigma^2(x)$ (solid) in Cases 1–4, based on the sample sizes $n = 400$

Table 3. Empirical coverage frequencies of the SCB in (2.8), the SCB in the complete case (SCB-CC), and the infeasible SCB in (2.7) (SCB*), as well as their corresponding average widths (inside parentheses), with 1,000 replications when the missingness mechanism $\pi_1^{\dagger}(y)$ is misspecified.

n	$1-\alpha$	Case 1			Case 2		
		SCB	SCB-CC	SCB*	SCB	SCB-CC	SCB*
200	0.95	0.901(1.56)	0.831(1.21)	0.924(1.58)	0.896(0.96)	0.803(0.76)	0.902(0.96)
	0.99	0.970(1.95)	0.940(1.51)	0.977(1.96)	0.972(1.19)	0.935(0.94)	0.980(1.19)
400	0.95	0.954(1.32)	0.870(0.94)	0.953(1.25)	0.946(0.83)	0.873(0.62)	0.940(0.79)
	0.99	0.992(1.63)	0.972(1.17)	0.990(1.54)	0.990(1.01)	0.970(0.75)	0.992(0.96)
600	0.95	0.955(1.21)	0.871(0.81)	0.955(1.09)	0.955(0.75)	0.847(0.53)	0.948(0.69)
	0.99	0.993(1.48)	0.967(0.99)	0.990(1.34)	0.998(0.91)	0.957(0.65)	0.998(0.84)
800	0.95	0.969(1.04)	0.884(0.73)	0.968(0.97)	0.960(0.68)	0.811(0.48)	0.946(0.62)
	0.99	0.994(1.34)	0.976(0.89)	0.994(1.18)	0.998(0.83)	0.961(0.59)	0.997(0.75)
n	$1-\alpha$	Case 3			Case 4		
		SCB	SCB-CC	SCB*	SCB	SCB-CC	SCB*
200	0.95	0.881(1.37)	0.794(1.12)	0.895(1.42)	0.868(0.76)	0.822(0.67)	0.892(0.77)
	0.99	0.956(1.71)	0.907(1.40)	0.965(1.78)	0.969(0.95)	0.945(0.84)	0.978(0.97)
400	0.95	0.919(1.13)	0.820(0.84)	0.928(1.11)	0.934(0.61)	0.859(0.52)	0.942(0.60)
	0.99	0.975(1.41)	0.928(1.05)	0.981(1.39)	0.988(0.76)	0.974(0.64)	0.993(0.74)
600	0.95	0.940(1.01)	0.790(0.71)	0.951(0.94)	0.949(0.53)	0.901(0.44)	0.947(0.51)
	0.99	0.990(1.25)	0.927(0.88)	0.990(1.16)	0.994(0.65)	0.971(0.55)	0.996(0.63)
800	0.95	0.958(0.96)	0.789(0.63)	0.959(0.85)	0.952(0.47)	0.879(0.40)	0.952(0.46)
	0.99	0.988(1.19)	0.924(0.78)	0.990(1.05)	0.999(0.59)	0.980(0.49)	0.996(0.56)

(left-sided) and $n = 800$ (right-sided), under the missingness mechanism $\pi_1(y)$. The estimate fits better and the SCB becomes narrower for $n = 800$ than for $n = 400$. The plots for other cases are similar, and hence are omitted.

5.2. Real-data analysis

In this section, we apply the proposed method to data from a 2010/2011 youth student survey. The survey was sponsored by Health Canada, and the target sampling group was pan-Canadian youth students in grades 6–12 between October 2010 and June 2011. Using the survey, Health Canada aimed to provide schools, provinces, communities, and parents with timely and reliable information on tobacco, alcohol, and drug use, as well as other related issues about Canadian students. More details can be found in the 2010–2011 YSS Student Survey Data Codebook and `https://uwaterloo.ca/canadian-student-tobacco-alcohol-drugs-survey`.

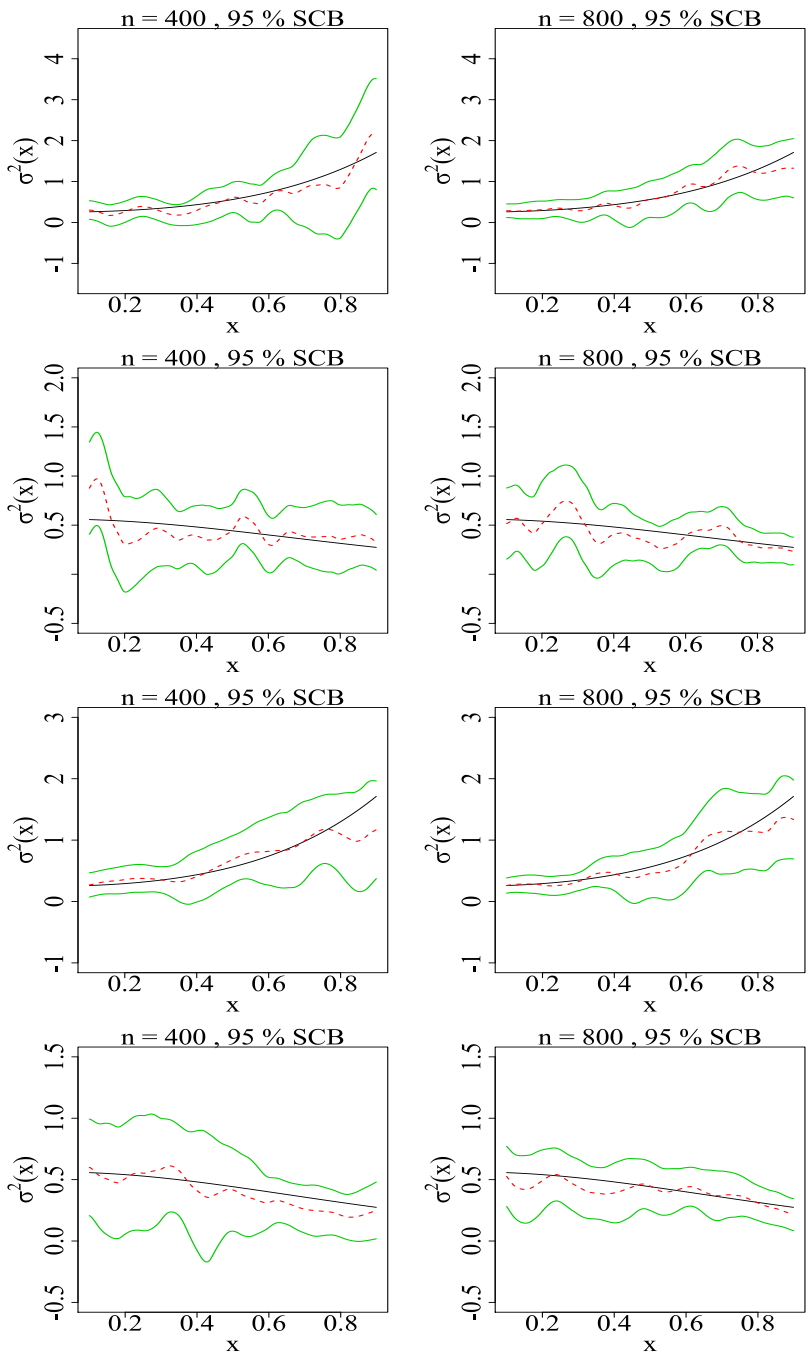

Figure 1. Plots of the variance estimate $\hat{\sigma}^2_{\mathrm{SLL}}(x)$ (dashed) and the 95% SCB (thick solid) for $\sigma^2(x)$ (solid) in Cases 1–4 from the first row to the fourth row, with $n = 400$ (left) and $n = 800$ (right), under missingness mechanism $\pi_1(y)$.

Table 4. Empirical coverage frequencies of the SCB in (2.8), the SCB in the complete case (SCB-CC), and the infeasible SCB in (2.7) (SCB*), as well as their corresponding average widths (inside parentheses), with 1,000 replications when the missingness mechanism $\pi_2^{\dagger}(y)$ is misspecified.

n	$1-\alpha$	Case 1			Case 2		
		SCB	SCB-CC	SCB*	SCB	SCB-CC	SCB*
200	0.95	0.867(1.49)	0.814(1.22)	0.911(1.71)	0.859(0.94)	0.765(0.78)	0.905(1.04)
	0.99	0.958(1.86)	0.931(1.52)	0.976(2.13)	0.959(1.16)	0.918(0.96)	0.979(1.29)
400	0.95	0.926(1.24)	0.840(0.94)	0.933(1.34)	0.932(0.81)	0.826(0.62)	0.938(0.87)
	0.99	0.982(1.53)	0.948(1.17)	0.987(1.65)	0.979(0.99)	0.953(0.76)	0.987(1.07)
600	0.95	0.950(1.12)	0.821(0.81)	0.953(1.21)	0.941(0.71)	0.794(0.54)	0.948(0.75)
	0.99	0.985(1.38)	0.952(0.99)	0.987(1.48)	0.994(0.86)	0.946(0.66)	0.993(0.92)
800	0.95	0.948(0.99)	0.808(0.73)	0.964(1.06)	0.946(0.65)	0.765(0.49)	0.953(0.68)
	0.99	0.995(1.22)	0.963(0.89)	0.993(1.29)	0.992(0.78)	0.949(0.59)	0.998(0.82)
n	$1-\alpha$	Case 3			Case 4		
		SCB	SCB-CC	SCB*	SCB	SCB-CC	SCB*
200	0.95	0.840(1.29)	0.740(1.09)	0.877(1.44)	0.881(0.77)	0.823(0.68)	0.903(0.81)
	0.99	0.946(1.62)	0.882(1.37)	0.958(1.81)	0.967(0.96)	0.949(0.86)	0.980(1.02)
400	0.95	0.914(1.07)	0.768(0.84)	0.929(1.21)	0.925(0.60)	0.869(0.52)	0.932(0.62)
	0.99	0.973(1.33)	0.910(1.04)	0.985(1.50)	0.988(0.75)	0.973(0.65)	0.991(0.78)
600	0.95	0.916(0.92)	0.724(0.71)	0.933(1.03)	0.933(0.52)	0.843(0.44)	0.937(0.54)
	0.99	0.980(1.15)	0.881(0.88)	0.983(1.27)	0.988(0.64)	0.955(0.55)	0.990(0.67)
800	0.95	0.936(0.82)	0.688(0.63)	0.945(0.91)	0.946(0.46)	0.862(0.40)	0.946(0.48)
	0.99	0.990(1.02)	0.871(0.77)	0.995(1.13)	0.994(0.57)	0.968(0.49)	0.995(0.59)

Mental and physical development during adolescence, as is well known, is a factor that may affect quality of life in adulthood. Many existing works in psychology investigate the physical effect of a person's body mass index (BMI) on his or her mental development of self-esteem; see, for example, Agarwal et al. (2013) and Al Ahmari et al. (2019). Here, we concentrate on studying the variance function of self-esteem with BMI. A subset of the data is used, focusing on white female youth students in grades 6 to 12. The data contain a variable for self-esteem, measured by a score ranging from 0 to 12, and one for BMI, ranging from 10 to 50, computed as weight in kilograms over height in meters squared. The sample contains data on 5,343 students, all of whom have complete self-esteem scores, but only 3,565 provide their BMI data (about 33% missing). The missingness of BMI may be closely related to self-esteem, which suggests that it is unlikely to be missing completely at random. However, it is plausible to assume that given self-esteem, the missingness mechanism is no longer dependent

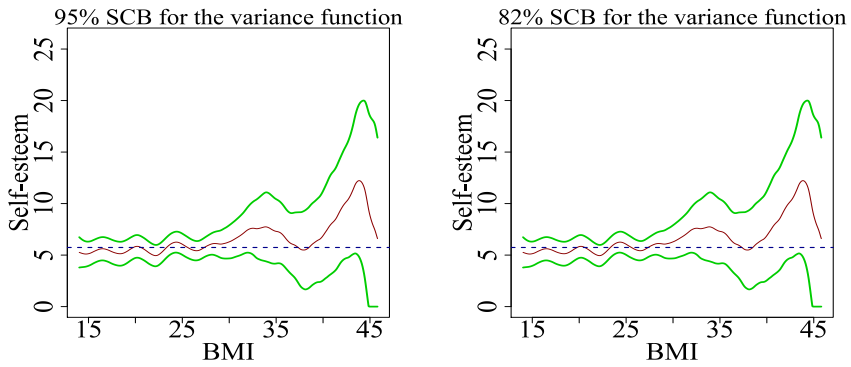

Figure 2. Plots of the variance estimate $\hat{\sigma}^2_{\mathrm{SLL}}(x)$ (solid), SCBs (thick solid), and null hypothesis curve $\sigma^2(x) \equiv n^{-1}\sum_{i=1}^{n}(\delta_i/\hat{\pi}_i)\hat{R}_i$ (dashed) for the youth student survey data.

on BMI itself (at least approximately). In other words, the dependence on X of the missingness mechanism may be expressed through Y, leading to MAR. The Hosmer–Lemeshow goodness-of-fit test was applied to check the linear logistic regression for the selection probabilities. The resulting p-value was 0.17, indicating no evidence of poor fit for the missingness mechanism. Thus, the linear logistic regression was used to model the selection probabilities.

Figure 2 plots the proposed two-step weighted estimator (solid line) and the 95% SCB (thick solid line) in (2.8) for the variance function. The SCB was applied to test the homoscedasticity of the data $H_0 : \sigma^2(x) \equiv \sigma^2$. Under this hypothesis, a consistent estimator $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(\delta_i/\hat{\pi}_i)\hat{R}_i$ for σ^2 was used, as shown in Figure 2 (dashed horizontal line), where $\hat{R}_i = (Y_i - \hat{g}_p(X_i))^2$, for $1 \le i \le n$, are the squared residuals from the weighted spline estimator in (2.3). The null hypothesis curve is completely covered by the 95% SCB, implying that the homoscedasticity of the data cannot be rejected at the significance level of 0.05. Simple calculations conclude that the minimum confidence level containing the null curve is 82%; see Figure 2 on the right panel. Hence, the p-value for the test is 0.18.

6. Conclusion

We have proposed a new two-step bias-corrected spline kernel estimator for the variance function when the covariates are MAR. The estimation procedure synthesizes a spline regression and kernel smoothing to take advantage of the fast computation speed of the former and the uniform convergence property of the latter. The two-step estimator was justified to be oracle-efficient in the sense that it is as efficient as the ideal estimator obtained from using the true mean

function and the true selection probability function. Applying the oracle efficiency and the uniform convergence property of the kernel regression, an asymptotic accurate SCB was constructed for the variance function. Our theoretical findings were supported by finite-sample simulation studies. An analysis of youth student survey data illustrated the usefulness of the proposed confidence band. Further research problems include investigating whether these strategies can be extended to time series (see Fan and Yao (1998), Wang et al. (2014)) and functional data (see, e.g., Yao, Müller and Wang (2005), Cao, Yang and Todem (2012)), as well as to the more complex setting of functional linear regression for functional data.

Supplementary Material

The supplement contains technical proofs for the main results.

Acknowledgments

The authors thank the associate editor and two referees for their helpful comments and suggestions. This research was supported in part by the National Natural Science Foundation of China Award NSFC 11901521, the National Statistical Science Research Key Program 2021LZ01, the Zhejiang Province Universities' Basic Operating Expenses Special Fund XRK21004, the Characteristic & Preponderant Discipline of Key Construction Universities in Zhejiang Province (Zhejiang Gongshang University–Statistics), and the Simons Foundation Mathematics and Physical Sciences Program Award 499650.

References

Agarwal, S., Bhalla, P., Kaur, S. and Babbar, R. (2013). Effect of body mass index on physical self concept, cognition & academic performance of first year medical students. *Indian Journal of Medical Research* **138**, 515–522.

Al Ahmari, T., Alomar, A., Al Beeybe, J., Asiri, N., Al Ajaji, R., Al Masoud, R. et al. (2019). 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* **24**, 1199–1207.

Al-Sharadqah, A. and Mojirsheibani, M. (2019). A simple approach to construct confidence bands for a regression function with incomplete data. *AStA Advances in Statistical Analysis* **104**, 81–99.

Brown, L. and Levine M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics* **35**, 2219–2232.

Cai, L., Gu, L., Wang, Q. and Wang, S. (2021). Simultaneous confidence bands for nonparametric regression with missing covariate data. *Annals of the Institute of Statistical Mathematics* **73**, 1249–1279.

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

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

Cai, T. and Wang, L. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. *The Annals of Statistics* **36**, 2025–2054.

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

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association* **96**, 260–269.

de Boor, C. (2001). *A Practical Guide to Splines.* Springer, New York.

Degras, D. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica* **21**, 1735–1765.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Routledge, New York.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645–660.

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

Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean. *Journal of the Royal Statistical Society Series B (Methodological)* **51**, 3–14.

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

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

Hosmer, D. and Lemeshow, S. (2005). *Applied Logistic Regression.* 2nd Edition. John Wiley & Sons, New York.

Liang, H., Wang, S. and Carroll, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika* **94**, 185–198.

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

Ma, S., Yang, L. and Carroll, R. J. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statistica Sinica* **22**, 95–122.

Meng, X. (2000). Missing data: Dial M for ???. *Journal of the American Statistical Association* **95**, 1325–1330.

Müller, H. G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics* **15**, 610–625.

Müller, H.G. and Stadtmüller, U. (1993). On variance function estimation with quadratic forms. *Journal of Statistical Planning and Inference* **35**, 213–231.

Pérez-González, A., Vilar-Fernández, J. M. and González-Manteiga, W. (2010). Nonparametric variance function estimation with missing data. *Journal of Multivariate Analysis* **101**, 1123–1142.

Ruppert, D., Wand, M. P., Holst, U. and Hössjer, O. (1997). Local polynomial variance-function estimation. *Technometrics* **39**, 262–273.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester.

Silverman, B. (1998). *Density Estimation for Statistics and Data Analysis*. Routledge, New York.

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

Sun, Z. and Wang, S. (2019). Semiparametric estimation in regression with missing covariates using single-index models. *Annals of the Institute of Statistical Mathematics* **71**, 1201–1232.

Wang, J., Cao, G., Wang, L. and Yang, L. (2020). Simultaneous confidence band for stationary covariance function of dense functional data. *Journal of Multivariate Analysis* **176**, 104584.

Wang, J., Liu, R., Cheng, F. and Yang, L. (2014). Oracally efficient estimation of autoregressive error distribution with simultaneous confidence band. *The Annals of Statistics* **42**, 654–668.

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

Wang, L., Brown, L., Cai, T. and Levine, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics* **36**, 646–664.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **60**, 797–811.

Yao, F., Müller, H. and Wang, J. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

Yu, K. and Jones, M. (2004). Likelihood based-local linear estimation of the conditional variance function. *Journal of the American Statistical Association* **99**, 139–144.

Zhang, J. (2013). *Analysis of Variance for Functional Data*. Chapman and Hall, London.

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

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

Ziegelmann, F. A. (2002). Nonparametric estimation of volatility functions: The local exponential estimator. *Econometric Theory* **18**, 985–991.

Li Cai

School of Statistics and Mathematics and Collaborative Innovation Center of Statistical Data Engineering, Technology & Application, Zhejiang Gongshang University, Hangzhou, 310018, China.

E-mail: caili16@126.com

Suojin Wang

Department of Statistics, Texas A&M University, College Station, TX 77843, USA.

E-mail: sjwang@stat.tamu.edu

(Received February 2021; accepted May 2021)