Electronic Journal of Statistics
Vol. 9 (2015) 2793–2827
ISSN: 1935-7524
DOI: 10.1214/15-EJS1080

Variance function additive partial linear models

Yixin Fang

Department of Population Health
New York University School of Medicine
New York, NY 10016, USA
e-mail: yixin.fang@nyumc.org

Heng Lian

School of Mathematics and Statistics
University of New South Wales
Sydney Australia, 2052
e-mail: heng.lian@unsw.edu.au

Hua Liang*

Department of Statistics
George Washington University
801 22nd St. NW
Washington, D.C. 20052, USA
e-mail: hliang@gwu.edu

and

David Ruppert

Department of Statistical Science
Cornell University
Ithaca, New York 14853, USA
e-mail: dr24@cornell.edu

Abstract: To model heteroscedasticity in a broad class of additive partial linear models, we allow the variance function to be an additive partial linear model as well and the parameters in the variance function to be different from those in the mean function. We develop a two-step estimation procedure, where in the first step initial estimates of the parameters in both the mean and variance functions are obtained and then in the second step the estimates are updated using the weights based on the initial estimates. We use polynomial splines to approximate the additive nonparametric components in both the mean and variation functions and derive their convergence rates. The resulting weighted estimators of the linear coefficients in both the mean and variance functions are shown to be asymptotically normal and more efficient than the initial un-weighted estimators. Simulation experiments are conducted to examine the numerical performance of the proposed procedure, which is also applied to analyze the dataset from a nutritional epidemiology study.

*Corresponding author. Liang's research was partially supported by NSF grants DMS-1440121 and DMS-1418042, and by Award Number 11529101, made by National Natural Science Foundation of China.

AMS 2000 subject classifications: Primary 62G08; secondary 62G20, 62J02, 62F12.

Keywords and phrases: Efficiency, heteroscedasticity, generalized least squares, regression spline, variance function.

Received March 2015.

1. Introduction

Additive partial linear models (APLMs) are a generalization of multiple linear regression models, and at the same time they are a special case of generalized additive nonparametric regression models (Hastie and Tibshirani, 1990). As discussed in Liu et al. (2011), APLMs allow an easier interpretation of the effect of each variable. Also, they are preferable to completely nonparametric additive models, since they combine both parametric and nonparametric components when it is believed that the response variable depends on some variables in a linear way but is nonlinearly related to the remaining independent variables.

Estimation and inference for APLMs have been well studied in literature (Opsomer and Ruppert, 1997; Stone, 1985; Opsomer and Ruppert, 1999; Liang et al., 2008; Li, 2000; Liu et al., 2011). However, most existing work focuses on statistical inference for the mean function while variance function estimation has received much less attention. Although a wealth of work has been done to take heteroscedasticity into account for enhancing the efficiency of estimating the parameters in the mean function, estimating variance function is also of independent interest. For example, an appropriate estimator of the variance is needed when one derives confidence intervals/bands for the mean function (Ruppert et al., 2003; Cai and Wang, 2008). Other examples in which the variable function estimation plays an important role include a study of kinetic rate parameters (Box and Hill, 1974), quality control (Box and Meyer, 1986), and a study of social inequality (Western and Bloome, 2009). More recently, Thomas et al. (2012) demonstrated that individual variability in longitudinal measurements for an individual can be predictive of a health outcome, and Teschendorff and Widschwendter (2012) argued that differential variability can be as important as differential means for predicting disease phenotypes in cancer genomes.

In response to these demonstrations of the importance of variance function estimation, many flexible and efficient methods for variance function estimation have been proposed; Carroll (2003) and Carroll and Ruppert (1988) are nice surveys. Representative work on modeling heteroscedasticity in linear or nonlinear models includes Carroll and Härdle (1989), Carroll and Ruppert (1982), Carroll (1982), Hall and Carroll (1989) and Bickel (1978). Motivated by Davidian and Carroll (1987), Lian et al. (2015) studied the variance function partially linear single index models (VF-PLSIMs), in which the variance function is a function of the sum of linear and single index functions and the parameters in the variance function are allowed to be different from those in the mean function. They developed efficient and practical estimators for the parameters in the mean and

variance functions, and weighted the objective function to obtain more efficient estimators for the parameters in the mean function.

In this paper, we consider variance function additive partial linear models (VF-APLMs), a broad class of heteroscedastic regression models where the mean function is an additive partial linear model and the variance function depends upon a generalized additive partial linear model as well. Unlike the classic generalized additive partial linear model (Wang et al., 2011), here we do not insist that the variance function depends only upon the mean function. Suppose that $\{(\boldsymbol{X}_1, \boldsymbol{Z}_1, Y_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Z}_n, Y_n)\}$ is an i.i.d. random sample of size n from the following VF-APLM:

$$\begin{aligned} Y &= \boldsymbol{X}^{\mathcal{T}}\boldsymbol{\alpha} + \sum_{k=1}^{K} g_k(Z_k) + \varepsilon, \\ \varepsilon &= \phi\{\boldsymbol{X}^{\mathcal{T}}\boldsymbol{\beta} + \sum_{k=1}^{K} h_k(Z_k)\}\epsilon, \end{aligned} \tag{1}$$

where $\boldsymbol{X} = (1, \boldsymbol{X}^{*\mathcal{T}})^{\mathcal{T}} = (1, X_1, \ldots, X_d)^{\mathcal{T}}$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_K)^{\mathcal{T}}$ are the linear and nonparametric components, $g_1, \ldots, g_K$ are unknown smooth functions in the mean function, $h_1, \ldots, h_K$ are unknown smooth functions in the variance function, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_d)^{\mathcal{T}}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^{\mathcal{T}}$ are vectors of unknown parameters, and ϵ is independent of $\boldsymbol{X}$ and $\boldsymbol{Z}$ with $E(\epsilon) = 0$ and $E(|\epsilon|) = 1$. Here ϕ is a known function, and generally either $\phi(v) = v$ or $\phi(v) = \exp(v)$. However, using $\phi(v) = v$ will not guarantee that $\phi(v)$ will be positive in practice. Thus in all our numerical examples we will use $\phi(v) = \exp(v)$. To ensure identifiability of the nonparametric functions, we assume that $E\{g_k(Z_k)\} = 0$ and $E\{h_k(Z_k)\} = 0$ for $k = 1, \ldots, K$. We also assume that $E(\boldsymbol{X}^*) = \boldsymbol{0}$, which can be achieved in practice by centering, that is $\boldsymbol{X}_i^* - \sum_{j=1}^{n} \boldsymbol{X}_j^*/n$.

The challenge of investigating model (1), in both theoretical derivation and numerical implementation, is that there could be more than one nonparametric component in both the mean and variance functions. If there is only one nonparametric component in both the mean and variance function, it may use kernel method to estimate the nonparametric components as in VF-PLSIMs investigated by Lian et al. (2015). However, the kernel method cannot be applied directly to estimate the variance parameter in VF-APLMs.

For APLMs, Opsomer and Ruppert (1997) and Stone (1985) proposed a backfitting algorithm and Opsomer and Ruppert (1999) studied the asymptotic properties of the kernel-based backfitting estimators for the parameters in the mean function. Liang et al. (2008) suggested that a kernel-based estimation procedure is available for APLMs without an undersmoothing requirement. When there are multiple nonparametric terms, the kernel-based procedures are computationally inexpedient. Challenged by these demands, Liu et al. (2011) proposed to approximate the nonparametric components with splines. The resulting estimators for the linear components are easily calculated and, of most importance, still asymptotically normal.

In this paper, we use the polynomial-spline procedure (Xue, 2009; Xue and Yang, 2006a,b) for approximating the multiple nonparametric components in both the mean and variance functions. However, we face additional challenges in establishing asymptotic properties for the estimators of parameters in the variance function. It is also worthwhile to point that the development of theory with spline approximation in VF-APLMs is more difficult than for that in the VF-PLSIMs (Lian et al., 2015).

We organize the remaining as follows. In Section 2, we describe in detail the initial and updated estimation procedures for VF-APLMs. In Section 3, we present the main theoretical results and their implications. We examine numerical performance of the proposed method through simulation studies in Section 4 and by the analysis of a real dataset in Section 5. Some discussion is presented in Section 6 and all the technical assumptions and proofs of the theoretical results are placed in the Appendix.

2. Methods

2.1. Spline approximation

In model (1), let $g_0(\boldsymbol{z}) = g_{01}(z_1) + \cdots + g_{0K}(z_K)$ and $\boldsymbol{\alpha}_0$ be the true additive function and parameter for the mean, and let $h_0(\boldsymbol{z}) = h_{01}(z_1) + \cdots + h_{0K}(z_K)$ and $\boldsymbol{\beta}_0$ be the true additive function and parameter for the variance. For simplicity, we assume that the covariate Z_k is distributed on a compact interval $[a_k, b_k]$, $k = 1, \ldots, K$, and without loss of generality, we take all intervals $[a_k, b_k] = [0, 1]$, $k = 1, \ldots, K$. Under some smoothness assumptions, the g_{0k}'s and h_{0k}'s can be well-approximated by spline functions. Although in practice we could consider different sets of spline functions for g_0 and h_0 respectively, for notational simplicity, here we consider a same set of spline functions for both g_0 and h_0.

Let $\mathcal{S}_n$ be the space of polynomial splines on $[0, 1]$ of degree $\varrho \geq 1$. We introduce a knot sequence with J_n interior knots,

$$t_{-\varrho} = \ldots = t_{-1} = t_0 = 0 < t_1 < \ldots < t_{J_n} < 1 = t_{J_n+1} = \ldots = t_{J_n+\varrho+1},$$

where J_n increases with sample size n in some order. Equally spaced knots are used here for simplicity. However, other regular knot sequences can also be used, with similar asymptotic results. Then $\mathcal{S}_n$ consists of functions ξ satisfying

(i) ξ is a polynomial of degree ϱ on each of the subintervals $I_j = [t_j, t_{j+1})$, $j = 0, \ldots, J_n - 1$, $I_{J_n} = [t_{J_n}, 1]$;
(ii) for $\varrho \geq 2$, ξ is $\varrho - 1$ continuously differentiable on $[0, 1]$.

We consider additive spline estimate $\widehat{g}$ of g_0 in the mean and additive spline estimate $\widehat{h}$ of h_0 in the variance based on the independent random sample $(\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i)$, $i = 1, \ldots, n$. Let $\mathcal{A}_n$ be the collection of functions ξ with the additive form $\xi(\boldsymbol{z}) = \xi_1(z_1) + \cdots + \xi_K(z_K)$, where each component function $\xi_k \in \mathcal{S}_n$ and $\sum_{i=1}^n \xi_k(Z_{ik}) = 0$.

2.2. Initial estimator of the mean

The problem of estimating g_0 and $\boldsymbol{\alpha}_0$ in the mean has been already well established if the potential heteroscedasticity is ignored; for example, Liu et al. (2011). We would like to find a function $g \in \mathcal{A}_n$ and a value of $\boldsymbol{\alpha}$ that minimize the following sum of squared residuals function

$$L\left(g, \boldsymbol{\alpha}\right) = \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \left\{g(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\alpha}\right\}\right]^2, \quad g \in \mathcal{A}_n. \tag{2}$$

For the k-th covariate z_k, let $\{b_{j,k}\left(z_k\right) : j = -\varrho, \ldots, J_n\}$ be the B-spline basis functions of degree ϱ. For any $g \in \mathcal{A}_n$, one can write

$$g\left(\boldsymbol{z}\right) = \boldsymbol{\eta}^{\mathcal{T}}\boldsymbol{b}\left(\boldsymbol{z}\right), \tag{3}$$

where $\boldsymbol{b}\left(\boldsymbol{z}\right) = \left\{b_{j,k}\left(z_k\right), j = -\varrho, \ldots, J_n, k = 1, \ldots, K\right\}^{\mathcal{T}}$, and the spline coefficient vector $\boldsymbol{\eta} = \left\{\eta_{j,k}, j = -\varrho, \ldots, J_n, k = 1, \ldots, K\right\}^{\mathcal{T}}$. Thus the minimization problem in (2) is equivalent to finding $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ to minimize

$$\ell(\boldsymbol{\eta}, \boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \left\{\boldsymbol{\eta}^{\mathcal{T}}\boldsymbol{b}\left(\boldsymbol{Z}_i\right) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\alpha}\right\}\right]^2. \tag{4}$$

We denote the minimizer as $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\eta}} = \left\{\widehat{\eta}_{j,k}, j = -\varrho, \ldots, J_n, k = 1, \ldots, K\right\}^{\mathcal{T}}$. Then the spline estimator of g_0 is $\widehat{g}(\boldsymbol{z}) = \widehat{\boldsymbol{\eta}}^{\mathcal{T}}\boldsymbol{b}\left(\boldsymbol{z}\right)$, and the centered spline estimator of the component g_k is

$$\widehat{g}_k\left(z_k\right) = \sum_{j=-\varrho}^{J_n}\widehat{\eta}_{j,k}b_{j,k}\left(z_k\right) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=-\varrho}^{J_n}\widehat{\eta}_{j,k}b_{j,k}\left(Z_{ik}\right), \tag{5}$$

for $k = 1, \ldots, K$. The above estimation approach can be easily implemented with existing linear models in any statistics software.

2.3. Initial estimator of the variance

Davidian and Carroll (1987) developed some general methodology and theory for variance function estimation in the parametric case. They distinguished between methods based on squared residuals and those based on absolute residuals, the former being more efficient if the regressions errors ϵ_i's are normally distributed, but called this potential efficiency gain "tenuous" because it is less robust to outliers. Here we consider absolute residuals.

Define unobserved absolute residuals $R_i = |Y_i - \{g_0(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\alpha}_0\}|$ and $R = |Y - \{g_0(\boldsymbol{Z}) + \boldsymbol{X}^{\mathcal{T}}\boldsymbol{\alpha}_0\}|$, variation functions $\Phi_i = \phi\{h_0(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\beta}_0\}$ and $\Phi = \phi\{h_0(\boldsymbol{Z}) + \boldsymbol{X}^{\mathcal{T}}\boldsymbol{\beta}_0\}$, and their differences $e_i = R_i - \Phi_i$ and $e = R - \Phi$. Recall that $E(|\epsilon_i|) = E(|\epsilon|) = 1$, we have $E(e_i) = E(e) = 0$. Also define $D_i = I_{(\varepsilon_i > 0)} - I_{(\varepsilon_i \leq 0)} = \text{sign}(\varepsilon_i)$ and $D = \text{sign}(\varepsilon)$.

Define absolute residuals $\widehat{R}_i = |Y_i - \{\widehat{g}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\widehat{\boldsymbol{\alpha}}\}|$ and $\widehat{R} = |Y - \{\widehat{g}(\boldsymbol{Z}) + \boldsymbol{X}^{\mathcal{T}}\widehat{\boldsymbol{\alpha}}\}|$. Because $E(e) = 0$, approximately, $E\{\widehat{R}|\boldsymbol{X}, \boldsymbol{Z}\} \approx \phi\{h_0(\boldsymbol{Z}) + \boldsymbol{X}^{\mathcal{T}}\boldsymbol{\beta}_0\}$. A very quick way to estimate h_0 and $\boldsymbol{\beta}_0$ is to regress $\widehat{R}$ on $\phi\{h_0(\boldsymbol{Z}) + \boldsymbol{X}^{\mathcal{T}}\boldsymbol{\beta}_0\}$. We would like to find a function $h \in \mathcal{A}_n$ and a value of $\boldsymbol{\beta}$ that minimize the following sum of squared residuals function

$$L\left(h, \boldsymbol{\beta}\right) = \frac{1}{2}\sum_{i=1}^{n}\left[\widehat{R}_i - \phi\{h(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\beta}\}\right]^2, \quad h \in \mathcal{A}_n. \tag{6}$$

For any $h \in \mathcal{A}_n$, one can write

$$h\left(\boldsymbol{z}\right) = \boldsymbol{\gamma}^{\mathcal{T}}\boldsymbol{b}\left(\boldsymbol{z}\right), \tag{7}$$

where the spline coefficient vector $\boldsymbol{\gamma} = \{\gamma_{j,k}, j = -\varrho, \ldots, J_n, k = 1, \ldots, K\}^{\mathcal{T}}$. Thus the minimization problem in (6) is equivalent to finding $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to minimize

$$\ell(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\left[\widehat{R}_i - \phi\left\{\boldsymbol{\gamma}^{\mathcal{T}}\boldsymbol{b}\left(\boldsymbol{Z}_i\right) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\beta}\right\}\right]^2. \tag{8}$$

We denote the minimizer as $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}} = \{\widehat{\eta}_{j,k}, j = -\varrho, \ldots, J_n, k = 1, \ldots, K\}^{\mathcal{T}}$. Then the spline estimator of h_0 is $\widehat{h}(\boldsymbol{z}) = \widehat{\boldsymbol{\gamma}}^{\mathcal{T}}\boldsymbol{b}\left(\boldsymbol{z}\right)$, and the centered spline estimator of the component h_k is

$$\widehat{h}_k\left(z_k\right) = \sum_{j=-\varrho}^{J_n}\widehat{\gamma}_{j,k}b_{j,k}\left(z_k\right) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=-\varrho}^{J_n}\widehat{\gamma}_{j,k}b_{j,k}\left(Z_{ik}\right), \tag{9}$$

for $k = 1, \ldots, K$. The above estimation approach can also be easily implemented with existing linear models in any statistics software.

2.4. More efficient estimators

After the initial estimates of h_0 and $\boldsymbol{\beta}_0$ in the variance function are obtained, we can estimate g_0 and $\boldsymbol{\alpha}_0$ in the mean function more efficiently via generalized least squares. For this aim, let

$$\widehat{\Phi}_i = \phi\{\widehat{h}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\widehat{\boldsymbol{\beta}}\}. \tag{10}$$

Then g_0 and $\boldsymbol{\alpha}_0$ can be estimated more efficiently by the minimizers, $\widehat{g}_{\text{wls}}$ and $\widehat{\boldsymbol{\alpha}}_{\text{wls}}$, of the following sum of weighted squared residuals function

$$L_{\text{wls}}\left(g, \boldsymbol{\alpha}\right) = \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \left\{g(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\alpha}\right\}\right]^2/\widehat{\Phi}_i^2, \quad g \in \mathcal{A}_n. \tag{11}$$

Equivalently, if $\widehat{\boldsymbol{\eta}}_{\text{wls}}$ and $\widehat{\boldsymbol{\alpha}}_{\text{wls}}$ are the minimizers of

$$\ell_{\text{wls}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \left\{\boldsymbol{\eta}^{\mathcal{T}}\boldsymbol{b}\left(\boldsymbol{Z}_i\right) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\alpha}\right\}\right]^2/\widehat{\Phi}_i^2, \tag{12}$$

then $\widehat{g}_{\text{wls}}(\boldsymbol{z}) = \widehat{\boldsymbol{\eta}}_{\text{wls}}^{\mathcal{T}}\boldsymbol{b}\left(\boldsymbol{z}\right)$, whose components can be centered as in (5).

Consequently, the absolute residuals $\widehat{R}_i$ can be updated as $\widehat{R}_{i,\mathrm{wls}} = |Y_i - \{\widehat{g}_{\mathrm{wls}}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\widehat{\boldsymbol{\alpha}}_{\mathrm{wls}}\}|$. Then h_0 and $\boldsymbol{\beta}_0$ can be estimated more efficiently by the minimizers, $\widehat{h}_{\mathrm{wls}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{wls}}$, of the following sum of weighted squared residuals function

$$L_{\mathrm{wls}}(h, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\left[\widehat{R}_{i,\mathrm{wls}} - \phi\{h(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\beta}\}\right]^2 / \widehat{\Phi}_i^2, \quad h \in \mathcal{A}_n. \tag{13}$$

Equivalently, if $\widehat{\boldsymbol{\gamma}}_{\mathrm{wls}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{wls}}$ are the minimizers of

$$\ell_{\mathrm{wls}}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\left[\widehat{R}_{i,\mathrm{wls}} - \phi\left\{\boldsymbol{\gamma}^{\mathcal{T}}\boldsymbol{b}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\beta}\right\}\right]^2 / \widehat{\Phi}_i^2, \tag{14}$$

then $\widehat{h}_{\mathrm{wls}}(\boldsymbol{z}) = \widehat{\boldsymbol{\gamma}}_{\mathrm{wls}}^{\mathcal{T}}\boldsymbol{b}(\boldsymbol{z})$, whose components can be centered as in (9).

3. Theoretical Results

Let r be an integer and $\nu \in (0, 1]$, with $p = r + \nu > 1.5$. Let $\mathcal{H}_{r,\nu}$ be the collection of functions ξ on $[0, 1]$ whose rth derivative, $\xi^{(r)}$, exists and satisfies the Lipschitz condition of order ν:

$$\left|\xi^{(r)}(z') - \xi^{(r)}(z)\right| \le C\,|z' - z|^{\nu}, \text{ for } 0 \le z', z \le 1,$$

where and below c and C are generic positive constants. In order to derive theoretical results, we make the following assumptions.

(A1) Nonparametric functions $g_{0k} \in \mathcal{H}_{r,\nu}$ and $h_{0k} \in \mathcal{H}_{r,\nu}$, $k = 1, \ldots, K$.
(A2) The distribution of $\boldsymbol{Z}$ is absolutely continuous and its density f is bounded away from zero and infinity on $[0, 1]^K$.
(A3) The random vector $\boldsymbol{X}$ satisfies that for any vector $\boldsymbol{w} \in \mathbb{R}^{d+1}$,

$$c\,\|\boldsymbol{w}\|^2 \le \boldsymbol{w}^{\mathcal{T}} E\left\{\boldsymbol{X}^{\otimes 2} | \boldsymbol{Z} = \boldsymbol{z}\right\}\boldsymbol{w} \le C\,\|\boldsymbol{w}\|^2,$$

where $\|\cdot\|$ is the Euclidean norm.
(A4) The number of interior knots J_n satisfies: $n^{1/(4p)} \ll J_n \ll n^{1/4}$.
(A5) Function ϕ is twice continuously differentiable, with $c \le \phi\{h_0(\boldsymbol{Z}) + \boldsymbol{X}^{\mathcal{T}}\boldsymbol{\beta}_0\} \le C$ and $c \le \left|\phi^{(1)}\{h_0(\boldsymbol{Z}) + \boldsymbol{X}^{\mathcal{T}}\boldsymbol{\beta}_0\}\right| \le C$.

Let $\Gamma_0(\boldsymbol{z}) = E\{\boldsymbol{X}|\boldsymbol{Z} = \boldsymbol{z}\}$. As in Wang et al. (2011), let $\Gamma_0^{\mathrm{add}}(\boldsymbol{z}) = \sum_{k=1}^{K}\Gamma_{0k}^{\mathrm{add}}(z_k)$ be the projection of Γ_0 onto the Hilbert space of theoretically centered additive functions with inner product $\langle\zeta_1, \zeta_2\rangle = E\{\zeta_1(\boldsymbol{Z})\zeta_2(\boldsymbol{Z})\}$.

Let $\Phi_i^{(1)} = \phi^{(1)}\{h_0(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\beta}_0\}$ and $\Phi^{(1)} = \phi^{(1)}\{h_0(\boldsymbol{Z}) + \boldsymbol{X}^{\mathcal{T}}\boldsymbol{\beta}_0\}$. Denote $(\Phi^{(1)})^2$ as $\Phi^{(1)2}$. Let $\Gamma_1(\boldsymbol{z}) = E\{\Phi^{(1)2}\boldsymbol{X}|\boldsymbol{Z} = \boldsymbol{z}\}/E\{\Phi^{(1)2}|\boldsymbol{Z} = \boldsymbol{z}\}$ and let $\Gamma_1^{\mathrm{add}}(\boldsymbol{z}) = \sum_{k=1}^{K}\Gamma_{1k}^{\mathrm{add}}(z_k)$ be the projection of Γ_1 onto the Hilbert space of theoretically centered additive functions with inner product $\langle\zeta_1, \zeta_2\rangle_{1z} = E\{\Phi^{(1)2}\zeta_1(\boldsymbol{Z})\zeta_2(\boldsymbol{Z})\}$.

Let $\Gamma_2(\boldsymbol{z}) = E\{\boldsymbol{X}/\Phi^2|\boldsymbol{Z} = \boldsymbol{z}\}/E\{1/\Phi^2|\boldsymbol{Z} = \boldsymbol{z}\}$ and let $\Gamma_2^{\mathrm{add}}(\boldsymbol{z}) = \sum_{k=1}^K \Gamma_{2k}^{\mathrm{add}}(z_k)$ be the projection of Γ_2 onto the Hilbert space of theoretically centered additive functions with inner product $\langle \zeta_1, \zeta_2\rangle_{2z} = E\{\zeta_1(\boldsymbol{Z})\zeta_2(\boldsymbol{Z})/\Phi^2\}$.

Let $\Gamma_3(\boldsymbol{z}) = E\{\Phi^{(1)2}\boldsymbol{X}/\Phi^2|\boldsymbol{Z} = \boldsymbol{z}\}/E\{\Phi^{(1)2}/\Phi^2|\boldsymbol{Z} = \boldsymbol{z}\}$ and let $\Gamma_3^{\mathrm{add}}(\boldsymbol{z}) = \sum_{k=1}^K \Gamma_{3k}^{\mathrm{add}}(z_k)$ be the projection of Γ_3 onto the Hilbert space of theoretically centered additive functions with inner product $\langle \zeta_1, \zeta_2\rangle_{3z} = E\{\Phi^{(1)2}\zeta_1(\boldsymbol{Z})\zeta_2(\boldsymbol{Z})/\Phi^2\}$. Write $\widetilde{\boldsymbol{X}}_{i\backslash m} = \boldsymbol{X}_i - \Gamma_m^{\mathrm{add}}(\boldsymbol{Z}_i)$ and $\widetilde{\boldsymbol{X}}_{\backslash m} = \boldsymbol{X} - \Gamma_m^{\mathrm{add}}(\boldsymbol{Z})$, for $m = 0, 1, 2, 3$.

We also make the following assumption on the above centered additive projections.

(A6) The additive components in Γ_m^{add} satisfy that $\Gamma_{mk}^{\mathrm{add}} \in \mathcal{H}_{r,\nu}$ for $k = 1, \ldots, K$ and $m = 0, 1, 2, 3$.

Theorem 1. *Let $Q_\alpha = E\{\widetilde{\boldsymbol{X}}_{\backslash 0}^{\otimes 2}\}$. Under Assumptions (A1)–(A6),*

$$\|\widehat{g} - g_0\|_2 = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}, \tag{15}$$

where $\|\zeta\|_2^2 = E\{\zeta^2(\boldsymbol{Z})\}$ for any L^2-integrable function ζ on $[0,1]^K$, and

$$\sqrt{n}Q_\alpha(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i \widetilde{\boldsymbol{X}}_{i\backslash 0} + o_p(1). \tag{16}$$

Consequently, $\sqrt{n}Q_\alpha(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \to \mathrm{MVN}(\boldsymbol{0}, \Sigma_\alpha)$, where $\Sigma_\alpha = E\{\varepsilon^2 \widetilde{\boldsymbol{X}}_{\backslash 0}^{\otimes 2}\}$.

Theorem 2. *Let $Q_\beta = E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\backslash 1})^{\otimes 2}\}$. Under Assumptions (A1)–(A6), there exists a local maximizer $(\widehat{h}, \widehat{\boldsymbol{\beta}})$ of (6) such that*

$$\|\widehat{h} - h_0\|_2 = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}, \tag{17}$$

and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$. Further,

$$\sqrt{n}Q_\beta(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(e_i\Phi_i^{(1)}\widetilde{\boldsymbol{X}}_{i\backslash 1} - \varepsilon_i E\{D\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\backslash 1}\boldsymbol{X}^{\mathcal{T}}\}Q_\alpha^{-1}\widetilde{\boldsymbol{X}}_{i\backslash 0}\Big) + o_p(1). \tag{18}$$

Consequently, $\sqrt{n}Q_\beta(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to \mathrm{MVN}(\boldsymbol{0}, \Sigma_\beta)$, where

$$\Sigma_\beta = E\Big\{\Big(e\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\backslash 1} - \varepsilon E\{D\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\backslash 1}\boldsymbol{X}^{\mathcal{T}}\}Q_\alpha^{-1}\widetilde{\boldsymbol{X}}_{\backslash 0}\Big)^{\otimes 2}\Big\}.$$

Theorem 3. *Let $Q_{\alpha,\mathrm{wls}} = E\{(\widetilde{\boldsymbol{X}}_{\backslash 2}/\Phi)^{\otimes 2}\}$. Under Assumptions (A1)–(A6),*

$$\|\widehat{g}_{\mathrm{wls}} - g_0\|_2 = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}, \tag{19}$$

$$\sqrt{n}Q_{\alpha,\mathrm{wls}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{wls}} - \boldsymbol{\alpha}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i \widetilde{\boldsymbol{X}}_{i\backslash 2}/\Phi_i^2 + o_p(1). \tag{20}$$

Consequently, $\sqrt{n}(\widehat{\boldsymbol{\alpha}}_{\mathrm{wls}} - \boldsymbol{\alpha}_0) \to \mathrm{MVN}(\boldsymbol{0}, \sigma^2 Q_{\alpha,\mathrm{wls}}^{-1})$, where $Var(\epsilon) = \sigma^2$.

Theorem 4. *Let* $Q_{\beta,\mathrm{wls}} = E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 3}/\Phi)^{\otimes 2}\}$. *Under Assumptions (A1)–(A6), there exists a local maximizer* $(\widehat{h}_{\mathrm{wls}}, \widehat{\boldsymbol{\beta}}_{\mathrm{wls}})$ *of (13) such that*

$$\|\widehat{h}_{\mathrm{wls}} - h_0\|_2 = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}, \tag{21}$$

and $\|\widehat{\boldsymbol{\beta}}_{\mathrm{wls}} - \boldsymbol{\beta}_0\| = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$. *Further,*

$$\sqrt{n}Q_{\beta,\mathrm{wls}}(\widehat{\boldsymbol{\beta}}_{\mathrm{wls}} - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(e_i\Phi_i^{(1)}\widetilde{\boldsymbol{X}}_{i\setminus 3}/\Phi_i^2 - \varepsilon_i E\{D\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 3}\boldsymbol{X}^{\mathcal{T}}/\Phi^2\} Q_{\alpha,\mathrm{wls}}^{-1}\widetilde{\boldsymbol{X}}_{i\setminus 2}/\Phi_i^2\Big) + o_p(1). \tag{22}$$

Consequently, $\sqrt{n}Q_{\beta,\mathrm{wls}}(\widehat{\boldsymbol{\beta}}_{\mathrm{wls}} - \boldsymbol{\beta}_0) \to \mathrm{MVN}(\mathbf{0}, \Sigma_{\beta,\mathrm{wls}})$, *where*

$$\Sigma_{\beta,\mathrm{wls}} = E\Big\{\Big(e\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 3}/\Phi^2 - \varepsilon E\{D\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 3}\boldsymbol{X}^{\mathcal{T}}/\Phi^2\}Q_{\alpha,\mathrm{wls}}^{-1}\widetilde{\boldsymbol{X}}_{\setminus 2}/\Phi^2\Big)^{\otimes 2}\Big\}.$$

Remark 1. The convergence rate $O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$ enjoyed by the estimators of nonparametric components in both the mean and variance functions is natural. Similar assumption on J_n was made and similar convergence rate was obtained in Wang et al. (2014). If $J_n \asymp n^{1/(2p+1)}$, then we obtain an optimal convergence rate $n^{-2p/(2p+1)}$.

Remark 2. Following the routine proposed in Newey (1994) and theory developed in Bickel et al. (1993), we can show that, when ϵ is normally distributed, $\widehat{\boldsymbol{\alpha}}_{\mathrm{wls}}$ is the most efficient estimator in the sense of semiparametric efficiency.

Remark 3. Consider the estimators for $\boldsymbol{\alpha}_0$. The weighted estimator $\widehat{\boldsymbol{\alpha}}_{\mathrm{wls}}$ is more efficient than the initial estimator $\widehat{\boldsymbol{\alpha}}$. To see this, in this and next remarks, for simplicity, we ignore the factor n. The asymptotic variance of $\widehat{\boldsymbol{\alpha}}$ is $\sigma^2[E(\widetilde{\boldsymbol{X}}_{\setminus 0}^{\otimes 2})]^{-1}E\{(\Phi\widetilde{\boldsymbol{X}}_{\setminus 0})^{\otimes 2}\}[E(\widetilde{\boldsymbol{X}}_{\setminus 0}^{\otimes 2})]^{-1}$ and the asymptotic variance of $\widehat{\boldsymbol{\alpha}}_{\mathrm{wls}}$ is $\sigma^2[E\{(\widetilde{\boldsymbol{X}}_{\setminus 2}/\Phi)^{\otimes 2}\}]^{-1}$. Noting that

$$E\left\{\begin{pmatrix}\widetilde{\boldsymbol{X}}_{\setminus 2}/\Phi \\ \Phi\widetilde{\boldsymbol{X}}_{\setminus 0}\end{pmatrix}^{\otimes 2}\right\} = \begin{pmatrix} E\{(\widetilde{\boldsymbol{X}}_{\setminus 2}/\Phi)^{\otimes 2}\} & E(\widetilde{\boldsymbol{X}}_{\setminus 0}^{\otimes 2}) \\ E(\widetilde{\boldsymbol{X}}_{\setminus 0}^{\otimes 2}) & E\{(\Phi\widetilde{\boldsymbol{X}}_{\setminus 0})^{\otimes 2}\}\end{pmatrix} \geq 0,$$

we see that $E\{(\widetilde{\boldsymbol{X}}_{\setminus 2}/\Phi)^{\otimes 2}\} \geq E(\widetilde{\boldsymbol{X}}_{\setminus 0}^{\otimes 2})[E\{(\Phi\widetilde{\boldsymbol{X}}_{\setminus 0})^{\otimes 2}\}]^{-1}E(\widetilde{\boldsymbol{X}}_{\setminus 0}^{\otimes 2})$.

Remark 4. Consider the estimators for $\boldsymbol{\beta}_0$. The weighted estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{wls}}$ is more efficient than the initial estimator $\widehat{\boldsymbol{\beta}}$. To see this, for simplicity, we only consider the special case where ϵ is symmetric. In this special case, $E(D) = 0$ and therefore the second term in each of Σ_β and $\Sigma_{\beta,\mathrm{wls}}$ becomes zero. Noting that $Var(e/\Phi) = Var(|\epsilon| - 1) = \sigma^2 - 1$, the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ is $(\sigma^2-1)[E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 1})^{\otimes 2}\}]^{-1}E\{(\Phi\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 1})^{\otimes 2}\}[E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 1})^{\otimes 2}\}]^{-1}$ and the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{wls}}$ is $(\sigma^2-1)[E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 3}/G)^{\otimes 2}\}]^{-1}$. Noting that

$$E\left\{\begin{pmatrix}\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 3}/\Phi \\ \Phi\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 1}\end{pmatrix}^{\otimes 2}\right\} = \begin{pmatrix} E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 3}/\Phi)^{\otimes 2}\} & E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 1})^{\otimes 2}\} \\ E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 1})^{\otimes 2}\} & E\{(\Phi\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 1})^{\otimes 2}\}\end{pmatrix} \geq 0,$$

we see that

$$E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 3}/\Phi)^{\otimes 2}\} \geq E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 1})^{\otimes 2}\}[E\{(\Phi\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 1})^{\otimes 2}\}]^{-1}E\{(\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 3})^{\otimes 2}\}.$$

4. Simulation Experiments

To assess the finite sample performance of the proposed methods in Section 2, we simulate data from model (1) with $\phi(v) = \exp(v)$. Set $d = 4$ and $K = 2$. First, obtain $(X_{i1}, \ldots, X_{i6})$, $i = 1, \ldots, n$, generated from zero-mean multivariate Gaussian distribution with $Cov(X_{ij}, X_{ij'}) = 0.2^{|j-j'|}$. Then, set $\boldsymbol{X}_i^* = (X_{i1}, \ldots, X_{i4})^{\mathcal{T}}$, $Z_{i1} = G(X_{i5})$, and $Z_{i2} = G(X_{i6})$, where G is the cumulative distribution function of the standard normal distribution and consequently Z_{i1} and Z_{i2} are in $[0, 1]$. Let $g_{01}(x) = -6(x - 0.5)^2$, $g_{02}(x) = 6(x - 0.5)^2$, $h_{01}(x) = \sin(4x)$, $h_{02}(x) = \cos(4x)$ and generate responses from

$$\begin{aligned} Y_i = {} & \boldsymbol{X}_i^{*\mathcal{T}}\boldsymbol{\alpha}_0 + g_{01}(Z_{i1}) + g_{02}(Z_{i2}) \\ & + \exp\left\{1 + \boldsymbol{X}_i^{*\mathcal{T}}\boldsymbol{\beta}_0 + h_{01}(Z_{i1}) + h_{02}(Z_{i2})\right\}\varepsilon_i, \end{aligned}$$

with $\boldsymbol{\alpha}_0 = (1, -1, 1, -1)^{\mathcal{T}}$, $\boldsymbol{\beta}_0 = (-0.125, 0.25, -0.125, 0.25)^{\mathcal{T}}$ and $\varepsilon_i \sim N(0, \sigma^2)$.

Choose $n \in \{200, 400, 800\}$ and $\sigma \in \{0.1, 0.2, 0.4\}$. Letting $\epsilon_i = \varepsilon_i / E(|\varepsilon_i|)$, the above model is one example of model (1). In each case, 500 datasets are generated and fitted. We use cubic splines ($\rho = 3$) to approximate the nonparametric functions. The number of basic functions is set to be 5 for $n = 200, 400$ and 6 for $n = 800$. This choice applies for both mean functions and variance functions, and for both initial estimators and more efficient estimators. Although data-adaptive choice for the number of internal knots can be developed, as in Wang and Yang (2007), Fan et al. (2011) and Lian et al. (2013), it is found that fixed choice of the number of internal knots is much more convenient and indeed adopted in most studies using B-splines for function estimation, and for regression splines a small number of basis functions is typically used in numerical studies. For both the mean and variance functions, we consider three estimators: the initial estimators (4) and (8), the weighted estimators (12) and (14), and the infeasible estimators where $\widehat{\Phi}_i$ is replaced by Φ_i and $\widehat{R}_{i,\text{wls}}$ is replaced by R_i in (12) and (14).

First, we examine the estimation errors, $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$, $\|\widehat{g}_k - g_{0k}\|_2$, $\|\widehat{h}_k - h_{0k}\|_2$, $k = 1, 2$. We approximate $\|\widehat{g}_1 - g_{01}\|_2$ by $\sqrt{\sum_{i=1}^{100}\{\widehat{g}_1(t_i) - g_{01}(t_i)\}^2/100}$ by a grid $t_1 = 0 < t_2 < \cdots < t_{100} = 1$ on $[0, 1]$ and similarly for other nonparametric functions. The estimation errors averaged over 500 datasets for each of the nine parameter settings are reported in Table 1, with the standard deviations of the errors shown in brackets. We see that both mean and variance estimation improve with larger sample size. While errors in mean estimation increase with noise, errors in variance estimation remain almost the same with different noise levels. Most importantly, we see that the updated, weighted estimators significantly improve on the initial estimators in all situations. In Figures 1–3, we show the estimated nonparametric functions on 20 generated data sets

TABLE 1

Estimation errors (average errors with standard deviations inside brackets on 500 simulated datasets) for the simulated data sets

(n, σ)		$\boldsymbol{\alpha}$	g_1	g_2	$\boldsymbol{\beta}$	h_1	h_2
$(200, 0.1)$	Initial	0.119(0.048)	0.093(0.037)	0.127(0.059)	0.227(0.090)	0.268(0.101)	0.200(0.072)
	Weighted	0.032(0.013)	0.041(0.014)	0.077(0.039)	0.150(0.054)	0.217(0.060)	0.159(0.058)
	Infeasible	0.023(0.009)	0.028(0.011)	0.061(0.032)	0.114(0.044)	0.115(0.042)	0.118(0.039)
$(200, 0.2)$	Initial	0.237(0.097)	0.187(0.074)	0.255(0.119)	0.227(0.090)	0.268(0.101)	0.200(0.072)
	Weighted	0.065(0.026)	0.083(0.029)	0.155(0.078)	0.150(0.054)	0.216(0.060)	0.158(0.058)
	Infeasible	0.047(0.018)	0.056(0.022)	0.123(0.064)	0.114(0.044)	0.115(0.042)	0.118(0.039)
$(200, 0.4)$	Initial	0.475(0.195)	0.374(0.149)	0.510(0.237)	0.227(0.090)	0.268(0.101)	0.200(0.072)
	Weighted	0.130(0.053)	0.167(0.058)	0.311(0.157)	0.150(0.054)	0.215(0.060)	0.158(0.058)
	Infeasible	0.095(0.036)	0.114(0.044)	0.245(0.130)	0.114(0.044)	0.115(0.042)	0.118(0.039)
$(400, 0.1)$	Initial	0.083(0.036)	0.063(0.026)	0.086(0.042)	0.166(0.060)	0.182(0.065)	0.136(0.051)
	Weighted	0.018(0.007)	0.021(0.009)	0.044(0.021)	0.089(0.033)	0.091(0.034)	0.090(0.034)
	Infeasible	0.015(0.006)	0.019(0.008)	0.038(0.019)	0.077(0.028)	0.080(0.029)	0.078(0.028)
$(400, 0.2)$	Initial	0.167(0.072)	0.128(0.052)	0.172(0.084)	0.166(0.060)	0.182(0.065)	0.136(0.051)
	Weighted	0.037(0.014)	0.043(0.018)	0.089(0.042)	0.089(0.033)	0.091(0.034)	0.090(0.034)
	Infeasible	0.031(0.012)	0.039(0.016)	0.077(0.039)	0.077(0.028)	0.080(0.029)	0.078(0.028)
$(400, 0.4)$	Initial	0.333(0.144)	0.256(0.105)	0.345(0.168)	0.166(0.060)	0.182(0.065)	0.136(0.051)
	Weighted	0.074(0.028)	0.087(0.037)	0.179(0.084)	0.089(0.033)	0.091(0.034)	0.090(0.034)
	Infeasible	0.062(0.024)	0.078(0.032)	0.155(0.079)	0.077(0.028)	0.080(0.029)	0.078(0.028)
$(800, 0.1)$	Initial	0.057(0.022)	0.051(0.018)	0.069(0.032)	0.123(0.044)	0.138(0.049)	0.102(0.031)
	Weighted	0.011(0.004)	0.017(0.006)	0.037(0.019)	0.059(0.023)	0.065(0.022)	0.065(0.022)
	Infeasible	0.010(0.003)	0.016(0.006)	0.033(0.015)	0.055(0.019)	0.060(0.022)	0.060(0.019)
$(800, 0.2)$	Initial	0.115(0.045)	0.103(0.036)	0.139(0.065)	0.120(0.044)	0.138(0.049)	0.102(0.031)
	Weighted	0.022(0.009)	0.034(0.013)	0.075(0.039)	0.059(0.023)	0.065(0.022)	0.065(0.022)
	Infeasible	0.020(0.007)	0.032(0.013)	0.066(0.031)	0.055(0.020)	0.060(0.022)	0.060(0.019)
$(800, 0.4)$	Initial	0.230(0.090)	0.205(0.073)	0.277(0.131)	0.123(0.044)	0.138(0.049)	0.102(0.031)
	Weighted	0.045(0.018)	0.069(0.027)	0.151(0.078)	0.059(0.023)	0.065(0.022)	0.065(0.022)
	Infeasible	0.040(0.014)	0.065(0.026)	0.134(0.062)	0.055(0.020)	0.060(0.022)	0.060(0.019)

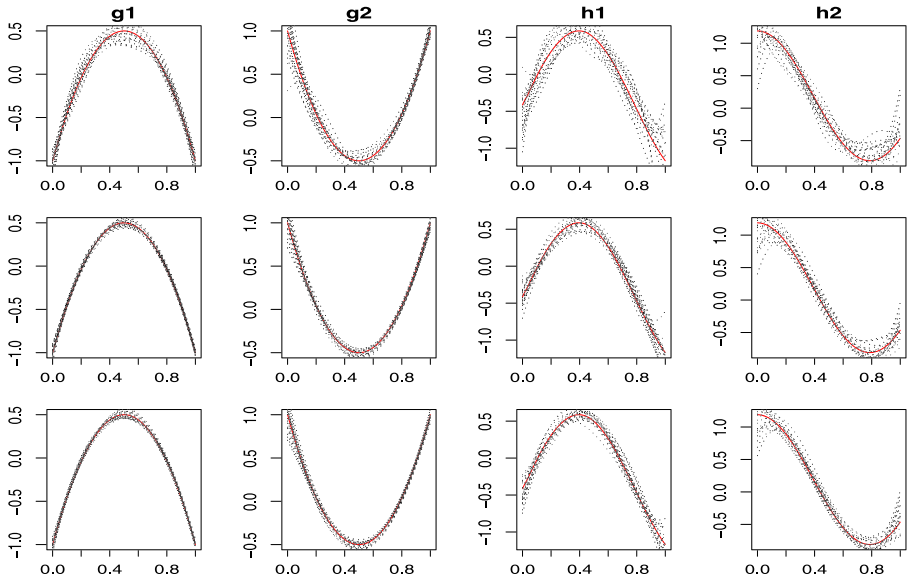

FIG 1. *Estimated nonparametric functions when* $n = 400$ *and* $\sigma = 0.1$. *The solid red curve is the true function. The three rows correspond to the initial estimators, the weighted estimators, and the infeasible estimators, respectively.*

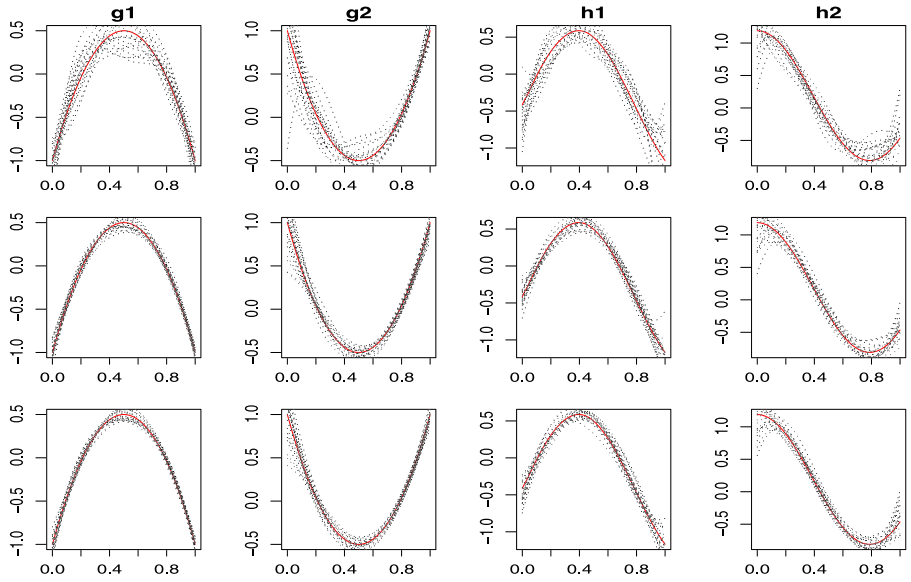

FIG 2. *Estimated nonparametric functions when* $n = 400$ *and* $\sigma = 0.2$. *The solid red curve is the true function. The three rows correspond to the initial estimators, the weighted estimators, and the infeasible estimators, respectively.*

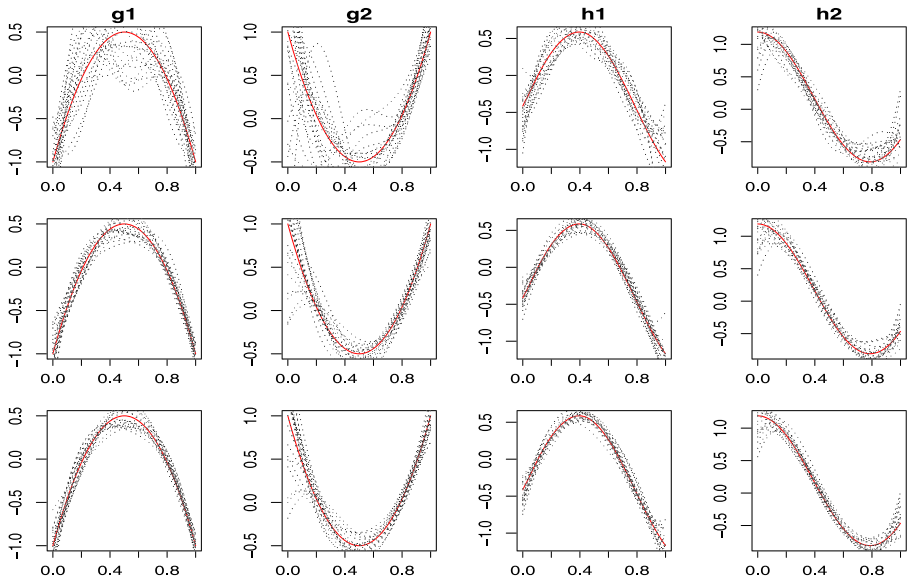

FIG 3. *Estimated nonparametric functions when* $n = 400$ *and* $\sigma = 0.4$. *The solid red curve is the true function. The three rows correspond to the initial estimators, the weighted estimators, and the infeasible estimators, respectively.*

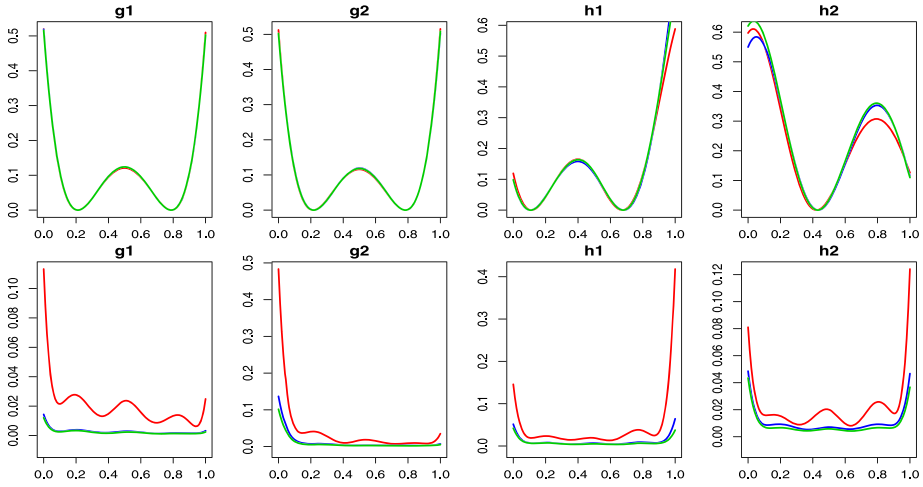

FIG 4. *Squared bias (first row) and variance (second row) for our simulation example when when* $n = 400$ *and* $\sigma = 0.2$. *The red, blue, green curves show the squared bias and variance for the initial estimator, the weighted estimator, and the infeasible estimation, respectively.*

when $n = 400$, for the three noise levels respectively. The weighted estimators are obviously better than the initial estimators, and are visually very similar to the infeasible estimators. For the case $n = 400$, $\sigma = 0.2$, we also show the squared bias and variance for the three estimators in Figure 4. We see that the improvement of the weighted estimator mostly come from the reduction of variance. Furthermore, there is relatively large bias and variance close to the boundary.

Second, we consider estimation of standard errors for the parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$. It is easy to obtain standard error estimates based on the asymptotic normality results, using the sandwich formula. On each generated dataset, we can get an estimate of standard errors, and the average of these over 500 datasets are reported in Table 2, on rows indicated by "s.e. (est)". The sample standard errors of the estimated parameter values on 500 datasets are reported on rows indicated by "s.e. (emp)". It is seen that the estimated standard errors are reasonably close to the empirical standard errors especially when the sample size is large.

Third, we consider the coverage of the pointwise confidence intervals for the nonparametric functions. Our construction of the pointwise confidence intervals is again based on the sandwich formula. More specifically, since the nonparametric functions are approximated by linear combinations of known basis functions, we regard the model as parametric with parameters $\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ and we find the estimated covariance matrix of these parameters as in parametric models by the standard sandwich formula. Note that this ignores the bias term caused by the series expansion of nonparametric functions. However, it is a difficult problem to construct bona fide confidence intervals in additive models. Thus we just use

TABLE 2

Estimated standard errors in the simulations. For each pair of (n, σ), the first two rows are the results of the initial estimators, and the next two rows are the results of the weighted estimators

case: (n, σ)		α_1	α_2	α_3	α_4	β_1	β_2	β_3	β_4
$(200, 0.1)$	s.e. (est)	0.057	0.061	0.057	0.060	0.101	0.103	0.102	0.099
	s.e. (emp)	0.059	0.069	0.060	0.066	0.120	0.123	0.122	0.121
	s.e. (est)	0.013	0.013	0.013	0.013	0.067	0.066	0.067	0.066
	s.e. (emp)	0.017	0.018	0.017	0.017	0.085	0.080	0.105	0.090
$(200, 0.2)$	s.e. (est)	0.115	0.122	0.114	0.122	0.101	0.103	0.102	0.099
	s.e. (emp)	0.119	0.138	0.1216	0.132	0.120	0.123	0.122	0.121
	s.e. (est)	0.027	0.027	0.027	0.027	0.067	0.066	0.067	0.066
	s.e. (emp)	0.035	0.036	0.035	0.034	0.084	0.080	0.105	0.090
$(200, 0.4)$	s.e. (est)	0.231	0.244	0.229	0.243	0.101	0.103	0.102	0.099
	s.e. (emp)	0.239	0.276	0.243	0.265	0.120	0.123	0.122	0.121
	s.e. (est)	0.054	0.055	0.055	0.055	0.067	0.066	0.067	0.066
	s.e. (emp)	0.071	0.072	0.071	0.069	0.084	0.080	0.105	0.090
$(400, 0.1)$	s.e. (est)	0.041	0.045	0.041	0.042	0.081	0.083	0.081	0.081
	s.e. (emp)	0.044	0.047	0.044	0.044	0.091	0.096	0.088	0.082
	s.e. (est)	0.008	0.008	0.008	0.008	0.046	0.046	0.046	0.044
	s.e. (emp)	0.010	0.010	0.010	0.009	0.049	0.050	0.048	0.048
$(400, 0.2)$	s.e. (est)	0.082	0.090	0.082	0.085	0.081	0.083	0.081	0.081
	s.e. (emp)	0.089	0.094	0.089	0.088	0.091	0.096	0.088	0.082
	s.e. (est)	0.017	0.017	0.017	0.016	0.046	0.046	0.046	0.044
	s.e. (emp)	0.020	0.020	0.020	0.019	0.049	0.050	0.048	0.048
$(400, 0.4)$	s.e. (est)	0.165	0.180	0.166	0.171	0.081	0.083	0.081	0.081
	s.e. (emp)	0.178	0.188	0.179	0.177	0.091	0.096	0.088	0.082
	s.e. (est)	0.034	0.034	0.034	0.033	0.046	0.046	0.046	0.044
	s.e. (emp)	0.040	0.040	0.040	0.039	0.049	0.050	0.048	0.048
$(800, 0.1)$	s.e. (est)	0.029	0.031	0.030	0.031	0.065	0.065	0.064	0.063
	s.e. (emp)	0.029	0.031	0.030	0.031	0.064	0.066	0.066	0.068
	s.e. (est)	0.005	0.005	0.005	0.005	0.031	0.031	0.031	0.030
	s.e. (emp)	0.006	0.006	0.006	0.005	0.031	0.032	0.031	0.031
$(800, 0.2)$	s.e. (est)	0.059	0.063	0.060	0.062	0.065	0.065	0.064	0.063
	s.e. (emp)	0.058	0.063	0.061	0.062	0.064	0.066	0.066	0.068
	s.e. (est)	0.011	0.011	0.011	0.011	0.031	0.031	0.031	0.030
	s.e. (emp)	0.012	0.012	0.012	0.011	0.031	0.032	0.031	0.031
$(800, 0.4)$	s.e. (est)	0.119	0.127	0.120	0.125	0.065	0.065	0.064	0.063
	s.e. (emp)	0.116	0.126	0.122	0.125	0.064	0.066	0.066	0.068
	s.e. (est)	0.023	0.023	0.022	0.022	0.031	0.031	0.031	0.030
	s.e. (emp)	0.024	0.024	0.025	0.023	0.031	0.032	0.031	0.031

this naive approach and mainly regard the intervals as "exploratory" in nature. For illustration we only used $n = 200, 400, 800$ with $\sigma = 0.2$. Figures 5–7 show the empirical coverage of the 95% pointwise confidence intervals for the three sample sizes, respectively. When $n = 200$ we see the coverage is not satisfactory, especially close to the boundary for some cases, although the coverage is always larger than 80% . For larger sample sizes, the coverage improved somewhat. Better construction of the pointwise confidence intervals or simultaneous confidence bands for the nonparametric part is an interesting problem to be investigated in the future.

Finally, we modify the above example in two ways. In the first modification, we change the mean functions to $g_{01}(x) = -\sin(15x)$, $g_{02}(x) = \sin(15x)$ while other aspects of the model remain the same. This is used to explore the case

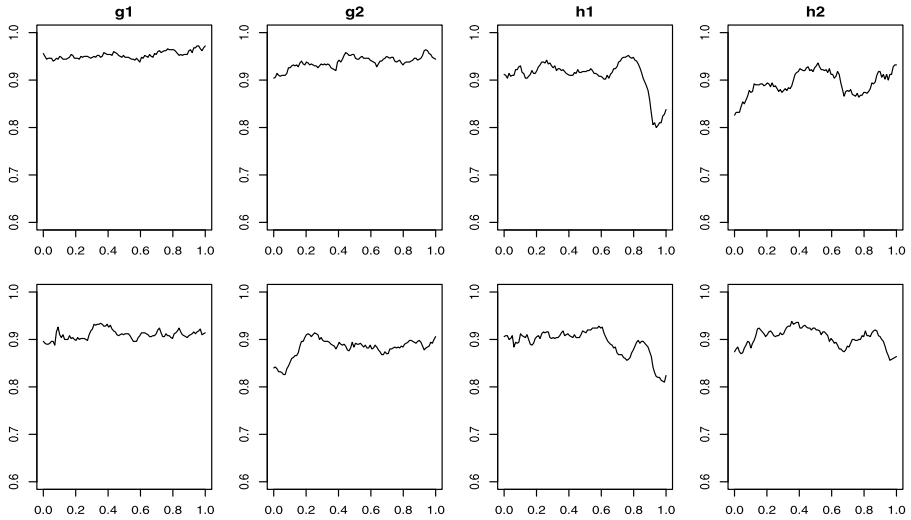

FIG 5. *Empirical coverage of the pointwise confidence intervals for the nonparametric functions, when* $n = 200$ *and* $\sigma = 0.2$. *The first row is for the initial estimators and the second row is for the weighted estimators.*

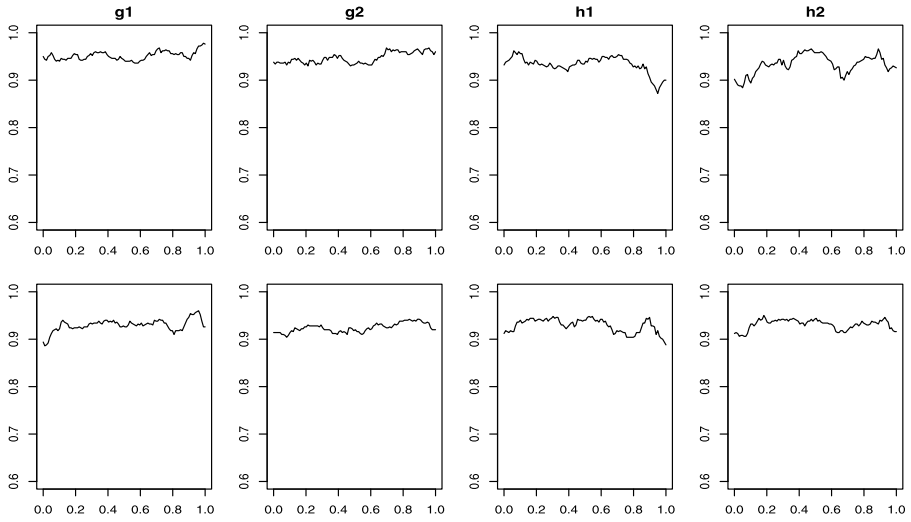

FIG 6. *Empirical coverage of the pointwise confidence intervals for the nonparametric functions, when* $n = 400$ *and* $\sigma = 0.2$. *The first row is for the initial estimators and the second row is for the weighted estimators.*

when the mean functions cannot be estimated well, whether the estimates of the variance functions will be seriously affected, and whether weighting will improve the estimation. In the second modification, we change the normal errors

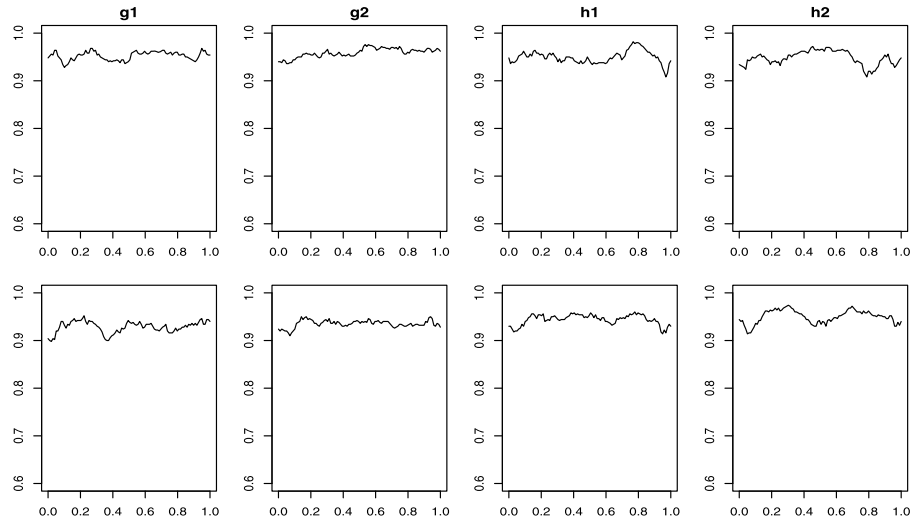

FIG 7. *Empirical coverage of the pointwise confidence intervals for the nonparametric functions, when* $n = 800$ *and* $\sigma = 0.2$. *The first row is for the initial estimators and the second row is for the weighted estimators.*

in the original example to Student's t errors with degrees of freedom 3 and scale parameter σ. We only report the results for $\sigma = 0.2$ with $n = 200, 400, 800$ for these two modified examples. The estimation errors are reported in Table 3 and 20 estimated nonparametric functions are illustrated in Figure 8 and Figure 9, respectively (for $n = 400$ only). From Figure 8, we see that when the mean function cannot be estimated well, this causes obvious bias in the variance function estimate. Table 3 shows that the second stage weighted estimate does not improve the estimation (actually for this particular case the second stage estimator is even worse). This indicates that using totally wrong weights could be worse than using equal weights, due to the extra variability of the estimated weights. Even using the correct weights does not help the seriously biased estimates for mean functions. In conclusion, reasonably accurate estimate of the mean function is a requirement for the proposed method to work well. The bad performance of this example is mainly due to that with highly oscillating true functions, the spline approximation using only 5 or 6 basis functions results in a large bias. To illustrate this point, we increase the number of basis functions to 9 and some estimated curves are shown in Figure 10. With a larger number of basis functions, the mean function estimation is now much more reasonable. However, a larger number of basis functions means the estimates for variance functions become more variable. In such examples, it is desirable that the number of basis functions is chosen adaptively according to (unknown) smoothness of each function, although this is outside of the scope of the current paper. For the second modification using heavy-tailed errors, we still see that the second stage estimator improves upon the initial estimator. With heavy-tailed errors, the estimates are obviously worse than those with normal errors.

TABLE 3

Estimation errors (average errors with standard deviations inside brackets on 500 simulated datasets) for the simulated data sets. The first block is for the case $g_{01}(x) = -\sin(15x)$ *and* $g_{02}(x) = \sin(15x)$*, and the second block reports results when the error follows a student's t distribution*

(n, σ)		$\boldsymbol{\alpha}$	g_1	g_2	$\boldsymbol{\beta}$	h_1	h_2
$(200, 0.2)$	Initial	0.269(0.108)	0.638(0.030)	0.667(0.049)	0.223(0.081)	0.289(0.081)	0.320(0.068)
	Weighted	0.179(0.068)	0.644(0.034)	0.690(0.062)	0.262(0.074)	0.316(0.073)	0.381(0.070)
	Infeasible	0.207(0.076)	0.674(0.045)	0.676(0.052)	0.114(0.044)	0.115(0.042)	0.118(0.039)
$(400, 0.2)$	Initial	0.187(0.079)	0.611(0.015)	0.627(0.027)	0.173(0.065)	0.254(0.056)	0.288(0.052)
	Weighted	0.121(0.050)	0.626(0.023)	0.654(0.039)	0.255(0.054)	0.300(0.044)	0.365(0.046)
	Infeasible	0.148(0.058)	0.672(0.037)	0.637(0.035)	0.077(0.028)	0.081(0.029)	0.078(0.028)
$(800, 0.2)$	Initial	0.129(0.049)	0.602(0.008)	0.612(0.013)	0.140(0.054)	0.284(0.036)	0.330(0.031)
	Weighted	0.082(0.033)	0.643(0.020)	0.731(0.049)	0.257(0.034)	0.307(0.032)	0.339(0.032)
	Infeasible	0.086(0.034)	0.734(0.028)	1.100(0.101)	0.055(0.020)	0.060(0.022)	0.060(0.019)
$(200, 0.2)$	Initial	0.384(0.166)	0.318(0.136)	0.425(0.199)	0.345(0.150)	0.416(0.150)	0.309(0.113)
	Weighted	0.137(0.057)	0.212(0.059)	0.285(0.143)	0.254(0.095)	0.347(0.105)	0.243(0.102)
	Infeasible	0.079(0.031)	0.095(0.038)	0.192(0.090)	0.229(0.083)	0.216(0.076)	0.218(0.083)
$(400, 0.2)$	Initial	0.275(0.109)	0.221(0.089)	0.285(0.122)	0.277(0.101)	0.311(0.112)	0.221(0.088)
	Weighted	0.086(0.033)	0.138(0.040)	0.201(0.079)	0.166(0.056)	0.205(0.067)	0.197(0.061)
	Infeasible	0.050(0.021)	0.066(0.027)	0.134(0.062)	0.139(0.053)	0.132(0.050)	0.132(0.056)
$(800, 0.2)$	Initial	0.197(0.069)	0.178(0.067)	0.252(0.112)	0.211(0.091)	0.260(0.080)	0.175(0.065)
	Weighted	0.055(0.019)	0.091(0.025)	0.152(0.065)	0.116(0.041)	0.231(0.039)	0.143(0.037)
	Infeasible	0.034(0.014)	0.051(0.020)	0.111(0.052)	0.095(0.034)	0.101(0.034)	0.102(0.036)

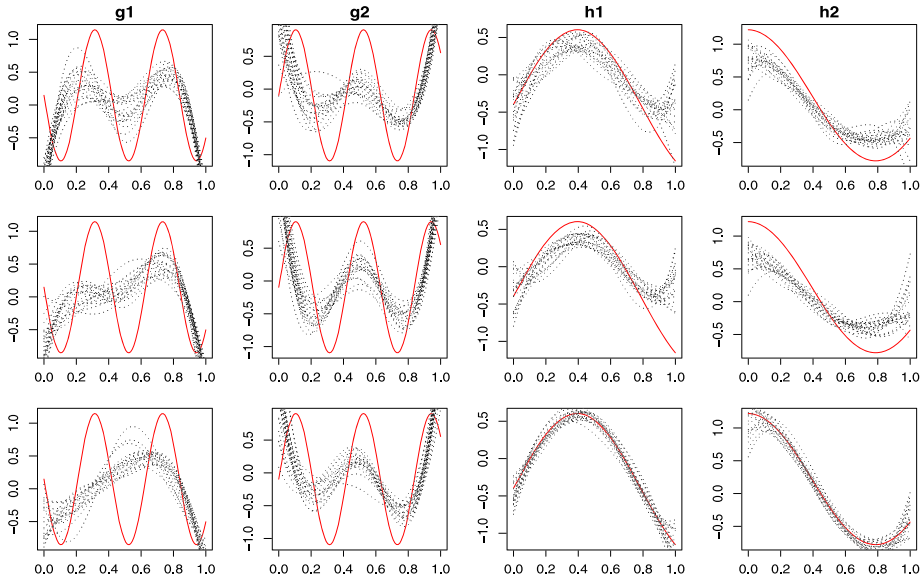

FIG 8. *Estimated nonparametric functions when* $n = 400$ *and* $\sigma = 0.2$ *when the model is changed to* $g_{01}(x) = -g_{02}(x) = -\sin(15x)$*. The solid red curve is the true function. The three rows correspond to the initial estimators, the weighted estimators, and the infeasible estimators, respectively.*

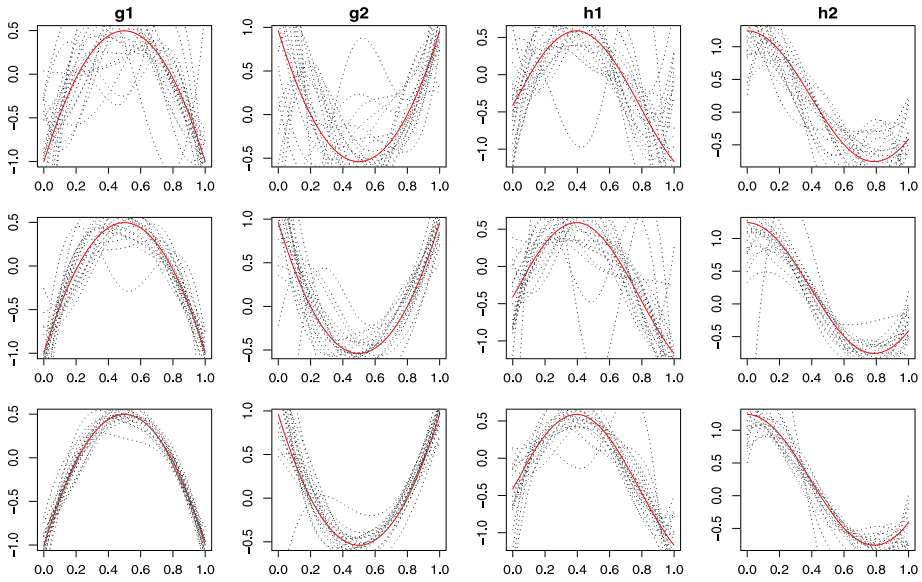

FIG 9. *Estimated nonparametric functions when* $n = 400$ *and* $\sigma = 0.2$ *when the error is changed to follow the Student's t distribution. The solid red curve is the true function. The three rows correspond to the initial estimators, the weighted estimators, and the infeasible estimators, respectively.*

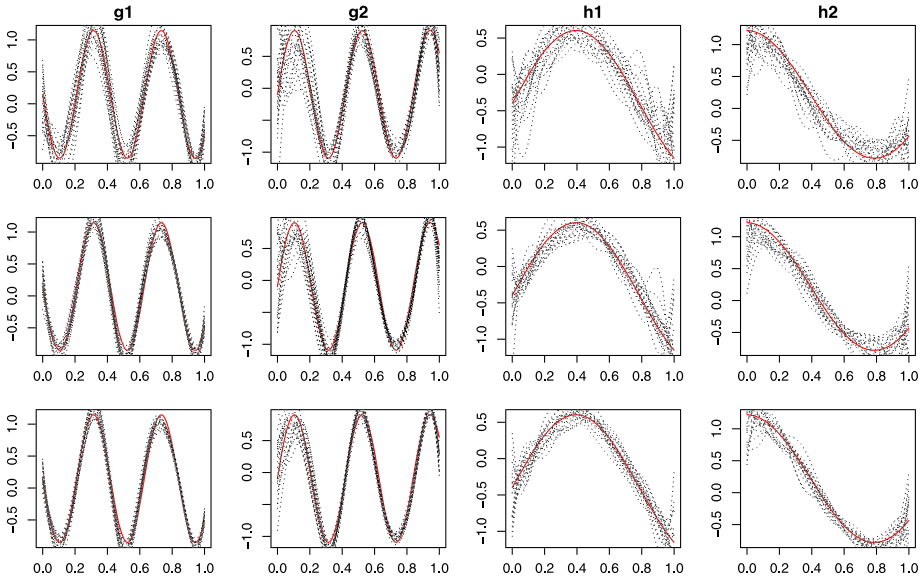

FIG 10. *Estimated nonparametric functions when* $n = 400$ *and* $\sigma = 0.2$ *when the model is changed to* $g_{01}(x) = -g_{02}(x) = -\sin(15x)$. *Here we use a larger number of basis* 9.

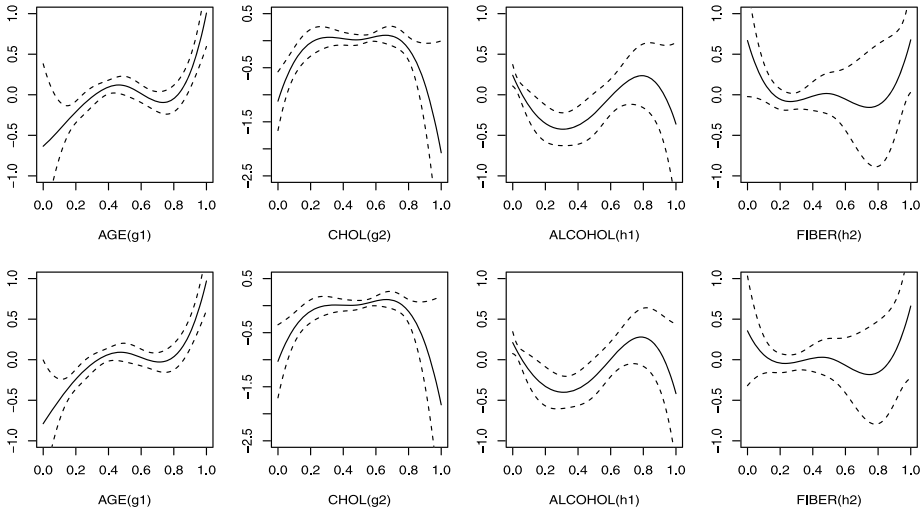

FIG 11. *Estimated nonparametric functions for the nutrition data with pointwise* 95% *CI. The first row shows the initial estimators and the second row shows the weighted estimators.*

5. Nutrition Data

We apply the proposed method to the dataset from a nutritional epidemiology study (Nierenberg et al., 1989), which attempted to investigate the relationships between the plasma beta-carotene levels and personal characteristics, including AGE, SEX, BMI, and other factors: CALORIES (number of calories consumed per day), FAT (grams of fat consumed per day), FIBER (grams of fiber consumed per day), ALCOHOL (number of alcoholic drinks consumed per week), CHOL (cholesterol consumed mg per day), BETADIET (dietary betacarotene consumed mcg per day), SMOKE2 (smoking status [1 = former smoker, 0 = never smoked]), and SMOKE3 (smoking status [1 = current smoker, 0 = never smoked]). We remove the predictor FAT since it is highly correlated with CALORIES. There is one extremely high leverage point in alcohol consumption that is deleted prior to analysis and thus the sample size of the dataset is $n = 314$. Similar to simulations, we used cubic splines with 5 basis functions. By first fitting an APLM that puts all continuous predictors in the nonparametric part and discrete predictors in the linear part, we find that AGE and CHOL seem to have nonlinear effect in the mean function, while ALCOHOL and FIBER seem to have nonlinear effect in the variance function. Thus we fit the model with AGE and CHOL in the nonparametric part in the mean and at the same time with ALCOHOL and FIBER in the nonparametric part in the variance. The shapes of the estimated nonparametric functions for both the initial estimators and the weighted estimators are shown in Figure 11. The nonparametric functions g_1(AGE) and g_2(CHOL) in the mean function are similar to that obtained in Liu et al. (2011). For the nonparametric functions h_1(ALCOHOL) and

TABLE 4
Estimated coefficients and standard errors (in brackets) for the real data.

Variable	Mean Estimation Initial	Weighted
ALCOHOL	0.105(0.160)	0.178(0.143)
BETADIET	0.425(0.265)	0.424(0.241)
BMI	−1.103(0.207)	−1.130(0.187)
CALORIES	−0.439(0.428)	−0.530(0.363)
FIBER	0.833(0.329)	0.632(0.277)
SEX	0.271(0.131)	0.335(0.120)
SMOKE2	−0.101(0.081)	−0.055(0.086)
SMOKE3	−0.316(0.114)	−0.296(0.086)
Variable	**Variance Estimation Initial**	**Weighted**
AGE	0.048(0.274)	0.166(0.280)
BETADIET	0.531(0.435)	0.463(0.399)
BMI	−0.116(0.296)	−0.029(0.275)
CALORIES	0.068(0.625)	−0.141(0.635)
CHOL	0.463(0.481)	0.348(0.413)
SEX	−0.085(0.170)	0.010(0.167)
SMOKE2	0.091(0.116)	0.107(0.120)
SMOKE3	−0.090(0.204)	−0.137(0.162)

h_2(CHOL) in the variance function, it shows there is nonlinear contribution of ALCOHOL to the variance while the effect of CHOL seems to be nonsignificant. The estimated coefficients and their standard errors are listed in Table 4. The variables BMI, FIBER, SEX and SMOKE3 have significant effects in the mean function at the 0.05 level, while none has significant effect in the variance function.

6. Discussion

The additive partial linear models have been well studied in the literature when the heteroscedasticity is ignored. In this paper we investigate a broad class of models, variance function additive partial linear models. The flexibility of such models comes from that the variance is not limited to be a known function of the mean. The models are useful for the settings where estimating the variance function is of its own interest. The models are also useful for the settings where estimating the mean function is of main interest, because taking into account the heteroscedasticity would improve the efficiency of estimating the mean function. Also, in cases of even moderate heteroscedasticity, prediction intervals will not have correct coverage probabilities unless the variance is modeled properly.

The polynomial-splines method we adopt for approximating the nonparametric components in both the mean and variance functions has at least three principal advantages. First, it avoids iterative algorithms and therefore it is computationally convenient. Second, the resulting estimators of the linear components in both the mean and variance functions are still asymptotically normal. Third, it is very easy to conduct variable selection on the linear components in both the mean and variance functions, which we discuss in detail in next. On

the other hand, the spline estimators tend to be less accurate at the boundary, compared to some other estimation methods such as local linear regression. As we show in our simulation, reasonably good mean function estimates is necessary for the weighted estimator to work well. Thus if boundary problem is a serious concern, one may consider local linear regression, at the cost of increased computational burden.

If the number of predictors is large and curse of high-dimensionality is a concern, we should consider variable selection. Fortunately, there is a wealthy of literature on the topic of variable selection in the past two decades and many existing variable selection procedures can be easily extended to our setting. For example, we can add some sparsity penalty terms such as LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001) to the objective functions (4) and (8), respectively.

In addition, the asymptotic properties of weighted estimators for the mean function have been studied in the literature by many authors and it is well known that in general weighted estimators are more efficient than unweighted estimators. However, the asymptotic properties of weighted estimators for the variance function draw much less attention. In this paper, we investigate the asymptotic properties of weighted estimators for the variance function for VF-APLMs and show that weighted estimators are more efficient than unweighted estimators.

Finally, we emphasize that a great deal of effort be put on making decision on which predictors should be the linear components of both the mean and variance functions. Scatterplots of the response variable versus those predictors or initial fitting using a fully additive model could help us make such decision for the mean function, as we demonstrate in Section 5. For the notational simplicity, in the main context we assume that the same subset of predictors are put in the linear components of both the mean and variance function. The model (1) can be easily extended to allow different subsets of predictors to be put in the linear components of the mean and variance functions, as we demonstrate in Section 5.

Appendix

Let $\|\cdot\|$ be the Euclidean norm. For matrix $\mathbf{A}$, denote its L_2 norm as $\|\mathbf{A}\|_2 = \sup_{\|\mathbf{u}\|\neq 0} \|\mathbf{A}\mathbf{u}\| / \|\mathbf{u}\|$. Let $\|\xi\|_\infty = \sup_z |\xi(z)|$ be the supremum norm of a function ξ on $[0,1]$.

Following Stone (1985) and Huang (2003), for any measurable functions ζ_1, ζ_2 on $[0,1]^K$, we take the empirical inner product and the corresponding norm to be

$$\langle \zeta_1, \zeta_2 \rangle_n = n^{-1} \sum_{i=1}^{n} \zeta_1(\boldsymbol{Z}_i)\, \zeta_2(\boldsymbol{Z}_i), \quad \|\zeta\|_n^2 = n^{-1} \sum_{i=1}^{n} \zeta^2(\boldsymbol{Z}_i),$$

where $\{\boldsymbol{Z}_i\}$ is a sample from density f. If ζ_1 and ζ_2 are L^2-integrable, take the

inner product

$$\langle \zeta_1, \zeta_2 \rangle = \int_{[0,1]^K} \zeta_1(\boldsymbol{z})\, \zeta_2(\boldsymbol{z})\, f(\boldsymbol{z}) d\boldsymbol{z},$$

with the corresponding induced norm $\|\zeta\|_2^2 = \int_{[0,1]^K} \zeta^2(\boldsymbol{z})\, f(\boldsymbol{z}) d\boldsymbol{z}$. The empirical and theoretical norm of a univariate function ξ on $[0,1]$ are to be

$$\|\xi\|_{nk}^2 = n^{-1} \sum_{i=1}^{n} \xi^2(Z_{ik}), \quad \|\xi\|_{2k}^2 = \int_0^1 \xi^2(z_k)\, f_k(z_k)\, dz_k,$$

where f_k is the density of Z_k for $k = 1, \ldots, K$. Define the centered version spline basis

$$b_{j,k}^*(z_k) = b_{j,k}(z_k) - \frac{E(b_{j,k})}{E(b_{1,k})} b_{1,k}(z_k), \ k = 1, \ldots, K, \ j = -\varrho + 1, \ldots, J_n,$$

with the standardized version given by, for any $k = 1, \ldots, K$,

$$B_{j,k}(z_k) = \frac{b_{j,k}^*(z_k)}{\left\| b_{j,k}^* \right\|_{2k}}, \ j = -\varrho + 1, \ldots, J_n. \tag{A.1}$$

Notice that finding the $(\boldsymbol{\eta}, \boldsymbol{\alpha})$ that minimizes (4) is equivalent to finding the $(\boldsymbol{\eta}, \boldsymbol{\alpha})$ that minimizes

$$\frac{1}{2} \sum_{i=1}^{n} \left[Y_i - \left\{ \boldsymbol{\eta}^{\mathcal{T}} \boldsymbol{B}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}} \boldsymbol{\alpha} \right\} \right]^2,$$

where $\boldsymbol{B}(\boldsymbol{z}) = \{B_{j,k}(z_k), \ j = -\varrho + 1, \ldots, J_n, k = 1, \ldots, K\}^{\mathcal{T}}$. Then the spline estimator of g_0 is $\widehat{g}(\boldsymbol{z}) = \widehat{\boldsymbol{\eta}}^{\mathcal{T}} \boldsymbol{B}(\boldsymbol{z})$ and the centered spline estimator of a component function is

$$\widehat{g}_k(z_k) = \sum_{j=-\varrho+1}^{J_n} \widehat{\eta}_{j,k} B_{j,k}(z_k) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=-\varrho+1}^{J_n} \widehat{\eta}_{j,k} B_{j,k}(Z_{ik}), \ k = 1, \ldots, K.$$

Similarly, finding the $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ that minimizes (8) is equivalent to finding the $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ minimizing

$$\frac{1}{2} \sum_{i=1}^{n} \left[\widehat{R}_i - \phi \left\{ \boldsymbol{\gamma}^{\mathcal{T}} \boldsymbol{B}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}} \boldsymbol{\beta} \right\} \right]^2.$$

Then the spline estimator of h_0 is $\widehat{h}(\boldsymbol{z}) = \widehat{\boldsymbol{\gamma}}^{\mathcal{T}} \boldsymbol{B}(\boldsymbol{z})$ and the centered components are

$$\widehat{h}_k(z_k) = \sum_{j=-\varrho+1}^{J_n} \widehat{\gamma}_{j,k} B_{j,k}(z_k) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=-\varrho+1}^{J_n} \widehat{\gamma}_{j,k} B_{j,k}(Z_{ik}), \ k = 1, \ldots, K.$$

In practice, basis $\{b_{j,k}, j = -\varrho, \ldots, J_n, k = 1, \ldots, K\}^{\mathcal{T}}$ is used for data-analytic implementation, and basis (A.1) is convenient for asymptotic analysis.

A.1. Proof of Theorem 1

The proof of Theorem 1 is similar to that of Theorem 1 in Liu et al. (2011), except that in proving their theorem, Liu et al. (2011) assumed that the intercept α_0 is zero and $\Gamma_0 = \Gamma_0^{\text{add}}$ in their Assumption (C5). Here we partition the predictor vector into $\boldsymbol{X} = (1, \boldsymbol{X}^{*\mathcal{T}})^{\mathcal{T}}$ to relax their zero-intercept assumption. And, we define $\widetilde{\boldsymbol{X}}_{\setminus 0}$ as $\boldsymbol{X} - \Gamma_0^{\text{add}}(\boldsymbol{Z})$ to relax their assumption that $\Gamma_0 = \Gamma_0^{\text{add}}$; in Liu et al. (2011), they defined $\widetilde{\boldsymbol{X}}_{\setminus 0}$ as $\boldsymbol{X} - \Gamma_0(\boldsymbol{Z})$. □

A.2. Proof of Theorem 2

According to the result of de Boor (2001, page 149), for any function $\xi \in \mathcal{H}_{r,\nu}$ and $n \geq 1$, there exists a function $\widetilde{\xi} \in \mathcal{S}_n$ such that $\|\widetilde{\xi} - \xi\|_\infty \leq C J_n^{-p}$. For h_0 satisfying (A1), we can find $\widetilde{\boldsymbol{\gamma}} = \{\widetilde{\gamma}_{j,k}, j = -\varrho + 1, \ldots, J_n, k = 1, \ldots, K\}^{\mathcal{T}}$ and an additive spline function $\widetilde{h} = \widetilde{\boldsymbol{\gamma}}^{\mathcal{T}} \boldsymbol{B}(\boldsymbol{z}) \in \mathcal{A}_n$, such that

$$\|\widetilde{h} - h_0\|_\infty = O(J_n^{-p}). \tag{A.2}$$

In the following, let

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{i=1}^{n} \left[\widehat{R}_i - \phi\{\widetilde{h}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}} \boldsymbol{\beta}\} \right]^2. \tag{A.3}$$

Let $\boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{Z})$. Write $m_0(\boldsymbol{T}) = h_0(\boldsymbol{Z}) + \boldsymbol{X}^{\mathcal{T}} \boldsymbol{\beta}_0$, $m_{0i} = m_0(\boldsymbol{T}_i) = h_0(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}} \boldsymbol{\beta}_0$, $\widetilde{m}_0(\boldsymbol{T}) = \widetilde{h}(\boldsymbol{Z}) + \boldsymbol{X}^{\mathcal{T}} \boldsymbol{\beta}_0$, and $\widetilde{m}_{0i} = \widetilde{m}_0(\boldsymbol{T}_i) = \widetilde{h}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}} \boldsymbol{\beta}_0$.

Lemma 1. *Under Assumptions (A1)–(A5), there exists a local minimizer $\widetilde{\boldsymbol{\beta}}$ of (A.3) such that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$, which can be further shown that*

$$\sqrt{n} \widetilde{Q}_\beta (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to MVN(\boldsymbol{0}, \widetilde{\Sigma}_\beta),$$

where $\widetilde{Q}_\beta = E\{(\Phi^{(1)} \boldsymbol{X})^{\otimes 2}\}$ *and* $\widetilde{\Sigma}_\beta = E\{(e\Phi^{(1)} \boldsymbol{X} - \varepsilon E\{D\Phi^{(1)} \boldsymbol{X}\boldsymbol{X}^{\mathcal{T}}\} Q_\alpha^{-1} \widetilde{\boldsymbol{X}}_{i\setminus 0})^{\otimes 2}\}$.

Proof of Lemma 1. Let $\widehat{\boldsymbol{\delta}} = \sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. According to (A.3), $\widehat{\boldsymbol{\delta}}$ minimizes

$$\widetilde{l}_n(\boldsymbol{\delta}) = \frac{1}{2} \sum_{i=1}^{n} \left[\left\{ \widehat{R}_i - \phi\left(\widetilde{m}_{0i} + n^{-1/2} \boldsymbol{\delta}^{\mathcal{T}} \boldsymbol{X}_i \right) \right\}^2 - \left\{ \widehat{R}_i - \phi(\widetilde{m}_{0i}) \right\}^2 \right].$$

By expansion, we have

$$\begin{aligned} \widetilde{l}_n(\boldsymbol{\delta}) = & -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\widehat{R}_i - \phi(\widetilde{m}_{0i})\} \phi^{(1)}(\widetilde{m}_{0i}) \boldsymbol{X}_i^{\mathcal{T}} \boldsymbol{\delta} + \frac{1}{2} \boldsymbol{\delta}^{\mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \phi^{(1)2}(\widetilde{m}_{0i}) \boldsymbol{X}_i \boldsymbol{X}_i^{\mathcal{T}} \right. \\ & \left. - \frac{1}{n} \sum_{i=1}^{n} \{\widehat{R}_i - \phi(\widetilde{m}_{0i})\} \phi^{(2)}(\widetilde{m}_{0i}) \boldsymbol{X}_i \boldsymbol{X}_i^{\mathcal{T}} \right] \boldsymbol{\delta} + o_p(1). \end{aligned}$$

The first term on the right-hand-side of the above can be further expressed as

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{R}_i - R_i)\Phi_i^{(1)}\boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\delta} - \frac{1}{\sqrt{n}}(R_i - \Phi_i)\Phi_i^{(1)}\boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\delta} + \frac{1}{\sqrt{n}}\{\phi(\widetilde{m}_{0i}) - \Phi_i\}\Phi_i^{(1)}\boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\delta} + o_p(1). \tag{A.4}$$

Consider the first summation in (A.4). Using an identity in Knight (1998, p. 758), we have

$$\widehat{R}_i - R_i = -\widehat{S}_i\{I_{(\varepsilon_i>0)} - I_{(\varepsilon_i\le 0)}\} + 2\int_0^{\widehat{S}_i}\{I_{(\varepsilon_i\le s)} - I_{(\varepsilon_i\le 0)}\}ds, \tag{A.5}$$

where $\widehat{S}_i = \{\widehat{g}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}\} - \{g_0(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_0\}$. Then the first summation in (A.4) equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\widehat{g}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)\}\Phi_i^{(1)}\boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\delta} \tag{A.6}$$

$$+\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^{\mathcal{T}} \times \Big\{\frac{1}{n}\sum_{i=1}^{n}\Phi_i^{(1)}D_i\boldsymbol{X}_i\boldsymbol{X}_i^{\mathcal{T}}\Big\}\boldsymbol{\delta} \tag{A.7}$$

$$-\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\widehat{S}_i}\{I_{(\varepsilon_i\le s)} - I_{(\varepsilon_i\le 0)}\}ds \times \Phi_i^{(1)}\boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\delta}. \tag{A.8}$$

By Lemma A.5 in Liu et al. (2011), term (A.6) equals $o_p(1)$. By Lemma 7 of Stone (1986) and Theorem 1, $\|\widehat{g} - g_0\|_\infty \le CJ_n^{1/2}\|\widehat{g} - g_0\|_2 = o_p(n^{-1/6})$. Hence, $\max_{1\le i\le n}\widehat{S}_i = o_p(n^{-1/6})$. Let $a_n = n^{1/6}$. Following the proof of Theorem 1 in Knight (1998), we can show that

$$\frac{a_n}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\int_0^{C/a_n}\{I_{(\varepsilon_i\le s)} - I_{(\varepsilon_i\le 0)}\}ds + \int_0^{-C/a_n}\{I_{(\varepsilon_i\le s)} - I_{(\varepsilon_i\le 0)}\}ds\Big\} = O_p(1). \tag{A.9}$$

Noting that $\Phi_i^{(1)}\boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\delta}$ is bounded, we can see that term (A.8) equals $O_p(a_n^{-1})$, which equals $o_p(1)$. In term (A.7), $\frac{1}{n}\sum_{i=1}^{n}\Phi_i^{(1)}D_i\boldsymbol{X}_i\boldsymbol{X}_i^{\mathcal{T}} = E\{\Phi^{(1)}D\boldsymbol{X}\boldsymbol{X}^{\mathcal{T}}\} + o_p(1)$. Combining these results of (A.6)–(A.8), we have

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{R}_i - R_i)\Phi_i^{(1)}\boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\delta} = \boldsymbol{\delta}^{\mathcal{T}}E\{D\Phi^{(1)}\boldsymbol{X}\boldsymbol{X}^{\mathcal{T}}\} \times \sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p(1) \tag{A.10}$$

We can easily see that the second summation in (A.4) equals $-n^{-1/2}\sum e_i\Phi_i^{(1)}\boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\delta}$. By Taylor expression, we can also easily see that the third summation in (A.4) equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\widetilde{g}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)\}\Phi_i^{(1)2}\boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\delta} + o_p(1),$$

which, again by Lemma A.5 in Liu et al. (2011), is equal to $o_p(1)$. In addition, we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\phi^{(1)2}(\widetilde{m}_{0i})\boldsymbol{X}_i\boldsymbol{X}_i^{\mathcal{T}}-\frac{1}{n}\sum_{i=1}^{n}\{\widehat{R}_i-\phi(\widetilde{m}_{0i})\}\phi^{(2)}(\widetilde{m}_{0i})\boldsymbol{X}_i\boldsymbol{X}_i^{\mathcal{T}}=Q_\beta+o_p(1).$$

Together, we have

$$\widetilde{l}_n(\boldsymbol{\delta})=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_i\Phi_i^{(1)}\boldsymbol{X}_i^{\mathcal{T}}\boldsymbol{\delta}\boldsymbol{\delta}^{\mathcal{T}}+\sqrt{n}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)^{\mathcal{T}}E\{D\Phi^{(1)}\boldsymbol{X}\boldsymbol{X}^{\mathcal{T}}\}\boldsymbol{\delta}+\frac{1}{2}\boldsymbol{\delta}^{\mathcal{T}}Q_\beta\boldsymbol{\delta}+o_p(1).$$

By Assumption (A5), $Q_\beta>0$. Then for any $\tau>0$, there exists a large constant C such that

$$\text{Prob}\Big\{\sup_{\|\boldsymbol{\delta}\|=C}\widetilde{l}_n(\boldsymbol{\delta})>\widetilde{l}_n(\mathbf{0})\Big\}\geq 1-\tau.$$

This implies with probability at least $1-\tau$ that there exists a local minimum in the ball $\{\boldsymbol{\beta}_0+n^{-1/2}\boldsymbol{\delta}:\|\boldsymbol{\delta}\|\leq C\}$. Hence, there exists a local minimizer such that $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(n^{-1/2})$. By the expression of $\sqrt{n}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)$ in (16), the second part of the lemma can be derived easily. □

The next lemma is Lemma A.2 in Liu et al. (2011). Let

$$\mathbf{V}_n=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\{\boldsymbol{B}(\boldsymbol{Z}_i)\}^{\otimes 2}, & \boldsymbol{B}(\boldsymbol{Z}_i)\boldsymbol{X}_i^{\mathcal{T}}\\ \boldsymbol{X}_i\boldsymbol{B}^{\mathcal{T}}(\boldsymbol{Z}_i), & \boldsymbol{X}_i^{\otimes 2}\end{pmatrix}. \tag{A.11}$$

Lemma 2. *Under Assumptions (A1)–(A4), there exists a positive constant C such that*

$$\|\mathbf{V}_n\|_2\leq C\quad\text{and}\quad\|\mathbf{V}_n^{-1}\|_2\leq C,\quad\text{a.s.}.$$

In the following, take $\boldsymbol{\theta}=(\boldsymbol{\gamma}^{\mathcal{T}},\boldsymbol{\beta}^{\mathcal{T}})^{\mathcal{T}}$, $\widetilde{\boldsymbol{\theta}}=(\widetilde{\boldsymbol{\gamma}}^{\mathcal{T}},\widetilde{\boldsymbol{\beta}}^{\mathcal{T}})^{\mathcal{T}}$, $\widehat{\boldsymbol{\theta}}=(\widehat{\boldsymbol{\gamma}}^{\mathcal{T}},\widehat{\boldsymbol{\beta}}^{\mathcal{T}})^{\mathcal{T}}$, $\widehat{l}_n(\boldsymbol{\theta})=\ell(\boldsymbol{\gamma},\boldsymbol{\beta})$, and $\widetilde{m}_i\equiv\widetilde{m}(\boldsymbol{T}_i)=\widetilde{h}(\boldsymbol{Z}_i)+\boldsymbol{X}_i^{\mathcal{T}}\widetilde{\boldsymbol{\beta}}=\widetilde{\boldsymbol{\gamma}}^{\mathcal{T}}\boldsymbol{B}(\boldsymbol{Z}_i)+\boldsymbol{X}_i^{\mathcal{T}}\widetilde{\boldsymbol{\beta}}$.

Lemma 3. *Under Assumptions (A1)–(A4), there exists a local minimizer $\widehat{\boldsymbol{\theta}}$ of $\widehat{l}_n(\boldsymbol{\theta})$ such that*

$$\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|=O_p\{(J_n/n)^{1/2}+J_n^{-p}\}.$$

Proof of Lemma 3. Let $a_n=(J_n/n)^{1/2}+J_n^{-p}$ and $\boldsymbol{\delta}=(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})/a_n$. Let $\boldsymbol{W}_i=\big(\boldsymbol{B}(\boldsymbol{Z}_i)^{\mathcal{T}},\boldsymbol{X}_i^{\mathcal{T}}\big)^{\mathcal{T}}$. Then

$$\widehat{l}_n(\boldsymbol{\theta})-\widehat{l}_n(\widetilde{\boldsymbol{\theta}})=\frac{1}{2}\sum_{i=1}^{n}\Big[\widehat{R}_i-\phi\{\widetilde{m}_i+a_n\boldsymbol{\delta}^{\mathcal{T}}\boldsymbol{W}_i\}\Big]^2-\frac{1}{2}\sum_{i=1}^{n}\Big[\widehat{R}_i-\phi(\widetilde{m}_i)\Big]^2,$$

which is equal to

$$-a_n\sum_{i=1}^{n}\big[\widehat{R}_i-\phi(\widetilde{m}_i)\big]\phi^{(1)}(\widetilde{m}_i)\boldsymbol{W}_i^{\mathcal{T}}\boldsymbol{\delta} \tag{A.12}$$

$$+ a_n^2 \boldsymbol{\delta}^{\mathcal{T}} \Big[\sum_{i=1}^{n} \phi^{(1)2}(\widetilde{m}_i) \boldsymbol{W}_i \boldsymbol{W}_i^{\mathcal{T}} - \sum_{i=1}^{n} \big[\widehat{R}_i - \phi(\widetilde{m}_i) \big] \phi^{(2)}(\widetilde{m}_i) \boldsymbol{W}_i \boldsymbol{W}_i^{\mathcal{T}} \Big] \boldsymbol{\delta} \quad \text{(A.13)}$$

$\{1 + o_p(1)\}$.

By expansion, term (A.12) equals

$$-a_n \sum_{i=1}^{n} \Big[\big(\widehat{R}_i - R_i \big) + \big(R_i - \Phi_i \big) - \big(\phi(\widetilde{m}_i) - \Phi_i \big) \Big] \Phi_i^{(1)} \boldsymbol{W}_i^{\mathcal{T}} \boldsymbol{\delta} \{1 + o_p(1)\}.$$

Now we will show that

$$\Big\| \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{R}_i - R_i \big) \Phi_i \boldsymbol{B}(\boldsymbol{Z}_i) \Big\| = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}, \quad \text{(A.14)}$$

$$\Big\| \frac{1}{n} \sum_{i=1}^{n} e_i \Phi_i \boldsymbol{B}(\boldsymbol{Z}_i) \Big\| = O_p\{(J_n/n)^{1/2}\}, \quad \text{(A.15)}$$

$$\Big\| \frac{1}{n} \sum_{i=1}^{n} \big(\phi(\widetilde{m}_i) - \Phi_i \big) \Phi_i \boldsymbol{B}(\boldsymbol{Z}_i) \Big\| = O_p\{(J_n/n)^{1/2}\}. \quad \text{(A.16)}$$

By Theorem 1, $\|\widehat{g} - g_0\|_2 = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$ and $\|\widehat{g} - g_0\|_n = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$. Then

$$\begin{aligned}
&\Big\| \frac{1}{n} \sum_{i=1}^{n} \big\{ \widehat{g}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i) \big\} D_i \Phi_i \boldsymbol{B}(\boldsymbol{Z}_i) \Big\|^2 \\
&\quad \le n^{-2} \lambda_{\max}(\mathbf{B}'\mathbf{B}) \sum_{i=1}^{n} \Big[\big\{ \widehat{g}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i) \big\} D_i \Phi_i \Big]^2 \\
&\quad = n^{-2} \lambda_{\max}(\mathbf{B}\mathbf{B}') \sum_{i=1}^{n} \Big[\big\{ \widehat{g}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i) \big\} D_i \Phi_i \Big]^2 \\
&\quad = n^{-2} \times O_p(n) \times O_p(J_n + n J_n^{-2p}) = O_p(J_n/n + J_n^{-2p}),
\end{aligned}$$

where $(J_n + \varrho) \times n$ matrix $\mathbf{B} = \big(\boldsymbol{B}(\boldsymbol{Z}_1) : \ldots : \boldsymbol{B}(\boldsymbol{Z}_n) \big)$ and $\lambda_{\max}(\mathbf{B}\mathbf{B}') = O_p(n)$ by Lemma 2. Similarly,

$$\Big\| \frac{1}{n} \sum_{i=1}^{n} (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^{\mathcal{T}} \boldsymbol{X}_i D_i \Phi_i \boldsymbol{B}(\boldsymbol{Z}_i) \Big\|^2 = O_p(J_n/n + J_n^{-2p}).$$

Together, by (A.5), we obtain (A.14). By

$$\Big\| - \frac{1}{n} \sum_{i=1}^{n} e_i \Phi_i \boldsymbol{B}\left(\boldsymbol{Z}_i\right) \Big\| = \Big[\sum_{k=1}^{K} \sum_{j=-\varrho+1}^{J_n} \Big\{ \frac{1}{n} \sum_{i=1}^{n} e_i \Phi_i B_{j,k}\left(Z_{ik}\right) \Big\}^2 \Big]^{1/2},$$

$$E\Big[\sum_{k=1}^{K} \sum_{j=-\varrho+1}^{J_n} \Big\{ \frac{1}{n} \sum_{i=1}^{n} e_i \Phi_i B_{j,k}\left(Z_{ik}\right) \Big\}^2 \Big] \le C J_n / n,$$

we obtain (A.15). By (A.2) and Lemma 1, we obtain (A.16). Similarly, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{R}_i - R_i\right)\Phi_i \boldsymbol{X}_i\right\| = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}, \tag{A.17}$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n} e_i \Phi_i \boldsymbol{X}_i\right\| = O_p(n^{-1/2}), \tag{A.18}$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\phi(\widetilde{m}_i) - \Phi_i\right)\Phi_i \boldsymbol{X}_i\right\| = O_p(n^{-1/2}). \tag{A.19}$$

Combining (A.14)–(A.19), we see that term (A.12) equals $O_p(na_n^2)\|\boldsymbol{\delta}\|$. By Assumption (A5) and Lemma 2, we see that term (A.13) equals $O_p(na_n^2)\|\boldsymbol{\delta}\|^2$. Therefore, term (A.13) dominates term (A.12) for large $\|\boldsymbol{\delta}\|$. Then for any $\tau > 0$, there exists a large constant C such that

$$\text{Prob}\Big\{\sup_{\|\boldsymbol{\delta}\|=C} \widehat{l}_n(\widetilde{\boldsymbol{\theta}} + a_n\boldsymbol{\delta}) - \widehat{l}_n(\widetilde{\boldsymbol{\theta}})\Big\} \geq 1 - \tau.$$

This implies there exists a local minimizer such that $\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\| = O_p(a_n)$. □

Lemma 4. *Under Assumptions (A1)–(A5),* $\|\widehat{h} - h_0\|_2 = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$, $\|\widehat{h} - h_0\|_n = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$, $\|\widehat{h}_k - h_{0k}\|_{2k} = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$ *and* $\|\widehat{h}_k - h_{0k}\|_{nk} = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$, *for* $k = 1, \ldots, K$.

Proof of Lemma 4. According to Lemma 2,

$$\left\|\widehat{h} - \widetilde{h}\right\|_2^2 = (\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}})^{\mathcal{T}} \left(\langle B_{j,k}, B_{j',k'}\rangle\right)_{\substack{-\varrho+1\leq j,j'\leq J_n,\\ 1\leq k,k'\leq K}} (\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}) \leq C\left\|\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}\right\|_2^2.$$

Then by Lemma 3, we have $\|\widehat{h} - \widetilde{h}\|_2 = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$, and

$$\left\|\widehat{h} - h_0\right\|_2 \leq \left\|\widehat{h} - \widetilde{h}\right\|_2 + \left\|\widetilde{h} - h_0\right\|_2 = O_p\{(J_n/n)^{1/2} + J_n^{-p}\},$$

where the last equality is from $\left\|\widetilde{h} - h_0\right\|_2 = O_p(J_n^{-p})$. By Lemma 1 of Stone (1985), $\left\|\widehat{h}_k - h_{0k}\right\|_{2k} = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$, $1 \leq k \leq K$.

By Lemma A.8 in Wang and Yang (2007), we have

$$A_n \equiv \sup_{\zeta_1,\zeta_2\in\mathcal{A}_n}\left|\frac{\langle\zeta_1,\zeta_2\rangle_n - \langle\zeta_1,\zeta_2\rangle}{\|\zeta_1\|_2\,\|\zeta_2\|_2}\right| = O\left\{\left(J_n \log n/n\right)^{1/2}\right\},\ a.s. \tag{A.20}$$

Then equation (A.20) implies that $\|\widehat{h} - h_0\|_n = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$ and $\|\widehat{h}_k - h_{0k}\|_{nk} = O_p\{(J_n/n)^{1/2} + J_n^{-p}\}$, for $k = 1, \ldots, K$. □

Lemma 5. *Under Assumptions (A1)–(A6),*

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{g}\left(\boldsymbol{Z}_i\right) - g_0\left(\boldsymbol{Z}_i\right)\right\} D_i \Phi_i^{(1)} \widetilde{\boldsymbol{X}}_{i\backslash 1} = o_p\left(n^{-1/2}\right), \tag{A.21}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{h}\left(\boldsymbol{Z}_i\right)-h_0\left(\boldsymbol{Z}_i\right)\right\}\Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\backslash 1}=o_p\left(n^{-1/2}\right), \tag{A.22}$$

$$\frac{1}{n}\sum_{i=1}^{n}\Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\backslash 1}\Gamma_1^{\mathrm{add}}(\boldsymbol{Z}_i)^{\mathcal{T}}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)=o_p\left(n^{-1/2}\right). \tag{A.23}$$

Proof of Lemma 5. By the same arguments as in the proof of Lemma A.5 in Liu et al. (2011), we can show (A.21). Now we consider (A.22). Let $s\left(\boldsymbol{Z}_i,h\right)=h(\boldsymbol{Z}_i)\Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\backslash 1}$. Note that

$$E\left\{s(\boldsymbol{Z}_i,\widehat{h})-s\left(\boldsymbol{Z},h_0\right)\right\}^2=E\left\{\left(\widehat{h}-h_0\right)\left(\boldsymbol{Z}_i\right)\Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\backslash 1}\right\}^2\leq O\left(\|\widehat{h}-h_0\|_2^2\right).$$

By Lemma A.2 of Huang (1999), the logarithm of the ε-bracketing number of the class of functions $\mathcal{A}_1(\delta)=\left\{s(\cdot,h)-s(\cdot,h_0):h\in\mathcal{A}_n,\|h-h_0\|_2\leq\delta\right\}$ is $c\left\{(J_n+\varrho)\log\left(\delta/\varepsilon\right)+\log\left(\delta^{-1}\right)\right\}$, so the corresponding entropy integral $J_{[]}(\delta,\mathcal{A}_1(\delta),\|\cdot\|_2)\leq c\delta\{(J_n+\varrho)^{1/2}+\log^{1/2}(\delta^{-1})\}$. According to Lemma 7 of Stone (1986) and Lemma 4, $\|\widehat{h}-h_0\|_\infty\leq cJ_n^{1/2}\|\widehat{h}-h_0\|_2=O_p\{J_nn^{-1/2}+J_n^{1/2-p}\}$. Lemma 3.4.2 of van der Vaart and Wellner (1996) implies that, for $a_n=(J_n/n)^{1/2}+J_n^{-p}$,

$$\begin{aligned}
&E\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{h}\left(\boldsymbol{Z}_i\right)-h_0\left(\boldsymbol{Z}_i\right)\right\}\Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\backslash 1}-E\left[\left\{\widehat{h}\left(\boldsymbol{Z}\right)-h_0\left(\boldsymbol{Z}\right)\right\}\Phi^{(1)2}\widetilde{\boldsymbol{X}}_{\backslash 1}\right]\right|\\
&\quad\leq n^{-1/2}Ca_n\left\{\left(J_n+\varrho\right)^{1/2}+\log^{1/2}\left(a_n^{-1}\right)\right\}\\
&\qquad\times\left[1+\frac{ca_n\left\{\left(J_n+\varrho\right)^{1/2}+\log^{1/2}\left(a_n^{-1}\right)\right\}}{a_n^2\sqrt{n}}C_0\right]\\
&\quad\leq O(1)n^{-1/2}a_n\left\{\left(J_n+\varrho\right)^{1/2}+\log^{1/2}\left(a_n^{-1}\right)\right\}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&E\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{h}\left(\boldsymbol{Z}_i\right)-h_0\left(\boldsymbol{Z}_i\right)\right\}\Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\backslash 1}-E\left[\left\{\widehat{h}\left(\boldsymbol{Z}\right)-h_0\left(\boldsymbol{Z}\right)\right\}\Phi^{(1)2}\widetilde{\boldsymbol{X}}_{\backslash 1}\right]\right|\\
&\quad=o\left(n^{-1/2}\right).
\end{aligned}$$

By the definition of $\widetilde{\boldsymbol{X}}_{\backslash 1}$ and Γ_1^{add}, for any measurable function ζ, $E\left\{\zeta\left(\boldsymbol{Z}\right)\Phi^{(1)2}\widetilde{\boldsymbol{X}}_{\backslash 1}\right\}=0$. Hence (A.22) holds.

Finally, consider (A.23). Again, by the definition of $\widetilde{\boldsymbol{X}}_{\backslash 1}$, $\Gamma_1(\boldsymbol{Z})$ and $\Gamma_1^{\mathrm{add}}(\boldsymbol{Z})$, we have

$$\begin{aligned}
E\left\{\Phi^{(1)2}\widetilde{\boldsymbol{X}}_{\backslash 1}\Gamma_1^{\mathrm{add}}(\boldsymbol{Z})\right\}&=E\left\{\Phi^{(1)2}\left[\boldsymbol{X}-\Gamma_1(\boldsymbol{Z})\right]\Gamma_1^{\mathrm{add}}(\boldsymbol{Z})\right\}\\
&\quad+E\left\{\Phi^{(1)2}\left[\Gamma_1(\boldsymbol{Z})-\Gamma_1^{\mathrm{add}}(\boldsymbol{Z})\right]\Gamma_1^{\mathrm{add}}(\boldsymbol{Z})\right\}\\
&=0,
\end{aligned}$$

which implies $n^{-1/2}\sum \Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\backslash 1}\Gamma_1^{\mathrm{add}}(\boldsymbol{Z}_i) = O_p(1)$. Then, (A.23) follows from

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p\{(J_n/n)^{1/2} + J_n^{-p}\},$$

which is implied by Lemmas 1 and 3. □

Now we are ready to prove Theorem 2.

Proof of Theorem 2. According to Assumption (A6), $\Gamma_1^{\mathrm{add}}(\boldsymbol{z}) = \sum_{k=1}^K \Gamma_{1k}^{\mathrm{add}}(z_k)$, where $\Gamma_{1k}^{\mathrm{add}} \in \mathcal{H}_{r,\nu}$. By the result of de Boor (2001, page 149), there exists an empirically centered function $\widetilde{\Gamma}_{1k}^{\mathrm{add}} \in \mathcal{S}_n$, such that $\|\widetilde{\Gamma}_{1k}^{\mathrm{add}} - \Gamma_{1k}^{\mathrm{add}}\|_\infty = O_p(J_n^{-p})$, $k = 1, \ldots, K$. As $\widetilde{\Gamma}_1^{\mathrm{add}}(\boldsymbol{z}) = \sum_{k=1}^K \widetilde{\Gamma}_{1k}^{\mathrm{add}}(z_k)$, $\widetilde{\Gamma}_1^{\mathrm{add}} \in \mathcal{A}_n$. Define a class of functions

$$\mathcal{M}_n = \left\{m(\boldsymbol{x}, \boldsymbol{z}) = h(\boldsymbol{z}) + \boldsymbol{x}^{\mathcal{T}}\boldsymbol{\beta} : h \in \mathcal{A}_n\right\}. \tag{A.24}$$

For any $\boldsymbol{v} \in \mathbb{R}^{(d+1)}$, let $\widehat{m}(\boldsymbol{x}, \boldsymbol{z}) = \widehat{h}(\boldsymbol{z}) + \boldsymbol{x}^{\mathcal{T}}\widehat{\boldsymbol{\beta}}$ and $\widehat{m}_{\boldsymbol{v}} = \widehat{m}(\boldsymbol{x}, \boldsymbol{z}) + \boldsymbol{v}^{\mathcal{T}}\{\boldsymbol{x} - \widetilde{\Gamma}_1^{\mathrm{add}}(\boldsymbol{z})\}$. Then $\widehat{m}_{\boldsymbol{v}} = \{\widehat{h}(\boldsymbol{z}) - \boldsymbol{v}^{\mathcal{T}}\widetilde{\Gamma}_1^{\mathrm{add}}(\boldsymbol{z})\} + (\widehat{\boldsymbol{\beta}} + \boldsymbol{v})^{\mathcal{T}}\boldsymbol{x} \in \mathcal{M}_n$. Note that $\widehat{m}_{\boldsymbol{v}}$ minimizes the function

$$\widehat{l}_n(m) = \frac{1}{2}\sum_{i=1}^n \left[\widehat{R}_i - \phi\{m(\boldsymbol{X}_i, \boldsymbol{Z}_i)\}\right]^2,$$

for all $m \in \mathcal{M}_n$ when $\boldsymbol{v} = \boldsymbol{0}$, thus $\frac{\partial}{\partial \boldsymbol{v}}\widehat{l}_n(\widehat{m}_{\boldsymbol{v}})\Big|_{\boldsymbol{v}=\boldsymbol{0}} = \boldsymbol{0}$. Write $\widehat{m}_i \equiv \widehat{m}(\boldsymbol{X}_i, \boldsymbol{Z}_i)$. Then

$$\begin{aligned}
\boldsymbol{0} \equiv \frac{\partial}{\partial \boldsymbol{v}}\widehat{l}_n(\widehat{m}_{\boldsymbol{v}})\Big|_{\boldsymbol{v}=\boldsymbol{0}} &= -\sum_{i=1}^n \left[\widehat{R}_i - \phi(\widehat{m}_i)\right]\phi^{(1)}(\widehat{m}_i)\left\{\boldsymbol{X}_i - \widetilde{\Gamma}_1^{\mathrm{add}}(\boldsymbol{Z}_i)\right\} \\
&= -\sum_{i=1}^n \left[\widehat{R}_i - \phi(\widehat{m}_i)\right]\phi^{(1)}(\widehat{m}_i)\widetilde{\boldsymbol{X}}_{i\backslash 1} + o_p(n^{1/2}) \\
&= -\sum_{i=1}^n \left[\widehat{R}_i - \phi(\widehat{m}_i)\right]\Phi_i^{(1)}\widetilde{\boldsymbol{X}}_{i\backslash 1} + o_p(n^{1/2}).
\end{aligned}$$

Noting that $\widehat{R}_i - \phi(\widehat{m}_i) = (\widehat{R}_i - R_i) + (R_i - \Phi_i) - [\phi(\widehat{m}_i) - \Phi_i]$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n (\widehat{R}_i - R_i)\Phi_i^{(1)}\widetilde{\boldsymbol{X}}_{i\backslash 1} + \frac{1}{n}\sum_{i=1}^n e_i\Phi_i^{(1)}\widetilde{\boldsymbol{X}}_{i\backslash 1} - \frac{1}{n}\sum_{i=1}^n \{\phi(\widehat{m}_i) - \Phi_i\}\Phi_i^{(1)}\widetilde{\boldsymbol{X}}_{i\backslash 1} \\
&\qquad = o_p(n^{-1/2}).
\end{aligned} \tag{A.25}$$

Consider the first term in (A.25). By (A.9), we have

$$\frac{1}{n}\sum_{i=1}^n \int_0^{\widehat{S}_i} \{I_{(\varepsilon_i \le s)} - I_{(\varepsilon_i \le 0)}\}ds \times \Phi_i^{(1)}\widetilde{\boldsymbol{X}}_{i\backslash 1} = o_p(n^{-1/2}).$$

Then, by the expression of $\widehat{R}_i - R_i$ in (A.5), the first term in (A.25) equals

$$-\frac{1}{n}\sum_{i=1}^n D_i\Phi_i^{(1)}\widetilde{\boldsymbol{X}}_{i\backslash 1}\boldsymbol{X}_i^{\mathcal{T}}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) - \frac{1}{n}\sum_{i=1}^n \{\widehat{g}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)\}D_i\Phi_i^{(1)}\widetilde{\boldsymbol{X}}_{i\backslash 1}.$$

Hence, by (A.21), the first term in (A.25) further equals

$$\frac{1}{n}\sum_{i=1}^{n} E\{D\Phi^{(1)}\widetilde{\boldsymbol{X}}_{\setminus 1}\boldsymbol{X}^{\mathcal{T}}\}Q_{\alpha}^{-1}\varepsilon_i\widetilde{\boldsymbol{X}}_{i\setminus 0}+o_p(n^{-1/2}).$$

Now consider the third term in (A.25). By expansion, it can be expressed as

$$-\frac{1}{n}\sum_{i=1}^{n}\{\widehat{h}(\boldsymbol{Z}_i)-h_0(\boldsymbol{Z}_i)\}\Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\setminus 1}-\frac{1}{n}\sum_{i=1}^{n}\Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\setminus 1}\boldsymbol{X}_i^{\mathcal{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+o_p(n^{-1/2}).$$

By (A.22), it is equal to

$$-\frac{1}{n}\sum_{i=1}^{n}\Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\setminus 1}\widetilde{\boldsymbol{X}}_{i\setminus 1}^{\mathcal{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)-\frac{1}{n}\sum_{i=1}^{n}\Phi_i^{(1)2}\widetilde{\boldsymbol{X}}_{i\setminus 1}\Gamma_1^{\mathrm{add}}(\boldsymbol{Z}_i)^{\mathcal{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+o_p(n^{-1/2}).$$

Hence, by (A.23), the third term in (A.25) equals $Q_{\beta}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$. Combining the final expressions of those three terms in (A.25), we complete the proof of Theorem 2. □

A.3. Proof of Theorem 3

According to Assumption (A6), $\Gamma_2^{\mathrm{add}}(\boldsymbol{z})=\sum_{k=1}^{K}\Gamma_{2k}^{\mathrm{add}}(z_k)$, where $\Gamma_{2k}^{\mathrm{add}}\in\mathcal{H}_{r,\nu}$. By the result of de Boor (2001, page 149), there exists an empirically centered function $\widetilde{\Gamma}_{2k}^{\mathrm{add}}\in\mathcal{S}_n$, such that $\|\widetilde{\Gamma}_{2k}^{\mathrm{add}}-\Gamma_{2k}^{\mathrm{add}}\|_{\infty}=O_p(J_n^{-p})$, $k=1,\ldots,K$. As $\widetilde{\Gamma}_2^{\mathrm{add}}(\boldsymbol{z})=\sum_{k=1}^{K}\widetilde{\Gamma}_{2k}^{\mathrm{add}}(z_k)$, $\widetilde{\Gamma}_2^{\mathrm{add}}\in\mathcal{A}_n$. Define a class of functions

$$\mathcal{W}_n=\left\{w(\boldsymbol{x},\boldsymbol{z})=g(\boldsymbol{z})+\boldsymbol{x}^{\mathcal{T}}\boldsymbol{\alpha}:g\in\mathcal{A}_n\right\}. \tag{A.26}$$

For any $\boldsymbol{u}\in\mathbb{R}^{(d+1)}$, let $\widehat{w}(\boldsymbol{x},\boldsymbol{z})=\widehat{g}_{\mathrm{wls}}(\boldsymbol{z})+\boldsymbol{x}^{\mathcal{T}}\widehat{\boldsymbol{\alpha}}_{\mathrm{wls}}$ and $\widehat{w}_{\boldsymbol{u}}=\widehat{w}(\boldsymbol{x},\boldsymbol{z})+\boldsymbol{u}^{\mathcal{T}}\{\boldsymbol{x}-\widetilde{\Gamma}_2^{\mathrm{add}}(\boldsymbol{z})\}$. Then $\widehat{w}_{\boldsymbol{u}}=\{\widehat{g}(\boldsymbol{z})-\boldsymbol{u}^{\mathcal{T}}\widetilde{\Gamma}_2^{\mathrm{add}}(\boldsymbol{z})\}+(\widehat{\boldsymbol{\alpha}}_{\mathrm{wls}}+\boldsymbol{u})^{\mathcal{T}}\boldsymbol{x}\in\mathcal{W}_n$. Note that $\widehat{w}_{\boldsymbol{u}}$ minimizes the function

$$\widehat{l}_{n,\mathrm{wls}}(w)=\frac{1}{2}\sum_{i=1}^{n}\left[Y_i-w(\boldsymbol{X}_i,\boldsymbol{Z}_i)\right]^2/\widehat{\Phi}_i,$$

for all $w\in\mathcal{W}_n$ when $\boldsymbol{u}=\boldsymbol{0}$, thus $\frac{\partial}{\partial\boldsymbol{u}}\widehat{l}_{n,\mathrm{wls}}(\widehat{w}_{\boldsymbol{u}})\Big|_{\boldsymbol{u}=\boldsymbol{0}}=\boldsymbol{0}$. Write $\widehat{w}_i\equiv\widehat{w}(\boldsymbol{X}_i,\boldsymbol{Z}_i)$. Then

$$\begin{aligned}\boldsymbol{0}\equiv\frac{\partial}{\partial\boldsymbol{u}}\widehat{l}_n(\widehat{w}_{\boldsymbol{u}})\Big|_{\boldsymbol{u}=\boldsymbol{0}}&=-\sum_{i=1}^{n}(Y_i-\widehat{w}_i)\left\{\boldsymbol{X}_i-\widetilde{\Gamma}_2^{\mathrm{add}}(\boldsymbol{Z}_i)\right\}/\widehat{\Phi}_i^2\\&=-\sum_{i=1}^{n}(Y_i-\widehat{w}_i)\widetilde{\boldsymbol{X}}_{i\setminus 2}/\widehat{\Phi}_i^2+o_p(n^{1/2})\\&=-\sum_{i=1}^{n}(Y_i-\widehat{w}_i)\widetilde{\boldsymbol{X}}_{i\setminus 2}/\Phi_i^2+o_p(n^{1/2}),\end{aligned}$$

where the last equality is by expansion and Theorem 2. Then we have

$$\sum_{i=1}^{n} \varepsilon_i \widetilde{\boldsymbol{X}}_{i\backslash 2}/\Phi_i^2 - \sum_{i=1}^{n} \{\widehat{g}_{\text{wls}}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)\}\widetilde{\boldsymbol{X}}_{i\backslash 2}/\Phi_i^2 - \sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_{i\backslash 2}\boldsymbol{X}_i^{\mathcal{T}}(\widehat{\boldsymbol{\alpha}}_{\text{wls}} - \boldsymbol{\alpha}_0) = o_p(n^{1/2}). \tag{A.27}$$

Following similar arguments as in the proof of Lemma 5, we can show that the second term of (A.27) is $o_p(n^{1/2})$. The third term of (A.27) equals

$$\sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_{i\backslash 2}\widetilde{\boldsymbol{X}}_{i\backslash 2}^{\mathcal{T}}(\widehat{\boldsymbol{\alpha}}_{\text{wls}} - \boldsymbol{\alpha}_0) + \sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_{i\backslash 2}\Gamma_2^{\text{add}}(\boldsymbol{Z}_i)^{\mathcal{T}}(\widehat{\boldsymbol{\alpha}}_{\text{wls}} - \boldsymbol{\alpha}_0),$$

where, following similar arguments as in the proof of Lemma 5, the second term is $o_p(n^{1/2})$. Hence, combining final expressions of the terms in (A.27), we prove Theorem 3. □

A.4. Proof of Theorem 4

According to Assumption (A6), $\Gamma_3^{\text{add}}(\boldsymbol{z}) = \sum_{k=1}^{K} \Gamma_{3k}^{\text{add}}(z_k)$, where $\Gamma_{3k}^{\text{add}} \in \mathcal{H}_{r,\nu}$. By the result of de Boor (2001, page 149), there exists an empirically centered function $\widetilde{\Gamma}_{1k}^{\text{add}} \in \mathcal{S}_n$, such that $\|\widetilde{\Gamma}_{3k}^{\text{add}} - \Gamma_{3k}^{\text{add}}\|_\infty = O_p(J_n^{-p})$, $k = 1, \ldots, K$. As $\widetilde{\Gamma}_3^{\text{add}}(\boldsymbol{z}) = \sum_{k=1}^{K} \widetilde{\Gamma}_{3k}^{\text{add}}(z_k)$, $\widetilde{\Gamma}_3^{\text{add}} \in \mathcal{A}_n$.

Consider the same class of functions defined in (A.24), $\mathcal{M}_n$. For any $\boldsymbol{v} \in \mathbb{R}^{(d+1)}$, let $\widehat{m}_{\text{wls}}(\boldsymbol{x}, \boldsymbol{z}) = \widehat{h}_{\text{wls}}(\boldsymbol{z}) + \boldsymbol{x}^{\mathcal{T}}\widehat{\boldsymbol{\beta}}_{\text{wls}}$ and $\widehat{m}_{\boldsymbol{v},\text{wls}} = \widehat{m}_{\text{wls}}(\boldsymbol{x}, \boldsymbol{z}) + \boldsymbol{v}^{\mathcal{T}}\{\boldsymbol{x} - \widetilde{\Gamma}_3^{\text{add}}(\boldsymbol{z})\}$. Then $\widehat{m}_{\boldsymbol{v},\text{wls}} = \{\widehat{h}_{\text{wls}}(\boldsymbol{z}) - \boldsymbol{v}^{\mathcal{T}}\widetilde{\Gamma}_3^{\text{add}}(\boldsymbol{z})\} + (\widehat{\boldsymbol{\beta}}_{\text{wls}} + \boldsymbol{v})^{\mathcal{T}}\boldsymbol{x} \in \mathcal{M}_n$. Note that $\widehat{m}_{\boldsymbol{v},\text{wls}}$ minimizes the function

$$\widehat{l}_{n,\text{wls}}(m) = \frac{1}{2}\sum_{i=1}^{n} \left[\widehat{R}_{i,\text{wls}} - \phi\{m(\boldsymbol{X}_i, \boldsymbol{Z}_i)\}\right]^2,$$

for all $m \in \mathcal{M}_n$ when $\boldsymbol{v} = \boldsymbol{0}$, thus $\frac{\partial}{\partial \boldsymbol{v}}\widehat{l}_{n,\text{wls}}(\widehat{m}_{\boldsymbol{v},\text{wls}})\Big|_{\boldsymbol{v}=\boldsymbol{0}} = \boldsymbol{0}$. Write $\widehat{m}_{i,\text{wls}} \equiv \widehat{m}_{\text{wls}}(\boldsymbol{X}_i, \boldsymbol{Z}_i)$. Then

$$\begin{aligned}
\boldsymbol{0} &\equiv \frac{\partial}{\partial \boldsymbol{v}}\widehat{l}_{n,\text{wls}}(\widehat{m}_{\boldsymbol{v},\text{wls}})\Big|_{\boldsymbol{v}=\boldsymbol{0}} \\
&= -\sum_{i=1}^{n} \left[\widehat{R}_{i,\text{wls}} - \phi(\widehat{m}_{i,\text{wls}})\right] \phi^{(1)}(\widehat{m}_{i,\text{wls}}) \left\{\boldsymbol{X} - \widetilde{\Gamma}_3^{\text{add}}(\boldsymbol{Z}_i)\right\}/\widehat{\Phi}_i^2 \\
&= -\sum_{i=1}^{n} \left[\widehat{R}_{i,\text{wls}} - \phi(\widehat{m}_{i,\text{wls}})\right] \phi^{(1)}(\widehat{m}_{i,\text{wls}})\widetilde{\boldsymbol{X}}_{i\backslash 3}/\widehat{\Phi}_i^2 + o_p(n^{1/2}) \\
&= -\sum_{i=1}^{n} \left[\widehat{R}_{i,\text{wls}} - \phi(\widehat{m}_{i,\text{wls}})\right] \Phi_i^{(1)}\widetilde{\boldsymbol{X}}_{i\backslash 3}/\Phi_i^2 + o_p(n^{1/2}),
\end{aligned}$$

where the last equality is by expansion, together with Theorem 2 and the consistency of $(\widehat{h}_{\mathrm{wls}}, \widehat{\boldsymbol{\beta}}_{\mathrm{wls}})$. The consistency of $(\widehat{h}_{\mathrm{wls}}, \widehat{\boldsymbol{\beta}}_{\mathrm{wls}})$ can be shown following similar arguments as in the proof of Lemma 4. Noting that $\widehat{R}_{i,\mathrm{wls}} - \phi(\widehat{m}_{i,\mathrm{wls}}) = (\widehat{R}_{i,\mathrm{wls}} - R_i) + (R_i - \Phi_i) - [\phi(\widehat{m}_{i,\mathrm{wls}}) - \Phi_i]$, we have

$$\sum_{i=1}^{n} (\widehat{R}_{i,\mathrm{wls}} - R_i)\Phi_i^{(1)} \widetilde{\boldsymbol{X}}_{i\setminus 3}/\Phi_i^2 + \sum_{i=1}^{n} e_i \Phi_i^{(1)} \widetilde{\boldsymbol{X}}_{i\setminus 3}/\Phi_i^2 - \sum_{i=1}^{n} \{\phi(\widehat{m}_{i,\mathrm{wls}}) - \Phi_i\}\Phi_i^{(1)} \widetilde{\boldsymbol{X}}_{i\setminus 3}/\Phi_i^2 = o_p(n^{1/2}). \tag{A.28}$$

Consider the first term in (A.28). By similar arguments as in the proof of (A.9), we have

$$\sum_{i=1}^{n} \int_0^{\widehat{S}_{i,\mathrm{wls}}} \{I_{(\varepsilon_i \le s)} - I_{(\varepsilon_i \le 0)}\} ds \times \Phi_i^{(1)} \widetilde{\boldsymbol{X}}_{i\setminus 3}/\Phi_i^2 = o_p(n^{1/2}),$$

where $\widehat{S}_{i,\mathrm{wls}} = \{\widehat{g}_{\mathrm{wls}}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathcal{T}} \widehat{\boldsymbol{\alpha}}_{\mathrm{wls}}\} - \{g_0(\boldsymbol{Z}_i) - \boldsymbol{X}_i^{\mathcal{T}} \boldsymbol{\alpha}_0\}$. Then, by a similar expression of $\widehat{R}_{i,\mathrm{wls}} - R_i$ as the one of $\widehat{R}_i - R_i$ in (A.5), the first term in (A.28) equals

$$-\sum_{i=1}^{n} D_i \Phi_i^{(1)} \widetilde{\boldsymbol{X}}_{i\setminus 3} \boldsymbol{X}_i^{\mathcal{T}} (\widehat{\boldsymbol{\alpha}}_{\mathrm{wls}} - \boldsymbol{\alpha}_0)/\Phi_i^2 - \sum_{i=1}^{n} \{\widehat{g}_{\mathrm{wls}}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)\} D_i \Phi_i^{(1)} \widetilde{\boldsymbol{X}}_{i\setminus 3}/\Phi_i^2.$$

Hence, by similar arguments as in the proof (A.21), the first term in (A.28) equals

$$\sum_{i=1}^{n} E\{D\Phi^{(1)} \widetilde{\boldsymbol{X}}_{\setminus 3} \boldsymbol{X}^{\mathcal{T}}/\Phi^2\} Q_{\alpha,\mathrm{wls}}^{-1} \varepsilon_i \widetilde{\boldsymbol{X}}_{i\setminus 2}/\Phi_i^2 + o_p(n^{1/2}).$$

Now consider the third term in (A.28). By expansion, it can be expressed as

$$-\sum_{i=1}^{n} \{\widehat{h}_{\mathrm{wls}}(\boldsymbol{Z}_i) - h_0(\boldsymbol{Z}_i)\} \Phi_i^{(1)2} \widetilde{\boldsymbol{X}}_{i\setminus 3}/\Phi_i^2 - \sum_{i=1}^{n} \Phi_i^{(1)2} \widetilde{\boldsymbol{X}}_{i\setminus 3} \boldsymbol{X}_i^{\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{\mathrm{wls}} - \boldsymbol{\beta}_0)/\Phi_i^2 + o_p(n^{1/2}).$$

By similar arguments as in the proof of (A.22), it is equal to

$$-\sum_{i=1}^{n} \Phi_i^{(1)2} \widetilde{\boldsymbol{X}}_{i\setminus 3} \widetilde{\boldsymbol{X}}_{i\setminus 3}^{\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{\mathrm{wls}} - \boldsymbol{\beta}_0)/\Phi_i^2 - \sum_{i=1}^{n} \Phi_i^{(1)2} \widetilde{\boldsymbol{X}}_{i\setminus 3} \Gamma_3^{\mathrm{add}}(\boldsymbol{Z}_i)^{\mathcal{T}} (\widehat{\boldsymbol{\beta}}_{\mathrm{wls}} - \boldsymbol{\beta}_0)/\Phi_i^2 + o_p(n^{1/2}),$$

where, by similar arguments as in the proof of (A.23), the second term is $o_p(n^{1/2})$. Hence, combining final expressions of those terms in (A.28), we complete the proof of Theorem 4. □

Acknowledgements

The authors sincerely thank the editor Professor George Michailidis, the AE and two reviewers for their insightful comments which have greatly improved the manuscript.

References

Bickel, P. (1978). Using residuals robustly i: Tests for heteroscedasticity, nonlinearity, *The Annals of Statistics* **6**: 266–291. MR0474625

Bickel, P. J., Klaasen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, The Johns Hopkins University Press. MR1245941

Box, G. and Hill, W. (1974). Correcting inhomogeneity of variance with power transformation weighting, *Technometrics* **16**: 385–389. MR0356380

Box, G. and Meyer, D. (1986). An analysis for unreplicated fractional factorials, *Technometrics* **28**: 11–18. MR0824728

Cai, T. T. and Wang, L. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression, *The Annals of Statistics* **36**: 2025–2054. MR2458178

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models, *The Annals of Statistics* **10**: 1224–1233. MR0673657

Carroll, R. J. (2003). Variances are not always nuisance parameters, *Biometrics* **59**(2): 211–220. MR1987387

Carroll, R. J. and Härdle, W. (1989). Second order effects in semiparametric weighted least squares regression, *Statistics* **2**: 179–186. MR0996861

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedasticity linear models, *The Annals of Statistics* **10**: 429–441. MR0653518

Carroll, R. and Ruppert, D. (1988). *Transformation and Weighting in Regression*, Chapman& Hall, New York. MR1014890

Davidian, M. and Carroll, R. J. (1987). Variance function estimation, *Journal of the American Statistical Association* **82**: 1079–1091. MR0922172

de Boor, C. (2001). *A Practical Guide to Splines*, Vol. 27 of *Applied Mathematical Sciences*, revised edn, Springer-Verlag, New York. MR1900298

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models, *Journal of the American Statistical Association* **106**: 544–557. MR2847969

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**: 1348–1360. MR1946581

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**(1): 3–14. MR0984989

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London. MR1082147

Huang, J. (1999). Efficient estimation of the partially linear additive Cox model, *The Annals of Statistics* **27**: 1536–1563. MR1742499

Huang, J. (2003). Local asymptotics for polynomial spline regression, *The Annals of Statistics* **31**: 1600–1625. MR2012827

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions, *The Annals of Statistics* **26**(2): 755–770. MR1626024

Li, Q. (2000). Efficient estimation of additive partially linear models, *International Economic Review* **41**: 1073–1092. MR1790072

Lian, H., Lai, P. and Liang, H. (2013). Partially linear structure selection in cox models with varying coefficients, *Biometrics* **69**: 348–357. MR3071053

Lian, H., Liang, H. and Carroll, R. J. (2015). Variance function partially linear single-index models, *Journal of the Royal Statistical Society, Series B* **77**: 171–194. MR3299404

Liang, H., Thurston, S. W., Ruppert, D., Apanasovich, T. and Hauser, R. (2008). Additive partial linear models with measurement errors, *Biometrika* **95**(3): 667–678. MR2443182

Liu, X., Wang, L. and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models, *Statistica Sinica* **21**: 1225–1248. MR2827522

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators, *Econometrica* **62**: 1349–1382. MR1303237

Nierenberg, D., Stukel, T., Baron, J., Dain, B. and Greenberg, E. (1989). Determinants of plasma-leverls of beta-carotene and retinol, *American Journal of Epidemiology* **130**: 511–521.

Opsomer, J. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression, *The Annals of Statistics* **25**: 186–211. MR1429922

Opsomer, J. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling, *Journal of Computational and Graphical Statistics* **8**: 715–732.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Springer. MR1998720

Stone, C. J. (1985). Additive regression and other nonparametric models, *The Annals of Statistics* **13**: 689–705. MR0790566

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models, *The Annals of Statistics* **14**: 590–606. MR0840516

Teschendorff, A. E. and Widschwendter, M. (2012). Differential variability improves the identification of cancer risk markers in dna methylation studies profiling precursor cancer lesions, *Bioinformatics* **28**: 1487–1494.

Thomas, L., Stefanski, L. A. and Davidian, M. (2012). Measurement error model methods for bias reduction and variance estimation in logistic regression with estimated variance predictors, *Technical report*, North Carolina State University.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* **58**: 267–288. MR1379242

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak convergence and empirical processes*, Springer Series in Statistics, Springer-Verlag, New York. With applications to statistics. MR1385671

WANG, L., LIU, X., LIANG, H. and CARROLL, R. (2011). Estimation and variable selection for generalized additive partial linear models, *The Annals of Statistics* **39**: 1827–1851. MR2893854

WANG, L., XUE, L., QU, A. and LIANG, H. (2014). Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates, *The Annals of Statistics* **42**: 592–624. MR3210980

WANG, L. and YANG, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model, *The Annals of Statistics* **35**: 2474–2503. MR2382655

WESTERN, B. and BLOOME, D. (2009). Variance function regressions for studying inequality, *Sociological Methodology* **39**: 293–326.

XUE, L. (2009). Consistent variable selection in additive models, *Statistica Sinica* **19**: 1281–1296. MR2536156

XUE, L. and YANG, L. (2006a). Additive coefficient modeling via polynomial spline, *Statistica Sinica* **16**: 1423–1446. MR2327498

XUE, L. and YANG, L. (2006b). Estimation of semi-parametric additive coefficient model, *Journal of Statistical Planning and Inference* **136**(8): 2506–2534. MR2279819