Varying Index Coefficient Models

Shujie MA and Peter X.-K. SONG

It has been a long history of using interactions in regression analysis to investigate alterations in covariate-effects on response variables. In this article, we aim to address two kinds of new challenges arising from the inclusion of such high-order effects in the regression model for complex data. The first kind concerns a situation where interaction effects of individual covariates are weak but those of combined covariates are strong, and the other kind pertains to the presence of nonlinear interactive effects directed by low-effect covariates. We propose a new class of semiparametric models with varying index coefficients, which enables us to model and assess nonlinear interaction effects between grouped covariates on the response variable. As a result, most of the existing semiparametric regression models are special cases of our proposed models. We develop a numerically stable and computationally fast estimation procedure using both profile least squares method and local fitting. We establish both estimation consistency and asymptotic normality for the proposed estimators of index coefficients as well as the oracle property for the nonparametric function estimator. In addition, a generalized likelihood ratio test is provided to test for the existence of interaction effects or the existence of nonlinear interaction effects. Our models and estimation methods are illustrated by simulation studies, and by an analysis of child growth data to evaluate alterations in growth rates incurred by mother's exposures to endocrine disrupting compounds during pregnancy. Supplementary materials for this article are available online.

KEY WORDS: B-splines; Interaction; Oracle property; Profile estimation; Semiparametric regression; Two-step estimation.

1. INTRODUCTION

Being an important generalization of the classical linear model, varying coefficient models (VCMs) proposed by Hastie and Tibishirani (1993) have been widely applied in real data applications. See also, for example, Cai, Fan, and, Li (2000) and Fan and Zhang (2008) , among others. An important feature of the VCM is that the coefficients of covariates are allowed to change with some other variables through smooth functions, so nonlinear interactions may be assessed. We consider a VCM of the form

$$Y = \sum_{l=1}^{d} m_l\,(Z)\,X_l + \varepsilon, \qquad (1)$$

where Y is the response variable, $(Z, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$ is a vector of predictors consisting of a scalar Z and a d-dimensional vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^{\mathrm{T}}$ with $X_1 = 1$, ε is the error term with mean 0, and $m_l\,(\cdot), l = 1, \ldots, d$, are unknown smooth functions. Such specification of VCM in (1) may be inadequate to address two types of challenges in the analysis of complex data structures. First, variable Z is of low effect (e.g., exposure to a certain pesticide contained in food), so the interaction effect between Z and X_l is hardly detectable. Second, as in our motivating example, variable $\mathbf{Z}$ is multidimensional (e.g., simultaneous exposure to many chemical components), so estimation of the coefficient function $m_l\,(\mathbf{Z})$ will be cumbered by the curse of dimensionality. To overcome such challenges and achieve both dimension reduction and sensible model interpretation, we propose a class of varying index coefficient models (VICM) given as

$$Y = m\,(\mathbf{Z}, \mathbf{X}, \boldsymbol{\beta}) + \varepsilon = \sum_{l=1}^{d} m_l(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_l)X_l + \varepsilon, \qquad (2)$$

where $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{lp})^{\mathrm{T}}$ are the coefficient vectors that vary across different covariates X_l, with β_{lk} being the loading weight for the kth component Z_k of $\mathbf{Z}$. As discussed in Section 2, such varying $\boldsymbol{\beta}_l$ in model (2) differentiate the VICM substantially from the existing models in the literature.

The development of model (2) is motivated by one of our collaborative projects in environmental health sciences. In this study, 214 children with age of 8.1 to 13.8 years old are sampled to assess the impact of in utero exposure to mixtures of endocrine disrupting compounds (EDCs) such as bisphenol a (BPA) and phthalates on child growth and weight status from birth through adolescence. Exposure to 10 EDC agents is measured from mother's blood samples collected during the third trimester of pregnancy. The central statistical task is to investigate whether or not, and if so in which form, fetal exposure to these EDCs at sensitive life stages could modify growth velocity throughout childhood and adolescence. Phthalates are a diverse class of high-production industrial chemicals that are widely used as plasticizers to make plastics more flexible, while bisphenol a (BPA) is a high-production chemical that is popularly used in the manufacture of polycarbonate plastics, epoxy resins, and thermal paper. In the U.S., both BPA and phthalates are still in use, and humans are constantly exposed. According to Meeker (2012), there is great public health concern regarding the potential developmental and reproductive effects resulting from the near ubiquitous environmental exposure to known or suspected EDCs currently experienced among pregnant women and children. It is known that EDCs may affect tempo of physical growth (i.e., weight status) across sensitive periods of development in childhood, which in itself is related to chronic disease risk (Grun and Blumberg 2009; Hatch et al. 2010; La Merrill and

Shujie Ma is Assistant Professor, Department of Statistics, University of California-Riverside, Riverside, CA 92521 (E-mail: shujie.ma@ucr.edu). Peter X.-K. Song is Professor, Department of Biostatistics, University of Michigan, Ann Arbor , MI 48109-2029 (E-mail: pxsong@umich.edu). Ma's research was partially supported by NSF grant DMS 1306972. Song's research was partially supported by NSF grant DMS 1208939. The authors are grateful to the Editor, the Associate Editor, and three anonymous reviewers for their constructive comments that helped us improve the article substantially. Part of the research was conducted by the second author during his visit at the Department of Statistics and Applied Probability, National University of Singapore. He is thankful to the department for computational and other logistic support.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jasa.

Journal of the American Statistical Association
March 2015, Vol. 110, No. 509, Theory and Methods
DOI: 10.1080/01621459.2014.903185

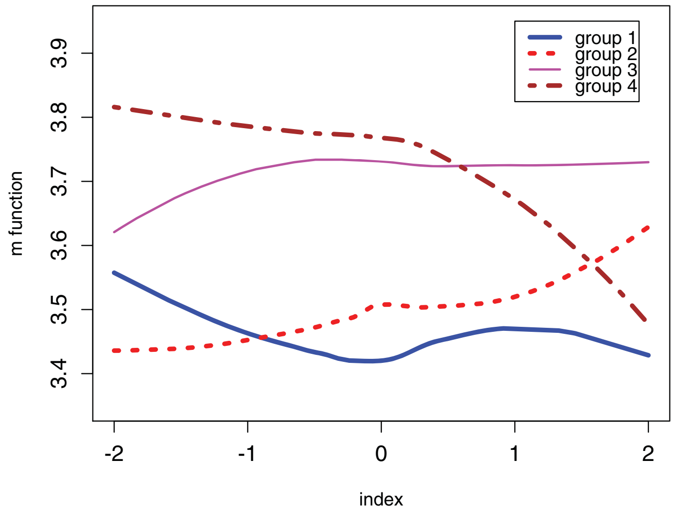

Figure 1. Plots of the estimated m functions against the index $\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}$ scaled on $[-2, 2]$ for the four groups: (1) age$\leq$ 10, gender=boy (solid line); (2) age$\leq$ 10, gender=girl (dotted line); (3) age$>$ 10, gender=boy (thin line); (4) age $>$ 10, gender=girl (dashed dotted line).

Birnbaum 2011) as well as to timing and tempo of sexual maturation. In reality, pregnant women and children are exposed to complex mixtures of chemicals in the environment, and thus, it has become of great importance to study health impacts related to exposure to mixtures of EDCs. Several studies showed that mixtures of reproductive toxicants may disrupt complex signaling pathways and result in cumulative effects on child's growth (Rider et al. 2010).

During the stage of data cleaning, two EDCs agents are removed, resulting in eight EDCs used in our analysis. In the data analysis, we need to answer three important questions: (i) Does exposure to the mixture of the eight EDCs modify the pattern of growth rate? (ii) If so, which EDC components are responsible for the modification? (iii) In which form (linear or nonlinear) does the mixture of EDCs modify the growth pattern? To explore how the mixture of the eight EDCs possibly modifies the association between weight and age as well as that between weight and gender, in a preliminary analysis we stratify the children into four groups: age $\leq$10 or $>$ 10, and gender=0 (boy) or 1 (girl), where age 10 is regarded as the beginning of puberty for girls. For each group, we run regression analysis using the single-index model $E(Y|\mathbf{Z}) = m(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta})$, where Y is the logarithm of weight at current age, $\mathbf{Z}$ is a vector of the eight log-transformed EDC agents. Figure 1 displays four estimated m functions against index $\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}$ scaled on $[-2, 2]$, each for one group. These estimated curves demonstrate clearly different patterns; for example, girl's weight growth appears to be more affected by the exposure, so does the weight growth during puberty (age 10–14). Such preliminary evidence suggests nonlinear interaction effects of the EDC exposure with age and gender. Arguably, the proposed VICM (2) is specified to capture alterations in growth rate patterns directed by the EDC exposure $\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_l$, where covariates are as gender and age.

The proposed VICM (2) is flexible, which encompasses various existing semiparametric models, and the details about the relationship of the VICM to the existing models are discussed in Section 2. In the rest of this section, we focus on the aspect of our contributions to statistical methodology. To address the three questions in the data analysis, we are interested in estimation and inference on both the loading coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_d^{\mathrm{T}})^{\mathrm{T}}$ and the nonparametric functions $m_l(\cdot)$. For the $\boldsymbol{\beta}$, we develop a profile least squares estimation (PLSE) procedure in which each $m_l(\cdot)$ is approximated by B-spline basis functions (de Boor 2001). An important methodological merit of our approach is the ease of simultaneously approximating multiple nonparametric functions to create a single objective function for $\boldsymbol{\beta}$, so that the PLSE can be established in a straightforward manner. In the literature, another version of the PLSE has been considered in the context of single-index models via kernel smoothing by Liang et al. (2010) and Cui, Härdle, and Zhu (2011), among others. However, their kernel-based PLSE may become very complicated to deal with the issue of simultaneously handling multiple nonparametric functions, unless some iterative procedures such as backfitting (Hastie and Tibshirani 1990; Mammen, Linton, and Nielsen 1999; Opsomer and Ruppert 1997) are invoked. A consequence of using iterative procedures is that the profile estimation is no longer available. Moreover, some other commonly used kernel-based methods in the single-index model, such as the backfitting approach proposed by Carroll et al. (1997) and the minimum average variance estimation (MAVE) developed by Xia and Li (1999), Xia, Tong, and Li (1999), and Xia and Härdle (2006), cannot be directly applied in the VICM. On the other hand, the spline estimation approach is also known to be computationally faster than kernel smoothing in semiparametric models (Ma, Song, and Wang 2013; Wang and Yang 2009a; Wang et al. 2011). For the proposed PLSE approach to estimation and inference of parameter $\boldsymbol{\beta}$, this article has made the following contributions: (i) we establish root-n consistency and asymptotic normality of the PLSE for $\boldsymbol{\beta}$. Because the PLSE is to minimize a single objective function, unlike iterative methods (e.g., backfitting), PLSE does not require root-n consistent initial estimators for the large sample properties. (ii) We derive an asymptotic formula for the gradient of the objective function, which makes the PLSE very easy to be implemented and computationally fast through nonlinear optimization. (iii) Since the PLSE of $\boldsymbol{\beta}$ implicitly involves the spline estimates of the nonparametric functions with a divergent number of nuisance parameters, the classical asymptotic theory cannot be directly applied in our setting. We provide a new pathway to establish the asymptotic normality for the PLSE of $\boldsymbol{\beta}$. Subsequently, we devise a Wald chi-square testing procedure for $\boldsymbol{\beta}$ based on the asymptotic distribution of the estimator.

In regard to estimation and inference for the nonparametric $m_l(\cdot)$ functions, although the one-step spline approximation can give a quick estimation of the multiple nonparametric functions, according to Stone (1985), no asymptotic distribution is available for the resulting estimators. To overcome this, we propose to update these splines estimators by the means of local linear smoothing, and show that the resulting estimators enjoy the oracle property; that is, they have the same asymptotic distribution as that of the univariate oracle estimators under the assumption that all the other nonparametric functions were known. This two-step estimation approach is also used in Wang and Yang (2007), Wang and Yang (2009b), and Liu and Yang (2010), in which they use piecewise constant or linear splines in the first step of splines estimation. In this article, we derive the uniform oracle efficiency without restricting the order of splines used in the first step. As a result, our method provides greater flexibility in estimation and inference.

The rest of this article is organized as follows. Section 2 states relationships between the VICM and some important existing models. Section 3 introduces the PLSE and presents asymptotic properties of the proposed estimators. Section 4 discusses the two-step estimation for the nonparametric function $m_l(\cdot)$ and inference for the parameter $\boldsymbol{\beta}$ and $m_l(\cdot)$. In Section 5, we describe the procedure of implementation. In Section 6, we evaluate finite sample properties of the proposed estimation and inference procedures via simulation studies. Section 7 illustrates the proposed model and method through the analysis of child growth data. Some concluding remarks are given in Section 8. All technical details including detailed proofs are provided in the Appendix and the online supplemental materials.

2. RELATIONSHIP TO THE EXISTING MODELS

We begin by noting that in the VICM (2), the loading coefficient vectors $\boldsymbol{\beta}_l$ vary with X-covariates as opposed to a common loading coefficient vector assumed in the single-index coefficient model (SICM) proposed by Xia and Li (1999), Fan, Yao, and Cai (2003), and Xue and Wang (2012). Because of such a difference in model specification, these two classes of models behave rather differently to characterize interaction effects, which are of central interest in our motivating examples. Some of the key differences are summarized as follows:

1. Allowing different loading vectors $\boldsymbol{\beta}_l$ in the VICM enables to engage different components of $\mathbf{Z}$ to modify slope functions of X_l. This is particularly useful to address the second question in our data analysis: which components of EDC agents are responsible for modifying covariate effects. In practice, it seems natural to start with a model with all components of $\mathbf{Z}$ in each function $m_l(\cdot)$, and then let data at hand to pick up an important subset $\mathbf{Z}_l$ of $\mathbf{Z}$ interacting with X_l by the means of, for example, a hypothesis testing procedure. Thus, a VICM (2) used for interpretation may take the form

$$Y = m(\mathbf{Z}, \mathbf{X}, \boldsymbol{\beta}) = \sum_{l=1}^{d} m_l\left(\mathbf{Z}_l^{\mathrm{T}}\boldsymbol{\beta}_l\right) X_l + \varepsilon,$$

   where $\mathbf{Z}_l$ in different indices may be completely or partially overlapped, or completely exclusive. Clearly, the VICM provides flexibility of practical importance for proper interpretation of nonlinear interaction effects. On the contrary, the SICM does not have such flexibility and thus loses desirable model fitting and interpretation. More details may be found in Section 7 of the data application.
2. It is interesting to observe that the VICM can be used to assess nonlinear interactions but the SICM cannot. Consider the case of $m_l(\cdot)$ being linear functions. In the VICM, the linear function $m_l(u) = a_l + u$, turns model (2) into

$$Y = a_1 + \sum_{l=2}^{d} a_l X_l + \sum_{k=1}^{p} \beta_{1k} Z_k + \sum_{l=2}^{d}\sum_{k=1}^{p} \beta_{lk} Z_k X_l + \varepsilon.$$

   In the SICM, the linear function $m_l(u) = a_l + b_l u$, leads the SICM to the form

$$Y = a_1 + \sum_{l=2}^{d} a_l X_l + \sum_{k=1}^{p} b_1 \beta_k Z_k + \sum_{l=2}^{d}\sum_{k=1}^{p} b_l \beta_k Z_k X_l + \varepsilon,$$

   which, apparently, is a linear model with ill-defined interaction effects because usually interaction effects do not satisfy $\beta_{lk} = b_l \beta_k$. Thus, the SICM is short of proper interpretability, and does not allow to test regular interaction effects between each Z_k and X_l.
3. When the $\boldsymbol{\beta}_l$ vectors are given, the SICM and the VICM give rise to different nonparametric models; the former is a varying-coefficient model that technically involves one nonparametric function, and the latter is an additive model that contains multiple nonparametric functions in estimation and inference.

In the literature, the varying-coefficient single-index model (VCSIM, Wong, Ip, and Zhang 2008) is another popular semiparametric model whose specification appears to be similar to that of the VICM. A VCSIM takes the form

$$Y = m(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}) + \sum_{l=1}^{d} \alpha_l(U) X_l + \varepsilon,$$

which is an extension of the partially linear single-index model (PLSIM, Carroll et al. 1997), where coefficients of covariates X_l vary with a scalar variable U. As a matter of fact, the VCSIM and VICM are clearly distinct. The VCSIM does not suit for the purpose of assessing alterations in effects of X_l directed by a set of multiple variables $Z_1, \ldots, Z_p$.

Albeit the aforementioned differences, technically both SICM and VCSIM may be regarded as special cases of the VICM by forcing common $\boldsymbol{\beta}$ for the SICM and using one single variable U (or $p = 1$) in the index for the VCSIM. The class of VICM models specified by (2) is quite general. Besides the SICM and the VCSIM, it encompasses many other existing models as special cases, such as the linear regression model when $m_l(\cdot)$ are assumed to be constant or linear function; the single-index model when $d = 1$; the partial linear single-index model (PLSIM, Carroll et al. 1997; Liang et al. 2010; Lu et al. 2006; Xia, Tong, and Li 1999) when $m_l(\cdot)$ are set as constant for $l \geq 2$; the additive index models when $X_l \equiv 1$ for all $1 \leq l \leq d$; the additive model (Hastie and Tibshirani 1990; Wang and Yang 2007) when $\boldsymbol{\beta}_l$ are given and $X_l \equiv 1$; the partially linear additive model (PLAM, Ma and Yang 2011; Wang et al. 2011) when some of $m_l(\cdot)$ are specified as constant; and the varying coefficient model (Härdle, Hall, and Ichimura 1993) when one variable is included in the nonparametric functions.

3. PROFILE LEAST SQUARES ESTIMATION

Denote an index by $U_l\left(\boldsymbol{\beta}_l\right) = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_l$, which is assumed to be confined in a compact set $[a, b]$, and without loss of generality, set $[a, b] = [0, 1]$. For the error term ε, we assume $E(\varepsilon\,|\mathbf{Z}, \mathbf{X}) = 0$ and $\mathrm{var}(\varepsilon\,|\mathbf{Z}, \mathbf{X}) = \sigma^2(\mathbf{Z}, \mathbf{X})$. For the sake of identifiability, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_d^{\mathrm{T}})^{\mathrm{T}}$ belong to the parameter space:

$$\Theta = \left\{\boldsymbol{\beta} = \left(\boldsymbol{\beta}_l^{\mathrm{T}} : 1 \leq l \leq d\right)^{\mathrm{T}} : \left\|\boldsymbol{\beta}_l\right\|_2 = 1, \beta_{l1} > 0, \boldsymbol{\beta}_l \in R^p\right\},$$

where $\|\cdot\|_2$ denotes the L_2 norm of a vector such that $\|\boldsymbol{\zeta}\|_2 = (|\zeta_1|^2 + \cdots + |\zeta_s|^2)^{1/2}$ for any vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_s)^{\mathrm{T}} \in R^s$. Here, we assume $\beta_{l1} > 0$ for all $l = 1, \ldots, d$ for identifiability. In practice, we can let $\beta_{lk_l} > 0$ for any $1 \le k_l \le p$. Suppose $(Y_i, \mathbf{Z}_i, \mathbf{X}_i, \mathbf{U}_i(\boldsymbol{\beta}))$, $1 \le i \le n$, are the iid realizations of $(Y, \mathbf{Z}, \mathbf{X}, \mathbf{U}(\boldsymbol{\beta}))$, where $\mathbf{U}(\boldsymbol{\beta}) = \left(U_1(\boldsymbol{\beta}_1), \ldots, U_d(\boldsymbol{\beta}_d)\right)^{\mathrm{T}}$ and $\beta \in \Theta$. We propose an estimation of parameter $\boldsymbol{\beta}$ by a profile least square procedure. Fixing $\boldsymbol{\beta}$, we estimate nonparametric functions $m_l(u_l)$ by the means of B-splines described as follows. Let $\mathcal{G}_n$ denote the space of polynomial splines of order q. Consider a knot sequence with $N \equiv N_n$ interior knots, denoted by

$$\xi_1 = \cdots = 0 = \xi_q < \xi_{q+1} < \cdots < \xi_{q+N} < 1 = \xi_{N+q+1} = \cdots = \xi_{N+2q},$$

where N increases along with the number of subjects n. Space $\mathcal{G}_n$ consists of functions, say ϖ, satisfying (i) ϖ is a polynomial of degree $q-1$ on each of subintervals $I_s = [\xi_s, \xi_{s+1})$, $s = 0, \ldots, N_n - 1$, and $I_{N_n} = [\xi_{N_n}, 1]$; (ii) for $q \ge 2$, function ϖ is $q-2$ times continuously differentiable on $[0, 1]$. For $0 \le s \le N_n$, let $H_s = \xi_{s+1} - \xi_s$ be the distance between neighboring knots and let $H = \max_{0 \le s \le N_n} H_s$. Following Zhou, Shen, and Wolfe (1998), to study asymptotic properties of the spline estimator of $m_l(\cdot)$, we assume that $\max_{0 \le s \le N_n - 1} |H_{s+1} - H_s| = o(N^{-1})$ and $H / \min_{0 \le s \le N_n} H_s \le M$, where $M > 0$ is a predetermined constant. Such an assumption assures that $M^{-1} < N_n H < M$, which is necessary for numerical implementation. Let $J_n = N_n + q$. Denote the qth order B spline basis for $\mathcal{G}_n$ (de Boor 2001, p. 89) as $\mathbf{B}_q(u) = (B_{s,q}(u) : 1 \le s \le J_n)^{\mathrm{T}}$, $u \in [0, 1]$, with some $q \ge 2$. Then, the nonparametric functions $m_l(u_l)$, $l = 1, \ldots, d$, are estimated by the spline functions

$$\widehat{m}_l(u_l, \boldsymbol{\beta}) = \sum_{s=1}^{J_n} B_{s,q}(u_l) \widehat{\lambda}_{s,l}(\boldsymbol{\beta}) = \mathbf{B}_q(u_l)^{\mathrm{T}} \widehat{\boldsymbol{\lambda}}_l(\boldsymbol{\beta}), \tag{3}$$

where $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = (\widehat{\boldsymbol{\lambda}}_1(\boldsymbol{\beta})^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\lambda}}_d(\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}}$, with $\widehat{\boldsymbol{\lambda}}_l(\boldsymbol{\beta}) = (\widehat{\lambda}_{s,l}(\boldsymbol{\beta}) : 1 \le s \le J_n)^{\mathrm{T}}$, is given by

$$\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \operatorname{argmin}_{\boldsymbol{\lambda} \in R^{dJ_n}} \sum_{i=1}^{n} \left\{ Y_i - \sum_{l=1}^{d} \sum_{s=1}^{J_n} B_{s,q}\left(U_{il}(\boldsymbol{\beta}_l)\right) \lambda_{s,l} X_{il} \right\}^2. \tag{4}$$

Denote $D_i(\boldsymbol{\beta}) = (D_{i,sl}(\boldsymbol{\beta}_l), 1 \le s \le J_n, 1 \le l \le d)^{\mathrm{T}}$ with $D_{i,sl}(\boldsymbol{\beta}_l) = B_{s,q}(U_{il}(\boldsymbol{\beta}_l))X_{il}$ and $\mathbf{D}(\boldsymbol{\beta}) = [(D_1(\boldsymbol{\beta}), \ldots, D_n(\boldsymbol{\beta}))^{\mathrm{T}}]_{n \times J_n d}$. Thus, the solution to (4) is expressed as

$$\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \{\mathbf{D}(\boldsymbol{\beta})^{\mathrm{T}} \mathbf{D}(\boldsymbol{\beta})\}^{-1} \mathbf{D}(\boldsymbol{\beta})^{\mathrm{T}} \mathbf{Y}, \tag{5}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. The estimation procedure of $\boldsymbol{\beta}$ requires estimates of both m_l and its first-order derivative $\dot{m}_l$. According to de Boor (2001, p. 116), $\dot{m}_l$ can be approximated by the spline functions of one order lower than that of m_l. That is, a spline estimator of $\dot{m}_l$ is given by

$$\widehat{\dot{m}}_l(u_l, \boldsymbol{\beta}) = \sum_{s=1}^{J_n} \dot{B}_{s,q}(u_l) \widehat{\lambda}_{s,l}(\boldsymbol{\beta}) = \sum_{s=2}^{J_n} B_{s,q-1}(u_l) \widehat{\omega}_{s,l}(\boldsymbol{\beta}), \tag{6}$$

where

$$\widehat{\omega}_{s,l}(\boldsymbol{\beta}) = (q-1)\{\widehat{\lambda}_{s,l}(\boldsymbol{\beta}) - \widehat{\lambda}_{s-1,l}(\boldsymbol{\beta})\}/(\xi_{s+q-1} - \xi_s),$$

for $2 \le s \le J_n$. In addition, $\widehat{\dot{m}}_l(u_l, \boldsymbol{\beta})$ can be reexpressed as $\widehat{\dot{m}}_l(u_l, \boldsymbol{\beta}) = \mathbf{B}_{q-1}(u_l)^{\mathrm{T}} \mathbf{D}_1 \widehat{\boldsymbol{\lambda}}_l(\boldsymbol{\beta})$, where $\mathbf{B}_{q-1}(u_l) = (B_{s,q-1}(u_l) : 2 \le s \le J_n)^{\mathrm{T}}$ and

$$\mathbf{D}_1 = (q-1) \times \begin{pmatrix} \frac{-1}{\xi_{q+1} - \xi_2} & \frac{1}{\xi_{q+1} - \xi_2} & 0 & \cdots & 0 \\ 0 & \frac{-1}{\xi_{q+2} - \xi_3} & \frac{1}{\xi_{q+2} - \xi_3} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{-1}{\xi_{N+2q-1} - \xi_{N+q}} & \frac{1}{\xi_{N+2q-1} - \xi_{N+q}} \end{pmatrix}_{(J_n-1) \times J_n}. \tag{7}$$

In the estimation of $\boldsymbol{\beta}$, to ensure identifiability, we exclude the first component β_{l1} of $\boldsymbol{\beta}_l$ by setting $\beta_{l1} = (1 - \|\boldsymbol{\beta}_{l,-1}\|_2^2)^{1/2}$, where $\boldsymbol{\beta}_{l,-1} = (\beta_{l2}, \ldots, \beta_{lp})^{\mathrm{T}}$, for all $1 \le l \le d$ (see Cui, Härdle, and Zhu 2011), and reformulate the parameter space of $\boldsymbol{\beta}_l$, $l = 1, \ldots, d$, as

$$\Theta_{-1} = \left[\left\{\left(1 - \|\boldsymbol{\beta}_{l,-1}\|_2^2\right)^{1/2}, \beta_{l2}, \ldots, \beta_{lp}\right\}^{\mathrm{T}} : \|\boldsymbol{\beta}_{l,-1}\|_2^2 < 1\right].$$

Let $\boldsymbol{\beta}_{l,-1} = (\beta_{l2}, \ldots, \beta_{lp})^{\mathrm{T}}$ and let $\mathbf{J}_l = \partial \boldsymbol{\beta}_l / \partial \boldsymbol{\beta}_{l,-1}^{\mathrm{T}}$ be the Jacobian matrix of size $p \times (p-1)$, which is $\mathbf{J}_l = \binom{-\boldsymbol{\beta}_{l,-1}^{\mathrm{T}} / \sqrt{1 - \|\boldsymbol{\beta}_{l,-1}\|_2^2}}{\mathbf{I}_{p-1}}$. Denote the estimator of $\boldsymbol{\beta}_{-1} = (\boldsymbol{\beta}_{1,-1}^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_{d,-1}^{\mathrm{T}})^{\mathrm{T}}$ by $\widehat{\boldsymbol{\beta}}_{-1} = (\widehat{\boldsymbol{\beta}}_{1,-1}, \ldots, \widehat{\boldsymbol{\beta}}_{d,-1})^{\mathrm{T}}$, which can be obtained by $\widehat{\boldsymbol{\beta}}_{-1} = \arg\min_{\boldsymbol{\beta}_{-1} \in \Theta_{-1}} L_n(\boldsymbol{\beta})$, where

$$L_n(\boldsymbol{\beta}) = 2^{-1} \sum_{i=1}^{n} \left\{ Y_i - \sum_{l=1}^{d} \sum_{s=1}^{J_n} B_{s,l}\left(U_{il}(\boldsymbol{\beta}_l)\right) \widehat{\lambda}_{s,l}(\boldsymbol{\beta}) X_{il} \right\}^2, \quad \boldsymbol{\beta}_{-1} \in \Theta_{-1}. \tag{8}$$

Moreover, one can obtain $\widehat{\boldsymbol{\beta}}_{-1}$ as the solution of the following estimation equations:

$$\begin{aligned} \partial L_n(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}_{-1} &= -\sum_{i=1}^{n} \left\{ Y_i - \sum_{l=1}^{d} \sum_{s=1}^{J_n} B_{s,l}\left(U_{il}(\boldsymbol{\beta}_l)\right) \widehat{\lambda}_{s,l}(\boldsymbol{\beta}) X_{il} \right\} \\ &\quad \times \begin{bmatrix} \{\widehat{\dot{m}}_1(U_{i1}(\boldsymbol{\beta}_1), \boldsymbol{\beta}) X_{i1} \mathbf{J}_1^{\mathrm{T}} \mathbf{Z}_i + (\partial \widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta})^{\mathrm{T}} / \partial \boldsymbol{\beta}_{1,-1}) D_i(\boldsymbol{\beta})\} \\ \vdots \\ \{\widehat{\dot{m}}_d(U_{id}(\boldsymbol{\beta}_d), \boldsymbol{\beta}) X_{id} \mathbf{J}_d^{\mathrm{T}} \mathbf{Z}_i + (\partial \widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta})^{\mathrm{T}} / \partial \boldsymbol{\beta}_{d,-1}) D_i(\boldsymbol{\beta})\} \end{bmatrix} \\ &= 0, \end{aligned} \tag{9}$$

where $\widehat{\dot{m}}_l(, \boldsymbol{\beta})$ is given in (6). Now, define the space $\mathcal{M}$ as a collection of functions with finite L_2 norm on $[0, 1]^d \times R^d$ by

$$\mathcal{M} = \left\{ g(\mathbf{u}, \mathbf{x}) = \sum_{l=1}^{d} g_l(u_l) x_l,\ E g_l(U_l)^2 < \infty \right\},$$

where $\mathbf{u} = (u_1, \ldots, u_d)^{\mathrm{T}}$ and $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$. To study the large-sample properties of parameter estimators, let $\boldsymbol{\beta}^0 = \{(\boldsymbol{\beta}_1^0)^{\mathrm{T}}, \ldots, (\boldsymbol{\beta}_d^0)^{\mathrm{T}}\}^{\mathrm{T}}$ with $\boldsymbol{\beta}_l^0 = \{\beta_{l1}^0, (\boldsymbol{\beta}_{l,-1}^0)^{\mathrm{T}}\}^{\mathrm{T}}$ and $\boldsymbol{\beta}_{l,-1}^0 = (\beta_{l2}^0, \ldots, \beta_{lp}^0)^{\mathrm{T}}$ for $1 \le l \le d$ be the true parameters in model

(2). For $1 \le k \le p$, define g_k^0 as the one satisfying:

$$\mathbb{P}(Z_k) = g_k^0(\mathbf{U}(\boldsymbol{\beta}^0), \mathbf{X}) = \sum_{l=1}^{d} g_{l,k}^0\left(U_l\left(\boldsymbol{\beta}_l^0\right)\right) X_l = \arg\min_{g \in \mathcal{M}} E\{Z_k - g(\mathbf{U}(\boldsymbol{\beta}^0), \mathbf{X})\}^2. \tag{10}$$

Let $\mathbb{P}(\mathbf{Z}) = \{\mathbb{P}(Z_1), \ldots, \mathbb{P}(Z_p)\}^{\mathrm{T}}$, $\widetilde{\mathbf{Z}} = \mathbf{Z} - \mathbb{P}(\mathbf{Z})$ and

$$\Phi(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}^0) = \left[\left\{\dot{m}_l\left(U_l\left(\boldsymbol{\beta}_l^0\right), \boldsymbol{\beta}^0\right) X_l \mathbf{J}_l^{\mathrm{T}} \widetilde{\mathbf{Z}}\right\}^{\mathrm{T}}, 1 \le l \le d\right]^{\mathrm{T}}. \tag{11}$$

Here, $\Phi(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}^0)$ is a vector with $(p-1)\,d$ elements. For any matrix $\mathbf{A}$, denote $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. For any positive numbers a_n and b_n, let $a_n \ll b_n$ denote that $a_n/b_n = o(1)$. Let r with $r \ge 2$ be the smoothness order of the coefficient functions $m_l(\cdot)$ as given in Condition (C2) in the Appendix.

Theorem 1. Under Conditions (C1)–(C5) in the Appendix, and $n^{1/(2r+2)} \ll N \ll n^{1/4}$, we have (i) (consistency) $\|\widehat{\boldsymbol{\beta}}_{-1} - \boldsymbol{\beta}_{-1}^0\|_2 = O_p(n^{-1/2})$; (ii) (asymptotic normality) as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{-1} - \boldsymbol{\beta}_{-1}^0\right) = \left\{n^{-1}\sum_{i=1}^{n} \Phi\left(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}^0\right)^{\otimes 2}\right\}^{-1} \times \left\{n^{-1/2}\sum_{i=1}^{n}(Y_i - m(\mathbf{Z}_i, \mathbf{X}_i))\,\Phi\left(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}^0\right)\right\} + o_p(1).$$

Moreover $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{-1} - \boldsymbol{\beta}_{-1}^0) \xrightarrow{d} \mathcal{N}_{d(p-1)}(\mathbf{0}, \Sigma)$, as $n \to \infty$, where

$$\Sigma = [E\{\Phi(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}^0)^{\otimes 2}\}]^{-1}[E\{\sigma^2(\mathbf{Z}, \mathbf{X})\Phi(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}^0)^{\otimes 2}\}] \times [E\{\Phi(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}^0)^{\otimes 2}\}]^{-1}. \tag{12}$$

Remark 1. If we assume homoscedasticity to the random noise ε in model (2), that is, $\sigma^2(\mathbf{Z}, \mathbf{X}) = \sigma^2$ for some constant $\sigma^2 > 0$, then the asymptotic variance matrix given in (12) is reduced to

$$\Sigma = \sigma^2[E\{\Phi(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}^0)^{\otimes 2}\}]^{-1}. \tag{13}$$

Let $\mathbb{J} = \oplus_{l=1}^{d} \mathbf{J}_l = \mathrm{diag}(\mathbf{J}_1, \ldots, \mathbf{J}_d)$ be the direct sum of Jacobian matrices $\mathbf{J}_1, \ldots, \mathbf{J}_d$ and its dimension is $dp \times d(p-1)$. For $1 \le l \le d$, β_{l1} is estimated by $\widehat{\beta}_{l1} = (1 - \sum_{k=2}^{p} \widehat{\beta}_{lk}^2)^{1/2}$. Let $\widehat{\boldsymbol{\beta}}_l = (\widehat{\beta}_{l1}, \ldots, \widehat{\beta}_{lp})^{\mathrm{T}}$. Both consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \ldots . \widehat{\boldsymbol{\beta}}_d^{\mathrm{T}})^{\mathrm{T}}$ follow directly from Theorem 1 with an application of the multivariate delta method. Thus, we obtain $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} \mathcal{N}_{dp}(0, \mathbb{J}\,\Sigma\,\mathbb{J}^{\mathrm{T}})$, $n \to \infty$.

Next, we consider the spline estimator of the nonparametric function $m_l(\cdot)$ given as follows:

$$\widehat{m}_l(u_l, \widehat{\boldsymbol{\beta}}) = \sum_{s=1}^{J_n} B_{s,q}(u_l)\,\widehat{\lambda}_{s,l}(\widehat{\boldsymbol{\beta}}) = \mathbf{B}_q(u_l)^{\mathrm{T}}\,\widehat{\boldsymbol{\lambda}}_l(\widehat{\boldsymbol{\beta}}), \tag{14}$$

where $\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\lambda}}_1(\widehat{\boldsymbol{\beta}})^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\lambda}}_d(\widehat{\boldsymbol{\beta}})^{\mathrm{T}})^{\mathrm{T}}$ with $\widehat{\boldsymbol{\lambda}}_l(\widehat{\boldsymbol{\beta}}) = (\widehat{\lambda}_{s,l}(\widehat{\boldsymbol{\beta}}) : 1 \le s \le J_n)^{\mathrm{T}}$ given by (5) in which $\boldsymbol{\beta}$ is replaced with $\widehat{\boldsymbol{\beta}}$. The following theorem provides the convergence rate of $\widehat{m}_l(u_l, \widehat{\boldsymbol{\beta}})$.

Theorem 2. Under Conditions (C1)–(C5) in the Appendix, and $n^{1/(2r+2)} \ll N \ll n^{1/4}$, we have for each $1 \le l \le d$, $|\widehat{m}_l(u_l, \widehat{\boldsymbol{\beta}}) - m_l(u_l)| = O_p(\sqrt{N/n} + N^{-r})$ uniformly for any $u_l \in [0, 1]$.

Remark 2. The order assumptions regarding N, that is, $n^{1/(2r+2)} \ll N \ll n^{1/4}$, in Theorem 2 implies that $N \asymp n^{1/(2r+1)}$, which is the optimal order for the number of interior knots needed to estimate the nonparametric functions. The resulting convergence rate is then $O_p(n^{-r/(2r+1)})$. For example, when $r = 2$, the optimal convergence rate is $O_p(n^{-2/5})$.

Remark 3. To estimate the asymptotic covariance matrix Σ given in (12), we need to estimate $\Phi(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}^0)$ given by (11). There, $\widetilde{\mathbf{Z}}$ can be estimated by $\widehat{\mathbf{Z}} = \mathbf{Z} - \mathbb{P}_n(\mathbf{Z})$, with $\mathbb{P}_n(\mathbf{Z}) = \{\mathbb{P}_n(Z_1), \ldots, \mathbb{P}_n(Z_p)\}^{\mathrm{T}}$ and $\mathbb{P}_n(Z_k) = \sum_{l=1}^{d} \widehat{g}_{l,k}^0(U_l(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\beta}})X_l$, where $\widehat{g}_{l,k}^0(\cdot, \widehat{\boldsymbol{\beta}})$ is the spline estimate of $g_{l,k}^0(\cdot)$ obtained by carrying out the same procedure as for $\widehat{m}_l(\cdot, \widehat{\boldsymbol{\beta}})$ with the response Y replaced by Z_k. Thus, $\Phi(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}^0)$ is estimated by

$$\widehat{\Phi}(\mathbf{X}, \mathbf{Z}, \widehat{\boldsymbol{\beta}}) = \left[\left\{\widehat{\dot{m}}_l(U_l(\widehat{\boldsymbol{\beta}}_l), \widehat{\boldsymbol{\beta}}) X_l \mathbf{J}_l^{\mathrm{T}} \widehat{\mathbf{Z}}\right\}^{\mathrm{T}}, 1 \le l \le d\right]^{\mathrm{T}},$$

and the resulting estimate of Σ defined in (12) is given by

$$\widehat{\Sigma} = n\left\{\sum_{i=1}^{n} \widehat{\Phi}(\mathbf{X}_i, \mathbf{Z}_i, \widehat{\boldsymbol{\beta}})^{\otimes 2}\right\}^{-1} \times \left\{\sum_{i=1}^{n} \widehat{e}^2(\mathbf{Z}_i, \mathbf{X}_i)\,\widehat{\Phi}(\mathbf{X}_i, \mathbf{Z}_i, \widehat{\boldsymbol{\beta}})^{\otimes 2}\right\} \times \left\{\sum_{i=1}^{n} \widehat{\Phi}(\mathbf{X}_i, \mathbf{Z}_i, \widehat{\boldsymbol{\beta}})^{\otimes 2}\right\}^{-1}, \tag{15}$$

where $\widehat{e}(\mathbf{X}_i, \mathbf{Z}_i) = Y_i - \sum_{l=1}^{d} \widehat{m}_l(\mathbf{Z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_l, \widehat{\boldsymbol{\beta}})X_{il}$. For the homoscedasticity case, Σ in (13) is estimated by

$$\widehat{\Sigma} = \widehat{\sigma}^2 n\left\{\sum_{i=1}^{n} \widehat{\Phi}\left(\mathbf{X}_i, \mathbf{Z}_i, \widehat{\boldsymbol{\beta}}\right)^{\otimes 2}\right\}^{-1}, \tag{16}$$

where $\widehat{\sigma}^2 = \sum_{i=1}^{n} \widehat{e}^2(\mathbf{X}_i, \mathbf{Z}_i)/\{n - d(J_n + p)\}$.

4. INFERENCE

4.1 Oracle Property of a Two-Step Estimation for $m_l(\cdot)$

In Theorem 2, we show that the spline estimator $\widehat{m}_l(\cdot, \widehat{\boldsymbol{\beta}})$ obtained from the profile estimation procedure in (14) is a consistent estimator of $m_l(\cdot)$. The asymptotic distribution of $\widehat{m}_l(\cdot, \widehat{\boldsymbol{\beta}})$, however, is not available. Thus, no measure of confidence can be established in statistical inference. To overcome this, we consider a two-step spline backfitted local linear (SBLL) estimation for the nonparametric function $m_l(\cdot)$, for which the spline estimate $\widehat{m}_l(\cdot, \widehat{\boldsymbol{\beta}})$ given in (14) will be used as the initial estimate. Here, we establish the asymptotic normality for the SBLL estimators. The SBLL estimation proceeds as follows. Without loss of generality, we focus on the estimation of the first nonparametric function $m_1(\cdot)$. The spline estimates $\widehat{m}_l(\cdot, \widehat{\boldsymbol{\beta}})$, $l \ge 2$, given in (14) are used as the initial estimates and held fixed in the estimation of $m_1(\cdot)$, and all the other functions can be estimated in a similar fashion. When $m_l(\cdot)$ for $l \ge 2$ were known, we could define the oracle pseudoresponse $Y_{i,1} = Y_i - \sum_{l=2}^{d} m_l(\mathbf{Z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_l)X_{il} = m_1(\mathbf{Z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_1)X_{i1} + \varepsilon_i$, where $\widehat{\boldsymbol{\beta}}_l$ are the PLSE given in Section 3. For each given u_1, $m_1(u_1)$ is estimated by the means of local linear fitting, namely $\widetilde{m}_{\mathrm{LL},1}(u_1, \widehat{\boldsymbol{\beta}}) = \widehat{a}(\widehat{\boldsymbol{\beta}})$, where $\widehat{a}(\widehat{\boldsymbol{\beta}})$ and $\widehat{b}(\widehat{\boldsymbol{\beta}})$ minimize the following local kernel objective function:

$$\sum_{i=1}^{n} \{Y_{i,1} - aX_{i1} - b(U_{i1}(\widehat{\boldsymbol{\beta}}_1) - u_1)X_{i1}\}^2 K_{h_1}(U_{i1}(\widehat{\boldsymbol{\beta}}_1) - u_1).$$

Here, $K_{h_1}(u) = K(u/h_1)/h_1$ is a symmetric kernel function and h_1 is a bandwidth. Let

$$\mathbf{C}(u_1, \widehat{\boldsymbol{\beta}}_1) = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ X_{11}\{(U_{11}(\widehat{\boldsymbol{\beta}}_1) - u_1) & \cdots & X_{1n}\{(U_{1n}(\widehat{\boldsymbol{\beta}}_1) - u_1) \\ /h_1\} & & /h_1\} \end{bmatrix}^{\mathrm{T}},$$

$$\mathbf{W}(u_1, \widehat{\boldsymbol{\beta}}_1) = \mathrm{diag}\{K_{h_1}(U_{11}(\widehat{\boldsymbol{\beta}}_1) - u_1), \ldots, K_{h_1}(U_{1n}(\widehat{\boldsymbol{\beta}}_1) - u_1)\},$$

and $\mathbf{Y}_1 = (Y_{1,1}, \ldots, Y_{n,1})^{\mathrm{T}}$. Then, we have

$$\widehat{a}(\widehat{\boldsymbol{\beta}}) = (1,0)\{\mathbf{C}(u_1, \widehat{\boldsymbol{\beta}}_1)^{\mathrm{T}}\mathbf{W}(u_1, \widehat{\boldsymbol{\beta}}_1)\mathbf{C}(u_1, \widehat{\boldsymbol{\beta}}_1)\}^{-1} \times \mathbf{C}(u_1, \widehat{\boldsymbol{\beta}}_1)^{\mathrm{T}}\mathbf{W}(u_1, \widehat{\boldsymbol{\beta}}_1)\mathbf{Y}_1. \tag{17}$$

Because $m_l(u_l)$ for $l \geq 2$ are actually unknown, we modify (17) by replacing $m_l(u_l)$ with their spline estimators $\widehat{m}_l(u_l, \widehat{\boldsymbol{\beta}})$ given in (14), which is equivalent to replacing $\mathbf{Y}_1$ in (17) by $\widehat{\mathbf{Y}}_1$, where $\widehat{\mathbf{Y}}_1 = (\widehat{Y}_{1,1}, \ldots, \widehat{Y}_{n,1})^{\mathrm{T}}$ and $\widehat{Y}_{i,1} = Y_i - \sum_{l=2}^{d} \widehat{m}_l(\mathbf{Z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_l, \widehat{\boldsymbol{\beta}})X_{il}$. The resulting SBLL estimator is denoted by $\widehat{m}_{\mathrm{SBLL},1}(u_1, \widehat{\boldsymbol{\beta}})$. Denote $\mu_2(K) = \int u^2 K(u)\,du$ and $\|K\|_2^2 = \int K^2(u)\,du$.

Theorem 3. Under Conditions (C1)–(C6) in the Appendix, $h_1 \asymp n^{-1/5}$, and $n^{1/(2r+2)} \ll N \ll n^{1/4}$, as $n \to \infty$, for any $u_1 \in [h_1, 1-h_1]$, we have

$$\sup_{u_1 \in [h_1, 1-h_1]} |\widetilde{m}_{\mathrm{LL},1}(u_1, \widehat{\boldsymbol{\beta}}) - m_1(u_1)| = O_p(\sqrt{\log(n)/(nh_1)}) = O_p(n^{-2/5}\sqrt{\log(n)}),$$

and

$$\sqrt{nh_1}\left\{\widetilde{m}_{\mathrm{LL},1}\left(u_1, \widehat{\boldsymbol{\beta}}\right) - m_1(u_1) - b_1(u_1)h_1^2\right\} \xrightarrow{d} \mathcal{N}(0, v_1(u_1)),$$

where

$$b_1(u_1) = \mu_2(K)\ddot{m}_1(u_1)/2,$$
$$v_1(u_1) = \left\{E\left(X_1^2 | u_1\right)\right\}^{-2} f_1^{-1}(u_1)\|K\|_2^2 E\left\{X_1^2\sigma^2(\mathbf{Z}, \mathbf{X}) | u_1\right\}.$$

Here $\ddot{m}_1(\cdot)$ is the second-order derivative of m_1 and $f_1(\cdot)$ is the density function of $\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_1^0$.

Theorem 4 presents the uniform oracle efficiency of the SBLL estimator $\widehat{m}_{\mathrm{SBLL},1}(u_1, \widehat{\boldsymbol{\beta}})$ such that the absolute difference between $\widehat{m}_{\mathrm{SBLL},1}(u_1, \widehat{\boldsymbol{\beta}})$ and $\widetilde{m}_{\mathrm{LL},1}(u_1, \widehat{\boldsymbol{\beta}})$ is of order $o_p(n^{-2/5})$ uniformly. As a result, $\widehat{m}_{\mathrm{SBLL},1}(u_1, \widehat{\boldsymbol{\beta}})$ has the same asymptotic distribution as $\widetilde{m}_{\mathrm{LL},1}(u_1, \widehat{\boldsymbol{\beta}})$.

Theorem 4. Under Conditions (C1)–(C6) in the Appendix, and $\max\{n^{1/(2r+2)}, n^{2/(5r)}\} \ll N \ll n^{1/4}$, we have

$$\sup_{u_1 \in [0,1]} |\widehat{m}_{\mathrm{SBLL},1}(u_1, \widehat{\boldsymbol{\beta}}) - \widetilde{m}_{\mathrm{LL},1}(u_1, \widehat{\boldsymbol{\beta}})| = O_p(n^{-1/2} + N^{-r}) = o_p(n^{-2/5}).$$

Corollary 1. Under Conditions (C1)–(C6) in the Appendix, $h_1 \asymp n^{-1/5}$, and $\max\{n^{1/(2r+2)}, n^{2/(5r)}\} \ll N \ll n^{1/4}$, for any $u_1 \in [h_1, 1-h_1]$, as $n \to \infty$, we have

$$\sqrt{nh_1}\{\widehat{m}_{\mathrm{SBLL},1}(u_1, \widehat{\boldsymbol{\beta}}) - m_1(u_1) - b_1(u_1)h_1^2\} \xrightarrow{d} \mathcal{N}(0, v_1(u_1)).$$

Remark 4. When $m_l(\cdot)$ for $2 \leq l \leq d$ are rth order smooth functions with $r \geq 3$, under the assumption of N given in Corollary 1, the same optimal order $N \asymp n^{1/(2r+1)}$ as given in Remark 2 can be applied in the first step of spline estimation. Let r equal to the spline order q as given in Zhou, Shen, and Wolfe (1998), we have $N \asymp n^{1/(2q+1)}$ for $q \geq 3$. When $q = 2$ such that linear splines are used in the first step, then an undersmoothing procedure is needed with N satisfying $n^{1/5} \ll N \ll n^{1/4}$. In the second step of SBLL estimation, the bandwidth satisfies the optimal order $h_1 \asymp n^{-1/5}$.

4.2 Inference for Loading Parameter $\boldsymbol{\beta}$

With the availability of asymptotic normality in Theorem 1, we can easily derive a Wald chi-square testing procedure to test whether a subset of $\boldsymbol{\beta}_l = (\beta_{2l}, \ldots, \beta_{pl})$, $l = 1, \ldots, d$, equals to zero. Let K be an integer satisfying $2 \leq K \leq p$, and let $(k_1, \ldots k_K)$ be a subset of indices in $\{2, \ldots, p\}$. The null hypothesis of interest is: $H_0 : \beta_{k_1 l} = \beta_{k_2 l} = \cdots = \beta_{k_K l} = 0$ for the lth loading coefficients. From Theorem 1, a Wald test statistic takes the form $\chi_W^2 = (\widehat{\boldsymbol{\beta}}_{Kl} - \mathbf{0}_K)^{\mathrm{T}}\{\widehat{\mathrm{V}}(\widehat{\boldsymbol{\beta}}_{Kl})\}^{-1}(\widehat{\boldsymbol{\beta}}_{Kl} - \mathbf{0}_K)$, where $\widehat{\boldsymbol{\beta}}_{Kl} = (\widehat{\beta}_{k_1 l}, \widehat{\beta}_{k_2 l}, \ldots, \widehat{\beta}_{k_K l})^{\mathrm{T}}$ is the PLSE of $\boldsymbol{\beta}_{Kl} = (\beta_{k_1 l}, \beta_{k_2 l}, \ldots, \beta_{k_K l})^{\mathrm{T}}$, and $\{\widehat{\mathrm{V}}(\widehat{\boldsymbol{\beta}}_{Kl})\}^{-1}$ is the inverse of the estimated asymptotic variance–covariance matrix of $\widehat{\boldsymbol{\beta}}_{Kl}$. Under H_0, χ_W^2 follows asymptotically the central chi-square distribution with K degrees of freedom.

4.3 Inference for Nonparametric Function $m_l(\cdot)$

For a given $1 \leq l \leq d$, both main and interaction effects of X_l are related to the nonparametric function $m_l(\cdot)$. To test whether $m_l(\cdot)$ has a specific parametric form, we set up the hypothesis testing as: $H_0 : m_l(\cdot) = m_{\boldsymbol{\theta},l}(\cdot)$ versus $H_a : m_l(\cdot) \neq m_{\boldsymbol{\theta},l}(\cdot)$, where $m_{\boldsymbol{\theta},l}(\cdot)$ is a certain given parametric function with the p_θ-dimensional parameter vector $\boldsymbol{\theta}$. For example, setting $m_{\boldsymbol{\theta},l}(u_l) \equiv \theta_{l0}$ (constant), we aim to test whether there exist any interaction effects, while setting $m_{\boldsymbol{\theta},l}(u_l) = \theta_{l1} + \theta_{l2}u_l$ (a linear function), we attempt to test whether there exists a linear interaction effect between U_l and X_l. Following Fan, Zhang, and Zhang (2001) and Liang et al. (2010), we construct generalized likelihood ratio (GLR) statistics based on the SBLL estimator $\widehat{m}_{\mathrm{SBLL},l}(u_l, \widehat{\boldsymbol{\beta}})$ given in Section 4.1. First, we construct a GLR statistic and establish its asymptotic distribution by using the local linear estimator $\widetilde{m}_{\mathrm{LL},l}(u_l, \widehat{\boldsymbol{\beta}})$ assuming that all the other nonparametric functions $m_{l'}(\cdot)$ for $l' \neq l$ were known. Because of Theorem 4, the same asymptotic distribution will be satisfied by the GLR statistic by plugging in the SBLL estimates.

Take the case of $l = 1$ as an example. Under H_0, we estimate $m_{\boldsymbol{\theta},1}(u_1)$ by minimizing $\sum_{i=1}^{n}\{Y_{i,1} - m_{\boldsymbol{\theta},1}(U_{i1}(\widehat{\boldsymbol{\beta}}_1), \widehat{\boldsymbol{\beta}})X_{i1}\}^2$, denoted as $\widetilde{m}_{\widehat{\boldsymbol{\theta}},1}(u_1, \widehat{\boldsymbol{\beta}})$, where $\widehat{\boldsymbol{\theta}}$ is the least squares estimator of the parameter vector $\boldsymbol{\theta}$ under the null hypothesis, and the resulting residual sum of squares under the null and alternative hypotheses are given as

$$\mathrm{RSS}_{\mathrm{LL},1}(H_0)=\sum_{i=1}^{n}\{Y_{i,1}-\widetilde{m}_{\widehat{\boldsymbol\theta},1}(U_{i1}(\widehat{\boldsymbol\beta}_1),\widehat{\boldsymbol\beta})X_{i1}\}^2,$$
$$\mathrm{RSS}_{\mathrm{LL},1}(H_1)=\sum_{i=1}^{n}\{Y_{i,1}-\widetilde{m}_{\mathrm{LL},1}(U_{i1}(\widehat{\boldsymbol\beta}_1),\widehat{\boldsymbol\beta})X_{i1}\}^2,$$

where $\widehat{\boldsymbol\beta}$ and $\widetilde{m}_{\mathrm{LL},1}(u_1,\widehat{\boldsymbol\beta})$ are the profile and local linear estimates of $\boldsymbol\beta$ and $m_1(u_1)$, respectively. It follows that a GLR statistic is defined by

$$\mathcal{T}_{\mathrm{LL},1}=\frac{n\{\mathrm{RSS}_{\mathrm{LL},1}(H_0)-\mathrm{RSS}_{\mathrm{LL},1}(H_1)\}}{2\,\mathrm{RSS}_{\mathrm{LL},1}(H_1)}.$$

Let

$$\Gamma_1(u_1)=E\left(X_1^2|U_1=u_1\right)f_1(u_1),$$
$$\Gamma_1^*(u_1)=E\left\{X_1^2\sigma^2(\mathbf{Z},\mathbf{X})|U_1=u_1\right\}f_1(u_1).$$

Corollary 2. Assume that Conditions (C1)–(C7) in the Appendix hold, $h_1\asymp n^{-1/5}$, and $n^{1/(2r+2)}\ll N\ll n^{1/4}$.

(i) Consider $H_0: m_{\boldsymbol\theta,1}(\cdot)$ is linear such that $m_{\boldsymbol\theta,1}(u_1)=\theta_{11}+\theta_{12}u_1$. Then, under H_0, $\tau_K\mathcal{T}_{\mathrm{LL},1}$ has an asymptotic χ^2 distribution with df_n degrees of freedom, where

$$\tau_K=\left\{K(0)-0.5\int K^2(u)\,du\right\}\Big/\int\left\{K(u)-0.5\int K*K(u)\,du\right\}^2du,$$
$$df_n=\tau_K\left\{K(0)-0.5\int K^2(u)\,du\right\}/h,$$

and $K*K(u)$ denotes the convolution of K; (ii) Consider $H_0: m_{\boldsymbol\theta,1}(\cdot)$ is a constant such that $m_{\boldsymbol\theta,1}(u_1)=\theta_{10}$. Then under H_0, $\widetilde{\tau}_K\mathcal{T}_{\mathrm{LL},1}$ has an asymptotic χ^2 distribution with $\widetilde{df}_n$ degrees of freedom, where

$$\widetilde{\tau}_K=\tau_K E\{\sigma^2(\mathbf{Z},\mathbf{X})\}\left\{\int\left(\Gamma_1^*(u_1)\Gamma_1^{-1}(u_1)\right)du_1\right\}\times\left\{\int\left(\Gamma_1^*(u_1)\Gamma_1^{-1}(u_1)\right)^2du_1\right\}^{-1},$$
$$\widetilde{df}_n=\tau_K c_K h^{-1}\left\{\int\left(\Gamma_1^*(u_1)\Gamma_1^{-1}(u_1)\right)du_1\right\}^2\times\left\{\int\left(\Gamma_1^*(u_1)\Gamma_1^{-1}(u_1)\right)^2du_1\right\}^{-1},$$

where $c_K=K(0)-0.5\|K\|_2^2$.

Results (i) and (ii) in Corollary 2 can be proved by following the same reasoning given in the proofs of Theorems 5 and 9 in Fan, Zhang, and Zhang (2001) as well as the proofs of Theorem 5 given by Liang et al. (2010). Now, we construct a sample version of the GLR statistic by using the SBLL estimator $\widehat{m}_{\mathrm{SBLL},1}(u_1,\widehat{\boldsymbol\beta})$. Similarly, denote by $\widehat{m}_{\widehat{\theta},1}(u_1,\widehat{\boldsymbol\beta})$ the least squares estimator that minimizes $\sum_{i=1}^{n}\{\widehat{Y}_{i,1}-m_{\theta,1}(U_{i1}(\widehat{\boldsymbol\beta}_1),\widehat{\boldsymbol\beta})X_{i1}\}^2$. Then, a GLR statistic is defined by

$$\mathcal{T}_{\mathrm{SBLL},1}=\frac{n\left\{\mathrm{RSS}_{\mathrm{SBLL},1}(H_0)-\mathrm{RSS}_{\mathrm{SBLL},1}(H_1)\right\}}{2\mathrm{RSS}_{\mathrm{SBLL},1}(H_1)},\tag{18}$$

where

$$\mathrm{RSS}_{\mathrm{SBLL},1}(H_0)=\sum_{i=1}^{n}\left\{\widehat{Y}_{i,1}-\widehat{m}_{\widehat{\boldsymbol\theta},1}\left(U_{i1}(\widehat{\boldsymbol\beta}_1),\widehat{\boldsymbol\beta}\right)X_{i1}\right\}^2,$$
$$\mathrm{RSS}_{\mathrm{SBLL},1}(H_1)=\sum_{i=1}^{n}\left\{\widehat{Y}_{i,1}-\widehat{m}_{\mathrm{SBLL},1}\left(U_{i1}(\widehat{\boldsymbol\beta}_1),\widehat{\boldsymbol\beta}\right)X_{i1}\right\}^2.$$

By the oracle property given in Theorem 4, under Conditions (C1)–(C7) and the order requirements of h_1 and N given in Corollary 1, it is easy to show that the previous test statistic $\mathcal{T}_{\mathrm{SBLL},1}$ in (18) has the same asymptotic distribution as that of $\mathcal{T}_{\mathrm{LL}}$ established in Corollary 2. The implementation of such GLR test is carried out by the bootstrap method as suggested by Fan and Jiang (2007).

5. IMPLEMENTATION

5.1 Computational Algorithm

The estimator of the parameter vector $\boldsymbol\beta$ is obtained through minimizing the objective function $L_n(\boldsymbol\beta)$ given in (8). We use the "constrOptim.nl" package in R software with constraint that the norm of $(\beta_{l2},\ldots,\beta_{ld})^{\mathrm T}$ is less than 1. To use the "constrOptim.nl" package, we specify the gradient of the objective function as

$$\partial L_n(\boldsymbol\beta)/\partial\boldsymbol\beta_{-1}\approx-\sum_{i=1}^{n}\left\{Y_i-\sum_{l=1}^{d}\sum_{s=1}^{J_n}B_{s,l}\left(U_{il}(\boldsymbol\beta_l)\right)\widehat{\lambda}_{s,l}(\boldsymbol\beta)X_{il}\right\}\times\left[\widehat{\dot m}_l(U_{il}(\boldsymbol\beta_l),\boldsymbol\beta)X_{il}\mathbf{J}_l^{\mathrm T}\widehat{\mathbf{Z}}_i\right]_{l=1}^{d},$$

which is derived in the Appendix, where $\widehat{\dot m}_l(\cdot)$ is a spline estimator of $\dot m_l(\cdot)$ given in (6) and $\widehat{\mathbf{Z}}_i$ is provided in Remark 1 in Section 3. To start the search, we suggest using initial values obtained by assuming linearity of each coefficient function following Carroll et al. (1997) and Xia et al. (2002) in single-index models. The details of generating initial values can be found in Section S.1 of the online supplemental materials.

5.2 Smoothing Parameter Selection

In the PLSE of $\boldsymbol\beta$, the nonparametric functions $m_l(\cdot)$ are approximated by cubic spline ($q=4$), where the number of interior knots is set as $N=[2n^{1/(2q+1)}]+1=[2n^{1/9}]+1$, which satisfies the optimal order of N as discussed in Remark 4. Here, $[a]$ denotes the closest integer to a. After we obtain an estimate of $\boldsymbol\beta$, each $m_l(\cdot)$ is estimated by the B-splines $\widehat{m}_l(\cdot,\widehat{\boldsymbol\beta})$ with the number of interior knots selected by minimizing following the BIC criterion on the range $[n^{1/9}]\le N\le[2n^{1/9}]+1$:

$$\mathrm{BIC}(N)=\log\left[n^{-1}\sum_{i=1}^{n}\{Y_i-\widehat{m}(\mathbf{Z},\mathbf{X})\}^2\right]+\frac{\log n}{n}d(N+q),$$

where $\widehat{m}(\mathbf{Z},\mathbf{X})=\sum_{l=1}^{d}\widehat{m}_l(\mathbf{Z}^{\mathrm T}\widehat{\boldsymbol\beta}_l,\widehat{\boldsymbol\beta})\mathbf{X}_l$. Then, the optimal number of interior knots is given by $\widehat{N}=\mathrm{argmin}_{N\in I_N}\mathrm{BIC}(N)$. In the second step, the SBLL estimation for $m_1(\cdot)$ is performed with the optimal bandwidth $h_{1,\mathrm{opt}}$, which minimizes the total asymptotic mean integrated squared errors (AMISE):

$$\mathrm{AMISE}\left(\widehat{m}_{\mathrm{SBLL},1}\right)=\int\left[\left\{b_1(u_1)h_1^2\right\}^2+v_1(u_1)/(nh_1)\right]f_1(u_1)\,du_1.$$

Section S.2 of the online supplementary materials presents the detailed procedure of obtaining an estimate of the optimal bandwidth $h_{1,\mathrm{opt}}$.

6. SIMULATION EXPERIMENTS

In this section, we conduct several simulation studies to evaluate the performance of the proposed methodology. We consider the following VICM:

$$Y_i = m\left(\mathbf{Z}_i, \mathbf{X}_i, \boldsymbol{\beta}\right) + \varepsilon_i = m_1\left(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_1\right) X_{i1} + m_2\left(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right) X_{i2} + m_3\left(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_3\right) X_{i3} + \varepsilon_i, \tag{19}$$

with $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})^{\mathrm{T}}$, where X_{i1} is generated from Bernoulli (0.5)−0.5, and $(X_{i2}, X_{i3})^{\mathrm{T}}$ is drawn from a bivariate normal distribution with mean 0, variance 1, and covariance 0.2. To generate $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3})^{\mathrm{T}}$, we first sample $(Z_{i1}^*, Z_{i2}^*, Z_{i3}^*)^{\mathrm{T}}$ from a multivariate normal with mean 0, variance 1, and covariance 0.2, and then let $Z_{ik} = \Phi(Z_{ik}^*) - 0.5$, $k = 1, 2, 3$, where $\Phi(\cdot)$ is the CDF of the standard normal. The true loading parameters are set as $\boldsymbol{\beta}_1 = \frac{1}{\sqrt{14}}(2, 1, 3)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = \frac{1}{\sqrt{14}}(3, 2, 1)^{\mathrm{T}}$, and $\boldsymbol{\beta}_3 = \frac{1}{\sqrt{14}}(2, 3, 1)^{\mathrm{T}}$. Set $m_l(u_l) = m_l^*(u_l) - E\{m_l^*(u_l)\}$, $l = 1, 2, 3$, where $m_1^*(u_1) = 10\exp(5u_1)/\{1 + \exp(5u_1)\}$, $m_2^*(u_2) = 5\sin(\pi u_2)$, and $m_3^*(u_3) = 3\{\sin(\pi u_3) + \cos(2\pi u_3 - 4\pi/3)\}$, and their shapes may be seen in Figure 2.

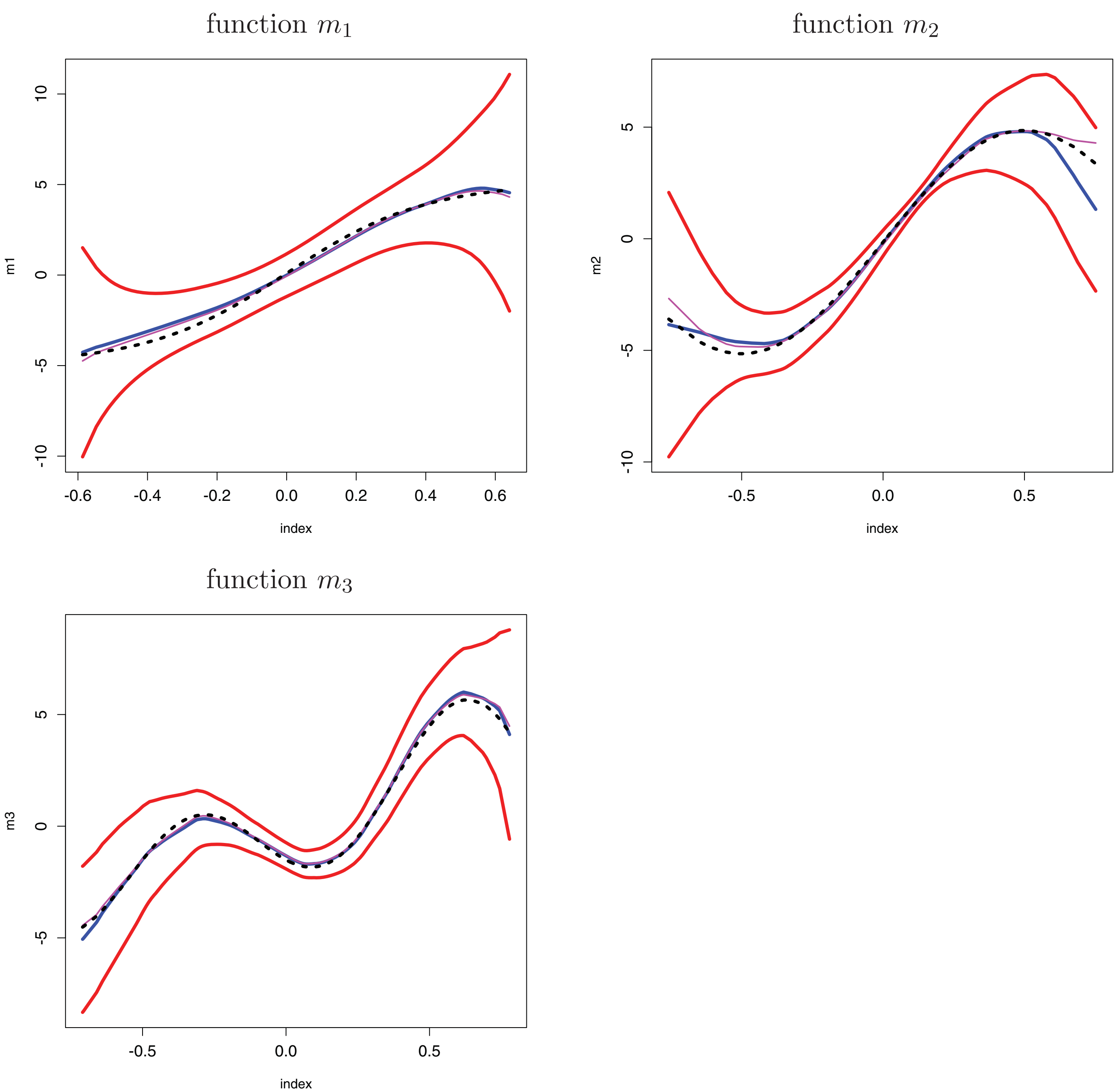

Figure 2. Plots of the two-step SBLL estimator $\widehat{m}_{\mathrm{SBLL},l}(\cdot)$ (thick line), the upper and lower 95% pointwise confidence intervals (upper and lower thick lines), the oracle estimator $\widetilde{m}_{\mathrm{LL},l}(\cdot)$ (thin line) and the true function $m_l(\cdot)$ (dashed line) for $l = 1, 2, 3$ based on one sample with $n = 200$.

Table 1. The empirical coverage rates of the 95% confidence intervals for individual loading parameters $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})^{\mathrm{T}}$, and $\boldsymbol{\beta}_3 = (\beta_{31}, \beta_{32}, \beta_{33})^{\mathrm{T}}$ for sample size $n = 200, 500, 1000$

n	β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}	β_{31}	β_{32}	β_{33}
200	0.920	0.906	0.926	0.928	0.928	0.922	0.926	0.916	0.904
500	0.958	0.940	0.952	0.948	0.942	0.946	0.946	0.944	0.934
1000	0.946	0.944	0.952	0.950	0.954	0.950	0.948	0.944	0.944

Finally, $Y_i, 1 \le i \le n$, are generated from the VICM (19), where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \boldsymbol{\beta}_3^{\mathrm{T}})^{\mathrm{T}}$, and errors ε_i follow $N(0, \sigma^2(\mathbf{Z}_i, \mathbf{X}_i))$ with $\sigma^2(\mathbf{Z}_i, \mathbf{X}_i) = \{100 - m(\mathbf{Z}_i, \mathbf{X}_i, \boldsymbol{\beta})\}/\{100 + m(\mathbf{Z}_i, \mathbf{X}_i, \boldsymbol{\beta})\}$.

The sample size takes $n = 200, 500, 1000$, respectively, and 500 simulation replications are run to draw summary statistics. Table 1 shows the empirical coverage rates of the 95% confidence intervals for individual loading parameters β_{lk}, $l, k = 1, 2, 3$, where standard errors are calculated according to the asymptotic formula given in (16). It is clear that all coverage rates approach to the 95% nominal level as the sample size increases. This result is confirmatory to the asymptotic normals of the loading parameter estimators established in Theorem 1.

Table 2 presents the biases of the PLSE for individual loading parameters β_{lk}, $l, k = 1, 2, 3$ over 500 replications. It is easy to see that all biases are close to 0 in the cases considered. This result confirms the consistency of the PLSE given in Theorem 1. It is interesting to note that estimation consistency is achieved with a relatively small sample size of $n = 200$. Table 3 shows the average asymptotic standard error (ASE) calculated according to Theorem 1 and the empirical standard error (ESE) among 500 replications. Apparently, both ASE and ESE become smaller as n increases, due to the fact of the PLSE being root-n consistent. More importantly, it is evident that the ASE and the corresponding ESE are very comparable in all cases, which presents an assurance for the use of the asymptotic covariance matrix in practice.

Now, we turn to the nonparametric part. To evaluate the performance of the two-step SBLL estimator $\widehat{m}_{\mathrm{SBLL},l}(\cdot)$ for a given l, we consider the mean integrated squared error (MISE) as the average of the following measure:

$$\mathrm{ISE}(\widehat{m}_{\mathrm{SBLL},l}) = n^{-1} \sum_{i=1}^{n} \{\widehat{m}_{\mathrm{SBLL},l}(U_{il}(\widehat{\boldsymbol{\beta}}_l), \widehat{\boldsymbol{\beta}}) - m_l(U_{il})\}^2$$

over the 500 replications. The MISE for the oracle estimator $\widetilde{m}_{\mathrm{LL},l}(\cdot)$ takes the same form. Table 4 shows the MISE for the two-step SBLL estimator $\widehat{m}_{\mathrm{SBLL},l}(\cdot)$ and the oracle estimator $\widetilde{m}_{\mathrm{LL},l}(\cdot)$ for $m_l(\cdot)$, $l = 1, 2, 3$. We can observe that the MISE values get closer to those of the oracle estimators as n increases, which demonstrates that the SBLL estimator is a reliable and desirable estimator. Moreover, the MISEs of both $\widehat{m}_{\mathrm{SBLL},l}(\cdot)$ and $\widetilde{m}_{\mathrm{LL},l}(\cdot)$ decrease as n increases.

To visualize the estimated functions, in Figure 2, we display the SBLL estimator $\widehat{m}_{\mathrm{SBLL},l}(\cdot)$ (thick line), with the upper and lower 95% pointwise confidence bands (two thick lines), and the oracle estimator $\widetilde{m}_{\mathrm{LL},l}(\cdot)$ (thin line) and the true function $m_l(\cdot)$ (dashed line) for $n = 200$. It is evident that the proposed SBLL estimators perform well.

Now we report the finite-sample performance of the Wald test statistic χ^2_W proposed in Section 4.2. We stick to the same model (19), except for now setting the true parameters as $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \frac{1}{\sqrt{6}}(1, 1, 1, 1, 1, 1)^{\mathrm{T}}$ and $\boldsymbol{\beta}_3 = \frac{1}{\sqrt{3+3c^2}}(1, 1, 1, c, c, c)^{\mathrm{T}}$, where c ranges from 0 to 0.2 with an increment of 0.02 to evaluate the power of the test. Under the null hypothesis $H_0 : \beta_{34} = \beta_{35} = \beta_{36} = 0$, the statistic χ^2_W approximately follows the chi-square distribution with three degrees of freedom (DF) as given in Section 4.2 . The left panel of Figure 3 displays the power function of the test χ^2_W at significance level 0.05 versus the c values for $n = 200$ and 500 based on 500 simulation replications. At $c = 0$, that is, the null hypothesis H_0 is true, the empirical sizes are 0.060 for $n = 200$ and 0.054 for $n = 500$, respectively, which are close to the nominal Type I error 0.05. It is easy to visualize that the empirical size rises up to 1 as the value of c increases, and the rate of rising-up becomes faster in the case with a larger sample size. These results demonstrate that the proposed GLR test performs well and serves as a reasonable approach to identify significant components in $\mathbf{Z}$ interacting with X_l.

In addition, we examine the performance of the GLR test statistic $\mathcal{T}_{\mathrm{SBLL},l}$ given in (18) to identify the functional form of interactions. To proceed, in model (19), we set $m_3(u_3) = m_3^*(u_3) - E\{m_3^*(u_3)\}$ with

$$m_3^*(u_3) = 6u_3 + \lambda\{\sin(\pi u_3) + \cos(2\pi u_3 - 4\pi/3)\},$$

where λ ranges from 0 to 1 with an increment of 0.2, and the other two m function specifications remain the same. The null hypothesis is $H_0 : m_3(u_3) = \theta_0 + \theta_1 u_3$, where θ_0 and θ_1 are two unknown constants. The null distribution of the GLR statistic (18) is obtained by the bootstrap procedure as suggested by Fan and Jiang (2007). The right panel of Figure 3 shows the power function of the test statistic $\mathcal{T}_{\mathrm{SBLL},3}$ at significance level 0.05 versus the λ values for $n = 200$ and 500 over 500 simulation replications. Once again, the empirical size is close

Table 2. The bias ($\times 10^{-2}$) of the estimators for $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})^{\mathrm{T}}$, and $\boldsymbol{\beta}_3 = (\beta_{31}, \beta_{32}, \beta_{33})^{\mathrm{T}}$ for $n = 200, 500, 1000$

n	β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}	β_{31}	β_{32}	β_{33}
200	−0.4221	−0.1225	−0.1880	−0.0205	−0.1714	0.1360	−0.0881	−0.4686	0.3328
500	−0.1615	−0.0415	−0.0586	−0.0251	0.0559	0.1043	−0.0922	0.0371	−0.0328
1000	0.0126	−0.0953	−0.0059	−0.0342	0.0126	0.0405	−0.0291	0.0152	−0.0395

Table 3. The average asymptotic standard error (ASE) ($\times 10^{-2}$) and empirical standard error (ESE) ($\times 10^{-2}$) of the estimators for $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})^{\mathrm{T}}$, and $\boldsymbol{\beta}_3 = (\beta_{31}, \beta_{32}, \beta_{33})^{\mathrm{T}}$ for $n = 200, 500, 1000$

n		β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}	β_{31}	β_{32}	β_{33}
200	ASE	4.4918	5.6051	3.0709	1.3448	2.0765	2.3631	2.5399	1.9011	2.4355
	ESE	5.1865	6.6102	3.3413	1.4157	2.2314	2.7094	3.5299	2.0633	3.0228
500	ASE	2.7761	3.5360	1.8952	0.8011	1.2421	1.4157	1.5095	1.1144	1.4494
	ESE	2.7910	3.6135	1.8993	0.8295	1.2408	1.4804	1.4906	1.1302	1.4703
1000	ASE	1.9397	2.5041	1.3259	0.5534	0.8544	0.9850	1.0439	0.7677	0.9877
	ESE	1.9432	2.5582	1.3224	0.5412	0.8540	0.9810	1.0363	0.7788	1.0220

to the nominal level 0.05, and the power escalates to 1 as the λ value deviates further from zero.

In summary, our proposed PLSE and SBLL estimators and the Wald and GLR tests perform satisfactorily in the simulation settings considered. It is worth pointing out that our proposed estimation procedure is computationally fast. The aforementioned simulation experiments are run in R software on an ordinary Macbook Pro with 2 GHz Intel Core, and the average operation time per dataset is 1.375, 2.429, and 4.068 s for sample size $n = 200, 500, 1000$, respectively, including the total running time of generating a dataset and computing both the PLSE of loading parameters $\boldsymbol{\beta}_l$, $l = 1, 2, 3$, and the SBLL estimation of nonparametric functions $m_l(\cdot)$, $l = 1, 2, 3$.

7. APPLICATION

This section presents the analysis of child growth data introduced in Section 1. Through data validation, we end up 214 children for the data analysis. The response variable Y is log-weight at current age, and $\mathbf{Z} = (Z_1, \ldots, Z_8)^{\mathrm{T}}$ consists of eight log-transformed measures of EDC agents from mother's blood samples during pregnancy. Covariates of interest include intercept ($X_1 = 1$), gender ($X_2 = 0$ for boy, 1 for girl), age (X_3, yrs), and child's weight at age 4 (X_4). To answer the three questions given in Section 1, we propose the following form:

$$Y = \sum_{l=1}^{4} m_l\left(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_l\right) X_l + \varepsilon, \tag{20}$$

where $m_l(\cdot)$ are unknown smooth functions and $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{l8})^{\mathrm{T}}$ are unknown loading parameters for $l = 1, \ldots, 4$. We normalize all variables in the analysis. Let $\widetilde{\boldsymbol{\beta}}_l = (\widetilde{\beta}_{l1}, \ldots, \widetilde{\beta}_{l8})^{\mathrm{T}}$ be the PLSE under the normalized values of $\mathbf{Z}$, and the resulting estimator of β_{lk} in the original scale is given by $\widehat{\beta}_{lk} = \widetilde{\beta}_{lk} \times \mathrm{SD}(Z_k)$, where $\mathrm{SD}(Z_k)$ is the sample standard deviation of variable Z_k, for $k = 1, \ldots, 8$, $l = 1, \ldots, 4$.

The initial values of the parameters are generated by the steps described in Section 5.1. In our analysis, the number of interior knots and the bandwidth are chosen based on the criteria discussed in Section 5.2. Fitting model (20) by the proposed methodology, we obtain the estimates (EST) of $\boldsymbol{\beta}_l$, $1 \le l \le 4$, their lower bound (LB) and upper bound (UB) of 95% confidence intervals (CI) with the standard errors calculated according to (15), as well as the p-values for testing significance of each EDC component. Table 5 lists the results.

Table 4. The MISE values for the two-step SBLL estimator $\widehat{m}_{\mathrm{SBLL},l}$ and the oracle estimators $\widetilde{m}_{\mathrm{LL},l}(\cdot)$ for $l = 1, 2, 3$

n	$\widehat{m}_{\mathrm{SBLL},1}$	$\widetilde{m}_{\mathrm{LL},1}$	$\widehat{m}_{\mathrm{SBLL},2}$	$\widetilde{m}_{\mathrm{LL},2}$	$\widehat{m}_{\mathrm{SBLL},3}$	$\widetilde{m}_{\mathrm{LL},3}$
200	0.1627	0.1138	0.0980	0.0755	0.2031	0.1426
500	0.0698	0.0605	0.0426	0.0373	0.0471	0.0436
1000	0.0407	0.0367	0.0210	0.0185	0.0237	0.0232

Table 5. The estimates (EST), lower bound (LB), and upper bound (UB) of 95% confidence intervals of β_l, and the p-values for testing significance of each component in β_l in model 20

		EST	LB	UB	p-value
		X_1 = intercept			
$\boldsymbol{\beta}_1$	Z_1	0.415	0.197	0.632	< 0.001
	Z_2	0.008	−0.108	0.123	0.895
	Z_3	−0.081	−0.321	0.159	0.507
	Z_4	0.379	0.229	0.529	< 0.001
	Z_5	0.432	0.271	0.593	< 0.001
	Z_6	< 0.000	−0.140	0.140	0.998
	Z_7	−0.730	−0.837	−0.624	−0.001
	Z_8	0.0200	−0.196	0.236	0.856
		X_2 = gender			
$\boldsymbol{\beta}_2$	Z_1	0.132	0.096	0.168	< 0.001
	Z_2	−0.038	−0.046	−0.031	< 0.001
	Z_3	−0.540	−0.566	−0.515	< 0.001
	Z_4	0.157	0.129	0.186	< 0.001
	Z_5	0.179	0.141	0.217	< 0.001
	Z_6	−0.398	−0.412	−0.385	< 0.001
	Z_7	0.759	0.721	0.796	< 0.001
	Z_8	0.198	0.181	0.216	< 0.001
		X_3 = age (yrs)			
$\boldsymbol{\beta}_3$	Z_1	0.147	−0.057	0.351	0.159
	Z_2	0.154	0.082	0.227	< 0.001
	Z_3	0.225	0.023	0.427	0.029
	Z_4	−0.080	−0.240	0.079	0.325
	Z_5	−0.307	−0.459	−0.156	< 0.001
	Z_6	−0.615	−0.715	−0.515	< 0.001
	Z_7	0.480	0.328	0.633	< 0.001
	Z_8	−0.575	−0.764	−0.385	< 0.001
		X_4 =weight at age 4			
$\boldsymbol{\beta}_4$	Z_1	0.042	−0.206	0.290	0.740
	Z_2	−0.121	−0.236	−0.005	0.040
	Z_3	0.054	−0.479	0.587	0.842
	Z_4	0.0532	−0.157	0.264	0.621
	Z_5	−0.312	−0.661	0.037	0.080
	Z_6	−0.317	−0.540	−0.095	0.005
	Z_7	0.846	0.681	1.012	< 0.001
	Z_8	−0.444	−0.680	−0.209	< 0.001

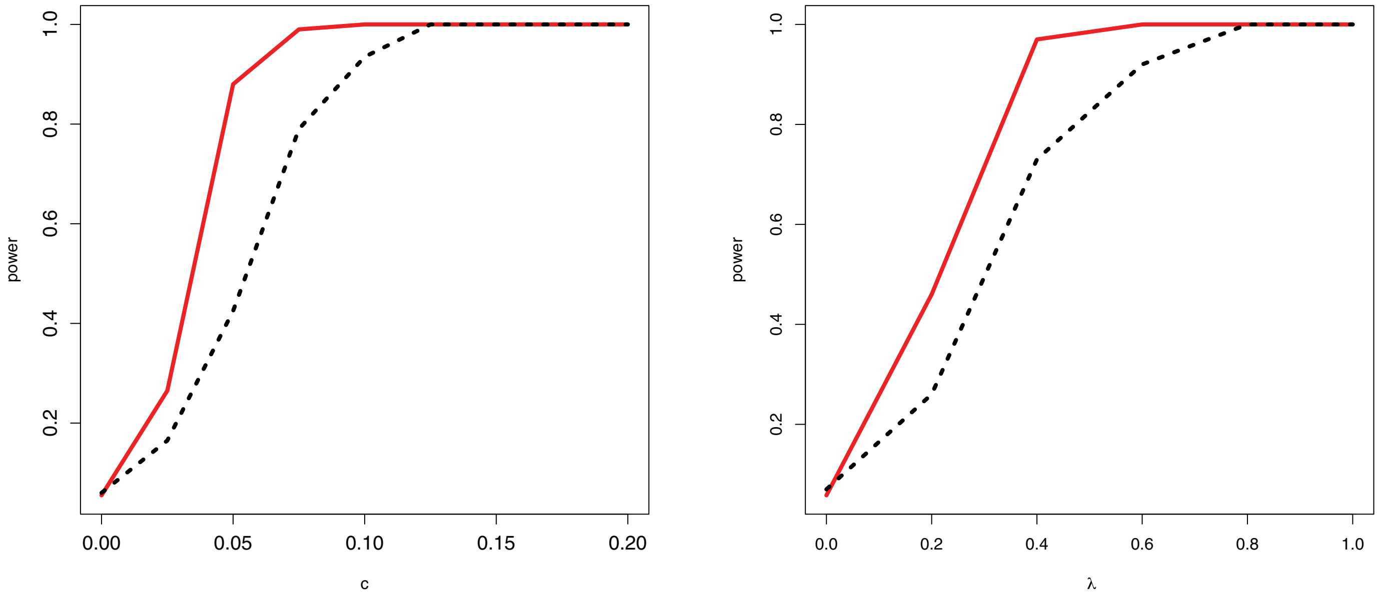

Figure 3. Plots of the Wald test statistic χ^2_W (left panel) and the GLR test statistic $\mathcal{T}_{\mathrm{SBLL},3}$ (right panel) at significance level 0.05 for $n = 200$ (dashed line) and $n = 500$ (solid line).

Statistical significance level $\alpha = 0.05$ is used in the following discussion. For X_1 =intercept, four loading parameters of Z_1, Z_4, Z_5, and Z_7 are significantly different from zero, suggesting that these four EDCs have significant main effects. For X_2 = gender, all of the eight EDCs are responsible for the alteration in the effect of gender. For X_3 =age, its effect is modified by a mixture of six EDCs, including Z_2, Z_3, Z_5, Z_6, Z_7, and Z_8. For X_4 = weight at age 4, a mixture of four EDCs, Z_2, Z_6, Z_7, and Z_8, alters the association between weight at current age and weight at age 4. In Table 5, the estimated loading coefficients appear to be very different in different indices implying that effects (or partial associations) of covariates X_l are modified by different configurations of EDC agents. This finding provides the answer to the question of which EDC agents are important, which however cannot be provided by the SICM under the assumption of a common vector $\boldsymbol{\beta}$ in all indices.

To achieve model simplicity, we further conduct the Wald chi-square test described in Section 4.2 to identify significant subsets of $\boldsymbol{\beta}_l$, $l = 1, 2, 3, 4$. For $\boldsymbol{\beta}_1$, we consider $H_0 : \beta_{12} = \beta_{13} = \beta_{16} = \beta_{18} = 0$, and obtain the p-value 0.977, implying that the set of four components (Z_2, Z_3, Z_6, Z_8) has no main effects. For $\boldsymbol{\beta}_3$, we consider $H_0 : \beta_{11} = \beta_{14} = 0$, and obtain the p-value 0.089, so EDC agents Z_1 and Z_4 are not contributing to the modification on the effect of age. For $\boldsymbol{\beta}_4$, we consider $H_0 : \beta_{11} = \beta_{13} = \beta_{14} = \beta_{15} = 0$, and obtain the p-value 0.042; then, we consider $H_0 : \beta_{11} = \beta_{13} = \beta_{14} = 0$, and obtain the p-value 0.957. This means that the set of (Z_1, Z_3, Z_4) has no significant impact on the altered effect of weight at age 4.

Summarizing the previous testing results, we reach a simplified model of the form:

$$Y = \sum_{l=1}^{4} m_l \left(\mathbf{Z}_l^{\mathrm{T}} \boldsymbol{\beta}_l\right) X_l + \varepsilon, \tag{21}$$

where $\mathbf{Z}_1 = (Z_1, Z_4, Z_5, Z_7)^{\mathrm{T}}$, $\mathbf{Z}_2 = (Z_1, \ldots, Z_8)^{\mathrm{T}}$, $\mathbf{Z}_3 = (Z_2, Z_3, Z_5, Z_6, Z_7, Z_8)^{\mathrm{T}}$, and $\mathbf{Z}_4 = (Z_2, Z_5, Z_6, Z_7, Z_8)^{\mathrm{T}}$.

Now we are ready to compare the full model (20) (FULL), the reduced model (21) (REDUCED) , and the partially linear single-index model (PLSIM)

$$Y = m_1(\mathbf{Z}^{\mathrm{T}} \boldsymbol{\beta}) + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \varepsilon, \tag{22}$$

the single-index coefficient model (SICM)

$$Y = \sum_{l=1}^{4} m_l(\mathbf{Z}^{\mathrm{T}} \boldsymbol{\beta}) X_l + \varepsilon, \tag{23}$$

and the varying coefficient model (VCM)

$$Y = \sum_{l=1}^{4} m_l(U^{\mathrm{PCA}}) X_l + \varepsilon, \tag{24}$$

where U^{PCA} is the first principle component obtained by a principle component analysis (PCA) on $\mathbf{Z}$. As pointed out in Section 2, the PLSIM, SICM, and VCM are special cases of the VICM, so that we can use the proposed PLSE to estimate the parameters in model (21), the PLSIM (22), and the SICM (21) with minor modifications. We perform the leave-one-out cross-validations for models (20)–(24), as well as two linear models by assuming constant and linear functions for $m_l(\cdot)$, respectively, with the estimated prediction error given as CVE$= n^{-1} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i^{-(i)})^2$, where $\widehat{Y}_i^{(-i)}$ is the predicted value for the ith response using the remaining $(n-1)$ observations.

Table 6 lists the estimated cross-validation prediction errors (CVE) and the relative prediction error (RCVE) to the smallest obtained by the reduced model (21). The next to the reduced model is the full VICM. It is interesting to observe that the SICM has 23.03% higher prediction error than the PLSIM, which further demonstrates that imposing common loading parameters fails to capture the nonlinear interactions directed by the EDCs. The fact that the VCM has a larger CV error than the SICM and PLSIM shows that the PCA method of allocating the loading weights for dimension reduction works poorly. Moreover, noting that the linear model with linear interactions has a larger CV

Table 6. Cross-validation errors (CVE) and the relative CVE (RCVE) to the smallest CVE for the full VICM (20) (FULL), reduced VICM (21) (REDUCED), PLSIM (22), SICM (23), VCM (24), and two linear models with constant (CONSTANT) and linear (LINEAR) functions for $m_l(\cdot)$

	REDUCED	FULL	PLSIM	SICM	VCM	CONSTANT	LINEAR
CVE	0.0152	0.0168	0.0207	0.0242	0.0284	0.0230	0.0297
RCVE	100%	110.53%	136.18%	159.21%	186.84%	151.32%	195.39%

error than the linear model without interactions, we conclude that in this data analysis, the classical linear interactions cannot properly capture the interplay between **Z** and **X**. Comparing the CV, error of the reduced VICM (REDUCED) with the error of the existing models, PLSIM, SICM, and VCM, we see that the proposed VICM improves the model prediction by 36.18%, 59.21%, and 86.84%, respectively.

To further examine if the outperformance of the VICM in the prediction observed in Table 6 is beyond the sampling errors, we take the following procedure based on difference of the CVE values. For illustration, let us focus on the comparison between the full VICM and the SICM. Denote the difference by $\mathrm{CVE}_{\mathrm{VICM}} - \mathrm{CVE}_{\mathrm{SICM}} = n^{-1}\sum_{i=1}^{n} D_i$, with $D_i = (Y_i - \widehat{Y}_i^{\mathrm{VICM},(-i)})^2 - (Y_i - \widehat{Y}_i^{\mathrm{SICM},(-i)})^2$, where $\mathrm{CVE}_{\mathrm{VICM}}$ and $\mathrm{CVE}_{\mathrm{SICM}}$ are the resulting CVE values, and $\widehat{Y}_i^{\mathrm{VICM},(-i)}$ and $\widehat{Y}_i^{\mathrm{SICM},(-i)}$ are the predicted values for the ith response obtained from the VICM and SICM, respectively. Since D_i and $D_{i'}$ are correlated, we take a deassociation transformation by letting $\widetilde{\mathbf{D}} = \left(\widetilde{D}_1, \ldots, \widetilde{D}_n\right)^T = \Sigma^{-1/2}\mathbf{D}$, where Σ is the covariance matrix of $\mathbf{D}$, which is estimated by the bootstrap resampling method with 500 replications. Obviously, $\mathrm{var}(\widetilde{\mathbf{D}}) = \mathbf{I}_{n\times n}$. Our calculations give $n^{-1}\sum_{i=1}^{n}\widetilde{D}_i = -0.102$ with the standard error 0.033. This leads to the p-value = 0.002 by the Z-test, and the p-value = 0.007 by the Wilcoxon rank test. Since both p-values are smaller than 0.05, the cross-validation prediction errors between VICM and SICM are significantly different.

To examine if there exists, and if so in which form, interactions between the EDCs and **X**, we conduct the GLR test proposed in Section 4.3. We obtain the p-values of the GLR test statistics all less than 0.05 in the following hypothesis tests. First, $H_0 : m_l(\cdot)$ is constant (or the absence of interaction) versus $H_a : m_l(\cdot)$ is not constant. Second, $H_0 : m_l(\cdot)$ is linear (or the existence of linear interactions) versus $H_a : m_l(\cdot)$ is nonlinear. These results suggest that there exist strong nonlinear main effects of exposure to a mixture of EDCs, and more importantly that exposure to mixture of these EDCs alters the effects of gender, age, and weight at age 4. Such findings are clearly supported by the graphic evidence in Figure 4.

Figure 4 displays the estimated curves obtained by the two-step SBLL method (middle solid line), the one-step spline estimate given in (14) (middle dashed line), and their 95% pointwise confidence intervals (lower and upper lines) of $m_l(\cdot)$, $1 \le l \le 4$. In addition, the estimates $\widehat{m}_{\theta,l} = \widehat{\theta}_{l0}$ (horizontal dashed lines) under the CONSTANT model and $\widehat{m}_{\theta,l} = \widehat{a}_l + \widehat{b}_l U_l\left(\widehat{\boldsymbol{\beta}}_l\right)$ (straight thin lines) under the LINEAR model are included for comparison.

The first plot for the intercept shows that the estimated function $\widehat{m}_1(\cdot)$ is a decreasing function of index $\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_1$, which indicates that exposure to the combination of EDCs has a negative effect on child's weight growth. The plot for covariate gender shows that the modification on the association of weight at current age with gender altered by the mixture of the EDCs is nonlinear. The plot for covariate age shows a decreasing trend, suggesting the velocity of weight growth becomes weaker as the exposure to the mixture of EDCs increases. The plot for covariate weight at age 4 again demonstrates that the effect of weight at age 4 is nonlinearly modified. These findings are of scientific importance and corroborative with the GLR test results. The two parametric models (CONSTANT and LINEAR) unfortunately missed the opportunity to capture those nonlinear features. Moreover, we can observe that the one-step spline method and the two-step SBLL method yield similar estimated curves for the nonparametric functions. Our collaborators are amazed by the novelty and power of these estimated nonlinear modifications to child's growth profiles and seeking for further scientific data to confirm these findings.

8. DISCUSSION

In this article, we propose a new class of semiparametric models with varying index coefficients, which allows us to study nonlinear interactive effects that are of scientific importance in the understanding of the response–covariate relationship. We demonstrate that regression coefficient of a covariate can be altered or directed by a nonlinear function of multiple other covariates. The proposed modeling framework gives rise to a rich class of regression models, including many popular semiparametric models as special cases. Using the least squares estimation approach, we develop a profile estimation procedure that is both conceptually simple and computationally efficient, and the resulting estimators are consistent and asymptotically normal.

We are currently involved in multiple collaborative projects studying effects of mother's and/or child's exposures to environmental pollutants (e.g., pesticides, BPA, and phthalates as well as heavy metals) on neurodevelopment of children in China and the somatic growth of children in the USA. As pointed by our science collaborators, being able to understand the discrepant interactive roles played by a Z variable with different X variables is a great scientific innovation, which has never been possibly done in the currently available statistical toolboxes. Based on our experience on the child growth data analysis, we are strongly encouraged by the flexibility of our VICM model, which provides a comprehensive way to understand interactions between environmental exposures and physiological variables in the study of human growth and diseases.

Our future work will be focused on the extension of the proposed VICM model for longitudinal data as well as on discrete or categorical response variables along the line of quasi-likelihood estimation inference. Since the proposed model may involve

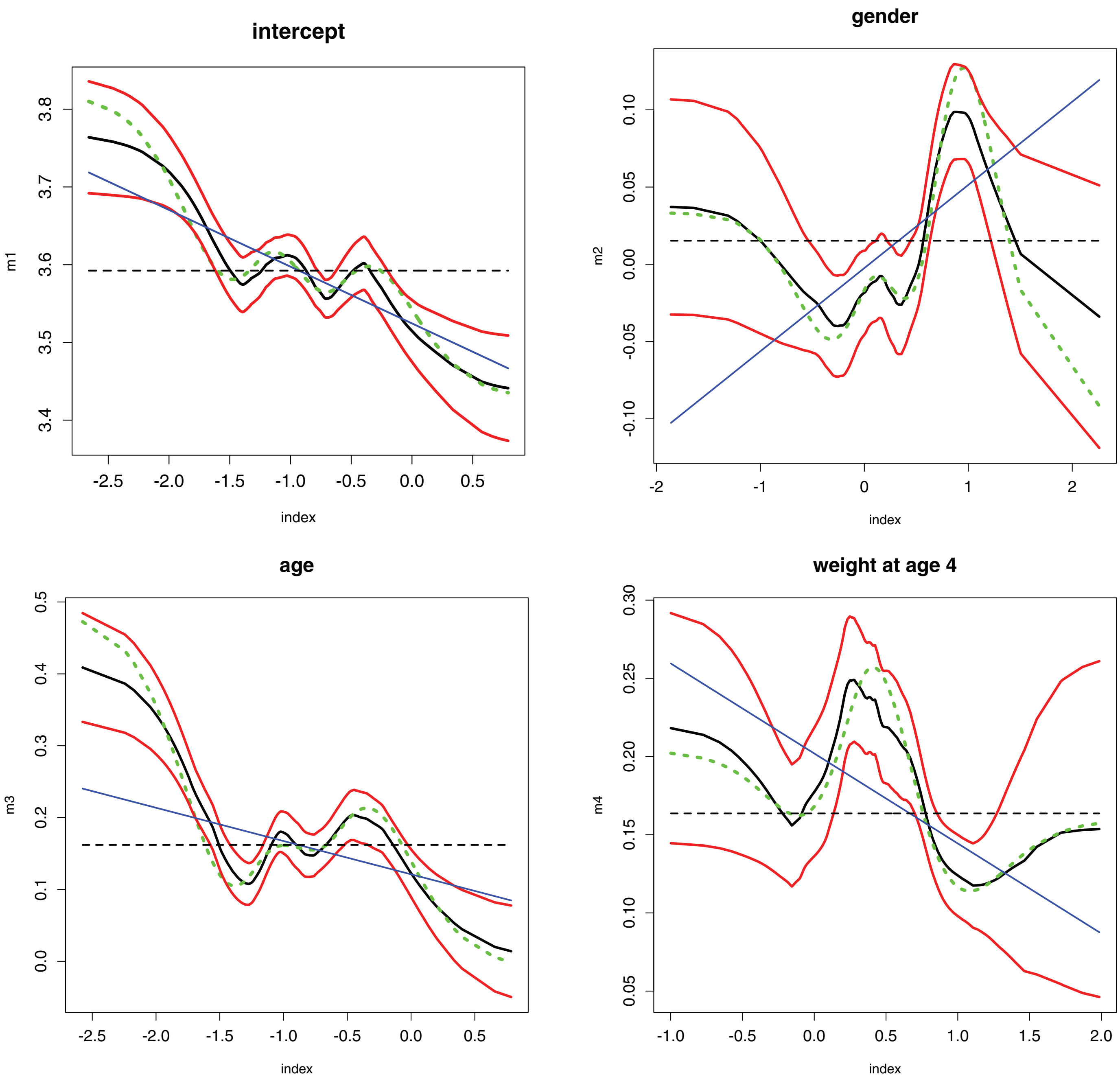

Figure 4. Plots of the SBLL estimator (middle solid line), the one-step spline estimator (middle dashed line), and the 95% pointwise confidence intervals (lower and upper lines) of $m_l(\cdot)$, $1 \le l \le 4$, as well as the estimates $\widehat{m}_{\boldsymbol{\theta},l} = \widehat{\theta}_{l0}$ (horizontal dashed lines) and $\widehat{m}_{\boldsymbol{\theta},l} = \widehat{\theta}_{l1} + \widehat{\theta}_{l2} U_l(\widehat{\boldsymbol{\beta}}_l)$ (straight thin lines).

a large number of parameters (e.g., loading coefficients) given that each coefficient function depends on different loading parameters, variable selection procedures via regularization will be investigated as future work to achieve model sparsity. Also a user-friendly R package for the implementation of the VICM in this article will be made available to the public.

APPENDIX

A.1 Assumptions

For positive numbers a_n and b_n, let $a_n \asymp b_n$ denote that $\lim_{n\to\infty} a_n/b_n = c$, where c is some nonzero constant. For any vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_s)^{\mathrm{T}} \in R^s$, denote $\|\boldsymbol{\zeta}\|_\infty = \max_{1\le l\le s} |\zeta_l|$. For any symmetric matrix $\mathbf{A}_{s\times s}$, denote its L_r norm as $\|\mathbf{A}\|_r = \max_{\boldsymbol{\zeta}\in\mathbf{s},\boldsymbol{\zeta}\neq\mathbf{0}} \|\mathbf{A}\boldsymbol{\zeta}\|_r \|\boldsymbol{\zeta}\|_r^{-1}$. For any matrix $\mathbf{A} = (A_{ij})_{i=1,j=1}^{s,t}$, denote $\|\mathbf{A}\|_\infty = \max_{1\le i\le s} \sum_{j=1}^{t} |A_{ij}|$.

We denote the space of rth order smooth function as $C^{(r)}[0,1] = \{\varphi | \varphi^{(r)} \in C[0,1]\}$. Let $C^{0,1}(\mathcal{X}_w)$ be the space of Lipschitz continuous function on $\mathcal{X}_w$, that is,

$$C^{0,1}(\mathcal{X}_w) = \left\{ \varphi : \|\varphi\|_{0,1} = \sup_{w\neq w', w, w' \in \mathcal{X}_w} \frac{|\varphi(w) - \varphi(w')|}{|w - w'|} < \infty \right\},$$

in which $\|\varphi\|_{0,1}$ is the $C^{0,1}$-norm of φ. To establish the consistency and asymptotic normality for the proposed estimators, we need the following regularity conditions:

(C1) For every $1 \le l \le d$, the density function $f_{U_l(\boldsymbol{\beta}_l)}(\cdot)$ of random variable $U_l(\boldsymbol{\beta}_l) = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_l$ is bounded away from 0 on $\mathcal{S}_w$ and $f_{U_l(\boldsymbol{\beta}_l)}(\cdot) \in C^{0,1}(\mathcal{S}_w)$ for $\boldsymbol{\beta}_l$ in the neighborhood of $\boldsymbol{\beta}_l^0$, where $\mathcal{S}_w = \{\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_l, \mathbf{Z} \in S\}$ and S is a compact support set of $\mathbf{Z}$. Without loss of generality, we assume $\mathcal{S}_w = [0, 1]$.

(C2) For every $1 \le l \le d$, the nonparametric function $m_l \in C^{(r)}[0, 1]$ for some integer $r \ge 2$, and the spline order q satisfies $q \ge r$.

(C3) The conditional variance function $\sigma^2(\mathbf{z}, \mathbf{x})$ is measurable and bounded above from C_σ, for some constant $0 < C_\sigma < \infty$.

(C4) There exist constants $0 < c_Q \le C_Q < \infty$, such that $c_Q \le Q(\mathbf{z}) = E(\mathbf{X}\mathbf{X}^{\mathrm{T}} | \mathbf{Z} = \mathbf{z}) \le C_Q$ for all $\mathbf{z} \in S$.

(C5) For $1 \le k \le p$ and $1 \le l \le d$, $g_{l,k}^0 \in C^{(1)}[0, 1]$.

(C6) The kernel function K is a symmetric probability density, supported on $[-1, 1]$ and $K \in C^{0,1}[-1, 1]$.

(C7) The functions $u^3K(u)$ and $u^3K'(u)$ are bounded and $\int u^4K(u)\,du < \infty$. $E|\varepsilon|^4 < \infty$.

It is noteworthy that Condition (C1) is the same as Condition (d) in Cui, Härdle, and Zhu (2011). Condition (C2) is given in Theorem 2.1 of Zhou, Shen, and Wolfe (1998). Condition (C3) is the same as Condition (C5) of Xue and Yang (2006). Condition (C4) is given in Condition (C2) of Xia and Härdle (2006) and Condition (C5) of Xue and Liang (2010). Condition (C5) gives the smoothness condition of functions $g_{l,k}^0$ defined in (10). Condition (C6) is a common assumption on the kernel function in the nonparametric smoothing literature. Condition (C7) is the same as Conditions (A3) and (A4) in Fan, Zhang, and Zhang (2001), which is used for obtaining the asymptotic distribution of the GLR statistic.

A.2 Proofs of Theorems 1 and 2

Denote $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $\mathbf{m} = \{m(\mathbf{Z}_1, \mathbf{X}_1, \boldsymbol{\beta}^0), \ldots, m(\mathbf{Z}_n, \mathbf{X}_n, \boldsymbol{\beta}^0)\}^{\mathrm{T}}$. By (5), $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ can be decomposed into $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \widehat{\boldsymbol{\lambda}}_m(\boldsymbol{\beta}) + \widehat{\boldsymbol{\lambda}}_e(\boldsymbol{\beta})$, where

$$\begin{aligned}\widehat{\boldsymbol{\lambda}}_m(\boldsymbol{\beta}) &= \left\{\mathbf{D}(\boldsymbol{\beta})^{\mathrm{T}}\mathbf{D}(\boldsymbol{\beta})\right\}^{-1}\mathbf{D}(\boldsymbol{\beta})^{\mathrm{T}}\mathbf{m},\\ \widehat{\boldsymbol{\lambda}}_e(\boldsymbol{\beta}) &= \left\{\mathbf{D}(\boldsymbol{\beta})^{\mathrm{T}}\mathbf{D}(\boldsymbol{\beta})\right\}^{-1}\mathbf{D}(\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{m}). \qquad \text{(A.1)}\end{aligned}$$

We first present three lemmas and one proposition which will be used in the proofs of Theorems 1 and 2. The detailed proofs of the lemmas are given in the online supplementary materials. Lemma A.2 is used for Lemma A.3, which is needed in the proof of Theorem 1. Define

$$\mathbf{V}(\boldsymbol{\beta}) = E\left(D_i(\boldsymbol{\beta})D_i(\boldsymbol{\beta})^{\mathrm{T}}\right), \widehat{\mathbf{V}}(\boldsymbol{\beta}) = n^{-1}\mathbf{D}(\boldsymbol{\beta})^{\mathrm{T}}\mathbf{D}(\boldsymbol{\beta}). \qquad \text{(A.2)}$$

Lemma A.1. Under Conditions (C1) and (C4), for any vector $\boldsymbol{\alpha} = \{(\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_d^{\mathrm{T}})^{\mathrm{T}}\}_{dJ_n \times 1}$ with $\boldsymbol{\alpha}_l = (\alpha_{s,l} : 1 \le s \le J_n)^{\mathrm{T}}$, there are constants $0 < c_V < C_V < \infty$, such that $\forall \boldsymbol{\beta} \in \Theta$ and for large enough n,

$$\begin{aligned}c_V J_n^{-1}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} &\le \alpha^{\mathrm{T}}\mathbf{V}(\beta)\alpha \le C_V J_n^{-1}\alpha^{\mathrm{T}}\boldsymbol{\alpha}, C_V^{-1}J_n\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} \le \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{V}(\beta)^{-1}\boldsymbol{\alpha}\\ &\le c_V^{-1}J_n\boldsymbol{\alpha}^{\mathrm{T}}\alpha. \qquad \text{(A.3)}\end{aligned}$$

$$\begin{aligned}&\sup_{1\le s,s'\le J_n, 1\le l\le d}\left|n^{-1}\sum_{i=1}^n D_{i,sl}(\boldsymbol{\beta}_l)D_{i,s'l}(\boldsymbol{\beta}_l) - E\left\{D_{i,sl}(\boldsymbol{\beta}_l)D_{i,s'l}(\boldsymbol{\beta}_l)\right\}\right|\\ &= O_{\text{a.s.}}\left(\sqrt{J_n^{-1}n^{-1}\log n}\right), \qquad \text{(A.4)}\\ &\sup_{1\le s,s'\le J_n, l\ne l'}\left|n^{-1}\sum_{i=1}^n D_{i,sl}(\boldsymbol{\beta}_l)D_{i,s'l'}(\boldsymbol{\beta}_l) - E\left\{D_{i,sl}(\boldsymbol{\beta}_l)D_{i,s'l'}(\boldsymbol{\beta}_l)\right\}\right|\\ &= O_{\text{a.s.}}\left(J_n^{-1}\sqrt{n^{-1}\log n}\right). \qquad \text{(A.5)}\end{aligned}$$

By Lemma A.1, one has with probability approaching 1, for large enough n, $\forall \boldsymbol{\beta} \in \Theta$,

$$\begin{aligned}c_V J_n^{-1}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} &\le \alpha^{\mathrm{T}}\widehat{\mathbf{V}}(\beta)\alpha \le C_V J_n^{-1}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}, C_V^{-1}J_n\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} \le \alpha^{\mathrm{T}}\widehat{\mathbf{V}}(\beta)^{-1}\boldsymbol{\alpha}\\ &\le c_V^{-1}J_n\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} \qquad \text{(A.6)}\end{aligned}$$

for any vector $\boldsymbol{\alpha} = \{(\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_d^{\mathrm{T}})^{\mathrm{T}}\}_{dJ_n\times 1}$ with $\boldsymbol{\alpha}_l = (\alpha_{s,l} : 1 \le s \le J_n)^{\mathrm{T}}$. By (A.3) and Demko (1986), it can be proved that $\forall \boldsymbol{\beta} \in \Theta$ and for large enough n, there is a constant $0 < C_V^* < \infty$ such that $\|\mathbf{V}(\beta)^{-1}\|_\infty \le C_V^* J_n$. Following this result, (A.4) and (A.5), it can be proved that $\forall \boldsymbol{\beta} \in \Theta$,

$$\left\|\widehat{\mathbf{V}}(\beta)^{-1}\right\|_\infty = O_p(J_n). \qquad \text{(A.7)}$$

Let $\mathbf{E} = \mathbf{Y} - \mathbf{m} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$.

Lemma A.2. Under Conditions (C1), (C3), and (C4), $\forall \boldsymbol{\beta} \in \Theta$, $\left\|n^{-1}\mathbf{D}(\boldsymbol{\beta})^{\mathrm{T}}\mathbf{E}\right\|_2 = O_p\left(n^{-1/2}\right)$.

Lemma A.3. Under Conditions (C1)–(C5), and $nN^{-4} \to \infty$ and $nN^{-2r-2} \to 0$, as $n \to \infty$,

$$\begin{aligned}\partial L_n\left(\boldsymbol{\beta}^0\right)/\partial\boldsymbol{\beta}_{-1} &= -\sum_{i=1}^n\left\{Y_i - \sum_{l=1}^d m_l\left(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_l^0\right)X_{il}\right\}\\ &\times\left[\dot{m}_l(U_{il}(\boldsymbol{\beta}_l^0), \boldsymbol{\beta}^0)X_{il}\mathbf{J}_l^{\mathrm{T}}\widetilde{\mathbf{Z}}_i\right]_{l=1}^d + o_p\left(n^{1/2}\right).\end{aligned}$$

The proposition presented next gives the convergence rate of the estimators $\widehat{m}_l(u_l, \boldsymbol{\beta}^0)$ and $\widehat{\dot{m}}_l(u_l, \boldsymbol{\beta}^0)$ for the nonparametric function $m_l(u_l)$ and its first derivative $\dot{m}_l(u_l)$, for $l = 1, \ldots, d$.

Proposition A.1. Under Conditions (C1)–(C4), and $N \to \infty$ and $nN^{-1} \to \infty$, as $n \to \infty$ one has (i) $|\widehat{m}_l(u_l, \boldsymbol{\beta}^0) - m_l(u_l)| = O_p(n^{-1/2}N^{1/2} + N^{-r})$ uniformly for any $u_l \in [0, 1]$; and (ii) under $N \to \infty$ and $nN^{-3} \to \infty$, as $n \to \infty$, $|\widehat{\dot{m}}_l(u_l, \boldsymbol{\beta}^0) - \dot{m}_l(u_l)| = O_p(n^{-1/2}N^{3/2} + N^{-r+1})$ uniformly for any $u_l \in [0, 1]$.

Proof. Let $\widehat{\boldsymbol{\lambda}}_e(\boldsymbol{\beta}) = \left\{\widehat{\boldsymbol{\lambda}}_{1,e}(\boldsymbol{\beta})^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\lambda}}_{d,e}(\boldsymbol{\beta})^{\mathrm{T}}\right\}^{\mathrm{T}}$, where $\widehat{\boldsymbol{\lambda}}_{l,e}(\boldsymbol{\beta}) = \left\{\widehat{\lambda}_{s,l,e}(\boldsymbol{\beta}) : 1 \le s \le J_n\right\}^{\mathrm{T}}$ and $\widehat{\boldsymbol{\lambda}}_m(\boldsymbol{\beta}) = \left\{\widehat{\boldsymbol{\lambda}}_{1,m}(\boldsymbol{\beta})^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\lambda}}_{d,m}(\boldsymbol{\beta})^{\mathrm{T}}\right\}^{\mathrm{T}}$, where $\widehat{\boldsymbol{\lambda}}_{l,m}(\boldsymbol{\beta}) = \left\{\widehat{\lambda}_{s,l,m}(\boldsymbol{\beta}) : 1 \le s \le J_n\right\}^{\mathrm{T}}$. Thus

$$\widehat{m}_l(u_l, \boldsymbol{\beta}) = \widehat{m}_{l,e}(u_l, \boldsymbol{\beta}) + \widehat{m}_{l,m}(u_l, \boldsymbol{\beta}), \qquad \text{(A.8)}$$

where

$$\widehat{m}_{l,e}(u_l, \boldsymbol{\beta}) = \mathbf{B}_r(u_l)^{\mathrm{T}}\widehat{\boldsymbol{\lambda}}_{l,e}(\boldsymbol{\beta}) \text{ and } \widehat{m}_{l,m}(u_l, \boldsymbol{\beta}) = \mathbf{B}_r(u_l)^{\mathrm{T}}\widehat{\boldsymbol{\lambda}}_{l,m}(\boldsymbol{\beta}). \qquad \text{(A.9)}$$

According to the result on p. 149 of de Boor (2001), for m_l satisfying Condition (C2), there is a function $m_l^0(u_l) = \mathbf{B}_r(u_l)^{\mathrm{T}}\boldsymbol{\lambda}_l \in G_n$, such that

$$\sup_{u_l\in[0,1]}\left|m_l^0(u_l) - m_l(u_l)\right| = O\left(J_n^{-r}\right). \qquad \text{(A.10)}$$

Let $\mathbb{B}_r(\mathbf{u}) = \begin{bmatrix}\mathbf{B}_r(u_1)^{\mathrm{T}} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \mathbf{B}_r(u_d)^{\mathrm{T}}\end{bmatrix}_{d\times J_n d}$, where $\mathbf{u} = (u_1, \ldots u_d)^{\mathrm{T}}$. Thus $\widehat{m}_{l,e}(u_l, \boldsymbol{\beta}^0) = \mathbf{1}_l^{\mathrm{T}}\mathbb{B}_r(\mathbf{u})\widehat{\boldsymbol{\lambda}}_e(\boldsymbol{\beta}^0)$ and $\widehat{m}_{l,m}(u_l, \boldsymbol{\beta}^0) = \mathbf{1}_l^{\mathrm{T}}\mathbb{B}_r(\mathbf{u})\widehat{\boldsymbol{\lambda}}_m(\boldsymbol{\beta}^0)$, where $\mathbf{1}_l$ is the $d \times 1$ vector with the lth element as "1" and other elements as "0". Let $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_1^{\mathrm{T}}, \ldots, \boldsymbol{\lambda}_d{}^{\mathrm{T}}\}^{\mathrm{T}}$. By Berstein's inequality in Bosq (1998), it can be proved that $\|n^{-1}\mathbf{D}(\boldsymbol{\beta}^0)^{\mathrm{T}}\mathbf{1}_n\|_\infty = O_p(J_n^{-1})$. Thus, by (A.6), (A.7), and (A.10), for every $u_l \in [0, 1]$,

$$\begin{aligned}&\left|\widehat{m}_{l,m}(u_l, \boldsymbol{\beta}^0) - m_l^0(u_l)\right|\\ &= \left|n^{-1}\mathbf{1}_l^{\mathrm{T}}\mathbb{B}_r(\mathbf{u})\widehat{\mathbf{V}}(\boldsymbol{\beta}^0)^{-1}\mathbf{D}(\boldsymbol{\beta}^0)^{\mathrm{T}}\left\{\mathbf{m} - \mathbf{D}(\beta^{\mathbf{0}})\lambda\right\}\right|\\ &\le \left|\sum_{s=1}^{J_n}B_{s,r}(u_l)\right|\left\|\widehat{\mathbf{V}}(\boldsymbol{\beta})^{-1}\right\|_\infty\left\|n^{-1}\mathbf{D}(\boldsymbol{\beta}^0)^{\mathrm{T}}\mathbf{1}_n\right\|_\infty O\left(J_n^{-r}\right)\\ &= O_p(J_n)\,O_p\left(J_n^{-1}\right)O\left(J_n^{-r}\right) = O_p\left(J_n^{-r}\right). \qquad \text{(A.11)}\end{aligned}$$

Moreover, for every $u_l \in [0,1]$, by (A.1), (A.6), and Condition (C3), with probability approaching 1,

$$\begin{aligned} E\left\{\widehat{m}_{l,e}(u_l,\boldsymbol\beta^0)\,|\mathbf{X},\mathbf{Z}\right\}^2 &= n^{-2}\mathbf{1}_l^{\mathrm T}\mathbb{B}_r(\mathbf{u})\widehat{\mathbf{V}}(\boldsymbol\beta^0)^{-1}\mathbf{D}(\boldsymbol\beta^0)^{\mathrm T}E\left(\mathbf{E}\mathbf{E}^{\mathrm T}\,|\mathbf{X},\mathbf{Z}\right)\mathbf{D}(\boldsymbol\beta^0)\widehat{\mathbf{V}}(\boldsymbol\beta^0)^{-1} \\ &\quad\times\mathbb{B}_r(\mathbf{u})^{\mathrm T}\mathbf{1}_l \\ &\le n^{-1}C_\sigma\mathbf{1}_l^{\mathrm T}\mathbb{B}_r(\mathbf{u})\widehat{\mathbf{V}}(\boldsymbol\beta^0)^{-1}\mathbb{B}_r(\mathbf{u})^{\mathrm T}\mathbf{1}_l \\ &\le n^{-1}C_\sigma\left\|\mathbb{B}_r(\mathbf{u})^{\mathrm T}\mathbf{1}_l\right\|_2^2\left\|\widehat{\mathbf{V}}(\boldsymbol\beta^0)^{-1}\right\|_2 = O(J_n/n). \end{aligned} \tag{A.12}$$

Thus, by the weak law of large numbers, for every $u_l \in [0,1]$, $\widehat{m}_{l,e}(u_l,\boldsymbol\beta^0) = O_p(J_n^{1/2}n^{-1/2})$. Therefore, by (A.10), (A.11), and (A.12), $|\widehat{m}_l(u_l,\boldsymbol\beta^0) - m_l(u_l)| = O_p(J_n^{1/2}n^{-1/2} + J_n^{-r})$, uniformly for every $u_l \in [0,1]$. Results in (i) of Proposition A.1 are proved. Similarly, $\widehat{\dot m}_l(u_l,\boldsymbol\beta^0)$ can be written as $\widehat{\dot m}_{l,e}(u_l,\boldsymbol\beta^0) + \widehat{\dot m}_{l,m}(u_l,\boldsymbol\beta^0)$, where $\widehat{\dot m}_{l,e}(u_l,\boldsymbol\beta^0) = \mathbf{B}_{r-1}(u_l)^{\mathrm T}\mathbf{D}_1\widehat{\boldsymbol\lambda}_{l,e}(\boldsymbol\beta^0)$ and $\widehat{\dot m}_{l,m}(u_l,\boldsymbol\beta^0) = \mathbf{B}_{r-1}(u_l)^{\mathrm T}\mathbf{D}_1\widehat{\boldsymbol\lambda}_{l,m}(\boldsymbol\beta^0)$. It is easy to prove that $\|\mathbf{D}_1\|_\infty = O(J_n)$, where $\mathbf{D}_1$ is defined in (7). Following the similar reasoning as the proof for $\widehat{m}_l(u_l,\boldsymbol\beta^0)$, one can prove that

$$\widehat{\dot m}_l(u_l,\boldsymbol\beta^0) - \dot m_l(u_l) = O_p\left(J_n^{3/2}n^{-1/2} + J_n^{-r+1}\right),$$

uniformly for every $u_l \in [0,1]$. Thus, results in (ii) of Proposition A.1 are proved. □

Proof of Theorem 1. Under the conditions of Theorem 1, we follow similar arguments as presented in Ichimura (1993) to show that $\widehat{\boldsymbol\beta}_{-1}$ is a root-n consistent estimator of $\boldsymbol\beta^0_{-1}$, and thus the proof is omitted. By Lemma A.3, it is straightforward to prove that

$$\partial L_n(\boldsymbol\beta^0)/\partial\boldsymbol\beta_{-1}\partial\boldsymbol\beta_{-1}^{\mathrm T} = \sum_{i=1}^n\left[\left[\dot m_l(U_{il}(\boldsymbol\beta_l^0),\boldsymbol\beta^0)X_{il}\mathbf{J}_l^{\mathrm T}\widetilde{\mathbf{Z}}_i\right]_{l=1}^d\right]^{\otimes 2} + o_p(n).$$

By Taylor expansion, Lemma A.3, and the previous result,

$$\begin{aligned} \widehat{\boldsymbol\beta}_{-1} - \boldsymbol\beta^0_{-1} &= -\left\{\partial L_n(\boldsymbol\beta^0)/\partial\boldsymbol\beta_{-1}\partial\boldsymbol\beta_{-1}^{\mathrm T}\right\}^{-1}\{\partial L_n(\boldsymbol\beta^0)/\partial\boldsymbol\beta_{-1}\}\{1+o_p(1)\} \\ &= \left[E\left[\left\{\dot m_l(U_{il}(\boldsymbol\beta_l^0),\boldsymbol\beta^0)X_{il}\mathbf{J}_l^{\mathrm T}\widetilde{\mathbf{Z}}_i\right\}_{l=1}^d\right]^{\otimes 2}\right]^{-1} \\ &\quad\times n^{-1}\sum_{i=1}^n\varepsilon_i\left[\dot m_l(U_{il}(\boldsymbol\beta_l^0),\boldsymbol\beta^0)X_{il}\mathbf{J}_l^{\mathrm T}\widetilde{\mathbf{Z}}_i\right]_{l=1}^d + o_p\left(n^{-1/2}\right). \end{aligned}$$

Thus, Theorem 1 follows from the previous results and Lindeberg–Feller central limit theorem. □

Proof of Theorem 2. Since $\|\widehat{\boldsymbol\beta} - \boldsymbol\beta^0\|_2 = O_p(n^{-1/2})$, Theorem 2 follows from this result and Proposition A.1. □

A.3 Proofs of Theorems 3 and 4

Following the same techniques employed in Fan and Zhang (2008), it can be proved that the oracle estimator $\widetilde{m}_{\mathrm{LL},1}(u_1,\boldsymbol\beta^0)$ has the asymptotic distribution and convergence rate given in Theorem 3. The detailed proof is thus omitted. Since $\|\widehat{\boldsymbol\beta} - \boldsymbol\beta^0\|_2 = O_p(n^{-1/2})$, Theorem 3 is proved by Slutsky's theorem. We will focus on the proof of Theorem 4.

According to (17) and (A.8),

$$\begin{aligned} &\widehat{m}_{\mathrm{SBLL},1}(u_1,\boldsymbol\beta^0) - \widetilde{m}_{\mathrm{LL},1}\left(u_1,\boldsymbol\beta^0\right) \\ &\quad= -(1,0)\left\{\mathbf{C}\left(u_1,\boldsymbol\beta_1^0\right)^{\mathrm T}\mathbf{W}\left(u_1,\boldsymbol\beta_1^0\right)\mathbf{C}\left(u_1,\boldsymbol\beta_1^0\right)\right\}^{-1} \\ &\qquad\times\mathbf{C}\left(u_1,\boldsymbol\beta_1^0\right)^{\mathrm T}\mathbf{W}\left(u_1,\boldsymbol\beta_1^0\right) \\ &\qquad\times\left[\sum_{l=2}^d\left\{\widehat{m}_l\left(U_{il}\left(\boldsymbol\beta^0\right),\boldsymbol\beta^0\right) - m_l\left(U_{il}\right)\right\}X_{il}\right]_{i=1}^n \\ &\quad= -(1,0)\left\{n^{-1}\mathbf{C}\left(u_1,\boldsymbol\beta_1^0\right)^{\mathrm T}\mathbf{W}\left(u_1,\boldsymbol\beta_1^0\right)\mathbf{C}\left(u_1,\boldsymbol\beta_1^0\right)\right\}^{-1} \\ &\qquad\times\left\{\begin{pmatrix}\Psi_{v1}\left(u_1,\boldsymbol\beta^0\right)\\ \Psi_{v2}\left(u_1,\boldsymbol\beta^0\right)\end{pmatrix} + \begin{pmatrix}\Psi_{b1}\left(u_1,\boldsymbol\beta^0\right)\\ \Psi_{b2}\left(u_1,\boldsymbol\beta^0\right)\end{pmatrix}\right\}, \end{aligned}$$

where

$$\begin{aligned} \Psi_{v1}(u_1,\boldsymbol\beta^0) &= n^{-1}\sum_{i=1}^n\sum_{l=2}^d X_{i1}X_{il}K_{h_1}\left(U_{i1}(\boldsymbol\beta_1^0) - u_1\right)\widehat{m}_{l,\varepsilon}(U_{il},\boldsymbol\beta^0), \\ \Psi_{v2}(u_1,\boldsymbol\beta^0) &= n^{-1}\sum_{i=1}^n\sum_{l=2}^d\left\{\left(U_{i1}(\boldsymbol\beta_1^0) - u_1\right)/h_1\right\}X_{i1}X_{il}K_{h_1} \\ &\quad\times(U_{i1}(\boldsymbol\beta_1^0) - u_1)\widehat{m}_{l,\varepsilon}(U_{il},\boldsymbol\beta^0), \\ \Psi_{b1}(u_1,\boldsymbol\beta^0) &= n^{-1}\sum_{i=1}^n\sum_{l=2}^d X_{i1}X_{il}K_{h_1}\left(U_{i1}(\boldsymbol\beta_1^0) - u_1\right) \\ &\quad\times\left\{\widehat{m}_{l,m}(U_{il},\boldsymbol\beta^0) - m_l(U_{il})\right\}, \\ \Psi_{b2}(u_1,\boldsymbol\beta^0) &= n^{-1}\sum_{i=1}^n\sum_{l=2}^d\left\{\left(U_{i1}(\boldsymbol\beta_1^0) - u_1\right)/h_1\right\}X_{i1}X_{il} \\ &\quad\times K_{h_1}(U_{i1}(\boldsymbol\beta_1^0) - u_1)\left\{\widehat{m}_{l,m}(U_{il},\boldsymbol\beta^0) - m_l(U_{il})\right\}. \end{aligned}$$

In the following, we present two lemmas which will be used in the proofs of Theorem 4. The detailed proofs are given in the online, supplementary materials.

Lemma A.4. Under Conditions (C1), (C3), (C4) and (C6), and $N\to\infty$ and $nN^{-1}\to\infty$, as $n\to\infty$, one has $\sup_{u_1\in[0,1]}|\Psi_{v1}(u_1,\boldsymbol\beta^0)| + \sup_{u_1\in[0,1]}|\Psi_{v2}(u_1,\boldsymbol\beta^0)| = O_p(n^{-1/2})$.

Lemma A.5. Under Conditions (C1), (C4), and (C6), and $N\to\infty$, as $n\to\infty$, one has $\sup_{u_1\in[0,1]}\left|\Psi_{b1}\left(u_1,\boldsymbol\beta^0\right)\right| + \sup_{u_1\in[0,1]}\left|\Psi_{b2}\left(u_1,\boldsymbol\beta^0\right)\right| = O_p\left(J_n^{-r}\right)$.

Proof of Theorem 4. It is straightforward to prove that

$$\sup_{u_1\in[0,1]}\|\{n^{-1}\mathbf{C}(u_1,\boldsymbol\beta_1^0)^{\mathrm T}\mathbf{W}(u_1,\boldsymbol\beta_1^0)\mathbf{C}(u_1,\boldsymbol\beta_1^0)\}^{-1}\|_2 \le C$$

for some constants $0 < C < \infty$. Thus, by Lemmas A.4 and A.5, one has

$$\sup_{u_1\in[0,1]}|\widehat{m}_{\mathrm{SBLL},1}(u_1,\boldsymbol\beta^0) - \widetilde{m}_{\mathrm{LL},1}(u_1,\boldsymbol\beta^0)| = O_p(n^{-1/2} + J_n^{-r}).$$

Since $\|\widehat{\boldsymbol\beta} - \boldsymbol\beta^0\|_2 = O_p(n^{-1/2})$, $\sup_{u_1\in[0,1]}|\widehat{m}_{\mathrm{SBLL},1}(u_1,\widehat{\boldsymbol\beta}) - \widetilde{m}_{\mathrm{LL},1}(u_1,\widehat{\boldsymbol\beta})| = O_p(n^{-1/2} + J_n^{-r})$. Therefore, under the assumption that $nN^{-5r/2} = o(1)$ and $n^{-1}N = o(1)$, Theorem 4 is proved. □

SUPPLEMENTARY MATERIALS

The online supplementary materials contain the procedure of generating initial values in Section 5.1, estimation of optimal bandwidth $h_{1,\mathrm{opt}}$, and additional proofs.

[Received March 2013. Revised February 2014.]

REFERENCES

Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes*, New York: Springer-Verlag. [354]

Cai, Z., Fan, J., and Li, R. (2000), "Efficient Estimation and Inferences for Varying-Coefficient Models," *Journal of the American Statistical Association*, 95, 888–902. [341]

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489. [342,343,347]

Cui, X., Härdle, W., and Zhu, L. (2011), "The EFM Approach for Single-Index Models," *The Annals of Statistics*, 39, 1658–1688. [342,344,354]

de Boor, C. (2001), *A Practical Guide to Splines*, New York: Springer. [342,344,354]

Demko, S. (1986), "Spectral Bounds for $|a^{-1}|_\infty$," *Journal of Approximation Theory*, 48, 207–212. [354]

Fan, J., and Jiang, J. (2007), "Nonparametric Inference With Generalized Likelihood Ratio Tests," *Test*, 16, 409–444. [347,349]

Fan, J., Yao, Q., and Cai, Z. (2003), "Adaptive Varying-Coefficient Linear Models," *Journal of the Royal Statistical Society,* Series B, 1, 57–80. [343]

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29, 153–193. [346,347,354]

Fan, J., and Zhang, W. (2008), "Statistical Methods With Varying Coefficient Models," *Statistics and its Interface*, 1, 179–195. [341,355]

Grun, F., and Blumberg, B. (2009), "Minireview: The Case for Obesogens," *Molecular Endocrinology*, 23, 1127–1134. [341]

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models," *The Annals of Statistics*, 21, 157–178. [343]

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall. [342,343]

——— (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society,* Series B, 55, 757–796. [341]

Hatch, E. E., Nelson, J. W., Stahlhut, R. W., and Webster, T. F. (2010), "Association of Endocrine Disruptors and Obesity: Perspectives From Epidemiological Studies," *International Journal of Andrology*, 33, 324–332. [341]

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58, 71–120. [355]

La Merrill, M., and Birnbaum, L. S. (2011), "Childhood Obesity and Environmental Chemicals," *Mount Sinai Journal of Medicine*, 78, 22–48. [342]

Liang, H., Liu, X., Li, R., and Tsai, C. L. (2010), "Estimation and Testing for Partially Linear Single-Index Models," *The Annals of Statistics*, 38, 3811–3836. [342,343,346,347]

Liu, R., and Yang, L. (2010), "Spline-Backfitted Kernel Smoothing of Additive Coefficient Model," *Econometric Theory*, 26, 29–59. [342]

Lu, X., Chen, G., Singh, R., and Song, X.-K. P. (2006), "A Class of Partially Linear Single-Index Survival Models," *Canadian Journal of Statistics*, 34, 97–112. [343]

Ma, S., Song, Q., and Wang, L. (2013), "Variable Selection and Estimation in Marginal Partially Linear Additive Models for Longitudinal Data," *Bernoulli*, 19, 252–274. [342]

Ma, S., and Yang, L. (2011), "Spline-Backfitted Kernel Smoothing of Partially Linear Additive Model," *Journal of Statistical Planning and Inference*, 141, 204–219. [343]

Mammen, E., Linton, O., and Nielsen, J. (1999), "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions," *The Annals of Statistics*, 27, 1443–1490. [342]

Meeker, J. D. (2012), "Exposure to Environmental Endocrine Disrupters and Child Development," *Archives of Pediatrics and Adolescent Medicine*, 166, 1–7. [341]

Opsomer, J. D., and Ruppert, D. (1997), "Fitting a Bivariate Additive Model by Local Polynomial Regression," *The Annals of Statistics*, 25, 186–211. [342]

Rider, C. V., Furr, J. R., Wilson, V. S., and Gray, L. E., Jr, (2010), "Cumulative Effects of In Utero Administration of Mixtures of Reproductive Toxicants That Disrupt Common Target Tissues via Diverse Mechanisms of Toxicity," *International Journal of Andrology*, 33, 443–462. [342]

Stone, C. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [342]

Wang, L., and Yang, L. (2007), "Spline-Backfitted Kernel Smoothing of Nonlinear Additive Autoregression Model," *The Annals of Statistics*, 35, 2474–2503. [342,343]

——— (2009a), "Spline Estimation of Single Index Model," *Statistica Sinica*, 19, 765–783. [342]

Wang, J., and Yang, L. (2009b), "Efficient and Fast Spline-Backfitted Kernel Smoothing of Additive Regression Model," *Annals of the Institute of Statistical Mathematics*, 61, 663–690. [342]

Wang, L., Liu, X., Liang, H., and Carroll, R. J. (2011), "Estimation and Variable Selection for Generalized Additive Partial Linear Models," *The Annals of Statistics*, 39, 1827–1851. [342,343]

Wong, H., lp, W., and Zhang, R. (2008), "Varying-Coefficient Single-Index Model," *Computational Statistics & Data Analysis*, 52, 1458–1476. [343]

Xia, Y., and Härdle, W. (2006), "Semi-Parametric Estimation of Partially Linear Single-Index Models," *Journal of Multivariate Analysis*, 97, 1162–1184. [342,354]

Xia, Y., and Li, W. K. (1999), "On Single-Index Coefficient Regression Models," *Journal of the American Statistical Association*, 94, 1275–1285. [342,343]

Xia, Y., Tong, H., and Li, W. K. (1999), "On Extended Partially Linear Single-Index Models," *Biometrika*, 86, 831–842. [342,343]

Xia, Y., Tong, H., Li, W. K., and Zhu, L. (2002), "An Adaptive Estimation of Dimension Reduction Space" (with discussion), *Journal of the Royal Statistical Society,* Series B, 64, 363–410. [347]

Xue, L., and Liang, H. (2010), "Polynomial Spline Estimation for the Generalized Additive Coefficient Model," *Scandinavian Journal of Statistics*, 37, 26–46. [354]

Xue, L., and Yang, L. (2006), "Additive Coefficient Modeling via Polynomial Spline," *Statistica Sinica*, 16, 1423–1446. [354]

Xue, L., and Wang, Q. (2012), "Empirical Likelihood for Single-Index Varying-Coefficient Models," *Bernoulli*, 3, 836–856. [343]

Zhou, S., Shen, X., and Wolfe, D. A. (1998), "Local Asymptotics for Regression Splines and Confidence Regions," *The Annals of Statistics*, 26, 1760–1782. [344,346,354]